Studies in Second Language Acquisition **41** (2019), 911–935
doi:10.1017/S0272263119000019

Research Article

USING LEXICAL FEATURES TO INVESTIGATE SECOND LANGUAGE LEXICAL DECISION PERFORMANCE

Cynthia Berger*

Georgia State University

Scott Crossley

Georgia State University

Stephen Skalicky

Georgia State University

Abstract

A large dataset of word recognition behavior from nonnative speakers (NNS) of English was collected using an online crowdsourced lexical decision task. Lexical features were used to predict NNS lexical decision latencies and accuracies. Predictors of NNS latencies and accuracy included contextual diversity, age of acquisition, and contextual distinctiveness, while length moderated the impact of contextual diversity and neighborhood size on accuracy. Results have implications for second language word recognition and demonstrate that NNS behavioral data collected through large crowdsourcing projects can afford a rich source for SLA research.

INTRODUCTION

The ability to recognize a word's visual form and associate it with meaning constitutes a basis for higher-level processing at both the sentence- or discourse-level (Cortese & Balota, 2012; Stanovich, 1991). Accordingly, a number of researchers have examined the features that impact the speed with which words are accurately recognized. With exceptions (e.g., Dijkstra & Van Heuven, 2002; Fender, 2003; Portin, Lehtonen, & Laine, 2007; Segalowitz & Segalowitz, 1993), the majority of such work has relied exclusively on native speakers while significantly less work has examined how word features explain the speed and accuracy with which nonnative speakers (NNS) recognize visual words. Insight into NNS word recognition is important not only for theoretical and

*Correspondence concerning this article should be addressed to Cynthia Berger. E-mail: cindymayberger@gmail.com

Copyright © Cambridge University Press 2019

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

empirical investigations of NNS reading but also for broader inquiry into how words are acquired, accessed, and processed by NNS. A better understanding of NNS word recognition can also help identify which features contribute to the difficulty (or ease) with which NNS learn words.

The purpose of the current study is to examine NNS English word recognition in a large-scale dataset of lexical decision performance measures. We examine the degree to which lexical predictors attested in L1 studies explain NNS word recognition to better understand NNS lexical processing and determine which features are predictive of word recognition. Our study is motivated by the following question: To what degree do lexical variables predict word recognition latencies and accuracy for NNS subjects?

The goal of this project is to examine which factors impact NNS visual word recognition to better understand the intrinsic difficulty of English words for NNS learners. By *intrinsic difficulty*, we refer to the properties of words that are fundamental to their form or meaning that may make them more difficult to learn (cf. Ellis, 1997; Hemchua & Schmitt, 2006).

NNS LEXICAL ACCESS AND WORD RECOGNITION

Psycholinguistic approaches to NNS lexical knowledge investigate the manner in which words are organized and accessed from NNS's mental lexicons, as well as what factors may influence NNS lexical processing[1] (Dijkstra, 2005; Sunderman & Campbell, 2013). One approach to better understanding NNS lexical organization involves the examination of lexical retrieval, or access (Sunderman & Campbell, 2013), during word recognition. Visual word recognition is one of the fundamental lower-level processes of successful lexical decoding (Stanovich, 1991) and includes subprocesses involving recognition and identification of letters, formation of grapheme-phoneme correspondence, and association of words to their conceptual/semantic representations (Segalowitz, Poulsen, & Komoda, 1991). In skilled decoders, these subprocesses are automatized, thus freeing up cognitive resources for more higher-level processing, such as syntactic and discoursal-level processing.

A common technique for exploring word recognition is the lexical decision task (Meyer & Schvaneveldt, 1971), during which subjects are presented with a string of letters and asked to judge as quickly and accurately as possible whether the letters constitute a real word in the target language. Typically, subjects press a designated key or button to indicate their decision. Both the speed and accuracy of subjects' decisions are analyzed, under the presumption that the relative speed of accurate word recognition indexes ease of processing. However, with exceptions (e.g., Cook & Gor, 2015; Cook, Pandža, Lancaster, & Gor, 2016), word recognition in NNS research has received relatively little attention when compared to L1 studies (Koda, 2005). The majority of studies that have investigated NNS word recognition have examined the influence of the L1 on NNS word recognition (e.g., Elston-Güttler, Paulmann, & Kotz, 2005; Fender, 2003) or the bilingual lexical organization (e.g., De Groot, Borgwaldt, Bos, & van den Eijnden, 2002; Dijkstra & Van Heuven, 2002). Of those studies that have investigated within-language (vs. cross-linguistic) NNS word recognition, most have focused on the development of automaticity (e.g., Segalowitz & Segalowitz, 1993; Segalowitz,

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Segalowitz, & Wood, 1998), although at least one focused on a large number of within- and between-linguistic variables to predict NNS performance (Lemhöfer et al., 2008).

VARIABLES KNOWN TO IMPACT VISUAL WORD RECOGNITION BEHAVIOR

While a great deal of research has investigated the type of lexical features that impact L1 visual word recognition in lexical decision tasks, little NNS literature exists on the matter. In general, L1 research demonstrates that features such as word frequency, semantic features, and neighborhood effects explain word recognition behaviors. These effects are discussed in more detail in the following text.

Frequency

Word frequency is the most robust predictor of speed of response and accuracy in L1 word recognition tasks (Balota & Chumbley, 1990; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Chambers, 1973; Whaley, 1978). While word frequency effects are found in essentially any task that involves visual recognition, they are especially strong in lexical decision tasks (Rastle, 2009). An assumption for this effect is that language speakers' experience with a known word is encoded in the activation level of its orthographic representation (McClelland & Rumelhart, 1981). Words that occur more frequently in input are more readily activated than words that occur rarely. Accordingly, it is easier to reach a threshold of activation for frequent words than for less frequently encountered words, which have a higher level of activation (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger & Jacobs, 1996; McClelland & Rumelhart, 1981).

Word frequency has also been found to predict lexical decision (De Groot et al., 2002) and word identification task performance (Lemhöfer et al., 2008) in second language studies, a finding that aligns with evidence of frequency-based learning and word representation during second language acquisition (Ellis, 2002a, 2002b). While there is evidence to suggest that NNS subjects demonstrate a stronger frequency effect in lexical decision tasks than L1 subjects (Duyck, Vanderelst, Desmet, & Hartsuiker, 2008), Diependaele, Lemhöfer, and Brysbaert (2013) have demonstrated that stronger NNS frequency effects can be explained by differences in language-specific proficiency rather than qualitative differences in NNS processing due to language competition. This same argument—the Lexical Quality Hypothesis—has been used to explain differential frequency effects in L1 speakers as well (Perfetti, 2007).

Contextual Diversity

Traditionally, most word frequency counts are corpus-based values built on occurrence frequencies gathered from large reference corpora (e.g., Kučera & Francis, 1967), though spoken corpus-based frequencies are available as well (e.g., Brysbaert & New, 2009b). Because this manner of operationalizing frequency effectively treats language as a randomly ordered collection of words, some researchers have begun to argue that the diversity of contexts in which a word is encountered may offer a more psychologically valid explanation of the word frequency effect (Adelman, Brown, & Quesada, 2006;

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Brysbaert & New, 2009a; McDonald & Shillcock, 2001). For example, Adelman et al. (2006) found that contextual diversity measures (operationalized as the number of documents within a corpus that contain a given word) explained more variance in L1 lexical decision latencies than word frequency measures (the total number of times a given word occurs in a corpus).

Age of Acquisition

Another factor found to impact L1 word recognition performance is age of acquisition (AoA) (Brown & Watson, 1987; Ellis & Morrison, 1998), which refers to the estimated age at which a native speaker of a language first encounters a word. The theoretical assumption underlying the analysis of AoA is that earlier acquired words play a more central role in the lexicon (Steyvers & Tenenbaum, 2005) while also boasting greater cumulative frequency over time (Cortese & Balota, 2012). Past studies have demonstrated strong AoA effects in word recognition tasks independent of frequency (Morrison & Ellis, 1995), and studies have also demonstrated evidence of AoA effects on visual word recognition in languages other than English (e.g., Brysbaert, Lange, & Wijnendaele, 2000). For example, Brysbaert et al. (2000) found a significant effect of AoA on L1 lexical decision performance in L1 Dutch, even when stimuli were controlled for length and imageability.

While the age of learners' initial exposure to a second language is a strong determinant of nativelike acquisition (DeKeyser, 2013; Flege, Yeni-Komshian, & Liu, 1999; Johnson & Newport, 1989), particularly for pronunciation (Saito, 2015), given individual differences in learning experiences, it is difficult to estimate the average AoA for individual NNS words the way researchers have for L1 words and, to our knowledge, no such metrics exist. However, some researchers (e.g., Crossley & McNamara, 2012; Salsbury, Crossley, & McNamara, 2011) have successfully used L1 AoA ratings to investigate NNS lexical acquisition. These studies report that L1 AoA is a strong predictor of NNS acquisition (Salsbury, Crossley, & McNamara, 2011) and that NNS learners produce words with higher AoA than L1 writers (Crossley & McNamara, 2012)

Length

The length of a word has also been found to influence performance in word recognition (Coltheart et al., 2001), with some research suggesting that the impact of length is stronger for low-frequency words than high-frequency words (Balota et al., 2004) and that word length facilitates visual word recognition, but only for words ranging from three to five letters long (New et al., 2006). In a bilingual study, De Groot et al. (2002) found that NNS English lexical decision time—but not L1 Dutch lexical decision time—was significantly influenced by word length in Dutch-English bilingual subjects. The study also reported an interaction between length and frequency for NNSs of English such that subjects responded more slowly to longer words only if they were also less frequent.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Neighborhood Size

Orthographic Neighborhood An orthographic neighbor is formed by changing one letter of a real word to form another real word (Coltheart, Davelaar, Jonasson, & Besner, 1977). For example, the word *dumpster* has zero orthographic neighbors while the word *fire* has eight, including *fare*, *file*, *firm*, and *wire*. While orthographic neighborhood size has been found to facilitate lexical decision performance (Andrews, 1997; Balota et al., 2007), its impact may be modulated by other variables (Andrews, 1992), differences between word processing and word production, or the presence of nonsense or pseudowords[2] as distractors (Johnson & Pugh, 1994). For instance, Balota et al. (2004) found that while orthographic neighborhood size was facilitatory in younger L1 subjects (M = 20.5 years), it had an inhibitory effect for older subjects (M = 73.6 years) and slow-processing young subjects. NNS research analyzing orthographic neighborhood size has found that for Dutch L1 subjects highly proficient in English, increasing orthographic neighborhood density in English had a facilitatory effect on English word recognition but inhibitory effects on Dutch word recognition (Van Heuven et al., 1998).

Phonological Neighborhood While orthographical neighbors differ by *one* letter, phonological neighbors differ by one phoneme. For example, the word *tomb* has 16 phonological neighbors, including *room*, *time*, and *zoom*, while the word *stone* only has three phonological neighbors: *stoat*, *stole*, and *stove.* Previous L1 research suggests that phonological neighborhood size may facilitate lexical decision performance independent of orthographic neighborhood size (Yates, Locker, & Simpson, 2004). Investigations of NNS phonological neighborhood size have typically examined cross-language bilingual word recognition and suggest that orthographic and phonological neighborhood are likely to have a similar impact on bilingual processing Marian and Blumenfeld (2006).

Semantic Characteristics

Previous studies have also demonstrated that the meaning-based characteristics of a word (e.g., concreteness, meaningfulness, imageability) may impact how quickly and accurately a word is recognized (Balota et al., 2004; Balota, Ferraro, & Connor, 1991). In an L1 study on visual word recognition, Yap and Balota (2008) found semantic effects on word recognition even after controlling for other lexical characteristics (e.g., length, neighborhood size). Cortese and Balota (2012) note that the existence of relatively larger semantic effects on lexical decision performance (compared to other word recognition tasks, such as word naming) is not surprising given that the central task in lexical decision is to distinguish meaningful strings of letter from nonmeaningful strings. We discuss a number of these features in greater depth in the text that follows.

Concreteness

Concrete words evoke mental images quickly and can be contrasted with abstract words. For instance, a word like *tree* is more concrete than a highly abstract word like *solidarity*. Previous studies have demonstrated that concrete words are responded to faster in lexical decision tasks than abstract words (Paivio, 1991; Schwanenflugel & Shoben, 1983;

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Schwanenflugel & Stowe, 1991). Concreteness may also have a greater effect on lexical decision latencies for low-frequency than high-frequency words (Carlton, 1975). Concrete words are also translated faster and more accurately by NNS subjects (De Groot, Dannenburg, & Van Hell, 1994).

Hypernymy

The term *hypernym* refers to a word whose meaning subsumes more specific word-members within a semantic hierarchy. Words higher in a hypernymic hierarchy—such as *plant*—are typically considered to be less specific than words found lower on a hypernymic hierarchy—such as *succulent* (Lakoff, 1987). To our knowledge, no L1 study has used hypernym values to investigate lexical decision performance outside of category judgments or semantic priming. In a priming study that included both L1 and NNS subjects, Crossley (2013) found evidence of bidirectional priming (hyponym to hypernym, and vice versa) for L1 subjects but only unidirectional effects (hyponym to hypernym only) for NNS subjects.

Contextual Distinctiveness

More recently, researchers have begun to investigate the degree of semantic clustering between a word and other words in the lexicon (Steyvers & Tenenbaum, 2005) and its impact on visual word recognition. For example, Balota et al. (2004) found that semantic connectivity—operationalized as the number of associations produced in response to a given word in a free association task (Nelson, McEvoy, & Schreiber, 1998)—explained significant variance in L1 lexical decision latencies, with faster reaction times for words with greater connectivity. Semantic connectivity is subsumed within a broader construct referred to as *contextual distinctiveness.* Contextual distinctiveness refers to the constraints made on a word by its frequent lexical, associative, and semantic contexts, and the degree to which such constraints contribute to a word's meaning representation (Berger, Crossley, & Kyle, 2017; Crossley, Subtirelu, & Salsbury, 2013; McDonald & Shillcock, 2001). For example, McDonald and Shillcock found that a word's frequent lexical environment was more predictive of L1 lexical decision performance than occurrence frequency, with words occurring in more constrained (i.e., more distinct) lexical contexts resulting in longer decision latencies.

CURRENT STUDY

Traditionally, researchers who investigated relationships between lexical features and lexical decision tasks manipulated and controlled specific features in small sets of word recognition stimuli. More recently, researchers have begun to forego such small-scale factorial designs to investigate word recognition in favor of larger-scale item-analyses (e.g., Balota et al., 2004; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Spieler & Balota, 1997). In the latter, multiple regression techniques are used to determine the unique variance of lexical features of stimuli on word recognition performance.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

In this study, we examine NNS lexical decision performance (response latencies and accuracies) in visual word recognition for 3,318 words in English. We select a number of empirically attested predictors associated with L1 word recognition response behavior to examine NNS lexical decision behavior. Our goal is to determine which factors impact NNS visual word recognition to better understand the intrinsic difficulty of English words for NNS learners and better understand NNS lexical access. Analyses include the examination of relationships between measures (i.e., lexical measures) and outcomes (i.e., NNS lexical decision response times [RTs] and accuracies). Attested interactions among variables are investigated as well.

METHODS

NNS LEXICAL DECISION RESPONSE DATA

Participants

NNS lexical decision performance measures were collected from 1,315 self-identified NNS speakers of English residing within the United States using the online crowdsourcing platform Amazon Mechanical Turk (AMT). All participants were asked to complete a brief linguistic and demographic questionnaire, which included questions pertaining to their dominant language (the language that they use most often), other languages spoken besides English, confidence in English ability (scale of 1–4), hours per day spent using English, years spent learning English, years spent studying English, months lived in an English-speaking country, gender, and handedness.[3] The results of this questionnaire are reported in Appendix A. However, because the current study was item based and not subject based, participants' demographic and individual difference information could not be included in the data analyses.

AMT workers independently choose from among a list of available tasks on the AMT site to complete in exchange for monetary compensation. Studies of AMT workers have demonstrated that this population is just as reliable as traditional psychology subject-pool populations (Buhrmester, Kwang, & Gosling, 2011), and several attested psycholinguistic phenomena have been replicated with AMT workers using online measurement techniques (Enochson & Culbertson, 2015). However, because AMT workers complete the tasks online and anonymously, researchers must take care to ensure the data collected represents unique individuals who were paying attention during the task. To account for these challenges, we conducted participant management through the website TurkPrime (Litman, Robinson, & Abberbock, 2016), which blocks workers from repeating tasks multiple times and helps ensure their Internet connection originated from within the United States.

Stimuli

For the NNS lexical decision tasks, 3,318 English content words and 3,318 pseudowords were selected from the English Lexicon Project (ELP), a large publicly available behavioral and descriptive dataset collected by Balota et al. (2007).[4] The ELP dataset includes lexical decision performance behavior from 816 L1 subjects with approximately 3,400 responses per subjects. Latencies (i.e., how long it took for a subject to make

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

a decision regarding whether a word was a real word in English) were calculated in response to 80,962 real word and pseudoword stimuli (40,481 each), including both mono- and multisyllabic words. All subjects in the ELP were native speakers of English recruited from participant pools across six different universities in the United States. The 40,481 real words were selected to represent a range of frequencies (based on Kučera-Francis and CELEX norms) while the pseudowords were generated by changing one to two letters in selected real words. For each word in the ELP, mean item latencies, z-scores, and standard deviations (SDs) were calculated using accurate observations across subjects who encountered a particular item. Mean accuracies were calculated for each word as well.

Word selection for this study followed that of the original ELP study except that function words were excluded from the current study because function words convey relatively little meaning compared to content words (Cann, 2000) and were thus hypothesized to offer little predictive power in an analysis based on word representation (McDonald & Shillcock, 2001). In addition, proper nouns were excluded to lessen the impact of cultural knowledge on word recognition. The pseudowords used in this study were randomly selected from the ELP project's pseudowords. Sixty-four sublists were created to ensure that all 6,636 real/pseudowords were evenly distributed across test subjects and that at least 20 observations per word were obtained. Following Balota et al. (2007), all real word stimuli were organized into randomly assigned sublists consisting of adjacent alphabetized words. In this manner, the likelihood that a subject would be presented with words with the same morphemic stem was minimized. Sixty-three of the sublists contained 52 real words and 52 pseudowords each (104 total), while the 64th sublist contained 42 real words and 42 pseudowords (84 total). Analyses of variance (ANOVA) found no significant differences in length between real words and pseudowords in the 64 sublists.

Procedure

Data collection was facilitated using two online services: Qualtrics and Testable. Data from 256 participants (26,564 trials) was first collected using the Qualtrics Reaction Time Engine (QRTE; Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015), a JavaScript plug-in for the online survey service Qualtrics that allows for collection of reaction time data. Due to unanticipated changes in the Qualtrics platform, the authors of the QRTE engine were forced to retire the engine while we were still collecting data. We then collected the remaining 1,059 participants (109,796 trials) using Testable, a self-contained website that uses JavaScript to collect online RT data.

AMT workers first viewed a description of the experiment on the AMT website and decided whether to participate. The instructions specified that only NNS speakers of English were able to participate. Once agreeing, participants were redirected to the experiment website and were presented with the informed consent. Participants who agreed were then asked again to complete a brief linguistic and demographic questionnaire (see Appendix A). After completing these questions, the participants moved on to the lexical decision task.

Each participant completed 104 trials (52 real word, 52 pseudowords) in the online lexical decision task. Following Balota et al. (2007), a fixation point was presented for

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

250 ms prior to each target word. Subjects were instructed to press "Q" if the string of letters was a "nonsense word" and "P" if the string of letters was a real word in English and were told that capitalization of the letters did not matter. Participant reaction time (measured in milliseconds) and accuracy was recorded for each trial. Each of the 64 subexperiments was completed by a minimum of 20 participants, and a total of 136,360 observations were collected overall. On average, participants took 8.5 minutes to complete the entirety of the experiment and were paid $0.80USD for their participation.

Calculation of NNS Item Norms

Following Balota et al. (2007), a two-step outlier identification and removal process was followed. In the first step, 4.780 observations with RTs $\leq$ 200 ms were removed from analysis. Another 1,685 observations with RTs $\geq$ than 3,000 ms were excluded as well. In the second step, all RTs were standardized within subjects and any observation for which the RT was less than three SDs below the mean for a given subject or greater than three SDs above the mean for that subject were removed. One observation was removed that fell less than three SDs below the subject's mean, while 2,361 observations were removed that fell three SDs above the mean for given subjects. Following this two-step outlier process, 127,533 observations remained for analysis.

Next, mean item latencies, standardized item RTs (z-scores), and SDs for every real word were calculated based on accurate observations only (90.44% of the observations after outlier removal). Z-scores were calculated by subtracting individual subjects' mean RT for accurate observations from each raw RT and dividing by their SD. The z-scores are useful because they minimize individual differences in speed or variation during the lexical decision task. The average number of accurate observations per word was also calculated. The resulting real word NNS lexical decision performance measures were used as variables in the current study.

LEXICAL MEASURES

All measures investigated in the current study operationalize variables that have previously demonstrated a consistent influence on L1 and/or NNS lexical decision task performance. In Table 1 and in the following text, we offer further details regarding the specific measures[5] selected.

Word Frequency

A frequency index derived from written texts in the Corpus of Contemporary American English (COCA; Davies, 2009) was used as a measure of written word frequency. COCA contains more than 520 million words of American English collected since 1990. Written genres that comprise the corpus include fiction, magazine, newspaper, and academic genres. A COCA log-transformed measure was assessed in the current study. A frequency measure derived from the SUBTLEXus corpus (Brysbaert & New, 2009b) was also used to determine the spoken frequency of each word. SUBTLEXus frequency scores were derived from a corpus of American film and television subtitles approximating spoken language but excluding highly formal or academic written registers. A

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

TABLE 1. Lexical measures selected as predictors for regression analyses

	Measure name(s)		Explanation	Reference	Corpus
Word frequency	Frequency written (log)		Number of occurrences in written genres (log-transformed)	Davies (2009)	COCA (Davies, 2009)
	Frequency (spoken, log)		Number of occurrences in American films and television shows (log-transformed)	Brysbaert & New (2009)	SUBTLEXus (Brysbaert & New, 2009)
Contextual diversity	Contextual diversity (log)		Number of films or television shows in which word occurs (log-transformed)	Brysbaert & New (2009)	SUBTLEXus (Brysbaert & New, 2009)
Age of acquisition	Age of acquisition		Estimated age at which a word is learned, based on human judgments	Kuperman et al. (2012)	n/a
Length	Length		Number of letters in a word	n/a	n/a
Neighborhood size	Neighborhood size (OLD20)		String edit distance	Yarkoni & Balota (2008)	n/a
	Neighborhood size (orthographic)		Number of orthographic neighbors	Balota et al. (2007)	n/a
	Neighborhood size (phonological)		Number of phonological neighbors		
Semantic characteristics	Concreteness		Average concreteness rating, based on human judgments	Brysbaert, Warriner, & Kuperman (2014)	n/a
	Hypernymy		Number of hyponyms (more specific categories) a word subsumes	Fellbaum (1998)	n/a
	Contextual distinctiveness	Word associations, USF	University of South Florida stimuli counts: ease with which target comes to mind in word association tasks (target word = response)	Nelson, McEvoy, & Schreiber (1998)	n/a
		Word associations, EAT	Edinburgh Association Thesaurus response type counts: number of words elicited by target in a word association task (target word = cue)	Kiss et al. (1973)	n/a
		Variety of semantic contexts	Semantic ambiguity: Semantic variability of contexts in which word occurs	Hoffman, Ralph, & Rogers (2013)	Written British National Corpus (2007)
		Variety of lexical contexts	Amount of info conveyed about word by its context of use measure reports relative entropy)	McDonald & Shillcock (2001)	Spoken British National Corpus (2007)

word with a relatively higher frequency value is assumed to occur more often in language usage. A measure of log-transformed frequency was used in this study.

Contextual Diversity

A range score derived from the SUBTLEXus corpus (Brysbaert & New, 2009b) was used to measure contextual diversity. This measure reports the number of individual texts (in this case, film or television shows) in which a word occurs in the corpus. A high contextual diversity value means that a word occurs in a greater variety of usage contexts, while a low value is indicative of a word that is relatively less contextually diverse. A log-transformed contextual diversity was used in this study.

Age of Acquisition

We analyzed average L1 AoA ratings collected by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). These AoA scores are derived from human judgments compiled using web-based crowdsourcing. A high AoA value is reflective of a word that is typically acquired later in life than a word with a lower AoA value.

Length

The number of letters per word was counted to measure length due to the strength of orthographic length in predicting lexical decision latencies (Coltheart et al., 2001) independent of frequency (New et al., 2006).[6]

Neighborhood Size

We used measures of both orthographic neighborhood and phonological (excluding homophones) neighborhood size obtained from the ELP (Balota et al., 2007). A higher neighborhood size value means that a word has more orthographical or phonological neighbors. In addition to a classic measure of neighborhood size (the number of words derived by changing one letter in the target word at a time), we also analyzed a computer science metric based on string edit distance: orthographic Levenshtein distance 20, or OLD20 (Yarkoni & Balota, 2008).

Semantic Characteristics

Concreteness We used a measure reporting the average human judgments of a word's concreteness. The original judgments were collected by Brysbaert, Warriner, and Kuperman (2014) using web-based crowdsourcing. A word with a higher value on the selected concreteness measure can be considered to have a more concrete referent than a word with a lower value.

Hypernymy The hypernymy measure assessed in this study uses WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) to determine the mean value of noun hypernymy in each text of the learner corpus. Words scoring lower on the

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

hypernymy scale are less specific, while words receiving a higher score on the scale are considered to be more specific.

Contextual Distinctiveness Four measures were selected to operationalize the contextual distinctiveness of words. One way of characterizing contextual distinctiveness is through the number of associations a word has, that is, its degree of semantic connectivity (Balota et al., 2004). Two databases were used to quantify a word's associations: the Edinburgh Associative Thesaurus (EAT) (Kiss, Armstrong, Milroy, & Piper, 1973), which reports the number of word types produced in response to a given stimuli in a free association task; and the University of South Florida (USF) norms (Nelson et al., 1998), which work backward by reporting the number of stimuli words that resulted in production of the target word as an associate in free association tasks. EAT values measures the number of associates a word elicits, with a higher number indicating a greater number of semantic associates, while USF values may be interpreted as the ease with which a word comes to mind in response to a variety of different cues (Nelson et al., 1998).

A third measure was selected to operationalize a word's contextual distinctiveness based on the variability of contexts in which it occurs. This computer-derived latent semantic analysis (LSA) measure, originally calculated by Hoffman, Ralph, and Rogers (2013), reports the semantic similarities of 1,000-word contexts in which a word occurs in the written British National Corpus (British National Corpus, 2007). *Contextual distinctiveness (variety of semantic contexts)* quantifies the degree to which the contexts associated with a given word vary in meaning. A word with a high value occurs in a greater variety of semantic contexts than a word with a lower value.

Finally, a relative entropy measure originally calculated by McDonald and Shillcock (2001) was used to reflect the frequency distribution of the immediate lexical contexts in which a word occurs. This computational measure, derived from the spoken British National Corpus (British National Corpus, 2007) and referred to here as *contextual distinctiveness (variety of lexical contexts)*, reports a higher value for words that co-occur statistically in the company of relatively few words, and a lower value for words that occur in a variety of lexical contexts.

STATISTICAL ANALYSIS

NNS RTs and accuracies for the lexical decision tasks were used as dependent (i.e., outcome) variables in separate multiple regressions. Models 1 and 2 used the lexical measures described previously as independent variables to predict NNS response latencies and accuracies, respectively.[7] Because frequency effects have been demonstrated to be moderated by length in past analyses of word recognition reaction times (Balota et al., 2004), interactions between length and frequency were included in both models. In addition, interactions between length and neighborhood size were investigated, under the presumption that longer words would have fewer orthographic or phonological neighbors.

Prior to each regression, correlations among the independent variables were checked for instances of multicollinearity. If any two measures were highly collinear ($r > .80$), only the measure with the strongest relationship to the response behavior was retained. Remaining measures were standardized to reduce multicollinearity and associated

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

problems caused by the inclusion of interaction terms. Prior to final regressions, an initial model was obtained to determine significant interactions. All remaining measures and significant interaction terms were then entered into a subsequent model to accurately obtain results. Any predictor demonstrating suppression was excluded from entry in the final regression. The resulting models are reported in the following text.

RESULTS

DESCRIPTIVE STATISTICS

Table 2 lists the mean and SD for the NNS lexical decision performance variables. Visual inspection of histograms indicated that values were relatively normally distributed (with slight evidence of positive skewness), with the exception of accuracy, which was nonnormally distributed. Table 3 reports descriptive statistics for predictor variables and correlations between the selected variables and the RTs and accuracy data.

NNS STANDARDIZED ITEM RESPONSE TIMES

Ten variables and two interactions were manually entered in a multiple regression as independent variables with NNS standardized RTs as the dependent variable. Three predictors (hypernymy, length, and neighborhood size [orthographic]) demonstrated suppression effects and were excluded from further analysis. The two interactions did not demonstrate significance and were excluded from additional analysis as well. The seven remaining predictors were entered in a subsequent multiple regression. A significant regression equation for Model 1 was found (F (7, 3247) = 223.000, $p < .001$), with an adjusted R^2 of .323. Five variables contributed significantly to Model 1 ($p < .05$) (see Table 4).

Model 1 findings indicate that contextual diversity, AoA, and three operationalizations of contextual distinctiveness (two word association measures and a measure reporting the distinctiveness of lexical contexts) explained roughly 32% of the variance in average item RTs in the NNS lexical decision task. The coefficients for these measures (Table 6) indicate the size of each variable's effect on NNS RT z-scores as well as the direction of the effect. The findings indicate that NNS subjects more quickly recognized English words that occur in a variety of contexts (i.e., contextual diversity), have more associations in native-speaker word association tasks, and are acquired earlier.

TABLE 2. Descriptive statistics for lexical decision performance measures

Performance measure	Mean	SD
NNS mean response times	734.159	82.823
NNS Mean response times (z-scores)	–0.252	0.305
NNS Mean response times standard deviations	276.497	112.211
NNS mean accuracy	0.940	0.078

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

TABLE 3. Descriptive statistics and correlations for initial lexical measures

Measure	Mean	St. Dev.	Response Times (z-scores)	Accuracy
Frequency (written, log)	3.363	1.369	–0.516 ***	0.377 ***
Frequency (spoken, log)	2.977	0.694	–0.516 ***	0.332 ***
Contextual diversity (log)	2.748	0.572	–0.545 ***	0.373 ***
Age of acquisition	6.978	2.189	0.369 ***	–0.212 ***
Length	5.864	1.957	0.131 ***	0.142 ***
Neighborhood size (OLD20)	2.029	0.711	0.137***	0.110***
Neighborhood size (orthographic)	4.978	6.327	–0.082 ***	–0.107 ***
Neighborhood size (phonological)	11.659	14.190	–0.090 ***	–0.090 ***
Concreteness	3.547	1.067	0.000	–0.075 ***
Hypernymy	6.984	6.607	–0.211 ***	0.116 ***
CD (word associations, USF)	15.862	24.279	–0.335 ***	0.175 ***
CD (word associations, EAT)	40.226	10.679	–0.120 ***	0.142 ***
CD (variety of semantic contexts)	1.717	0.282	–0.228 ***	0.194 ***
CD (variety of lexical contexts)	1.373	0.702	0.424 ***	–0.282 ***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.10$

MODEL 2: NNS MEAN ITEM ACCURACIES

The same variables and interactions were entered in a multiple regression as independent variables with NNS mean item accuracy as the dependent variable. Three predictors (concreteness, hypernymy, and contextual distinctiveness [word associations, USF]) demonstrated suppression effects and were excluded from further analysis. The seven

TABLE 4. Model 1 regression results predicting NNS standardized RTs

Variables	β	Error	t	
(Intercept)	–0.251	0.004	–57.058	***
Contextual diversity (log)	–0.115	0.008	–15.294	***
Age of acquisition	0.046	0.006	7.572	***
Concreteness	–0.009	0.006	–1.573	
Contextual distinctiveness (word associations, USF)	–0.011	0.005	–2.088	*
Contextual distinctiveness (word associations, EAT)	–0.020	0.005	–4.174	***
Contextual distinctiveness (variety of semantic contexts)	–0.002	0.006	–0.378	
Contextual distinctiveness (variety of lexical contexts)	0.028	0.006	4.544	***

Goodness of fit	*R*-squared	adj. *R*-squared	Resid *SE*	DF	*F*	
Model 1	0.325	0.323	0.251	7, 3247	223.0	***

Note: β is standardized coefficient, *SE* is standard error, and no unstandardized coefficients reported because all predictors were standardized; hypernymy, length, and neighborhood size were removed from analysis due to suppression effects; interaction terms did not demonstrate significant in initial analysis and were removed from Model 1.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.10$

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

TABLE 5. Model 2 regression results predicting NNS mean item accuracy

Variables	β	Error	t	
(Intercept)	0.942	0.002	453.428	***
Contextual diversity (log) × length	–0.009	0.001	–7.234	***
Neighborhood size (orthographic) × length	0.006	0.003	2.174	*
Contextual diversity (log)	0.024	0.002	12.29	***
Length	0.020	0.003	7.782	***
Neighborhood size (orthographic)	–0.001	0.003	–0.401	
Age of acquisition	–0.013	0.002	–8.417	***
Contextual distinctiveness (word associations, EAT)	0.004	0.001	3.097	**
Contextual distinctiveness (variety of semantic contexts)	0.000	0.001	–0.017	
Contextual distinctiveness (variety of lexical contexts)	–0.005	0.002	–2.693	**

Goodness of fit	*R*-squared	adj. *R*-squared	Resid *SE*	DF	*F*
Model 2	0.238	0.235	0.068	9, 3245	112.3***

Note: β is standardized coefficient, *SE* is standard error, and no unstandardized coefficients reported because all predictors were standardized; concreteness, hypernymy, and contextual distinctiveness (word associations, USF) removed from analysis due to suppression.
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.10$

remaining indices and two interaction terms were then reentered into a final model (Model 2). A significant regression equation was reported for Model 2, (F (9, 3245) = 112.300, $p < .001$), with an adjusted R^2 of .235. Five measures contributed significantly to the model (Table 5), along with two interactions between length and contextual diversity (log) and length and neighborhood size (orthographic).

Model 2 findings indicate that contextual diversity, length, AoA, and two operationalizations of contextual distinctiveness (one word association measure and a measure reporting the distinctiveness of lexical contexts), as well as interactions between contextual diversity and length (Figure 1) and between neighborhood size and

TABLE 6. Comparing lexical decision performance between L1 subjects (Balota et al., 2007) and L2 subjects (current study)

Variable	r		t	
Mean response times	0.330	***	–53.028	***
Mean response times (z-scores)	0.489	***	–43.569	***
Mean response times standard deviations	0.038	*	–26.493	***
	ρ		U	
Mean accuracy	0.180[a]	***	4202399[a]	***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.10$
[a]Due to nonnormality, tie-corrected Spearman's rho (ρ) and Mann-Whitney (U) test are reported for mean accuracy.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

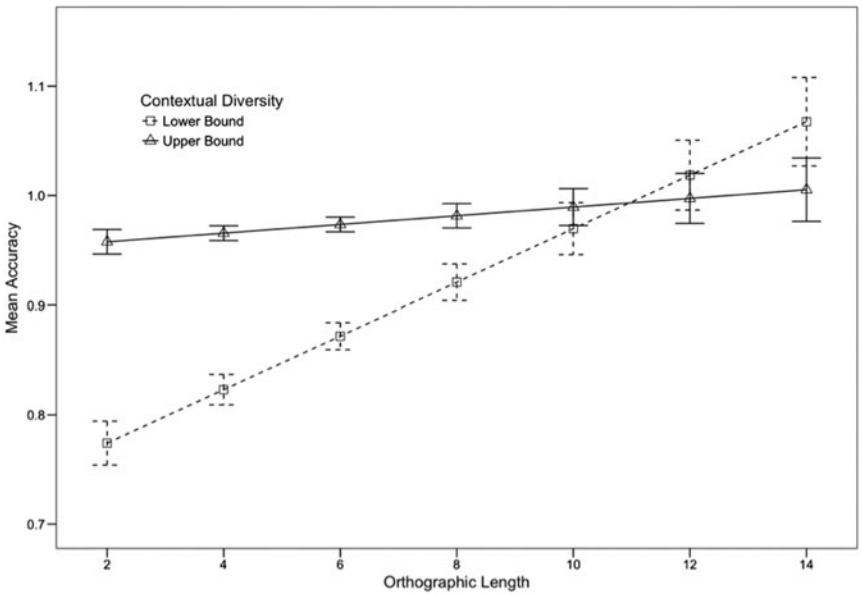


FIGURE 1. Demonstrates that the interaction between length and contextual diversity is strongest for shorter words, in that accuracy was significantly higher for shorter words if they had higher contextual diversity values (i.e., upper bound) compared to shorter words with lower contextual diversity values (i.e., lower bound). For longer words, differences in contextual diversity had less of an impact on accuracy. In other words, if a shorter word is encountered in fewer diverse contexts, it is more difficult to accurately recognize. However, this effect attenuates as length increases, suggesting that for a very long word, the diversity of contexts in which the word naturally appears has less of an effect on accurate word recognition.

length (Figure 2), explained roughly 24% of the variance in average item accuracy in the NNS lexical decision task.

The remaining coefficients suggest that NNS subjects were more likely to accurately recognize English words they are more contextually diverse, are longer, and have more associations in word association tasks. NNS subjects were less likely to accurately recognize English words that are acquired later in life and occur in more distinct lexical contexts.

DISCUSSION

While predictors of L1 word recognition are well attested, relatively little research has examined components of NNS word recognition. A better understanding of NNS word recognition is important not only for investigations of NNS lexical decoding but also for broader inquiry into how words are accessed and processed by NNS speakers. In particular, knowledge of within-language (vs. cross-linguistic) components of NNS word recognition can help identify which features contribute to the difficulty (or ease) with which NNS speakers from a variety of L1 backgrounds learn new words. Significant predictors of NNS latencies and accuracy included frequency, AoA, and contextual distinctiveness, with length moderating the impact of frequency and neighborhood size

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

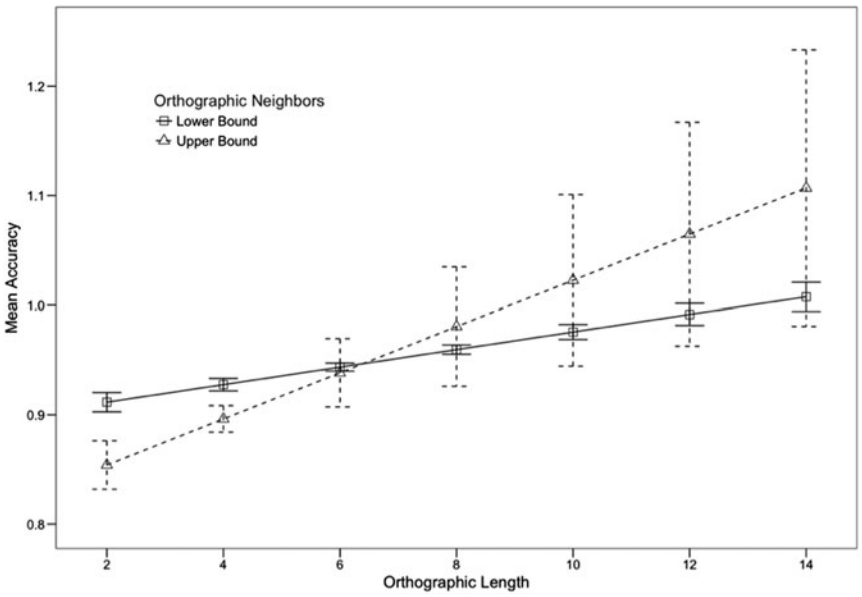


FIGURE 2. Represents the interaction between length and orthographic neighborhood size. For shorter words, those with fewer orthographic neighbors were recognized more accurately compared to shorter words with a greater number of orthographic neighbors. In other words, if a shorter word appears orthographically similar to several other shorter words, that word is more difficult to accurately recognize. While this interaction appears to reverse as length increases in Figure 2, confidence intervals suggest a great deal of variance in Model 2 predictions for words with a large number of orthographic neighbors. This is likely due to the fact that longer words generally have fewer (if any) orthographic neighbors, and also due to the relatively few number of words in our dataset at this range in length.

on accuracy. These predictors collectively explained approximately 32% of variance in NNS lexical decision item RTs and 24% of the variance in item accuracy. For comparison, Balota et al.'s (2004) L1 study demonstrated that lexical variables explained 49% and 52% of variance in ELP lexical decision and accuracy RTs, respectively. Because Balota et al. analyzed monosyllabic RTs, one would anticipate participants' decisions to multisyllabic words to exhibit greater variability. In addition, factors present in the current study but not analyzed—specifically L1 background and NNS proficiency—likely account for some of this unexplained variance.

Item RTs (Model 1) were predicted by contextual diversity, AoA, and three contextual distinctiveness measures (two word association measures and a measure reporting the distinctiveness of lexical contexts). Meanwhile, item accuracy (Model 2) was impacted by contextual diversity, AoA, and two contextual distinctiveness measures (one word association measure and a measure reporting the distinctiveness of lexical contexts). In addition, length was found to moderate the effect of contextual diversity and neighborhood size on accuracy of NNS word recognition (Figures 1 and 2).

Taken as a whole, these results demonstrate that the word recognition behavior of NNS speakers of English is impacted significantly by contextual diversity, (L1) AoA, and contextual distinctiveness, and that the impact of contextual diversity and

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

neighborhood size on NNS word recognition may be moderated by length. The relationship of each of these variables to NNS word recognition and lexical access is further discussed in the text that follows.

CONTEXTUAL DIVERSITY

The strongest predictor of both RTs and accuracy was contextual diversity. Like L1 subjects, NNS subjects respond more quickly and accurately to words that they are more frequently exposed to in a greater variety of contexts. Because of the high correlation between our contextual diversity measure and a word frequency measure ($r = .980$), it is reasonable to interpret the two as overlapping operationalizations of frequency. In this regard, our results support previous evidence suggesting a strong frequency effect in NNS word representation and recognition (De Groot et al., 2002; Ellis, 2002a).

AGE OF ACQUISITION

Another variable that predicted both RTs and accuracy was AoA. The AoA measure used here is based on human estimates of L1 age of word learning. While one would not assume that the age or sequence of learning for L1 children and NNS learners are the same, the unique variance explained by AoA in NNS lexical performance suggests that this measure may capture cumulative frequency or centrality of representation in the NNS lexicon, as it has been argued to do for the L1 lexicon (Cortese & Balota, 2012; Steyvers & Tenenbaum, 2005). Our findings also highlight the importance of including AoA in analyses to avoid falsely attributing word recognition effects to frequency or contextual diversity alone (Morrison & Ellis, 1995).

CONTEXTUAL DISTINCTIVENESS

Contextual distinctiveness also predicted lexical decision behavior for both RTs and accuracy. The two contextual distinctiveness variables retained in Models 1 and 2 measured number of word associations and the distinctiveness of a word's lexical context. This means that target words with a greater number of associations in native speakers' minds were recognized more quickly and accurately by NNS subjects. This finding aligns with previously attested L1 behavior (Balota et al., 2004), demonstrating that both L1 and NNS subjects are quicker to recognize words with greater semantic connectivity in the lexicon.

Words that occur in the company of a variety of different words (vs. in more distinct lexical contexts) were also recognized more quickly and with greater accuracy by NNS subjects. Following contextual diversity, this lexical contextual distinctiveness measure (McDonald & Shillcock, 2001) was the second strongest predictor of NNS decision times. Whereas contextual diversity reports the variety of global contexts (e.g., documents, newspapers, television shows) in which a word is encountered, the preceding contextual distinctiveness measure operationalizes context more locally as the immediate lexical environment in which a word frequently occurs (see Berger et al., 2017). The independent variance explained by both contextual diversity and a lexical contextual distinctiveness measure indicates that the constraints of both global *and* local context

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

impact NNS subjects' lexical access in a manner unique from raw frequency alone, a finding that has been demonstrated in NNS lexical access as well (McDonald & Shillcock, 2001).

LENGTH AS MODERATING VARIABLE

The observed interaction between length and contextual diversity (Figure 1) for NNS accuracy (i.e., differences in contextual diversity values have a reduced influence on accuracy as word length increases) may reveal a proficiency effect, whereby subjects respond more quickly and accurately to more contextually diverse words—regardless of length—because they know them better (De Groot et al., 2002). This impact of length on NNS accuracy runs parallel to a similar trend found in L1 studies, whereby length demonstrates a stronger effect for low-frequency words than high-frequency words (Balota et al., 2004; Weekes, 1997). Length was also found to moderate the effect of neighborhood size on accuracy of NNS word recognition (Figure 2) such that shorter words with fewer orthographic neighbors were recognized more accurately when compared to those with a greater number of orthographic neighbors. Taken as a whole, the two interactions demonstrated in Model 2 suggest that length's impact on accurate NNS word recognition is strongest for less contextually diverse words with a greater number of similarly spelled neighbors; conversely, the impact of length wanes for more contextually diverse words with fewer orthographic neighbors. It is also the case that longer words naturally have fewer orthographic neighbors.

Previously attested predictors of L1 word recognition including frequency (written or spoken), concreteness, hypernymy, and the variety of semantic contexts in which a word typically occurs (Balota et al., 2004; Balota, Ferraro, & Connor, 1991; De Groot et al., 2002) were not significant predictors in our models. It should be noted, however, that while these features were not predictive in our models, they all correlated significantly with NNS lexical decision performance variables (one exception was concreteness, a purely semantic variable that correlated significantly with NNS accuracy but not RTs). Thus, the findings suggest that NNS word recognition is not entirely distinct from L1 word recognition and that the two are influenced by similar lexical variables. To further test links between NNS and L1 word recognition, post hoc independent-samples t-test and Pearson's correlations were conducted to compare differences between mean item RTs, mean item standardized RTs (z-scores), mean item RT SD, and accuracies. Significant differences between all L1 and NNS performance variables ($p < .001$) were demonstrated, such that L1 recognition was faster (by ~100 ms) and more accurate (see Table 6). In addition, moderate and significant correlations were reported between L1 and NNS mean average RTs, mean RTs z-scores, and mean item accuracy. These post hoc analyses indicate that while L1 subjects recognized real words in English faster and more accurately, some shared variance (as evidenced by correlations) was observed for RTs and accuracies between L1 and NNS subjects. The effect sizes of these correlations suggest that at least a moderate amount of variance in the NNS lexical decision datasets—particularly for RTs—is systematically similar to L1 subjects.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

CONCLUSION

This study was motivated by the degree to which lexical variables could predict NNS subject data. In total, these analyses were conducted to better understand the intrinsic difficulty of English words for NNS speakers and how this difficulty may affect second language word recognition and acquisition. Findings indicated that contextual diversity, AoA, and contextual distinctiveness measures related to number of word associations and lexical context significantly predicted variance in both NNS RTs and accuracy. In addition, significant interactions were observed between length and contextual diversity, and between length and neighborhood size, for NNS accuracy.

This study demonstrates that contextual diversity, AoA, and contextual distinctiveness may be useful in understanding the intrinsic difficulty of English words for NNS users across a variety of L1 backgrounds. Measures derived from this study could also help mitigate concerns about using L1 baseline data to examine NNS phenomena (Ortega, 2016) in that NNS baseline norms would now be available for use alongside L1 norms. Ultimately, more studies involving NNS word norms will contribute to the cumulative development of knowledge regarding NNS lexical access specifically, and the processes of NNS word learning and SLA more broadly. Finally, our post hoc analysis found that while L1 and NNS subjects differed significantly in their lexical decision task performance, they demonstrated shared variance (moderate correlations) across most lexical decision measures, suggesting that differences in L1 and NNS lexical access are more a matter of degree than kind.

The current findings should be interpreted carefully in consideration of limitations. One limitation is our lack of control for L1 background and the potential for cross-linguistic influence and/or nonselective lexical access (cf. Bultena & Dijkstra, 2013; Van Heuven, Dijkstra, & Grainger, 1998). For example, it has been shown that NNS lexical decision performance can be facilitated by L1 cognates (Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010) and L1 orthography effects (Wang & Koda, 2005). A subject-based analysis of these data—specifically, one that includes research participants' L1 background—would be a crucial next step in exploring this limitation. Another constraint of the dataset is the inability to account for proficiency in NNS speakers. While all our participants were adult NNS English speakers in the United States that were advanced enough to navigate the AMT online platform, it is likely that a range of proficiencies were represented. Future studies of a similar nature would do well to control for both L1 background and NNS proficiency. Another limitation is the population. We had a greater percentage of L1 Spanish speakers who may demonstrate patterns that differ from other L1s (although see Lemhöfer et al. [2008] as a counterpoint). Additionally, many participants reported English as their dominant language. These participants, although NNS, may perform in a more nativelike way than NNS that do use English as their dominant language. Also, like all behavioral studies, we are not able to thoroughly control when a participant was potentially distractor or was not paying attention. We partially controlled for this by removing outliers, but some variance in responses may have been missed. Finally, differences in the method of data collection between the current study and Balota et al. (2007) should be acknowledged. Whereas Balota et al.'s research participants were seated in a physical lab and presumably monitored while taking the experiment, the AMT workers who participated in the current study were

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

unsupervised. Even though previous research suggests that AMT workers are reliable (Buhrmester, Kwang, & Gosling, 2011; Enochson & Culbertson, 2015), it is possible that the use of AMT workers introduced some additional variability in the data. Furthermore, the fact that Balota et al.'s participants were university students makes it likely that the age of their research subjects was less variable than the age of AMT workers who contributed to the current study's lexical decision response data.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263119000019

NOTES

[1]While evidence supports integrated and nonselective lexicons (Bultena & Dijkstra, 2013; Van Heuven, Dijkstra, & Grainger, 1998), some researchers argue that bilingual lexicons may not be integrated beyond cognates and interlingual homographs (Kroll & Stuart, 1995). Ultimately, the degree of cross-language activation is likely moderated by speakers' proficiency in the NNS (Bijeljac-Babic, Biardeau, & Grainger, 1997; Sunderman & Campbell, 2013).

[2]A *pseudoword* is a nonsense string of letters that is pronounceable by language speakers due to conformity to that language's phonotactic restrictions.

[3]Due to an oversight, participants' age was not collected.

4These 3,318 content words were selected based on the criteria that they appeared in the multiple datasets used in the current study (i.e., the lexical databases reported by TAALES, see following).

[5]All lexical measures were obtained from existing datasets; however, many are freely available as computational indices in the Tool for the Automatic Assessment of Lexical Sophistication (Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2017).

[6]Number of syllables was excluded due to collinearity with orthographic length (e.g., Pexman, Lupker, & Hino, 2002).

[7]ELP (L1) RTs and accuracies were also analyzed to compare L1 results to the models derived from NNS data. Due to space limitations, these L1 models are not reported in the current study. However, results of the analysis aligned with predictions from previous L1 research.

REFERENCES

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determined word-naming and lexical decision times. *Psychological Science*, *17*, 814–823. doi: 10.1111/j.1467-9280.2006.01787.x 16984300.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 234–254.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *41*, 439–461.

Balota, D. A., & Chumbley, J. I. (1990). What are the effects of frequency in visual word recognition tasks? Right where we said they were! *Journal of Experimental Psychology: General*, *133*, 231–237. doi: 10.1037/0096-3445.119.2.231.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316. doi: 10.1037/0096-3445.133.2.283.

Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 187–222). Hillsdale, NJ: Erlbaum.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. doi: 10.3758/bf03193014.

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Educational Technology and Society*, *20*, 201–212.

Bijeljac-Babic, R., Biardeau, A., & Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Memory and Cognition*, *25*, 447–457.

British National Corpus. (2007). *Version 3 (BNC XML Edition)*. Retrieved from http://www.natcorp.ox.ac.uk/.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word-frequency as predictors of word familiarity and word naming latency. *Memory and Cognition*, *15*, 208–216.

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*, 65–85.

Brysbaert, M., & New, B. (2009a). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. doi: 10.3758/brm.41.4.977.

Brysbaert, M., & New, B. (2009b). *Subtlexus: American word frequencies*. Retrieved from http:/subtlexus.lexique.org.

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

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, *6*(1), 3–5.

Bultena, S., & Dijkstra, T. (2013). Lexical access in bilingual visual word recognition. In C. A. Chappele (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Hoboken, NJ: Blackwell Publishing.

Cann, R. (2000). Functional versus lexical: A cognitive dichotomy. In R. D. Borsely (Ed.), *The nature and function of syntactic categories* (Vol. *32*). London, UK: Academic Press.

Carlton, J. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 130–136.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance (Vol. VI)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

Cook, S. V., & Gor, K. (2015). Lexical access in NNS: Representational deficit or processing constraint? *Mental Lexicon*, *10*, 247–270.

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*, 1–17.

Cortese, M. J., & Balota, D. A. (2012). Visual word recognition in skilled adult readers. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 159–185). New York, NY: Cambridge University Press.

Crossley, S. A. (2013). Assessing automatic processing of hypernymic relations in first language speakers and advanced second language learners: A semantic priming approach. *The Mental Lexicon*, *8*, 96–116. doi: 10.1075/ml.8.1.05cro.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*, 115–135.

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning. *Studies in Second Language Acquisition*, *35*, 727–755.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190.

De Groot, A. M., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language*, *47*, 91–124.

De Groot, A., Dannenburg, L., & Van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language*, *33*, 600.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*, 52–67.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*, 843–863.

Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. M. De Groots (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179–201). London UK: Oxford University Press.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*, 284–301.

Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*, 175–197.

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin and Review*, *15*, 850–855.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 515.

Ellis, N. C. (1997). Vocabulary acquisition: Word structure, collocation, word-class, and meaning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 122–139). Cambridge, UK: Cambridge University Press.

Ellis, N. C. (2002a). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188. doi: 10.1017/s0272263102002024.

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Second Language Acquisition*, *24*, 297–339. doi:10.1017/s0272263102002140.

Elston-Güttler, K. E., Paulmann, S., & Kotz, S. A. (2005). Who's in control? Proficiency and L1 influence on NNS processing. *Journal of Cognitive Neuroscience*, *17*, 1593–1610.

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS ONE*, *10*(3).

Fellbaum, C. (1998). *WordNet: An electronic lexical database*: Cambridge, MA: MIT Press.

Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics*, *24*, 289–315.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*, 78–104.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–635.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*, 718–730.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Johnson, N. F., & Pugh, K. R. (1994). A cohort model of visual word recognition. *Cognitive Psychology*, *26*, 240–346.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge, UK: Cambridge University Press.

Kučera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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/10.1017/S0272263119000019 Published online by Cambridge University Press

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

Kyle, K., Crossley, S., & Berger, C. M. (2017). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods.* Advance Online Publication.

Lakoff, G. (1987). *What categories reveal about the mind*. Chicago, IL: University of Chicago.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 12.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Marian, V., & Blumenfeld, H. K. (2006). Phonological neighborhood density guides: Lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries*, *2*, 3–35.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: An account of basic findings. *Psychological Review*, *88*, 375–407.

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*, 295–323.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, *3*, 235–244.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 116–153.

Nelson, D. L., McEvoy, C., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from http://www.usf.edu.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review*, *13*, 45–52.

Ortega, L. (2016). Multicompetence in second language acquisition: Inroads into the mainstream? In V. Cook & L. Wei (Eds.), *The Cambridge handbook of linguistic multicompetence* (pp. 50–76). Cambridge, UK: Cambridge University Press.

Paivio, A. (1991). *Images in mind: The evolution of a theory*. Hertfordshire, UK: Harvester Wheatsheaf.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin and Review*, *9*, 542–549.

Portin, M., Lehtonen, M., & Laine, M. (2007). Processing of inflected nouns in late bilinguals. *Applied Psycholinguistics*, *28*, 135–156.

Rastle, K. (2009). Visual word recognition. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 71–87). Oxford. UK: Oxford University Press.

Saito, K. (2015). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition*, *37*, 713–743.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343–360.

Schmitt, N., & Hemchua, S. (2006). An analysis of lexical errors in the English composition of Thai learners. *Prospect*, *21*(3), 3–25.

Schwanenflugel, P., & Stowe, R. W. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.), *The psychology of word meanings*. Hillsdale, NJ: Erlbaum.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 82–102. doi: 10.1037/0278-7393.9.1.82.

Segalowitz, N., Poulsen, C., & Komoda, M. (1991). Lower level components of reading skill in higher level bilinguals: Implications for reading instruction. *AILA review*, *8*, 15–30.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*, 369–385. doi: 10.1017/S0142716400010845.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*, 53–67. doi: 10.1017/s0142716400010572.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*, 411–416.

Stanovich, K. (1991). Word recognition: Changing perspectives. In R. Barr, M. Kamil, P. Mosenthal, & P. Pearson (Eds.), *Handbook of reading research* (Vol. *2*, pp. 418–452). Mahwah, NJ: Lawrence Erlbaum.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structures of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Sunderman, G., & Campbell, A. (2013). Psycholinguistic approaches to vocabulary. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Blackwell Publishing.

Van Heuven, W. J., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, *39*, 458–483.

Wang, M., & Koda, K. (2005). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning*, *55*, 71–98.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *50A*, 439–456.

Whaley, C. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*, 143–154.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, *15*, 971–979.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin and Review*, *11*, 452–457.

https://doi.org/10.1017/S0272263119000019 Published online by Cambridge University Press