Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 0 (0). 2024. 1-29. Published online: 30.07.2024
https://doi.org/10.14746/ssllt.38492
http://pressto.amu.edu.pl/index.php/ssllt

Relative complexity in a model of word difficulty: The role of loanwords in vocabulary size tests[1]

Derek N. Canning ✉
Seigakuin University, Ageo, Japan
https://orcid.org/0009-0004-5868-4502
dn_canning@seigakuin-univ.ac.jp

Stuart McLean
Momoyama Gakuin University, Osaka, Japan
https://orcid.org/0000-0002-7035-378X
stumc93@gmail.com

Joseph P. Vitta
Waseda University, Tokyo, Japan
https://orcid.org/0000-0002-5711-969X
vittajp@waseda.jp

Abstract

Recent studies have shown that the frequency effect, although long used as a guide to word difficulty, fails to explain all variance in learner word knowledge. As such, a "more than frequency" conclusion has been offered to explain how lexical sophistication accounts for word difficulty. This study presents a multiple regression model of word-learning difficulty from a data set of monolingual Japanese first language (L1) learners. Vocabulary Size Test (VST) scores of 2,999 L1 Japanese university students were converted to logit scores to determine the word-learning difficulty of 80 target words. Five lexical sophistication variables were found to correlate with word-learning difficulty (frequency,

[1] This article is based on data published in McLean et al. (2014).

cognate status, age of acquisition, prevalence, and polysemy) above a practical significance threshold. These were subsequently entered into a regression model with the logit scores as the dependent variable. The model (R^2 = .55) indicates that three lexical sophistication variables significantly predicted VST scores: frequency ($ß$ = -.28, p = .029), cognateness ($ß$ = -.24, p = .005), and prevalence ($ß$ = 0.22, p = .040). Despite suggestions that complexity studies be interpreted considering what is understood about the construct of linguistic complexity, researchers have rarely made explicit the differences between absolute and relative complexity variables. As some variables can be shown to vary in complexity according to the L1 population, these must be considered in discussions of test generalizability. Although frequency will continue to be the primary criterion for the selection of lexical items for teaching and testing, the cognate status of words can be used to predict the potential learning burden of the word more precisely for learners of different L1 backgrounds.

Keywords: lexical sophistication; loanwords; cognates; vocabulary size test (VST); frequency effect; L2 word difficulty

1. Introduction

Interest in lexical sophistication as the multidimensional predictive construct for second language (L2) word difficulty, or what makes one lexical item more or less easy to learn than any other word, has prompted recent research (De Wilde, 2023; Hashimoto, 2021; Hashimoto & Egbert, 2019; Vitta et al., 2023). These studies suggest that the frequency effect, although long used as a guide to word difficulty, fails to explain all variance in learner word knowledge and thus a "more than frequency" conclusion has been offered to explain how lexical sophistication accounts for word difficulty. It has been demonstrated that additional lexical variables, including but not limited to word neighborhood, cognateness, and the psychometric properties of lexical items explain as much of the learning burden of a word as does its frequency (De Wilde, 2023; Hashimoto, 2021; Hashimoto & Egbert, 2019; Laufer, 1989; Willis & Ohashi, 2012; Vitta et al., 2023). Precisely which of these variables are of the greatest importance to models of lexical difficulty and to what degree they contribute to predicted (or explained) variance is still a matter of debate.

Lexical sophistication is regarded as a sub-construct of lexical complexity, itself an element of the broader construct of linguistic complexity (Bulté & Housen, 2012). Despite the lack of a comprehensive theory of linguistic complexity, Bulté and Housen (2012) recommend that our understanding of operationalized variables would be well served by "a more explicit characterization of complexity" (2012, p. 22). They argue that all properties of language can be regarded as facets of either *relative complexity* (which they also term *difficulty*), or *absolute complexity*

(which they term simply *complexity*). Relative complexity includes those aspects of language acquisition that are learner-dependent, such as memory, motivation, or first language (L1) background. Absolute complexity refers to the number of different components, and the connections between them, of a language system. In linguistics, absolute complexity is characterized as an objective, theoretical characterization of the overall system, distinct from the difficulty involved in learning that system (Dahl, 2004).

Studies on word difficulty and lexical sophistication have included variables such as cognateness (De Wilde, 2023; De Wilde et al., 2020; Willis & Ohashi, 2012), frequency and range (Hashimoto & Egbert, 2019; Vitta et al., 2023) and psycholinguistic variables (Vitta et al., 2023). Despite suggestions that results be interpreted in light of what is understood about the construct of linguistic complexity (Bulté & Housen, 2012; Pallotti, 2015), researchers appear to have yet to explore these variables while explicitly considering the differences between absolute and relative complexity. A theoretical understanding of the variables underpinning lexical sophistication has ramifications for future studies in this area. Most importantly, as some variables can be shown to be relatively complex according to a given L1 population, these must be considered in discussions of generalizability. The majority of studies in this area have focused solely on absolute complexity lexical sophistication variables. However, a recent study (De Wilde, 2023) found that a relative complexity variable, cognateness, was a significant predictor of English L2 word knowledge for L1 Dutch speakers. The following review of recent literature highlights that lexical sophistication variables have not been sufficiently considered in terms of absolute and relative complexity, particularly in a context in which the languages are unrelated, namely, English and Japanese.

2. Review of literature

Although there is no agreed-upon set of factors that contribute to lexical sophistication, research in the field has converged on a limited set of variables that have been explored in several studies. Following the example of Vitta et al. (2023), lexical sophistication is used as a superordinate term for constructs and corresponding measurements that can account for word difficulty. This semantic labeling is not universal in the literature, however (e.g., Hashimoto & Egbert, 2019), but it is useful to organize trends in an area where clarity is lacking and debates are ongoing (see Kim et al., 2018). The following literature review provides an overview of how four broad categories of variables have been used in previous studies: usage- and learner-driven variables, psycholinguistic variables, semantic-network variables, and loanwords (or cognateness). It concludes with a

brief overview of studies that have modeled lexical sophistication in ways similar to the present study.

2.1. Usage- and learner-driven variables

Software applications such as the Tool for the Automatic Analysis of LExical Sophistication (TAALES) (Kyle & Crossley, 2015; Kyle et al., 2018) access databases of lexical sophistication variables. In particular, the introduction of TAALES in 2015 and TAALES 2.0 in 2018 has enabled researchers to quickly index various measures of lexical sophistication (Kyle & Crossley, 2015; Kyle et al., 2018). The TAALES software returns data on factors such as word frequency, word range across corpora, contextual distinctiveness, and L1 word recognition norms. What all of these have in common is that they represent aspects of how the word is used and/or processed by native speakers of the target language, or in the case of L2 prevalence, learners of the target language. How TAALES organizes lexical sophistication variables serves as a useful framework for a brief discussion of some of the theory underpinning these factors and how they fit into a taxonomy of lexical complexity. Similar categories have guided recent discussions of research in this area (Peters, 2020). These categories will guide the following brief overview of lexical sophistication variables, although in the present study, not all variables were indexed through TAALES. To demonstrate the importance of relative complexity in general and loanword status in particular to the construct of lexical complexity, it is necessary to examine the wide range of variables that have been argued to contribute to lexical complexity.

2.1.1. Frequency

Most researchers agree on the primacy of the frequency effect in learning vocabulary (Ellis, 2002; Nation, 2006). It has been argued that all language acquisition is dependent on learners attending to the relative frequency of form-function mappings at all levels of language use, from phonology to grammar (Ellis, 2002). Justifications for the structuring of vocabulary tests have been predicated on the frequency effect (Nation & Beglar, 2007). Despite awareness that frequency is not the only factor affecting vocabulary acquisition (Beglar, 2010; Nation & Beglar, 2007), it is often operationalized as a simple function of frequency (e.g., Siskova, 2012; Lu, 2012).

2.1.2. Range

The simplest measure of range is the number of corpus parts in which the lexical item(s) occur (Gries, 2020). Other measures of range account for other factors, including the relative sizes of different parts of the corpus. A word with a larger range score can be expected to occur in more general contexts than a word with a lower range score that is found in specialized contexts. In both validation studies of the TAALES software, the range of lexical items used by participants significantly explained holistic lexical sophistication in speaking tasks (Kyle & Crossley, 2015) and essay writing quality (Kyle & Crossley, 2016; Kyle et al., 2018). An additional study from Kyle and Crossley (2016) showed that the use of words with a more restricted range resulted in higher holistic scores on a writing task.

2.1.3. Prevalence

Word prevalence is a bottom-up measure of what words are known by a given population of L2 learners (Brysbaert et al., 2021). Prevalence considers word knowledge independently of frequency and acts as a type of learner-driven metric to complement the usage-based constructs and measurements presented in this category. Crowd-sourced surveys of what words were known by English learners from a list of more than 60,000 lemmas were compiled by Brysbaert et al. (2021). As these data have only recently been made available, they are not indexed in TAALES and have not yet been used in research on word difficulty such as the present study.

2.1.4. Contextual distinctiveness

Some of the factors thought to contribute to lexical sophistication have been drawn from corpus linguistics research. Among these is contextual distinctiveness, which is a measure of the probability a given word has of appearing in diverse contexts (McDonald & Shillcock, 2001). Contextual distinctiveness has been found to influence lexical decision reaction times in psycholinguistic research (McDonald & Shillcock, 2001), and semantic diversity has been shown to correlate with frequency and to predict semantic judgment tasks in aphasia patients (Hoffman et al., 2013). In the field of SLA, Hashimoto and Egbert (2019) showed that contextual distinctiveness significantly predicted vocabulary difficulty. In their validation study of TAALES 2.0, however, Kyle et al. (2018) did not find that contextual distinctiveness significantly predicted holistic measures of lexical sophistication in a post-secondary student writing task, although it was correlated with the dependent variable.

2.1.5. Word length

Another lexical sophistication variable that has been shown to influence vocabulary acquisition is word length. Ellis and Beaton (1993) provided evidence that word length negatively correlated with the productive translation of an L2 word. Willis and Ohashi (2012) also found that word length in phonemes helped predict the difficulty of L2 lexical acquisition. Longer words simply contain more opportunities to differ phonologically and orthographically from the L1. Not all considerations of lexical sophistication, however, have included word length as a metric (e.g., Kyle et al., 2018; Vitta et al., 2023).

2.1.6. Word neighborhood

Word neighborhood has also revealed a relationship with word difficulty. Phonological and orthographic neighborhoods are defined as the group of words that can be formed by switching out only one phoneme or letter, respectively, from a reference word (Adelman & Brown, 2007). Research in psycholinguistics has shown that phonological and phonographic neighbors (those closely related both phonologically and orthographically) have facilitative effects on L1 word naming latencies (Adelman & Brown, 2007). Hashimoto and Egbert (2019) found that orthographic neighbor density significantly predicted word difficulty.

2.2. Semantic network

Semantic network measures are those pertaining to a word's degree of polysemy, that is, the number of senses a word has (usually measured at the lemma level) and hypernymy, where words with higher hypernymy scores have more superordinate terms than words with lower scores (Kyle et al., 2018). Standing in contrast to usage-/learner-driven metrics, semantic network properties are determined "within" the nature of the word itself. Research in L2 vocabulary acquisition has shown that polysemy contributes to the difficulty involved in learning a word (Peters, 2020). In a longitudinal study of university students in Britain, Schmitt (1998) found that advanced learners of English tended to acquire only a single meaning of polysemous words as determined through a word-knowledge interview test. Hashimoto and Egbert (2019) combined measurements of polysemy and hypernymy and demonstrated that these measurements correlated with word difficulty and contributed to their final model of word difficulty derived from a yes/no vocabulary knowledge test. Polysemy and hypernymy indices were included only in TAALES Version 2.0. The validation

study of this version of the software uncovered correlations between polysemy and hypernymy, and lexical proficiency in essay-writing tasks (Kyle et al., 2018).

2.3. Psycholinguistic properties of words

Concreteness, imageability, familiarity, and meaningfulness ratings have long been researched in psycholinguistics, although they have more recently been recognized for their potential to explain L2 vocabulary learning (Crossley et al., 2011; Ellis & Beaton, 1993). Imageability and concreteness are measures of the ease with which a referent of a word can be brought to mind. Imageability and concreteness have been argued to influence recall performance equal to frequency (Christian et al., 1978). Word meaningfulness is an index of the subjective ease with which a word can be associated with another word (Ellis & Beaton, 1993; Toglia & Batting, 1978). Familiarity, however, is not a well-defined concept, but is the subjective ranking of a word's "ease of perception" (Tanaka-Ishii & Terada, 2011, p. 1) and is regarded as closely related to meaningfulness (Chumbley & Balota, 1984).

TAALES software makes use of two databases of the psycholinguistic properties of words: the MRC Psycholinguistic Database (Coltheart, 1981) and Brysbaert et al.'s (2013) list of concreteness ratings for 40,000 English lemmas. Second language acquisition research that has made use of these data includes Crossley et al. (2016), who demonstrated that spoken learner vocabulary output tended to become less concrete, or in the researcher's terminology, salient, over time, suggesting that higher concreteness correlates with higher lexical sophistication. In two validation studies for the TAALES software, correlations were observed between the dependent variable and familiarity, imageability, and meaningfulness in Kyle and Crossley (2015) and with familiarity and meaningfulness in Kyle et al. (2018).

An additional psycholinguistic index included in TAALES is age of acquisition. Age of acquisition measures are derived from subjective judgments by adult L1 speakers regarding the period in one's life that a particular word is learned (Kuperman et al., 2012). Age of acquisition has been proposed as a more reliable indication of word-naming latency than frequency in L1 psycholinguistic studies (Morrison & Ellis, 1995). Crossley et al. (2016) left this index out of their study on word concreteness, arguing that it does not reflect salience. Hashimoto and Egbert stated that they excluded this variable from their study of lexical complexity and word difficulty as it pertained only "to learners of a specific background" (2019, p. 850). The variable has been included here, as at least one study has shown that L2 word knowledge correlates with L1 age of acquisition (De Wilde et al., 2020).

One aspect of the psycholinguistic property of words that has not been explored regarding second language word-learning difficulty is valence, arousal, and

dominance (VAD). These three measurements, derived from human ratings, rank words according to positiveness and negativeness (valence) on one axis, on activeness and passiveness (arousal) on another, and dominance and submissiveness (dominance) on the third. It has been argued that these are the three most important dimensions of word meaning and that they can lead to an understanding of affect in language use (Mohammad, 2018).

2.4. Loanwords

Cognates or loanwords have been shown to contribute to vocabulary size test scores (Laufer & McLean, 2016), meaning recall test scores (Daulton, 1998), and word-learning difficulty in monolingual L1 groups in languages related to English (De Wilde et al., 2020; De Wilde, 2023) and in Japanese (Willis & Ohashi, 2012). These studies share a definition of loanwords as phonologically similar lexical items with similar meanings across two or more languages. As the subset of cognates between any two languages differs, they can serve as a readily identifiable aspect of relative complexity. If the learning of a set of lexical items in the L2 is facilitated by their status as loanwords in the L1, this raises concerns with our understanding of the construct of lexical sophistication and the generalizability of word difficulty studies.

2.5. Modeling lexical sophistication

In the last decade researchers have made attempts to model multiple measures of lexical sophistication and word difficulty, or vocabulary knowledge. Willis and Ohashi (2012) found that cognateness, frequency, and phonemic word length best predicted Japanese L1 vocabulary knowledge on the Vocabulary Size Test (VST), accounting for almost 50% of the variance in a multiple linear regression model. Cognateness was the most significant contributor to the model, followed by the log of frequency distribution, and finally by word length in phonemes. The results here are interesting considering Beglar's (2010) observation that some of the frequency bands on the VST were not clearly distinguished in terms of difficulty due to the presence of English to Japanese loanwords. In other words, despite their lower frequency, certain words on the VST were easier, in terms of Rasch logit scores, because they were known to the test takers as loanwords in their L1.

In validation studies done for the release of both versions of TAALES, researchers established that a collection of lexical sophistication measures including range and frequency, as well as familiarity and meaningfulness explained 51.7% (Kyle & Crossley, 2015) and 58% (Kyle et al. 2018) of the variance in holistic scores of vocabulary

proficiency in speaking and writing tasks, respectively. In contrast to the Willis and Ohashi (2012) investigation, which was conducted with participants from a single L1, Japanese, the TAALES validation studies were done with a corpus of speakers and writers drawn from different L1 backgrounds, including data from the TOEFL public use data set (Kyle & Crossley, 2015; 2016; Kyle et al., 2018).

More recent attempts to model lexical sophistication measures and vocabulary knowledge draw on data from non-monolingual L1 sources. Hashimoto and Egbert (2019) argue that there is little empirical evidence to show that frequency is the primary variable affecting vocabulary difficulty and the order of acquisition. They hoped to demonstrate which variables correlated with word knowledge on a yes/no vocabulary test given to speakers from 36 different L1 backgrounds. A best subsets regression model with nine final variables revealed that ranked frequency on the Corpus of Contemporary American English (COCA) (Davies, 2008) was the strongest predictor in the model and two measures of range, BNC fiction and COCA, were also included in their model. Three other lexical sophistication predictors: contextual distinctiveness, number of senses (polysemy), and the number of words in orthographic neighborhood together explained 37.2% of word difficulty.

Despite the advantage of having reliable access to a wide array of predictor variables, there is some concern that removing variables through stepwise and/or subset regression fails to prioritize theory in model constructions. In a conceptual replication of Hashimoto and Egbert (2019), a theoretically derived model of lexical sophistication explained 52% of variance in a yes/no word knowledge exam given to learners from two different L1 backgrounds: Japanese and Arabic (Vitta et al., 2023). Of this 52%, Pratt product measurements estimated that two non-frequency lexical sophistication variables, word naming reaction time and age of acquisition, both underpinned by L1 speaker norms, predicted approximately 34% of the variance. Despite the methodological differences between the Hashimoto and Egbert (2019) and Vitta et al. (2023) studies, there is agreement that lexical sophistication involves more than frequency. The Vitta et al. (2023) study, however, like the Hashimoto and Egbert (2019) paper, does not include cognateness, or differentiate between relative and absolute complexity. Although this is a necessary part of the research design to accommodate speakers of multiple first languages, the lack of discussion of relative complexity makes these conclusions difficult to generalize to speakers of a single L1. To date, researchers appear not to have explicitly addressed the role of absolute and relative complexity in studies of word difficulty. Studies that have looked at word difficulty data from multiple L1 sources have left out relative complexity variables such as loanwords or cognates (e.g., Hashimoto & Egbert, 2019; Vitta et al., 2023). Research that has included relative complexity variables has been small in scale and has not highlighted the distinction between relative

and absolute complexity (e.g., Laufer & McLean, 2016; Willis & Ohashi, 2012). Work that has been done on word difficulty and relative complexity variables has focused on related languages (e.g., De Wilde et al., 2020; De Wilde, 2023). To address this gap, the following research questions are explored in the present study:

1. To what extent does absolute multidimensional lexical sophistication predict word difficulty among Japanese EFL learners?
2. To what extent does relative multidimensional lexical sophistication predict word difficulty among Japanese EFL learners?

3. Methods

To answer the research questions for this study, vocabulary item difficulty scores were calculated from the VST (Nation & Beglar, 2007). Lexical sophistication indices were correlated with item difficulty scores in logits. Finally, a multiple regression model was constructed with lexical sophistication variables that were found to correlate with the dependent variable at $r \geq .30$, $p < .050$.

3.1. Participants

VST scores were taken from an existing data set consisting of 3,449 Japanese first-, second-, and third-year university students in Western Japan (McLean et al., 2014) from 25 universities. Only data from those explicitly listing Japanese as their L1 were used in the study. 113 (3.3%) listed a language other than Japanese as their first language, and 337 (9.8%) of the participants did not list their first language, leaving a total of 2,999 (86.9%) L1 Japanese participants (for more details on data collection, see McLean et al., 2014). Corresponding to recent calls for multisite research in ISLA research in general (Vitta & Al-Hoorie, 2021) and vocabulary research in particular (Vitta et al., 2022), a multisite dataset was selected, which has enhanced external validity over single site samples (Moranski & Ziegler, 2021).

3.2. Instruments and operationalization

3.2.1. VST and word difficulty

The VST (Nation & Beglar, 2007) is a vocabulary knowledge test of the first 8,000 most common words on the BNC. Ten words are sampled from each band of

1,000 words for a total of 80 words on the test. The test is a meaning-recognition multiple-choice test with the target item given in an English sentence. The item sentences were designed not to provide any semantic information about the word and were intended only as a guide to its part of speech. The correct description of the word is chosen from four possible answers. The VST has been shown to display a high degree of psychometric unidimensionality with high construct validity (Beglar, 2010). Item difficulty was operationalized through a dichotomous Rasch analysis of the VST scores in the McLean et al. (2014) data set.

As an existing data set was selected for this study, G*Power (Faul et al., 2007) was used to conduct sensitivity power analyses (for such use in L2 vocabulary research, see Vitta et al., 2022). Assuming two-tailed correlation testing with conventional Type I ($\alpha = .05$) and Type II ($\beta = .20$) thresholds, the sample of 80 target words was just large enough to detect $r \geq .30$, the a priori threshold for practical significance (referencing Hashimoto & Egbert, 2019; Vitta et al., 2023) and thus suitable for the intended bivariate screening process. Assuming the same thresholds and eight predictors (referencing the final multiple regression models of Hashimoto & Egbert, 2019 and Vitta et al., 2023), a sensitivity power analysis determined that the data set was large enough to detect $R^2 \geq .18$ which was acceptable given the smallest effect in Hashimoto and Egbert (2019) was larger, $R^2 = .24$.

3.2.2. Evidence of psychometric suitability for the use of VST data in current study

Rasch analysis was conducted in Winsteps Version 4.4.7 on the VST scores to determine item reliability and fit statistics. Rasch person reliability (separation) was .89 (2.89) and item reliability (separation) was 1.00 (26.66). Due to the large sample size, model fit was determined by infit mean-square values rather than standardized *z* scores (Linacre, 2002). Item infit mean square scores for all items were found to be between 0.77 (*vocabulary*) and 1.21 (*restore*). These scores are acceptable for a "run of the mill" multiple choice test (Wright & Linacre, 1994) and confirmed the VST as a valid representation of the comparative difficulty of the 80 words on the test.

3.2.3. Lexical sophistication variables

Lexical sophistication indices were calculated for the 80 words tested on the VST. These predictor variables were derived from Vitta et al. (2023) and Hashimoto and Egbert (2019), with some key differences, summarized in Appendix A, together with a key to the variable codes.

3.2.4. Range and frequency

The primary measures of range and frequency were taken from the COCA (Davies, 2008). These were used because it has been argued that the corpus texts chosen for L2 research should "make sense" regarding the population under study (Pinchbeck et al., 2022, p. 4). Given that the COCA corpus comprises sources from American English, it was thought to best represent the language to which learners had been exposed. In keeping with previous research in this area (Eguchi & Kyle, 2020; Kim & Crossley, 2018; Kim et al., 2018; Stewart et al., 2022; Vitta et al., 2023), all range and frequency measures were log transformed. As raw frequency measures of word frequency follow Zipfian distributions, log-transformed measures of these indices are better suited to linear regression analysis. The *z* scores of four COCA measurements for frequency and four measurements for range were aggregated, again in keeping with recommendations in Vitta et al. (2023). All variables were then screened for correlation with the dependent variable, discussed in section 3.2.16 dealing with correlational analysis.

3.2.5. Loanword status

The Japanese loanword status of the 80 words on the VST was established in two ways. First, the list of the first 70 words of the VST and their status as loanwords in Japanese were derived from Daulton's (2007) monograph on English cognates in Japanese. The remaining English VST words, representing the final ten items of the VST, were checked manually on three online Japanese dictionaries (https://dictionary.goo.ne.jp/; http://www.kotoba.ne.jp/; https://ejje.weblio.jp/). If two out of the three dictionaries returned a *katakana* equivalent for any of the VST words, that word was coded as having loanword status. Following Willis and Ohashi (2012), cognate status was coded as a binary variable.

A more detailed understanding of the degree of cognateness of a word would require research into the degree of semantic overlap between the words in English and their *katakana* counterparts in Japanese. This has been explored using bilingual perceptions of word meaning. Coupled with the degree of perceived semantic overlap, Allen and Conklin (2013) compiled bilingual perceptions of phonetic overlap. However, as only 198 words were rated in the Allen and Conklin (2013) paper, the data were not able to be applied in the present study. 26 (32.5%) of the 80 VST words were determined to be Japanese loanwords, in common katakana usage. A list of these words can be found in Appendix B.

3.2.6. Psycholinguistic variables

Concreteness was a predictor variable in Vitta et al. (2023), but was not included in Hashimoto and Egbert (2019). Indices for all concreteness ratings were taken directly from the data provided in Brysbaert et al. (2013). As in Vitta et al. (2023) and in contrast to Hashimoto and Egbert (2019), age of acquisition was included as a predictor variable in the present study and indexed through TAALES. The age of acquisition measurement for one word, *demography*, was not available, and the index for *demographic* was used in its place. TAALES did not return an index for the word "yoghurt," and the index for the American spelling of the word, *yogurt*, was used in its place. The word *bloc* was not indexed by TAALES. Using SPSS ver. 28, a missing value analysis was undertaken assuming a regression and normal distribution approach and the value for the word *bloc* was estimated.

3.2.7. Word length

Word length indices, including number of letters, number of phonemes, and number of syllables were derived from the MRC Psycholinguistics Database. None passed screening, as discussed below in section 3.2.16. (number of letters [NLET]: $r = .215$, $p = .056$, number of phonemes [NPHON]: $r = .17$, $p = .134$, number of syllables [NSYL]: $r = .24$, $p = .036$).

3.2.8. Semantic network

Due to technical issues with TAALES Version 2.0.3, polysemy indices were derived from TAALES Version 2.8.1 (beta). Indices were not included for hypernymy as it applies only to nouns and verbs. More than 10% of the word list consists of adjectives or words that could be construed as adjectives, and therefore only polysemy (content_poly) was included in the correlation screening. It was found to correlate with word difficulty at $r = -.532$, $p < .001$.

3.2.9. Word neighborhood

Word neighborhood is defined as the set of lexical items that can be formed by switching out one phoneme or letter from a reference word (Adelman & Brown, 2007). In total, 13 indices of word neighborhood (Freq_N, Freq_N_P, Freq_N_PH, OG_N, OG_N_H, OLD, OLDF, Ortho_N, PLD, PLDF, Phono_N, Phono_N_H) were operationalized in TAALES. None passed the bivariate correlation screening process.

3.2.10. Contextual distinctiveness

Two measures of contextual distinctiveness, semantic diversity (Sem_D), and McDonald co-occurrence probability (McD_CD), were operationalized through TAALES. Neither passed bivariate screening with the dependent variable (Sem_D: r = -.24, p = .031, McD_CD: r = -.10, p = .396).

3.2.11. Age of acquisition

Age of acquisition (Kup_AoA) is a measure of the age at which a native speaker of a language is thought to acquire a given word (Kuperman et al., 2012). It has been researched in L1 word naming and lexical decision paradigms and shown to influence these variables independent of a word's frequency (Morrison & Ellis, 1995). AoA was operationalized through TAALES and correlated with item difficulty (r = .44, p < .001).

3.2.12. Word recognition

Another psycholinguistic variable drawn from L1 research is word naming latencies. Following Vitta et al. (2023), and in the interests of succinctness, only two variables were operationalized through TAALES: word naming response accuracy (WN_Mean_Accuracy), and word naming response time (WN_Mean_RT). Neither variable was correlated significantly with word difficulty.

3.2.13. Concreteness

Concreteness measures derived from Brysbaert et al. (2013) were not available for 3 of the 80 words on the VST. Without imputing missing values, correlation with logit scores were low (r = -.10) and the variable was eliminated from consideration for the regression model.

3.2.14. Valence, arousal, and dominance

Indices of valence, arousal, and dominance were derived from Mohammed (2018). Only 76 of the 80 words contained on the VST were available. Without imputing missing values, the low correlation between all three indices and logit

scores (valence: r = -.12; arousal: r = .05, and dominance: r = -.14) eliminated these variables from consideration for the regression model.

3.2.15. Prevalence

Values for prevalence were derived from Brysbaert et al. (2021). The prevalence scores were log-transformed. The prevalence score for one word on the VST, *null*,was imputed assuming a normal distribution. Prevalence scores were found to correlate with the logit scores at r = .59, and this variable was included in the regression analysis.

3.2.16. Correlation analyses

Correlation analyses with pairwise deletion for missing data were run in JASP between the 34 lexical sophistication variables and the item difficulty logits derived from a Rasch analysis. As in Vitta et al. (2023), only variables that correlated with the item difficulty logits at $r \geq .30$, $p < .050$ were considered for the final model. This left six remaining variables: loanword status, prevalence, age of acquisition, polysemy, and the aggregated log-transformed scores of COCA frequency and range.

Variables found to correlate with the dependent variable were then checked for bivariate collinearity. The threshold for being removed from consideration for the final linear regression model was set at r = .90, following Hashimoto and Egbert (2019). None of the six predictor variables exceeded the bivariate collinearity threshold. The correlation matrix of variables retained for the regression can be seen in Table 1. A complete correlation table for all variables considered, as well as the data necessary to recreate the regression models can be found at: https://ur0.jp/4Lwjz.

Table 1 Correlation matrix of lexical sophistication indices and word difficulty in item logits

Variable	Word difficulty	Prevalence	Cognateness	Age of acquisition	Polysemy	Frequency
Word difficulty	—					
Prevalence	.59**	—				
Cognateness	-.39**	-.26**	—			
Age of acquisition	.48**	.47**	-.21	—		
Polysemy	-.53**	-.41**	.13	-.44**	—	
Frequency	-.63**	-.64**	.17	-.46**	.67**	—
Range	-.55**	-.38**	.08	-.32*	.57**	.74**

Note. * $p < .05$; ** $p < .01$

3.2.17. Multiple linear regression

The final list of six variables that correlated with item difficulty at $r \geq .30$, $p < .050$ were entered into a multiple linear regression analysis using JASP. The linear regression returned high VIF values for COCA range (2.34) and frequency (4.00), and so the variable with the lower correlation with logit scores, range (r = -.55) was removed. Following the regression analysis, residuals were checked for linearity and homoscedasticity. None violated the assumptions of multiple linear regression.

4. Results

4.1. Regression model

After removing range, the five lexical sophistication variables found to correlate with item difficulty at $r \geq .30$, $p < .05$ were entered into a multiple linear regression model in JASP. The overall regression was statistically significant (R^2 = .55, $F(5, 74)$ = 18.17, $p < .001$). The model showed that three of the predictors were statistically significant: cognateness ($ß$ = -.24, p = .005), prevalence ($ß$ = .22, p = .040) and frequency ($ß$ = -.28, p = .029). The other two predictors in the model: polysemy (content_poly) ($ß$ = -.17, p = .125) and age of acquisition ($ß$ = .13, p = .177) did not significantly predict logit score. The final regression model was: word difficulty = -.25 - .63*(cognateness) - .04*(polysemy) + .05*(age of acquisition) - .38*(frequency) + .51*(prevalence) + error. The model is shown in Table 2. Assumptions for the general linear model (using Vitta et al., 2023 as a model) were met after the removal of range: a – residuals were normally distributed; b – scatterplot between *z*-scored residuals (Y-axis) and predicted values (X-values) randomly and evenly straddled the Y = 0 line suggesting the assumptions of homoscedasticity and linearity were met; c – VIFs were under 2.50 or only marginally above in the case of frequency suggesting that multicollinearity was not a concern, especially as there was no sign switching resulting in negative Pratt values; and d – values for centered leverage and Cook's distance were under 1 suggesting that no case had undue influence.

4.2. Interaction analyses

After independent expert review, we conducted a post-hoc analysis of the interaction between cognateness and all absolute lexical sophistication predictors. The interaction variables displayed moderate and significant associations with the dependent variable, that is, word difficulty. The interaction predictors, however,

were inconsequential in multiple regression modeling (see Appendix C for further details). This finding supports the conclusion that cognateness did not not meaningfully interact with (absolute) lexical sophistication variables when accounting for L2 word difficulty in this current study.

Table 2 Model coefficients of lexical sophistication factors and word difficulty

Model		*B*	*SE B*	ß	*t*	*p*	95% CI		VIF	Pratt
							LL	*UL*		
H_o	(Intercept)	-2.500e -4	.14		-0.002	.999	-0.28	0.28		
H_1	(Intercept)	-2.05	.95		-2.16	.034	-3.95	-0.16		
	Cognateness	-.63	.21	-.24	-2.92	.005	-1.06	-0.20	1.08	9.22
	Prevalence	.51	.24	.22	2.10	.040	0.03	0.99	1.89	13.20
	Age of acquisition	.05	.04	.13	1.36	.177	-0.03	0.13	1.43	6.10
	Polysemy	-.04	.02	-.17	-1.55	.125	-0.08	0.01	1.91	8.88
	Frequency	-.38	.17	-.28	-2.23	.029	-0.72	-0.04	2.59	17.70

Note. Pratt, computed via ß (in the regression model) x *r* (bivariate association with the DV) denotes the amount each predictor contributes to the models R^2 (see use in Vitta et al., 2023).

5. Discussion

In answer to the first research question, the only absolute complexity variables found to significantly contribute to the model were frequency (*ß* = -.28, *p* = .029, accounting for 32.18% of the model's predictive variance [17.70% / 55.00% = 32.18%]) and prevalence (*ß* = .22, *p* = .040, 24.07% of the model's predicted variance). This corroborates evidence that frequency is a primary determinant of the difficulty involved in word learning (e.g., Hashimoto & Egbert, 2019; Vitta et al., 2023; Willis & Ohashi, 2012). Frequency can be regarded as a measure of absolute complexity in that it is ostensibly a measure of the amount of interaction a learner can be expected to have had with the target lexis. This objective measurement is independent of the subjective experiences of learners from disparate L1 backgrounds.

In answer to the second research question, loanword status as a measure of relative complexity played a significant role in predicting word difficulty (*ß* = -.24, *p* = .005, 16.76% of the model's predicted variance). Of other studies in this area, only Willis and Ohashi (2012), De Wilde et al. (2020), and De Wilde (2023) included loanword status as an independent variable, and the results of the present study support their findings. Cognateness as a significant factor has important ramifications for the validation of vocabulary size tests. In studies involving learners from multiple first languages, such as Hashimoto and Egbert (2019) and Vitta et al. (2023), loanword status introduces a confounding complexity variable that differs between populations sampled. Just as importantly, in monolingual L1 contexts, the loanword status of test items should be considered when considering the learning difficulty of words.

The findings of the current study in relation to cognateness raise important questions in relation to sampling. Studies such as De Wilde (2023) and the current study were done with homogeneous L1 populations, which enabled the modeling of L1-driven relative complexity variables such as cognateness and prevalence. On the other hand, other word difficulty studies such as Hashimoto and Egbert (2019) and Vitta et al. (2023) purposefully included learners with different L1s, rendering the modeling of L1-driven relative complexity variables difficult if not impossible due to measurement invariance concerns. Given the significant and substantial contribution of cognateness to the current study's model, future researchers should consider whether heterogeneity regarding L1 is a useful design feature or something that precludes the modeling of useful constructs to understand L2 word difficulty. Put differently, perhaps L2 word difficulty should be addressed controlling for L1 differences.

The findings in the present study are consistent with similar research into word difficulty. This includes those that have shown at least one index of frequency that significantly predicted word difficulty. However, as a metric of absolute complexity, frequency should be considered in conjunction with relative complexity variables. The results suggest a broad approach to studies of lexical complexity that includes explicit consideration of relative and absolute variables. At the theoretical level of Bulté and Housen's (2012) construct specification, systemic lexical complexity can be considered as the breadth of the target language lexis, an absolute measure of the complexity of the system and frequency as an estimated measure of exposure to that system. However, the subset of loanwords in an L1 that overlap with the target language contributes a subjective element of relative complexity to a model of L2 word learning. At an observational level, lexical complexity was manifested in test performance, which was then operationalized in the variables explored in this study and studies like it. The results here suggest that as loanword status is a significant predictor of lexical complexity, it needs to be considered alongside other independent variables. What lexical sophistication variables are of primary importance to word-learning difficulty is still an open question, and how they interact with learner- and context-specific variables is an area for future research. The results of the current study suggest that whatever variables are considered should be evaluated in terms of absolute or relative complexity.

This study is in line with the trend exemplified in De Wilde (2023), which included a relative complexity (more specifically, lexical sophistication) predictor alongside a suite of absolute complexity (lexical sophistication) predictors. The current study's findings, furthermore, add new insights to this discussion. First, relative complexity appears to have little association with absolute complexity predictors given the very low VIF observed, 1.08, which implies that the other

predictors together only accounted for 7.80% of cognateness's variance. Second, relative complexity complements absolute complexity as opposed to completely overtaking or replacing it in accounting for word difficulty. It is noteworthy that more than 80% of the variance predicted by the model, as summarized by the reported Pratt values reported in Table 2 where each value ascribes a discrete amount of the variance a predictor contributes to the model's R^2, can be attributed to absolute complexity variables. The inclusion of a relative complexity variable does not disturb the model's satisfying the assumptions of the general linear equation.

Researchers in vocabulary testing have long acknowledged that frequency is not the only factor in word difficulty (Beglar, 2010; Nation & Beglar, 2007). Stewart et al. (2022) have argued that vocabulary size tests are organized around frequency to facilitate efficient learning of the words that are most useful in English: namely, those that are the most frequent. It has recently been argued that drawing from knowledge-based vocabulary lists would better target the abilities of the test-takers (Schmitt et al., 2021). This task might be made more efficient still by first identifying which words are loanwords for the target population and thereby partly accounting for the relative difficulty of the items on the test.

6. Limitations and future research

There are limitations to this study. First, psycholinguistic variables were underrepresented in the correlation phase of the study. Psycholinguistic variables were found to correlate or predict word-learning difficulty and lexical sophistication in De Wilde et al. (2020), De Wilde (2023), and Vitta et al. (2023). Imageability and meaningfulness were not included in the present study as indices were not available for all items on the VST. It would be of interest to know what the inclusion of a broader range of psycholinguistic variables in a study of this kind would yield. Additionally, only one relative complexity variable, loanword status or cognateness, was included for analysis. In keeping with calls to define complexity more explicitly (Bulté & Housen, 2012), future research can be done to identify other relative complexity variables that are thought to contribute to lexical sophistication and to include them in studies of this type. One other limitation is that the research was carried out on an existing dataset. Future research into word difficulty should rely on carefully chosen sample populations, which would allow researchers to fully account for relative complexity factors. A final notable limitation is the coding of cognateness as a binary variable. Future research in this area might focus on the perceived degree of semantic overlap between Japanese loanwords and their English counterparts as in Allen and Conklin (2013), coupled with measurements edit or Levenshtein or edit distance. Research of this type on languages that share scripts is common (e.g., Dijkstra et al., 1999;

Schepens et al., 2011). However, given the difference in orthographic systems between Japanese and English, it may be necessary to analyze the weighted phonetic edit distance between the lexical items while also controlling for word length. Work in this area includes Kondrak (2000, 2003) and applying research of this type to English loanwords in Japanese is likely to yield valuable insights.

In further regard to future research, the response variable in the present study was calculated from VST scores, a test constructed from a frequency-based word list. Schmitt et al. (2021) have shown that knowledge-based word lists are not strongly associated with frequency. It would be instructive to compare the proportional contributions of relative and absolute complexity variables to knowledge-based word lists and frequency-based lists. Furthermore, the number of items on the VST is limited. Brysbaert (2019) has shown that to observe an effect size of d = .40 which equates to r = .2, the minimum number of participants, or in this case, vocabulary items, is 200. Research done on future tests can also control for the absolute and relative complexity factors affecting distractors, which was not done in the current study.

7. Conclusion and pedagogical implications

Findings from this study have corroborated research into the complex nature of lexical sophistication and show that frequency alone does not determine the difficulty of a word. Several categories of lexical sophistication can be demonstrated to correlate highly with word learning difficulty. Additionally, this study has shown that loanword status, as a measure of relative complexity, significantly predicts word difficulty. This has important implications for how vocabulary is tested and taught. Although vocabulary size tests such as the VST are structured around frequency bands, these bands do not necessarily reflect the relative difficulty of all the words contained within it.

These findings suggest two important pedagogical implications. Testing and teaching protocols would do well to consider the relative complexity factors when selecting or sequencing word lists for instruction. Words that are cognates in the learners' L1 may require less attention than other words of greater or similar frequency. When deriving word lists for instruction, cognates or loanwords require special attention. Those in the L1 that have high semantic overlap with the target language may have facilitative effects on acquisition and vice versa. The takeaway is that the relationship between learner prior knowledge and the target language must be accounted for. Indeed, there have been recent calls to structure vocabulary learning around knowledge-based vocabulary lists (Schmitt et al., 2021). This would enable curriculum designers and teaching practitioners to account

for relative factors affecting vocabulary learning. A second implication for the classroom involves assessment. This study has shown that cognates or loanwords are a key nexus of interaction between languages and therefore vocabulary tests should control for their effect. In all learning contexts, this involves a consideration of the relationship between languages. Closely related languages that share orthographies may require different testing protocols than more distantly related languages. Furthermore, the outcome of vocabulary tests that are administered to learner groups of variegated first languages are likely affected by the relative differences between the L1s and the target language, including the influence of loanwords or cognates. In short, there is a clear indication that vocabulary lists and tests should be crafted with an understanding of what knowledge the learners do or do not bring to the classroom.

References

Adelman, J. S., & Brown, G. D. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*(3), 455-459. https://doi.org/10.3758/BF03194088

Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education*, *19*(3). https://doi.org/10.1080/10691898.2011.11889620

Allen, D., & Conklin, K. (2013). Cross-linguistic similarity norms for Japanese-English translation equivalents. *Behavior Research Methods*, *46*(2), 540-563. https://doi.org/10.3758/s13428-013-0389-z

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118. https://doi.org/10.1177/0265532209340194

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911. https://doi.org/10.3758/s13428-013-0403-5

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*. *2*(1), 1-38. https://doi.org/10.5334/joc.72

Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, 37(2), 207-231. https://doi.org/10.1177/0267658320934526

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21-46). John Benjamins.

Christian, J., Bickley, W., Tarka, M., & Clayton, K. (1978). Measures of free recall of 900 English nouns: Correlations with imagery, concreteness, meaningfulness, and frequency. *Memory & Cognition*, *6*(4), 379-390. https://doi.org/10.3758/BF03197470

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition*, *12*(6), 590-606. https://doi.org/10.3758/BF03213348

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*(4), 497-505.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*(1), 182-193. https://doi.org/10.5054/tq.2010.244019

Crossley, S., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *Modern Language Journal*, *100*(3), 702-715. https://doi.org/10.1111/modl.12344

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity.* John Benjamins.

Daulton, F. E. (1998). Japanese loanword cognates and the acquisition of English vocabulary. *The Language Teacher, 22*(1), 17-25.

Daulton, F. E. (2007). *Japan's built-in lexicon of English-based loanwords*. Multilingual Matters.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. https://corpus.byu.edu/coca/

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure: How do word-related variables and proficiency influence receptive vocabulary learning? *Language Learning*, *70*(2), 349-381. https://doi.org/10.1111/lang.12380

De Wilde, V. (2023). The auditory picture vocabulary test for English L2: A spoken receptive meaning-recognition test intended for Dutch-speaking L2 learners of English. *Language Teaching Research*. https://doi.org/10.1177/13621688221147462

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, *41*(4), 496-518. https://doi.org/10.1006/jmla.1999.2654

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143-188. https://doi.org/10.1017/S0272263102002140

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*(4), 559-617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *Modern Language Journal*, *104*(2), 381-400. https://doi.org/10.1111/modl.12637

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Gries, S. T. (2020). Analyzing dispersion. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 99-118). Springer. https://doi.org/10.1007/978-3-030-46216-1_5

Hashimoto, B. J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*(2), 171-187. https://doi.org/10.1080/15434303.2020.1860058

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*(4), 839-872. https://doi.org/10.1111/lang.12353

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718-730. https://doi.org/10.3758/s13428-012-0278-x

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal*, *102*(1), 120-141. https://doi.org/10.1111/modl.12447

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 288-295.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities, 37*, 273-291. https://doi.org/10.1023/A:1025071200644

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

Laufer, B. (1989). A factor of difficulty in vocabulary learning: Deceptive transparency. In I. S. P. Nation & R. Carter (Eds.), *Vocabulary acquisition* (pp. 10-20). Free University Press.

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*(3), 202-217.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), p. 878.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*(2), 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322. https://doi.org/10.1177/00238309010440030101

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 47-55. https://doi.org/10.7820/vli.v03.2.mclean.et.al

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174-184. https://doi.org/10.18653/v1/p18-1017

Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, *71*(1), 204-242. https://doi.org/10.1111/lang.12434

Morrison, C., & Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 116-133.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

NLP tools for the social sciences. (2016). *TAALES 2.0 index description spreadsheet*. https://docs.google.com/spreadsheets/d/1axmeHlKE-aelPHX4L17WpHjC7Jn4yQlE/edit#gid=858394526

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134. https://doi.org/10.1177/0267658314536435

Pinchbeck, G. G., Brown, D., McLean, S., & Kramer, B. (2022). Validating word lists that represent learner knowledge in EFL contexts: The impact of the definition of word and the choice of source corpora. *System*, *106*, 1-14. https://doi.org/10.1016/j.system.2022.102771

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp.125-142). Routledge.

Schepens, J., Dijkstra, T., & Grootjen, F. (2011). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157-166. https://doi.org/10.1017/s1366728910000623

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*(2), 281-317. https://doi.org/10.1111/1467-9922.00042

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, *12*(4), e622. https://doi.org/10.1002/tesj.622

Siskova, Z. (2012). Lexical richness in EFL students' narratives. *University of Reading Language Studies Working Papers*, *4*, 26-36.

Stewart, J., Vitta, J. P., Nicklin, C., McLean, S., Pinchbeck, G. G., & Kramer, B. (2022). The relationship between word difficulty and frequency: A response to Hashimoto (2021). *Language Assessment Quarterly*, *19*(1), 90-101. https://doi.org/10.1080/15434303.2021.1992629

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96-116. https://doi.org/10.1111/j.1467-9582.2010.01176.x

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Lawrence Erlbaum.

Vitta, J. P., & Al-Hoorie, A. (2021). Measurement and sampling recommendations for L2 flipped learning experiments: A bottom-up methodological synthesis. *Journal of Asia TEFL*, *18*(2), 682-692. https://doi.org/10.18823/asiatefl.2021.18.2.23.682

Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, *44*(5), 1424-1448. https://doi.org/10.1017/s0272263121000541

Vitta, J. P., Nicklin, C., & Albright, S. W. (2023). Academic word difficulty and multidimensional lexical sophistication: An English-for-academic-purposes-focused conceptual replication of Hashimoto and Egbert (2019). *Modern Language Journal*, *107*(1), 373-397. https://doi.org/10.1111/modl.12835

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal, 40*(1), 125-137. https://doi.org/10.1080/09571736.2012.658232

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), p. 370.

APPENDIX A

List of variables used in the study

Variable	Variable Code	In-text name
logit score	logit_score	Word difficulty
Loan word status	COGNT	Cognateness
Valence	VAL	
Arousal	AROU	
Dominance	DOM	
Concreteness	CONC	
log transformation of prevalence score	PREV_log	Prevalence
orthographic neighborhood frequency	Freq_N	
PhongrapPhonographic Neighborhood Frequency Logarithm (homophones included)	Freq_N_OG	
Phonographic Neighborhood Frequency Logarithm (homophones excluded)	Freq_N_OGH	
Phonological Neighborhood Frequency (homophones included)	Freq_N_P	
Phonological Neighborhood Frequency (homophones excluded)	Freq_N_PH	
Age of acquisition	Kup_AoA	Age of Acquisition
Lexical Decision Accuracy	LD_Mean_Accuracy	
Lexical Decision Time	LD_Mean_RT	
McDonald Co-occurrence Probability AW	McD_CD	
Phonographic Neighbors (homophones excluded)	OG_N	
Phonographic Neighbors (homophones included)	OG_N_H	
Average Levenshtein Distance of closest orthographic neighbors	OLD	
Average log HAL frequency of closest orthographic neighbors	OLDF	
Orthographic Neighbors	Ortho_N	
Average Levenshtein Distance of closest phonological neighbors	PLD	
Average log HAL frequency of closest phonological neighbors	PLDF	
Phonological Neighbors (homonyms excluded)	Phono_N	
Phonological Neighbors (homonyms included)	Phono_N_H	
Hoffman et al. Semantic Distinctiveness CW	Sem_D	
Word Naming Response Accuracy	WN_Mean_Accuracy	
Word Naming Response Time	WN_Mean_RT	
LDA Age of Exposure (.40 cosine threshold)	aoe_index_above_40	
Polysemy (content words)	content_poly	Polysemy
Aggregated COCA frequency measures	COCAfreqZagg	Frequency
Aggregated COCA range measures	COCARangeZagg	Range
Number of syllables	NSYL	
Number of phonemes	NPHN	
Number of characters	NLET	

Note. TAALES 2.0 codes are from NLP tools for the social sciences. (2016). *TAALES 2.0 index description spreadsheet*. TAALES 2.8.1 (beta) codes are taken from NLP tools for the social sciences. (2016). *TAALES 2.2 Index Description Spreadsheet*. Both can be found at: https://docs.google.com/spreadsheets/d/1axmeHlKE-aelPHX4L17WpHjC7Jn4yQlE/edit#gid=858394526

APPENDIX B

Loanword status of VST items

Item #	Item	Loan word status	Item #	Item	Loan word status
1	see	0	41	deficit	0
2	time	1	42	weep	0
3	period	0	43	nun	0
4	figure	0	44	haunt	0
5	poor	0	45	compost	0
6	drive	1	46	cube	1
7	jump	1	47	miniature	1
8	shoe	1	48	peel	0
9	standard	1	49	fracture	0
10	basis	0	50	bacterium	1
11	maintain	0	51	devious	0
12	stone	0	52	premier	1
13	upset	0	53	butler	0
14	drawer	0	54	accessory	0
15	patience	0	55	threshold	0
16	nil	0	56	thesis	0
17	pub	1	57	strangle	0
18	circle	1	58	cavalier	0
19	microphone	0	59	malign	0
20	pro	1	60	veer	0
21	soldier	0	61	olive	1
22	restore	0	62	quilt	0
23	jug	0	63	stealth	0
24	scrub	0	64	shudder	0
25	dinosaur	0	65	bristle	0
26	strap	1	66	bloc	1
27	pave	0	67	demography	0
28	dash	1	68	gimmick	0
29	rove	0	69	azalea	0
30	lonesome	0	70	yoghurt	1
31	compound	0	71	erratic	0
32	latter	0	72	palette	1
33	candid	0	73	null	1
34	tummy	0	74	kindergarten	1
35	quiz	1	75	eclipse	1
36	input	1	76	marrow	0
37	crab	0	77	locust	0
38	vocabulary	1	78	authentic	1
39	remedy	0	79	cabaret	1
40	allege	0	80	mumble	0

Note. Adapted from Daulton (2007).

APPENDIX C

Post-hoc interaction analysis

Because of power concerns, we only tested interactions using the significant predictions in the multiple regression model: Cognateness, Frequency, and Prevalence. This process began by constructing two interaction variables: cognateness x frequency (COGNTxfreq) and cognateness x prevalence (COGNTxprev). According to best practice (Afshartous & Preston, 2011), the continuous variables were centered via z-scoring to avoid multicollinearity in the multivariable model. We then bivariate screened these two interaction variables with the three predictors. Because this was an additional model, we corrected alpha to .025 but retained the practical significance threshold at .3.

Table C1 Interaction effect correlations

	logit_score	COGNT	Zscore (Prev_log)	Zscore(COCAfreqZagg)	COGNTxfreq
Word difficulty	—				
COGNT	-.39**	—			
Zscore (Prev_log)	.59**	-.26*	—		
Zscore(COCAfreqZagg)	-.63**	.17	-.64**	—	
COGNTxfreq	-.41**	.20	-.31**	.58**	—
COGNTxprev	.40**	-.30**	.61**	-.30**	-.50**

**. Correlation is significant at the .01 level (2-tailed).

*. Correlation is significant at the .05 level (2-tailed).

Table C2 Model coefficients – interaction effects

Model		*B*	*SE B*	*ß*	*t*	*p*	95% CI		VIF
							LL	*UL*	
1	(Constant)	-1.629	1.290		-1.263	.211	-4.200	.941	
	Prevalence	.47	.32	.21	1.47	.150	-.17	1.12	3.17
	Cognateness	-.65	.23	-.24	-2.86	.010	-1.10	-.20	1.11
	Frequency	-.59	.19	-.43	-3.12	.003	-.96	-.21	2.97
	COGNTxfreq	-.02	.27	-.01	-.09	.930	-.56	.51	2.30
	COGNTxprev	.13	.27	.06	.49	.630	-.41	.68	2.51

The correlation matrix revealed that the interaction variables were not collinear with the predictors and had practically significant associations with the DV. In the regression model, however, the interaction variables had a null effect and thus there was evidence to reject the meaningful contributions of the interactions between cognateness and frequency with prevalence, respectively.