English Language Teaching; Vol. 13, No. 5; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

The Nativelikeness Problem in L2 Word-association Tasks: Examining Word Class and Trials

Boji P. W. Lam[1] & Li Sheng[2]

[1]Department of Audiology and Speech-Language Pathology, University of North Texas, Texas, United States of America

[2]Department of Communication Sciences and Disorders, University of Delaware, Newark, United States of America

Correspondence: Boji P. W. Lam, Department of Audiology and Speech-Language Pathology, University of North Texas, Denton, TX 76201, United States of America

Received: March 21, 2020 Accepted: April 26, 2020 Online Published: April 27, 2020

doi: 10.5539/elt.v13n5p125 URL: https://doi.org/10.5539/elt.v13n5p125

Abstract

Significant variation exists in how native speakers respond to word association tasks and challenges the usage of nativelikeness as a benchmark to gauge second language (L2) performance. However, the influence of word class and trials of elicitation is not sufficiently addressed in previous work. With controlled stimuli from multiple word classes, repeated elicitations, and analytic approaches aiming to tease apart their interactions, this study compared the extent to which native speaker controls and late L2 learners generated associates that converged to a large-scale association norm, and examined the influence of word class and trial on the likelihood to elicit idiosyncratic responses within the two language groups. During initial elicitation, only adjectives elicited greater convergence to the norm among native speakers than L2 learners. Furthermore, native speakers were more likely to generate synonyms whereas L2 learners were more likely to generate antonyms to adjectives in the initial elicitation. For nouns and verbs, 30% of associates produced by the native speaker controls failed to converge to the norm. In fact, the native speaker controls were not more "nativelike" than L2 learners for nouns and verbs until later elicitations. Finally, despite reports of significant variation among native speakers in previous work, the amount of response idiosyncrasy was consistently lower in native speakers than in L2 learners, regardless of word class or elicitation trial. By revealing the effects of word class and trials on association performance, findings from this study suggest potential means to ameliorate the issue with nativelikeness in L2 word association studies.

Keywords: second language learners, trial effects, word association tasks, word class

1. Introduction

Breadth and depth are two well-known dimensions in vocabulary assessments (Schmitt, 2014; Zhang & Koda, 2017). Breadth refers to the size of a person's vocabulary or, simply, the number of known words (Schmitt, 2014). Depth, which is more challenging to conceptualize and measure than breadth (Zhang & Koda, 2017), refers to how well the words are known. This study focused on a specific area of vocabulary depth in second language (L2) learners and native speakers, namely the network knowledge of words in the forms of word associates. For examples, for adult English-speaking native speakers, *spoon* is frequently the first word that comes to mind upon hearing *fork*, *picture* upon hearing *draw*, and *good* upon hearing *bad (*cf., Nelson, McEvoy & Schreiber, 2004). The high likelihood for *fork* to trigger *spoon* in native speakers suggest close relationships between the two concepts due to frequent co-occurrence in language usage and daily experience. We focused on the assessment of network knowledge of words because it reflects lexical-semantic experience shared by a language community (Nelson, McEvoy, & Dennis, 2000). As multiple scholars have claimed, an important feature of vocabulary development is to observe the conventions of word use and assimilate how words are used in the native-speakers' community (Adams & Bullock, 1987; Goldberg, 2019; Sheng, Bedore, Peña, & Taliancich-Klinger, 2013).

1.1 Word Association Tasks and the Nativelikeness Problems in Studies on Second Language Learners

The word associate format is frequently used to assess network knowledge of words in L2 learners (for a review, see Zhang & Koda, 2017) and native speakers (e.g., Entwistle, 1966; Ervin 1961; Nelson, 1977). In the L2

literature, the free word association task (WAT) has been widely used (e.g., Jiang 2002; Nissen & Henriksen, 2006; Schmitt, 1998; Wolter 2001; Zareva 2007; Zareva & Wolter, 2012). In a typical WAT, participants report the first single-word associate that comes to mind upon hearing or reading a lexical cue. The context-free WAT "taps into lexical knowledge acquired through world experience" (Nelson et al., 2004, p.402) and sheds light on the development of associative structure among words. The WAT is appealing to researchers and educators for its ease of administration. However, the lack of restrictions on participants' responses and straightforward scoring methods limit the potential of WAT in measuring vocabulary depth (Zhang & Koda, 2017). The associates produced by native speakers are frequently used as a benchmark to score the responses produced by L2 learners (Fitzpatrick, Playfoot, Wray, & Wright, 2015; Zhang & Koda, 2017). Recent studies question the assumption underlying this scoring approach, which posits that native speakers are coherent and reliable targets (Fitzpatrick 2007; Fitzpatrick et al., 2015; Nissen & Henriksen, 2006; Zareva & Wolter, 2012). The goal of this study was to further examine the issue of nativelikeness in L2 WAT studies, with a focus on the role of word class and elicitation trial on performance differences between native speakers and L2 learners. As described below, word class exerts noticeable influence on associative behaviors (Deese, 1962; Entwisle, 1966). In addition, a repeated elicitation procedure would likely result in reduced coherence in the group's responses in later trials. However, analyses in previous work insufficiently address how word class and trials of elicitation affect indices of nativelikeness and within-group coherence.

Schmitt (1998) argued, for assessments using the word associate format, "the attribute on which L2 responses are judged *must* logically be native-likeness" (p. 390). The assessment of nativelikeness refers to evaluating learners' associates against the "native speaker norms" (Fitzpatrick et al., 2015). Two common analytic approaches are to (1) *compare* native speakers and L2 learners with regard to the cue-associate relationships and (2) *match* participants' exact responses against a native speaker norm (for an overview, see Fitzpatrick et al., 2015; Zareva & Wolter, 2012). The first approach categorizes participants' responses into meaning-based versus form-based (e.g., Fitzpatrick, 2007) or paradigmatic versus syntagmatic relationships (e.g., Naemi, 2004; Nissen & Henriksen, 2006). Developmental studies show that mature native speakers often respond to WAT with a predominance of paradigmatic associates (e.g., dog-*cat*) as opposed to the less mature syntagmatic associates (e.g., dog-*bark*) or phonological associates (e.g., dog-*bog*) (Ervin, 1961; Nelson, 1977). Extending these developmental phenomena in native speakers, L2 studies often categorize participants' responses according to level of maturity, which are then compared against those produced by native speakers or participants' own L1 (e.g., Namei, 2004; Nissen & Henriksen, 2006; Sheng, McGregor, & Marian, 2006). However, this scoring practice has been challenged in recent studies. Nissen and Henriksen (2006) found that adult native speakers produced an unexpectedly high percentage of syntagmatic associates. A potential explanation is that late-acquired syntagmatic relationships, such as those existing in idioms and slangs, may be preferred by some mature native-speakers (Entwisle, 1966) and yet "tends to be overlooked in the literature" (Nissen & Henriksen, 2006, p.390). Fitzpatrick (2006) reported similar findings that over 25% of native speakers' responses would be categorized as the less mature, position-based associates. Furthermore, significant heterogeneity exists in native speakers that the percentage of position-based associates ranged from 3% to 80% of total responses. For nativelikeness to be useful in evaluating WAT performance in L2 learners, the performance of native speakers should present with defining characteristics that are sufficiently reliable and homogenous. Findings from Nissen and Henriksen (2006) and Fitzpatrick (2006) question the reliability of using developmental phenomenon in native speakers to evaluate L2 learners.

The use of pre-determined coding schemes has other methodological concerns. First, various coding schemes for cue-associate relationship exist in current literature. The validity of these coding schemes in operationalizing nativelikeness remains a matter of debate (Fitzpatrick et al., 2015). Second, coding participants' responses often involves raters' subjective interpretation of cue-associate relationship. Because cues and/or responses often belong to multiple parts of speech (e.g., horse – ride; ride can be a noun or a verb) or have multiple meaning senses, inter-rater reliability could be jeopardized. To overcome these potential problems, this study adopted a direct matching approach (e.g., Kruse, Pankhurst, & Sharwood, 1987; Wolter, 2002), which matched participants' exact responses against a native speaker norm, instead of coding participants' responses according to a particular developmental phenomenon (e.g., the syntagmatic-paradigmatic distinction). Furthermore, matching provides a direct means to operationalize nativelikeness by simply asking: do participants produce associates that are present in the norm?

1.2 The Needs to Examine the Effects of Word Class and Elicitation Trial in Word Association Tasks

Several previous studies using a direct-matching approach question the validity of nativelikeness in L2 WAT studies (Kruse et al., 1987; Wolter, 2002; Zareva & Wolter, 2012), Kruse et al. (1987) developed a scoring system using native-speaker norms and reported cases in which L2 learners outperformed native speakers. However,

Kruse and colleagues sampled associates to only ten cues but allowed participants to produce as many as 12 responses, which is cognitively challenging and may affect the reliability of the test (Wolter, 2002). Furthermore, Wolter (2002) argued that the scoring system adopted by Kruse and colleagues was rather complex and arbitrary. Like Kruse and colleagues, Wolter elicited multiple associates from participants but limited the number of elicitations to three. In addition to direct matching, Wolter assigned a stereotypy score for each response based on how many times an associate occurs in the normative data. For examples, for the cue "angry", "mad" will have a higher stereotypy score than "bird" because more native speakers generate "mad" as the associate. Wolter reported significant individual differences in native speakers, with some native speakers obtaining a lower stereotypy score than L2 learners did. However, this study sampled only verbs and performance on each elicitation was collapsed into an overall score. In a later study, Zareva and Wolter (2012) sampled associates to nouns, verbs, and adjectives. Their study elicited only one response to each cue, which does not inform changes in performance during initial and subsequent elicitations. In summary, analyses in previous work seldom teased apart the interactions between word class and elicitation trials. The extent to which word class and trials influence association patterns produced by native speakers and L2 learners is unclear.

With a repeated WAT, this study investigated the effects of word class (nouns, adjectives, and verbs) on eliciting associates from native speakers and late L2 learners that match the responses reported on the University of South Florida association norm (USF; Nelson et al., 1998, 2004), with a focus on the amount of *between-group convergence* and *within-group idiosyncrasy*. In this study, between-group convergence refers to the extent to which participants produce associates in the repeated WAT that resemble the responses on the USF. Within-group idiosyncrasy does not make any references to the USF but is a within-group analysis that focuses on the likelihood for group members to generate associates produced by only one participant. Between-group convergence and within-group idiosyncrasy are not interchangeable and should be separated in analyses. To be specific, a local community of native speakers with low idiosyncrasy (i.e., high within-group coherence) may not produce responses that converge to another community due to potential influence of geographical and cultural factors on word usage (Nelson et al., 2004). The USF was adopted as a "benchmark" of native speaker model in the analyses on between-group convergence; this large-scale database provides a culturally appropriate norm to our participants, who resided in the United States. USF contains approximately 750,000 associates to 5,019 stimulus words elicited from over 6,000 college student participants, with over 100 participants responding to each prompt (Nelson et al., 1998, 2004). This study elicited three responses to each cue (i.e., a total of 90 responses), which allowed us to examine changes in convergence and idiosyncrasy across trials with potentially reduced cognitive demands relative to previous work that invited an extensive set of responses (e.g., Kruse et al., 1987).

This study investigated the effects of word class on convergence and idiosyncrasy because word class exerts noticeable influence on associative behaviors (Deese, 1962; Entwisle, 1966; Nissen & Henriksen, 2006). Specifically, adjectives, nouns, and verbs are organized with distinct networking principles (Miller & Fellbaum, 1991). For examples, familiar adjectives elicit more polarized, antonymous associates than other word classes (Deese, 1964, 1965), and nouns elicit more associates under the same conceptual categories (e.g., dog, cat, horse) than verbs or adjectives (Nissen & Henriksen, 2006). Also, relative to nouns, adjectives comprise a small group of words so the latter word class may prompt more predictable associates. The antonymous nature of some adjectives may prompt participants to produce more stereotypical associates (e.g., dark-light) and thus more USF-converging responses and less within-group idiosyncrasy than nouns and verbs. In contrast, though nouns are likely to elicit associates from the same conceptual categories, the word choices are many (e.g., dog-animal/horse/cat/Shepherd). As a result, nouns may elicit words that show a lower between-group convergence to the USF norm and greater within-group idiosyncrasy than adjectives. However, differences between word classes may be influenced by elicitation trial. Since the word choices for antonyms are limited, the elevated tendency for adjectives to elicit USF-converging responses and reduced idiosyncrasy may be observed more readily during the initial than subsequent elicitations. Predictions for verbs are not straightforward due to its complex nature (Miller and Fellbaum, 1991). However, this study predicted lower USF-convergence and greater within-group idiosyncrasy for verbs than for adjectives similar to nouns because verbs are less antonymous.

Conflicting predictions exist regarding the comparisons between language groups. On the one hand, native-speaking proficiency and closer cultural backgrounds between native speaker controls and the respondents of USF should predict greater USF-convergence for native speaker controls than L2 learners. Native speaker controls should also exhibit a smaller within-group idiosyncrasy than L2 learners since they have had longer exposure to the English language, hence should be more entrenched in the convention of English word use (Goldberg, 2019; Meara, 1983). On the other hand, considerable variation among native speakers (e.g., Fitzpatrick, 2007; Wolter, 2002) may lead to inconsistent language-group differences. This study predicted that

language-group differences may be influenced by word class and trials. Specifically, since adjectives may elicit more converging and less idiosyncratic responses than nouns or verbs from native speakers, language-group differences may be more readily observed for adjectives than for other word classes, especially during the initial elicitation.

To summarize, this study investigated three research questions:

1. Using USF as a norm reference, do native speakers exhibit a higher likelihood to produce USF-converging associates than L2 learners?
2. Using USF as a norm reference, does the likelihood to produce USF-converging associates differ across word class (noun, verb, adjective) and trials of elicitation?
3. Do native speakers exhibit smaller within-group idiosyncrasy (i.e., a lower likelihood to generate idiosyncratic associates that are produced by only one participant among group members) than L2 learners?

2. Methods

2.1 Participants

The L2 learners were 30 Mandarin-English speakers (14 males and 16 females) between the age of 19 to 41. All L2 learners were students from the University of Texas at Austin in the United States or professionals working in the Austin area. This study recruited only L2 learners whose first language was Mandarin to assure that the cue words in English were not loanwords in their first language, or may be easily mistaken by mispronunciations. The L2 learners recruited in this study were of a wide range of English learning experience (seven years to 29 years) with a late onset of English learning as a group. Twenty-one of the participants were from Taiwan and nine were from Mainland China. The L2 learners met the following criteria 1) must be proficient in Mandarin Chinese and English; 2) must have lived in an English-speaking country for at least two years; 3) must be able to read Mandarin and English; 4) have no known cognitive impairments. Study procedures were approved by the Institutional Review Board of The University of Texas at Austin. Participants provided written informed consent and received monetary compensation.

All L2 learners completed a language use questionnaire (Kastenbaum et al., 2019). To estimate the amount of current daily language use, participants described a typical weekday and weekend by listing their daily activities for each hour of the day, and what language(s) they heard and spoke during those activities. Participants were also asked to rate their English and Mandarin proficiency using a five-point scale. Other background information such as age of immigration and onset of English learning was also gathered. The majority of the participants were considered late L2 learners with an average onset of English learning around 11 years of age (Table 1).

The native-speaker group consisted of 30 English-speaking monolinguals (15 females and 15 males) between the age of 18 and 33 (M=22.13, SD=3.33). They were students at the University of Texas at Austin and were born and raised in an English-speaking country. All participants in the native-speaker group reported speaking English only throughout their lives. None of the participants reported any cognitive impairments.

Table 1. Characteristics of the late L2 learners of English

	Age (years)	*Age of immigration (years)*	*Years of Stay*[a]	*Age of Onset (years)*[b]	*Years learning English*	*% English use*	*English Proficiency*[c]	*Mandarin Proficiency*[c]
Mean	25.2	16.7	8.1	10.7	14.5	60.0	3.9	4.7
(SD)	(5.4)	(7.4)	(5.3)	(3.5)	(5.9)	(15.0)	(0.5)	(0.6)
Minimum	19	2	2	0	7	26	3	3
Maximum	41	38	26	18	29	87	5	5

Note. [a] Years living in an English-speaking country; [b] Onset of English Learning; [c] Self-rated proficiency using a five-point scale (1= non-fluent, 5=native fluency).

2.2 Materials

The stimuli consisted of 30 early-acquired English cues previously used in Sheng et al. (2006) (See Appendix A). There were 10 adjectives, 10 nouns, and 10 verbs. To examine the effect of word class, we controlled the cues for lexical-semantic factors that may influence the extent to which participants generate converging or highly stereotypical associates. First, the adjectives, nouns, and verbs did not differ in word frequency because higher

word frequency predicts greater homogeneity in associates and thus greater convergence among participants (Fitzpatrick, 2007; Meara, 1983). Word frequency was obtained from Brysbaert and New (2009), a norm based on 51 million American words. According to Brysbaert and New (2009), this word frequency norm provided better predictability on psycholinguistic behavior than traditional norms, such as CELEX (Burnage, 1990). Second, all stimuli used in the current study were early-acquired words with comparable ages of acquisition across word classes (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), therefore posing few challenges for individuals with less English exposure to complete the task. Besides age of acquisition and word frequency, this study made reference to the USF and controlled for the set-size (i.e., the number of different single-word associates produced by more than two respondents to a cue) and the proportion of participants producing the most predominant associates to the cue. According to the USF, some cues exhibit a high likelihood to elicit highly stereotypical associate (e.g., bad → good) and a smaller number of different associates (Nelson et al., 1998). The rationale to control these two factors is that a cue of small set size and/or elicits highly predominant associate may elicit highly stereotypical associates. A native speaker of English with a standard American accent recorded the 30 cues in a sound-proof booth. The audio-recordings were segmented into individual words using the SoundEdit program.

2.3 Procedure

Participants were presented with the following instructions for the word association task by a trained research assistant:

"Sometimes when you hear one word you think of another. For example, when you hear the word "brother" you might think of the word "sister." Or when you hear the word "birthday" you might think of the word "cake". You will hear a list of words from a computer. Each time you hear a word, say the first word that comes to your mind. Try not to cough or make sounds like "uhmm" or "ah". If the word you want to say is "computer", don't say anything extra like "a computer" or "I think of a computer"; just say "computer" and nothing else. Sometimes you'll hear the same words that you heard before; if this happens, try to give me a different word, not the ones you gave me before. Let's do a short practice."

No participants had questions about the instructions. After the presentation of the instructions, the participants completed a training with "Moon____ Swim____ Green____ Swim____ Moon____" presented via a computer. The participants were reminded to respond as quickly as possible with the first word that came to mind upon hearing the stimulus before the word association task began. The pre-recorded cues for the word association task were presented on a computer one at a time via E-Prime (Version 2.0). Three pre-randomized word lists were created. Participants responded to all 30 cues in one list before they moved on to the next lists. This ensured that the same cues would not be presented to the participant consecutively. A trained examiner sat by the participant and wrote down the participant's response verbatim on a score sheet. If the participant repeated a response from an earlier elicitation, they were reminded to provide a novel response. After the participant provided a valid response, the examiner advanced the experiment to the next item by hitting the "Enter" key on the keyboard.

2.4 Data Coding and Statistical Analysis

To address the first and second research question, this study first matched participants against the associates recorded on the USF norm. Responses were coded as 1 if they were found on USF. All other responses were coded as 0. This study applied minimal lemmatization rule that concerns only inflectional morphemes that mark the plural forms of nouns (i.e., -s; e.g., gifts → gift). Similar to Nelson et al (2004), this study applied lemmatization reluctantly and only when justified because certain lemmatization rules may potentially lead to changes in word classes (e.g., exciting/excited → excite). Among the associates reported on USF, this study only considered those produced by at least two native speaker respondents. Idiosyncratic associates, which were produced by only one respondent, may not reflect lexical-semantic experience readily shared by a language community. As a result, these idiosyncratic associates are reported separately in the USF norm and not considered in this study.

A generalized linear mixed model (GLMER) was conducted in R (R Core Team, 2012) using the package lme4 (Bates, Maechler & Bolker, 2012). The critical dependent variable was match/mismatch, which is dichotomous (match = 1; mismatch =0). The analysis started with language group (native speakers, L2 learners of English), word classes (nouns, verbs, adjectives), trials of elicitation (one, two, three), and their interactions as fixed effects. By-participant and by-item intercepts were included in the model as random effects. The model was refined by removing, one at a time, factors that exhibited the highest p-value, while retaining the hierarchical rule of interactions. Likelihood ratio comparisons were performed to confirm that including a given factor did not improve the amount of variance explained (Baayen et al., 2008).

To address the third research question, this study located the idiosyncratic associates produced to particular cues within each language group and trial. Idiosyncrasy, which did not involve matching responses to USF, was defined as an associate that is produced by only one participant in the respective language community during a particular trial. This analysis was trial-specific because accessibility should be considered when gauging idiosyncrasy in associative performance. For examples, "small" is a highly dominant associate produced by many individuals to the cue "big" during the first trial, which suggests high accessibility of the cue-associate relationship and should not be idiosyncratic during initial elicitation. However, if there is a participant who does not say "small" until a much later elicitation (e.g., the 3rd trial) and is the only participant who does this within the respective language community, such behavior should be described as idiosyncratic. In this study, idiosyncratic associates were coded as 1 and all other responses coded as 0. For the example provided above, "small" would be coded as 1 during the third trial and 0 during the first trial. To summarize, this analysis compared the amount of idiosyncrasy associated with word classes inherent to native speakers and L2 learners across trials of elicitations. This analysis adopted the same analytic approach as for the first and second research question.

3. Results

Table 2. Between-group convergence in native speaker controls and late L2 learners.

	First elicitation			*Second elicitation*			*Third elicitation*		
	Adjective	*Noun*	*Verb*	*Adjective*	*Noun*	*Verb*	*Adjective*	*Noun*	*Verb*
L2 learners	67.0 (17.6)	65.0 (15.7)	71.3 (19.3)	42.0 (17.7)	45.3 (20.5)	38.0 (17.1)	34.0 (16.3)	40.0 (14.6)	27.3 (13.1)
Native speakers	80.7 (9.4)	69.3 (17.6)	71.7 (14.6)	53.3 (15.4)	56.0 (17.7)	46.7 (19.5)	34.3 (17.9)	42.0 (16.5)	36.7 (15.2)

Note. Between-group convergence in this table reported percent match between participants' responses and associates found on the University of South Florida association norm.

3.1 Likelihood to Produce USF-Converging Associates

Model comparison indicated a three-way Language group X Word class X Trial interaction [χ^2 (1) =12.1, p=.02] (see Figure 1, Table 2). The interaction was analyzed with the lsmeans package 2.30-0 to obtain the least-squares means and to test linear contrasts for linear and generalized mixed models (Lenth, 2016). During the first trial, only for adjectives did native speakers exhibit a higher likelihood to produce USF-converging responses than L2 learners (β=0.84, SE=0.22, p<.001). There were no language-group differences for nouns (p=.36) nor verbs (p=.93) (Figure 1A, Table 2). During the second trial, for all word classes, native speakers exhibited a higher likelihood to produce USF-converging than L2 learners (adjectives: β=0.48, SE=0.19, p=.01; nouns: β= 0.44, SE=0.19, p=.02; verbs: β=0.42, SE=0.19, p=.03) (Figure 1B). During the last trial, native speakers exhibited a higher likelihood to produce USF-converging only for verbs (β=0.46, SE=0.20, p=.02) but not for adjectives (p=.94) nor nouns (p=.64) (Figure 1C).

Examining the influence of word class on eliciting USF-converging associates, analyses *within* language groups revealed no word class effects in either native speakers or L2 learners in all trials of elicitations (range of p values = 0.06 to 0.99). However, when comparing the *relative* differences among word classes in native speakers to L2 learners, adjectives were more likely to elicit USF-converging associates than nouns (β=0.66, SE=0.27, p=.01) in native speakers than in L2 learners. The same was true for the comparison between adjectives and verbs (β=0.83, SE=0.27, p=.002) (Figure 1A).

The final analyses examined the changes in likelihood to elicit USF-converging associates across trials. Figure 1A to 1C suggested a lower likelihood for later trials to elicit USF-converging associates. L2 learners exhibited a relatively stable profile that the second trial elicited fewer USF-converging associates than the first trial for all word classes (adjectives: β=-1.08, SE=0.30, p<.001; nouns: β=-0.89, SE=0.29, p=.007; verbs: β=-1.54, SE=0.30, p<.001). Comparing the second to the third trial, there were no trial effects for all word classes (range of p values = 0.25 to 0.67). In native speakers, there was no trial effect for nouns that this word class exhibited stability in their likelihood of eliciting USF-converging associates across trials (first vs second trial: p=.08; second vs. third trial: p=.10). Verbs exhibited a decrease only between the first and second trial (β=-1.14, SE=0.30, p<.001). Finally, adjectives exhibited a continuing decrease across trials (first vs second trial: β=-1.45, SE=0.31, p<.001; second vs. third trial: β=-0.84, SE=0.29, p=.01).

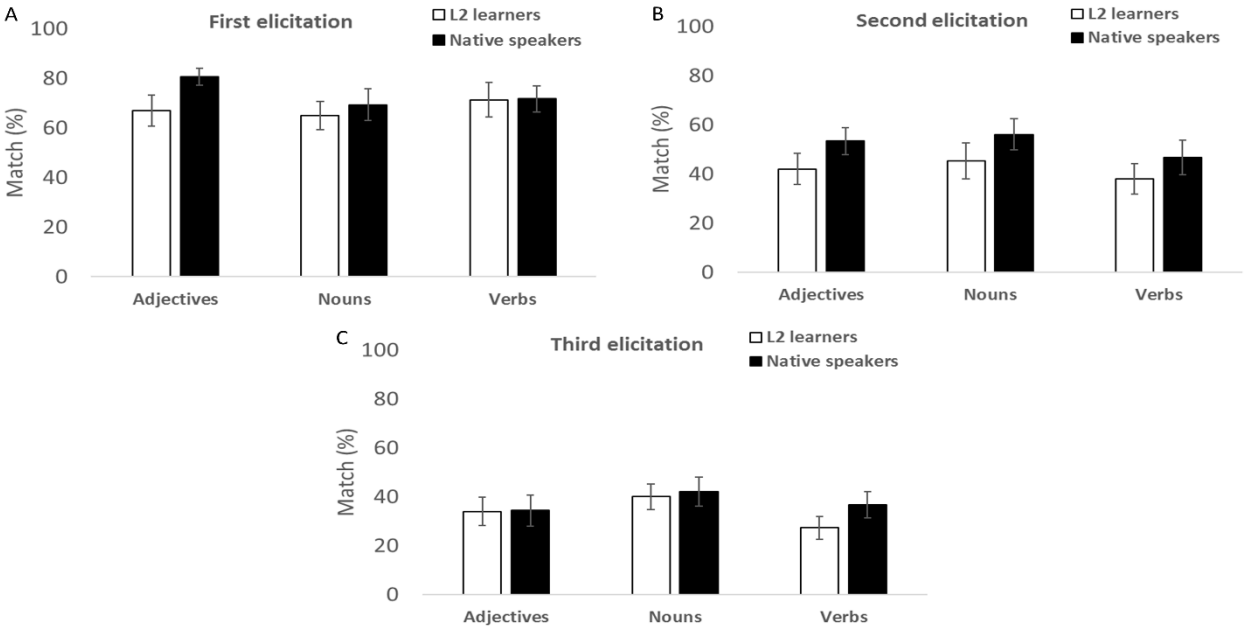

Figure 1. Percentage of associates generated for adjectives, nouns, and verbs that matched the responses reported on the University of South Florida Association Norm during (A) the first elicitation, (B) the second elicitation, and (C) the third elicitation. Error bars indicated 95% confidence intervals.

3.2 Likelihood to Produce Antonyms and Synonyms to Adjective Cues during the First Trial

Since native speakers and L2 learners differed in convergence only for adjectives during the first trial, this study performed a post-hoc analysis to compare compositions of USF-converging associates generated by the two language groups. The first step of this follow-up analysis focused on antonyms because this word category is strongly associated with adjectives than other word classes, such as nouns or verbs (Deese, 1964, 1965). The analysis questioned whether L2 speakers differed from native speakers in their sensitivity to the antonymous nature of adjectives and produced fewer antonyms during the first trial. All USF-converging associates to adjective cues during the first trial were coded as 1 if they were antonyms (e.g., happy →sad) and 0 for others. A GLMER was performed with language group (i.e., native speakers, L2 learners) as the main effect. The random effects and model selection procedure were the same as other analyses reported above.

Model comparison rejected our prediction and, in fact, revealed a much-elevated likelihood for these associates to be antonyms in L2 learners than in native speakers [χ^2 (1) =16.0, p<.001; β=1.39, SE=0.34, p<.001] (Figure 2). The second step of the follow-up analysis then examined whether L2 learners exhibited a lower likelihood to produce synonyms, or words from the same word class (i.e., adjectives) that have related meanings to the cue (e.g., sleepy → tired; thirsty → hungry; sick → ill). Synonymous associates are words that share related meanings with the cue. Synonyms are less stereotypical than antonyms (cf. Nelson et al., 1998), and may be more taxing to the language system (Roehm, Bornkessel-Schlesewsky, Rösler, & Schlesewsky, 2007). This follow-up analysis questioned whether L2 learners exhibited a lower likelihood in generating synonymous associates to adjectives than native speakers. All USF-converging associates to adjective cues during the first trial were coded as 1 if they were synonyms and 0 for others. A GLMER was performed with approaches described above. Model comparisons indicated a language-group effect [χ^2 (1) =16.8, p<.001] that native speakers exhibited a much-elevated likelihood to generate synonyms than L2 learners (β=1.63, SE=0.42, p<.001) (Figure 2). To conclude, while L2 learners generated more antonyms to adjective cues during the first trial than native speakers did, they exhibited a much lower likelihood to generate synonyms.

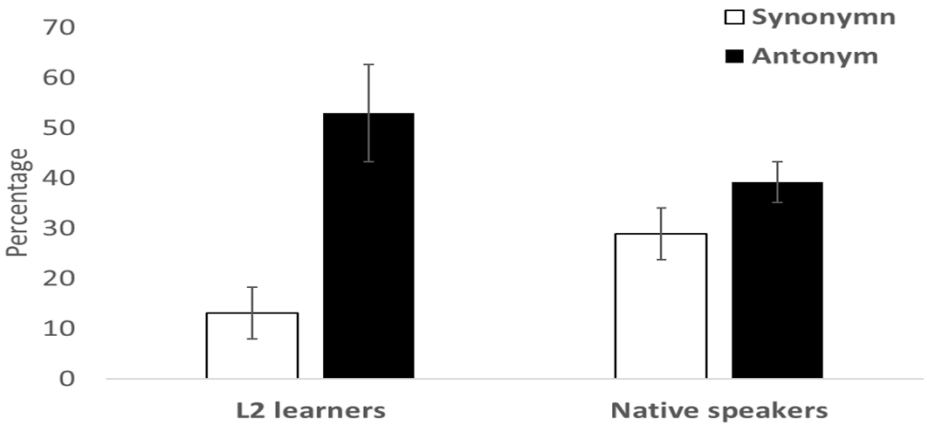

Figure 2. Percentage of synonymous or antonymous associates generated for adjectives that converge to the University of South Florida norm during the first trial. Error bars indicated 95% confidence intervals.

3.3 Likelihood to Produce Idiosyncratic Associates

Table 3 reports within-group idiosyncrasy in native speaker controls and late L2 learners. Model comparison indicated a language group effect [χ^2 (1) = 4.4, p=.04], a trial effect [χ^2 (1) = 73.2, p<.001], but no word class effect (p=.28) (see Table 3). The language group effect indicated native speakers consistently exhibited a lower likelihood to produce idiosyncratic associates than L2 learners (β=-0.19, SE=0.09, p=.03), regardless of word classes and trials. The trial effect indicated an increasing amount of idiosyncrasy in later elicitations (second vs. first β= 0.91, SE=0.13, p<.001; third vs. second: β=0.41, SE=0.13, p<.001). No interactions of any kinds were found (range of p values = 0.12 to 0.91)

Table 3. Within-group idiosyncrasy in native speaker controls and late L2 learners.

	First elicitation			*Second elicitation*			*Third elicitation*		
	Adjective	*Noun*	*Verb*	*Adjective*	*Noun*	*Verb*	*Adjective*	*Noun*	*Verb*
L2 learners	27.0 (7.9)	29.0 (10.3)	27.3 (12.4)	48.3 (16.3)	46.3 (17.5)	49.7 (10.8)	60.7 (19.7)	47.3 (13.3)	63.3 (12.4)
Native speakers	21.0 (14.7)	23.3 (9.4)	25.3 (15.3)	40.7 (14.8)	41.7 (11.9)	45.0 (10.6)	56.7 (13.2)	50.3 (19.3)	41.3 (11.4)

Note. Within-group idiosyncrasy in this table reported the percentage of participant's responses that were produced by only one participant in the respective language group and trial.

4. Discussion

Using controlled stimuli of multiple word classes, repeated elicitations, and analytic approaches that aimed at teasing apart the effects of word class and trials, this study yields three findings that extend recent studies on the potential issue of nativelikeness in L2 WAT studies. First, relative to nouns and verbs, adjectives were more effective in eliciting greater convergence to norm-like behaviors among native speakers than L2 learners during the first elicitation. Native speakers and L2 learners also differed in their likelihood to produce antonymous and synonymous associates to adjective cues during the initial elicitation. Second, when participants were prompted to produce additional associates during the second elicitation, language-group differences were found consistently for all word classes. Native speakers and L2 learners also exhibited different sensitivity to the interaction between word class and trials on eliciting norm-like associates. Finally, analyses on within-group idiosyncrasy showed highly consistent effects that the local community of native speakers was less idiosyncratic than L2 learners, regardless of word classes and trials.

4.1 The Problems with Nativelikeness in L2 Word Association Studies and the Need to Consider Word Class Effects

In L2 word association studies, native speakers are often assumed to exhibit "remarkedly stable patterns of word associations" (Read, 1993, p. 359) and thus a "benchmark" to evaluate L2 learners (Zhang & Koda, 2017). Previous work questioned this assumption (e.g., Fitzpatrick, 2007; Wolter, 2002). This study shows that this

assumption is more problematic for nouns and verbs than adjectives, especially if WAT is to elicit only a single response to each prompt. Using USF as the model of "nativelikeness", 30% of the associates produced for nouns and verbs by native speaker controls during the first elicitation failed to converge. Equally important, during the first elicitation, the native speaker controls were no more "nativelike" than late L2 learners for nouns and verbs. Fitzpatrick et al. (2015) raised concerns about the use of established associate norms in WAT studies; the potential influence of cohort characteristics, such as demographics, may impact the validity of associate norms. In this study, the amount of divergence between the native speaker controls and USF cannot be completely attributed to differences in geographical locations or age. Both Texas and Florida are the southernmost part of the United States, and the controls recruited for this study and those for USF were university students. Nevertheless, the lack of differentiation between native speakers and L2 learners for nouns and verbs during the first elicitation supports the concerns of using native norm lists as a benchmark in L2 WAT studies (Fitzpatrick, 2007; Fitzpatrick et al., 2015; Wolter, 2002).

An increased sampling of adjectives in WAT may ameliorate the issue of nativelikeness in L2 WAT studies. Not only may adjectives exhibit an improved effectiveness in eliciting convergence among native speakers during initial elicitation, comparing the relative composition of synonymous and antonymous associates may provide an index to distinguish native speakers and L2 learners. According to Miller and Fellbaum (1993), "the basic semantic relation among adjectives is antonymy". Because word choices that denote the polarity of a concept are often limited (e.g., left vs right; fat vs. slim), the antonymous nature of adjectives may facilitate the acquisition of lexical items that denote the two ends of polarity and thus strengthen the antonymous cue-associate relationships. As a result, despite a late onset of English acquisition, L2 learners exhibited a heightened likelihood to produce antonymous associates than native speakers. This response pattern may result from language learning experience –antonyms are paired up (e.g., cold-hot) and taught to L2 learners in an explicit manner (e.g., x versus not-x).

In contrast, synonyms require the development of a finer-grained differentiation of words with related meanings. For examples, "wet" is a consequence of "rainy", and "little" a feature of "young". The development and retrieval of synonymous cue-associate relationships may require deeper processing and thus are more challenging than the retrieval of antonymous associates or common collocations (e.g., rainy – day), to which participants may have more exposures in daily language usage. Relative to synonymous relationships, the processing of antonymous relationships is more automatic in neurotypical individuals (Roehm et al., 2007) and spared in patients affected by neurological disorders (Crutch, Williams, Ridgway, & Borgenicht, 2012). In this study, more than twenty-five percent of L2 learners did not produce synonyms to adjective cues at all during the first elicitation. Only one out of 30 native speakers (i.e., 3%) exhibited this associative behavior. While antonymous and synonymous associates are classified as the more mature responses under the traditional paradigmatic-syntagmatic distinction, these two semantic relationships may have distinct implications on evaluating L2 learners and should be separated in analyses.

Our suggestion of an increased sampling of adjectives in WAT resonates with Read's revised word association test (Read, 2013). Read's test focuses on adjectives and targets "more homogeneous subsets of vocabulary items so that greater consistency can be achieved in the semantic relationships among the words and in the pattern of responses elicited" (Read, 1993, p. 369). Note that we do not suggest a complete omission of nouns and verbs from the WAT. Doing so would lead to a missed opportunity to observe the distinct interactions between word class and trials in native speakers and L2 learners. As warned by Zhang and Koda (2017), targeting only particular word class may limit the scope of network knowledge and language competence measured by the WAT. Furthermore, given the antonymous nature of adjectives, the number of adjective prompts and their choices demand caution to avoid converting the WAT into an antonym generation task. Specifically, to adhere to the primary purpose of using the WAT as a measure of network knowledge, future studies may want to ensure that no more than half of the adjective cues (and verbs) have antonyms as the most dominant responses (e.g., easy-hard; stand-sit).

4.2 Potential Benefits of Repeated Elicitation and the Implications of Within-group Idiosyncrasy

Trial analyses suggest potential benefits of repeated elicitations over single elicitation procedures in L2 WAT studies. Prompts in WAT should be familiar to L2 learners to assure that the task measures vocabulary depth, not breadth. As a result, even learners with basic L2 proficiency could produce at least one acceptable, or even good, associate to most prompt words (Wolter, 2002). Therefore, Wolter warned against single elicitation procedures because they may not be sensitive in detecting performance differences among L2 learners of different proficiency levels. In this study, the absence of language-group differences for nouns and verbs in the first trial illustrates that the single elicitation procedure falls short of detecting performance differences between L2

learners and native speakers. When prompted to produce additional associates, L2 learners may be more likely to diverge from norm-like associative performance. Unlike native speakers, the amount of convergence to USF significantly declined in L2 learners during the second trial for all word classes. In contrast, nouns remain relatively stable in eliciting norm-like associates from native speaker controls across trials. Since USF prompted respondents to produce only one response, it is logical to expect fewer USF-converging associates during the second or third trial of elicitations. The fact that native speakers exhibited a weakened trial effect for nouns is surprising. A potential explanation is the high tendency for nouns to elicit paradigmatic associates (Nissen & Henriksen, 2006) that share the same theme as the cue (e.g., teacher, student, professor; ocean, sea, beach). The organizing principles underlying nouns may facilitate the generation of associates within bounded semantic categories and thus a higher resistance to the trial effect in native speakers. Note that the prompts were presented pseudo-randomly to participants during repeated elicitation in this study, which should have weakened the influence of prior responses on subsequent associates via a chaining effect. The susceptibility to the trial effect for all word classes in L2 learners may result from a less robust organization of associative structures in the L2 lexicon and/or knowing fewer vocabularies for particular semantic categories.

In short, repeated elicitations may ameliorate the issue of nativelikeness in L2 WAT studies. However, the determination of number of elicitations in repeated WAT is rather arbitrary in previous work, such as three in Wolter (2002), five in Randall (1980), and up to 12 in Kruse et al (1987). Our findings warn against over-elicitation. Not only may eliciting many responses from participants pose significant cognitive demands even for native speakers (Wolter, 2002), increases in within-group idiosyncrasy in subsequent elicitations may result in reduced reliability in group comparisons. This study shows that the concern about trial-related increases in within-group idiosyncrasy is comparably applicable to all word classes. Note that the lack of a word class effect on idiosyncrasy is unexpected: we had predicted that adjectives would elicit lower idiosyncrasy than nouns and verbs. On the one hand, the absence of word class effects may result from the control this study placed on the cues, such as set size of associates, which may moderate the influence on idiosyncrasy. On the other hand, despite the tendency for adjectives to elicit antonyms as stereotypical associates, the absence of word class effects on within-group idiosyncrasy may suggest the existence of natural variation inherent in associative behaviors within a language community that go beyond the constraints of word class. Under the framework of spreading activation, words are interconnected as nodes within a semantic memory network (Collins & Loftus, 1975). During WAT, the activation of a word (i.e., cue) spreads within the network and activates other words (i.e., associates) that co-occur frequently and overlap in meanings. An important feature of the model is that the spread is initially strong and diminishes as the activation propagates out of the center of the semantic neighborhood (Sheng et al., 2012). The initial activation may be more representative of lexical-semantic knowledge and experience shared by a language community; the associates far from the center of the semantic neighborhood may be more diffuse and thus more idiosyncratic. The systematic trial-related increases in idiosyncrasy suggests the amount of diffusion in subsequent activations may be comparable among adjectives, nouns, and verbs.

Previous work on word association highlights considerable variation in L2 learners (Meara, 1983) and native speakers (e.g., Fitzpatrick, 2007; Wolter, 2002). This study shows that the amount of idiosyncrasy in native speakers is consistently lower than in L2 learners for adjectives, nouns, and verbs, regardless of elicitation trials. Relative to native speakers, the semantic knowledge may be more diffuse in L2 learners, whose knowledge about the conventional usage of words is incomplete and influenced by learning experience that may vary greatly among learners. Interestingly, though the L2 learners were from a more diverse cultural background, with some participants originated from Taiwan while others from the mainland China, they did not exhibit a much-elevated trial-related increase in idiosyncrasy than the native speaker control. On the one hand, a lower within-group idiosyncrasy suggests that native speakers share a more tightly defined semantic representation space than L2 learners. On the other hand, the lack of a much-elevated trial-related increase in idiosyncrasy in L2 learners may point to the existence of natural variation inherent in human associative behaviors.

4.3 Educational Implications and Future Studies

The inconclusive findings in L2 WAT studies (Fitzpatrick et al., 2015) may result from the significant variation in native speakers' responses to WAT. To ameliorate the problematic nature of native norm lists in L2 WAT studies, Fitzpatrick (2007) suggested the use of individual profiling, which focuses on comparing how L2 learners respond to the same cues in L1 and L2 instead of comparing L2 learners to native speakers. This innovative approach may be problematic for several reasons. First, from a functional perspective, learners do not learn L2 to communicate with their L1-speaking community. As a result, profiling performance in L1 may have limited educational utility. Second, translated equivalents in L1 and L2 do not necessarily share the same word frequency and word collocation in respective languages, which have significant influence on word associates (Fitzpatrick, 2007; Meara,

1983). Furthermore, individual profiling requires language teachers to know the L1 of the L2 learners, which can be logistically challenging. Higginbotham' (2010) individual profiling does not sample performance of L1 or L2 but responses to L2 words of different lexical properties, such as word frequency. Both Fitzpatrick (2007) and Higginbotham (2010) highlight the needs to move away from treating learners (or native speakers) as homogenous groups, which this study supports. Findings from this study suggest potential benefits of increasing the number of adjectives cues, analyzing the composition of synonymous antonymous associates to adjective cues, and obtaining additional elicitations in WAT.

Besides these changes, a potential alternative to the problems of native norm lists in L2 WAT studies would be the use of *both* norm lists collected locally and established corpus. Large-scale norming database is valuable in L2 studies because its sampling size improves the reliability in indexing the boundary of nativelikeness. From a practical perspective, it is not feasible to develop a large-scale norming database just to evaluate a particular group of L2 learners. By comparing large-scale norming database to the norm list collected from native speaker controls, educators and researchers may better define natural variation in nativelike behaviors that allows a more precise evaluation of L2 learners. For example, this study suggests greater discrepancy for nouns and verbs than adjectives in native speakers' responses to WAT, which should be considered in evaluating L2 performance. The overlap between USF and the responses collected from the native speaker controls also highlights the associates that may be less influenced by individual, geographical, or cultural factors on word usage. This study is limited by recruiting only Mandarin-speaking learners of English and using self-reported measures of language proficiency. However, recruiting only Mandarin-speaking learners improves homogeneity in culture and ethnicity, which may influence word associates produced. The fact that native speaker controls are not more "nativelike" than L2 learners for nouns and verbs during the first elicitation despite distinct cultural backgrounds and language proficiency highlights the importance to integrate large-scale association norm and norm lists from local native speaker controls.

Though previous work highlights the importance of nativelikeness in L2 word association studies (Schmitt, 1998) and this study suggests multiple ways to ameliorate the nativelikeness problems in evaluating L2 learners, researchers and educators should maintain cautions in language-group comparisons; there are significant differences between the processes of native language and L2 acquisition. Language aptitude of native language and L2 (for a review, Skehan, 1991), motivation (Dörnyei, 2014; Gardner, 1985), learners' belief (Horwitz, 1987), and other factors (e.g., anxiety; Horwitz, Horwitz, Cope, 1986), contribute to individual differences in achievement of L2 proficiency (for a review, see Ellis, 2004). In vocabulary development, native language influences L2 learning; L2 learners often learn new words in L2 via translations (Kroll, Van Hell, Tokowicz, & Green, 2010). Differences in native language and L2 acquisition processes may impact the validity of direct comparisons between native speakers and L2 learners. This issue should always be considered when evaluating L2 performance with the model of native speakers.

5. Conclusion

Native speakers are not always easily distinguishable from L2 learners in tasks purported to measure semantic network knowledge. We demonstrated that by including a variety of word classes and additional elicitations, differentiation between groups can be achieved. Future studies should continue to refine the analytic approach, examine multiple L2 groups, and include objective measures of language proficiency as predictors in modelling to tease apart the interaction among language proficiency, word class, and trials in eliciting norm-like performance in word association task.

References

Adams, A. K. & Bullock, D. (1986). *Apprenticeship in word use: Social convergence processes in learning categorically related nouns.* In The development of word meaning (pp. 155-197). Springer New York. https://doi.org/10.1007/978-1-4612-4844-6_7

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Maechler, M. & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes.*

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and cognition, 12*(1), 3-11. https://doi.org/10.1017/S1366728908003477

Brysbaert, M. & New, B. (2009). 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*(4), 977-990. https://doi.org/10.3758/BRM.41.4.977

Burnage, G. (1990). *CELEX: A Guide for Users.* Nijmegen: SSN.

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review, 82*(6), 407-428. https://doi.org/10.1037/0033-295X.82.6.407

Crutch, S. J., Williams, P., Ridgway, G. R. & Borgenicht, L. (2012). The role of polarity in antonym and synonym conceptual knowledge: Evidence from stroke aphasia and multidimensional ratings of abstract words. *Neuropsychologia, 50*(11), 2636-44. https://doi.org/10.1016/j.neuropsychologia.2012.07.015

Deese, J. (1962). Form class and the determinants of association. *Journal of verbal learning and verbal behavior, 1*(2), 79-84. https://doi.org/10.1016/S0022-5371(62)80001-2

Deese, J. (1964). The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior, 3*(5), 347-357. https://doi.org/10.1016/S0022-5371(64)80001-3

Deese, J. (1965). *The structure of associations in language and thought*. Johns Hopkins University Press.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge. https://doi.org/10.4324/9781410613349

Ellis, R. (2004). Individual Differences in Second Language Learning. *The handbook of applied linguistics*, 525-551. https://doi.org/10.1002/9780470757000.ch21

Entwisle, D. R. (1966). Form class and children's word associations. *Journal of verbal learning and verbal behavior, 5*(6), 558-565. https://doi.org/10.1016/S0022-5371(66)80091-9

Ervin, S. M. (1961). Changes with age in the verbal determinants of word-association. *The American journal of psychology, 74*(3), 361-372. https://doi.org/10.2307/1419742

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EuroSLA Yearbook, 6*(1), 121-145. https://doi.org/10.1075/eurosla.6.09fit

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319-331. https://doi.org/10.1111/j.1473-4192.2007.00172.x

Fitzpatrick, T., Playfoot , D. , Wray , A. & Wright , M. J . (2015). Establishing the Reliability of Word Association Data for Investigating Individual and Group Differences. *Applied Linguistics, 36*, 23-50. https://doi.org/10.1093/applin/amt020

Gardner, R. (1985) *Social psychology and second language learning: the role of attitude and motivation*. London.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press. https://doi.org/10.2307/j.ctvc772nn

Higginbotham, G. (2010). Individual learner profiles from word association tests: The effect of word frequency. *System, 38*(3), 379-390. https://doi.org/10.1016/j.system.2010.06.010

Horwitz, E. (1987) *Surveying student beliefs about language learning. In A. Wenden & J. Rubin* (eds.), Learner strategies in language learning.

Horwitz, E. K., Horwitz, M. B. & Cope, J. (1986). Foreign language classroom anxiety. *The Modern language journal, 70*(2), 125-132. https://doi.org/10.2307/327317

Jiang, S. (2002). Chinese word associations for English speaking learners of Chinese as a Second Language. *Journal-Chinese Language Teachers Association, 37*(3), 55-70.

Kastenbaum, J., Bedore, L., Peña, E., Sheng, L., Mavis, I., Sebastian-Vaytadden, R., Rangamani, G., Vallila-Rohter, S. & Kiran, S. (2019). The influence of proficiency and language combination on bilingual lexical access. *Bilingualism: Language and Cognition, 22*, 300-330. https://doi.org/10.1017/S1366728918000366

Kroll, J. F., Van Hell, J. G., Tokowicz, N. & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition, 13*(3), 373-381. https://doi.org/10.1017/S136672891000009X

Kruse, H., Pankhurst, J. & Smith, M. S. (1987). A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9*(02), 141-154. https://doi.org/10.1017/S0272263100000449

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.

Lenth, R. V. (2016). Using lsmeans. *Journal Statistical Software, 69*, 1-33. https://doi.org/10.18637/jss.v069.i01

Meara, P. (1978). Learners' word associations in French. *Interlanguage Studies Bulletin*, 192-211.

Meara, P. (1983). Word associations in a foreign language. *Nottingham Linguistics Circular, 11*(2), 29-38.

Miller, G. A. & Fellbaum, C. (1991). Semantic networks of English. *Cognition, 41*(1), 197-229. https://doi.org/10.1016/0010-0277(91)90036-4

Namei, S. (2004). Bilingual lexical development: A Persian–Swedish word association study. *International Journal of Applied Linguistics, 14*(3), 363-388. https://doi.org/10.1111/j.1473-4192.2004.00070.x

Nelson, D. L., McEvoy, C. L. & Dennis, S. (2000). What is free association and what does it measure? *Memory & cognition, 28*(6), 887-899. https://doi.org/10.3758/BF03209337

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

Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 402-407. https://doi.org/10.3758/BF03195588

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: a review of research and theory. *Psychological bulletin, 84*(1), 93-116. https://doi.org/10.1037/0033-2909.84.1.93

Nissen, H. B. & Henriksen, B. (2006). Word class influence on word association test results 1. *International Journal of Applied Linguistics, 16*(3), 389-408. https://doi.org/10.1111/j.1473-4192.2006.00124.x

Randall, M. (1980). Word association behaviour in learners of English as a foreign language. *Polyglot, 2*(2), 1-26.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language testing, 10*(3), 355-371. https://doi.org/10.1177/026553229301000308

Read, J. (2013). *Validating a test to measure depth of vocabulary knowledge.* In Validation in language assessment (pp. 55-74). Routledge. https://doi.org/10.4324/9780203053768-9

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F. & Schlesewsky, M. (2007). To predict or not to predict: Influences of task and strategy on the processing of semantic relations. *Journal of Cognitive Neuroscience, 19*(8), 1259-1274. https://doi.org/10.1162/jocn.2007.19.8.1259

Schmitt, N. (1998). Quantifying word association responses: what is native-like? *System, 26*(3), 389-401. https://doi.org/10.1016/S0346-251X(98)00019-0

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language learning, 64*(4), 913-951. https://doi.org/10.1111/lang.12077

Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition, 13*(2), 275-298. https://doi.org/10.1017/S0272263100009979

Sheng, L., Bedore, L. M., Peña, E. D. & Taliancich-Klinger, C. (2013). Semantic convergence in Spanish–English bilingual children with primary language impairment. *Journal of Speech, Language, and Hearing Research, 56*(2), 766-777. https://doi.org/10.1044/1092-4388(2012/11-0271)

Sheng, L., McGregor, K. K. & Marian, V. (2006). Lexical–semantic organization in bilingual children: Evidence from a repeated word association task. *Journal of Speech, Language, and Hearing Research, 49*(3), 572-587. https://doi.org/10.1044/1092-4388(2006/041)

Sheng, L., Peña, E. D., Bedore, L. M. & Fiestas, C. E. (2012). Semantic deficits in Spanish–English bilingual children with language impairment. *Journal of Speech, Language, and Hearing Research, 55*(1), 1-15. https://doi.org/10.1044/1092-4388(2011/10-0254)

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope? *System, 30*(3), 315-329. https://doi.org/10.1016/S0346-251X(02)00017-9

Zareva, A. & Wolter, B. (2012). The 'promise'of three methods of word association analysis to L2 lexical research. *Second Language Research, 28*(1), 41-67. https://doi.org/10.1177/0267658311423452

Zhang, D. & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(1), 1. https://doi.org/10.1186/s40862-017-0024-0

Appendix A. Cues used in the repeated word association task and the associated lexical-semantic and associative properties according to established corpora and the University of South Florida Association Norm.

Word class	Cue	Most Dominant Associate (MDA)[1]	Proportion of respondents generating MDA[1]	Number of different associates produced by more than one respondent[1]	Word frequency of cue[2]	Age of acquisition[3]
Adjective	angry	mad	0.52	9	59.0	4.5
Adjective	easy	hard	0.58	8	265.7	5.4
Adjective	fat	skinny	0.23	16	79.4	5.2
Adjective	happy	sad	0.63	8	333.2	2.7
Adjective	rainy	day	0.23	11	3.8	4.0
Adjective	sick	ill	0.35	13	165.4	4.1
Adjective	sleepy	tired	0.66	10	8.6	3.4
Adjective	sticky	glue	0.18	16	5.7	4.1
Adjective	thirsty	water	0.31	13	12.3	3.9
Adjective	wide	narrow	0.18	15	23.8	5.8
Noun	car	auto	0.13	25	483.1	3.4
Noun	chair	table	0.31	14	49.2	3.4
Noun	corn	cob	0.33	14	14.2	4.6
Noun	eye	see	0.36	13	111.8	3.8
Noun	fork	spoon	0.44	8	8.8	3.6
Noun	horse	ride	0.26	16	92.9	4.2
Noun	juice	orange	0.65	10	26.9	4.4
Noun	ocean	sea	0.29	17	30.3	4.7
Noun	spider	web	0.25	12	10.1	3.4
Noun	teacher	student	0.19	12	55.7	4.6
Verb	bring	take	0.30	11	327.2	4.4
Verb	catch	fish	0.16	18	135.5	4.6
Verb	draw	picture	0.22	15	40.4	4.1
Verb	eat	food	0.41	11	251.9	2.8
Verb	jump	rope	0.23	12	69.8	2.8
Verb	pull	push	0.59	10	146.5	4.8
Verb	read	book	0.39	9	241.2	4.1
Verb	sell	buy	0.54	14	92.3	7.1
Verb	sing	song	0.46	17	97.6	3.5
Verb	stand	sit	0.53	10	226.2	4.4

1. University of South Florida Association Norm (Nelson et al., 1998, 2004).
2. Word frequency was defined as how often a word occurs per million words in a corpus of 51 million words based on television and film subtitles (Brysbaert & New, 2009).
3. Age of acquisition was based on the age-of-acquisition ratings for 30,000 English words (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)