Studies in Second Language Acquisition **40** (2018), 831–856
doi:10.1017/S0272263118000141

Research Article

STATISTICAL SENSITIVITY, COGNITIVE APTITUDES, AND PROCESSING OF COLLOCATIONS

Wei Yi*

University of Maryland, College Park

Abstract

Frequency and contingency (i.e., co-occurrence probability of words in multiword sequences [MWS]) are two driving forces of language acquisition and processing. Previous research has demonstrated that L1 and advanced L2 speakers are sensitive to phrasal frequency and contingency when processing larger-than-word units. However, it remains unclear whether such statistical sensitivity is robust across tasks and among subcategories of MWS. In addition, little is known about whether cognitive aptitudes can moderate such sensitivity. This study examined L1 and advanced L2 speakers' statistical sensitivity to phrasal frequency and contingency as well as cognitive aptitudes' moderating effects on such sensitivity when processing English adjective-noun collocations. Participants performed a phrasal acceptability judgment task (PJT). Meanwhile, their aptitude profiles were measured by six aptitude tests, which loaded separately onto implicit language aptitude, explicit language aptitude, and working memory capacity. Linear mixed-effects modeling revealed that both L1 and L2 English speakers were sensitive to phrasal frequency and contingency of collocations, although L2 speakers' sensitivity was much stronger than that of L1 speakers. None of the aptitudes was found to moderate language users' statistical sensitivity to either collocation frequency or contingency. Interestingly, disassociation patterns between the PJT performance and the involvement of implicit or explicit language aptitude among the L1 and L2 speakers were found. It was concluded that L1 and L2 speakers

This article is based on the author's qualifying paper, supported by the PhD program in Second Language Acquisition at the University of Maryland, College Park. I would like to express my gratitude to my supervisor Dr. Michael Long for his tremendous support for this project. I would also like to thank Drs. Nan Jiang, Robert DeKeyser, and Steven Ross for their insightful suggestions that greatly improved the research. I am immensely grateful to Dr. Gisela Granena for her advice on the use of the LLAMA test, and to Dr. Scott Kaufman for giving me the access to the serial reaction time task. The data collection was generously supported by Lars Bokander, Anxin Bai, and Wenbo Li, and I would like to thank them. I also thank Dr. Stefano Rastelli for his constructive comments, as well as Dr. Michael Long, Nicco Cooper, Jason Struck, and Zhiyuan Deng for their proofreading of the manuscript.

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://www.iris-database.org/iris/app/home/detail?id=york%3a934519&ref=search.

*Correspondence concerning this article should be addressed to Wei Yi, 3215E Jimenez Hall, University of Maryland, College Park, 20742, Maryland. E-mail: weiyi@umd.edu

Copyright © Cambridge University Press 2018

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

differed in terms of the way they processed the collocations, as well as the nature of their collocational knowledge.

BACKGROUND

Language consists of units of various sizes and can be represented at different levels. From the perspective of usage-based approaches (Bybee, 1998; Ellis, 2003; Goldberg, 1995; Langacker, 1987; Tomasello, 2003), constructions—which are form-meaning mappings that relate specific linguistic patterns to certain semantic, pragmatic, and discourse functions (Goldberg, 1995)—are the building blocks of language. Constructions include linguistic units of different grain sizes, and bridge the gap between words and rules, in that they can function as both lexical and grammatical units. Among various types of constructions, multiword sequences (MWS) have attracted an increasing amount of attention in recent years. MWS refer to word sequences that co-occur more frequently than by chance and cover a variety of linguistic phenomena, such as idioms (*kick the bucket*), phrasal verbs (*take off*), speech formulae (*What's up?*), irreversible binomials (*bride and groom*), collocations (*make progress*), and lexical bundles (*is one of the*). Although subcategories of MWS differ in terms of length, idiomaticity, or compositionality, they are ubiquitous (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Erman & Warren, 2000) and play a critical role in the development of nativelike L2 proficiency (Pawley & Syder, 1983).

Language is rich with different kinds of distributional information, including frequency, variability, and co-occurrence probability (Erickson & Thiessen, 2015). The human mind is sensitive to such statistics. By relying on distributional statistics, children and adults can decipher the underlying structural regularities of language (Ellis, 2006a, 2006b). Such a process is usually called "statistical learning" (Ellis, 2008; Frost, Armstrong, Siegelman, & Christiansen, 2015). Current research has revealed that statistical learning is a mechanism responsible for phonological learning (e.g., Maye, Weiss, & Aslin, 2008), word segmentation (e.g., Saffran, Aslin, & Newport, 1996), syntactic learning (e.g., Thompson & Newport, 2007), and category formation (e.g., Gómez & Gerken, 2000). Moreover, statistical learning functions in both children (e.g., Gómez & Gerken, 2000; Saffran et al., 1996) and adults (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Zuhurudeen & Huang, 2016), and in both first (e.g., Saffran et al., 1996) and second language acquisition (e.g., Frost et al., 2015; Hamrick, 2014; Rastelli, 2014).

Two types of statistical information are said to play important roles in the acquisition and processing of MWS, namely, phrasal frequency and contingency (Gries & Ellis, 2015). Phrasal frequency indicates how often a word sequence is used, and it can be operationalized as the number of occurrences of the word combination in corpora. Contingency refers to the co-occurring probability of words that construct the sequence, which is operationalized as the strength of the statistical association between constituent words within MWS. Current literature suggests that the human mind can acquire knowledge of statistical correlations between stimulus pairings, or the predictive relationships between stimuli and outcomes (Schmidt, 2012). As argued by

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Ellis (2006a, 2006b), language acquisition can be understood as contingency learning, in that learners must determine the reliability of form-meaning/function mappings or the strength of the statistical association between linguistic elements (Gries & Ellis, 2015). Using contingency information, language users can arrive at the interpretations that are most relevant to the context and predict what is most likely to be heard or seen next.

STATISTICAL SENSITIVITY TO PHRASAL FREQUENCY

Frequency is one of the most robust statistics to which language users are sensitive. Frequency indicates the likelihood of a construction to be experienced by language users. Furthermore, it determines the degree of entrenchment of a construction stored in the mind, as well as the degree of automaticity when it is retrieved (Gries & Ellis, 2015). Language users are intimately tuned to input frequency, and frequency effects exist in the processing of many aspects of language (Diessel, 2007; Ellis, 2002; Jurafsky, 2003).

Usage-based approaches hold that language usage shapes the acquisition of language at every level. Current research has demonstrated that high-frequency words are processed faster than low-frequency ones (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Duyck, Vanderelst, DEsmet, & Hartsuiker, 2008). Similar to language processing at the single word level, effects of MWS frequency (phrasal frequency) also has been found in L1 and L2 speakers, and across subcategories of word combinations. For instance, studies have found that L1 speakers are sensitive to the whole-string frequency of MWS including lexical bundles (Arnon & Snider, 2010; Tremblay & Baayen, 2010; Yi et al., 2017) and collocations (Durrant & Doherty, 2010). Arnon and Snider (2010) divided four-word English compositional expressions with varying levels of frequency into three-frequency bins. Analyses of the behavioral data collected from a phrasal acceptability judgment task (PJT) suggest that frequent MWS were processed significantly faster than less frequent control phrases. Moreover, such a processing advantage was observed in all frequency bins.

Sensitivity to phrasal frequency of MWS has also been observed in L2 speakers. Wolter and Gyllstad (2013) carried out a study in which Swedish learners of English were required to judge whether certain adjective-noun collocations presented on a computer screen exist in English. Unknown to the participants, the collocations fell into two categories: Some had word-by-word translations in Swedish (congruent collocations), while others (incongruent collocations) did not have direct word-by-word L1 translations. Advanced Swedish-English speakers were found to be sensitive to collocation frequency, as they responded faster to more frequent collocations than less frequent ones. Moreover, such frequency effects were independent of the congruency status of the collocations. In another study (Yi et al., 2017), L1 and advanced L2 speakers of Mandarin read 80 Chinese disyllabic adverbial sequences embedded in sentence contexts, with their eye movements recorded by an eye tracker. Statistical analyses based on fixation durations showed that both native and nonnative Chinese speakers were sensitive to the phrasal frequency of the Chinese adverbial sequences.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

STATISTICAL SENSITIVITY TO CONTINGENCY

Frequency plays a crucial role in the processing of MWS, but it is not the only determining factor (Ellis, 2008). MWS consist of multiple words in a sequence, and the co-occurrence of the constituent words is not by chance. As already mentioned, probabilistic relationships in MWS can be understood as contingency, and captured by different corpus metrics such as forward/backward transitional probability (McDonald & Shillcock, 2003; Tremblay & Baayen, 2010), mutual information (MI) (Church, Gale, Hanks, & Hindle, 1991; Durrant & Doherty, 2010; Ellis, Simpson-Vlach, & Maynard, 2008; Yi et al., 2017), *t*-score (Wolter & Gyllstad, 2011), and ΔP (Gries & Ellis, 2015). In the study to be reported here, MI was used as the measure of contingency, because MI has been documented as a robust measure to which language users are sensitive. In fact, sensitivity to contingency as measured by MI has been demonstrated in various subcategories of MWS, including lexical bundles (Ellis et al., 2008; Tremblay & Baayen, 2010; Yi et al., 2017) and collocations (Durrant & Doherty, 2010). Although computed based on frequency counts, contingency as measured by MI or other corpus metrics is distinct from phrasal frequency. Briefly, phrasal frequency refers to the likelihood that language users will experience certain MWS, while contingency illustrates the reliability of the co-occurrence patterns of MWS (Gries & Ellis, 2015). Therefore, one can expect highly frequent MWS with low MI, and vice versa (for examples, see Appendix 1 in the supporting information online). Statistically, contingency and phrasal frequency should be weakly correlated. For example, the current study extracted 8,340 adjective-noun word combinations (see the "Materials" section) from the British National Corpus, yet the correlation between MI scores and collocational frequency was only .09.

Many studies have demonstrated that L1 speakers are sensitive to the contingency information underlying language input. Saffran et al. (1996) found that 8-month-olds can detect word boundaries in an artificial language after only 2 minutes of exposure, by using the transitional probability information between syllables. Similar results have also been found with L1 adults. For instance, Gregory et al. (1999) showed that probable collocations were more often shortened in duration, compared to less probable ones. McDonald and Shillcock (2003) also found that forward and backward transitional probabilities were predictive of eye-fixation durations of native English speakers during natural reading.

Little has been done to investigate L2 speakers' sensitivity to contingency information during the processing of MWS. In one study, Ellis et al. (2008) validated the psychological reality of corpus-extracted academic English formulae (e.g., *the value of the*). A series of comprehension and production tasks were used with both native and advanced nonnative speakers of English. Multiple regression analyses revealed that only the L1 speakers of English, not the L2 speakers, were sensitive to MI when performing the tasks. Such findings are interesting yet questionable due to the limited sample size and lack of control over confounding variables (e.g., constituent word frequency).

Conversely, Ellis and colleagues did detect contingency effects among L2 speakers in another study (Ellis, O'Donnell, & Römer, 2014). The researchers examined the processing of English verb-argument constructions (VAC) using two free-association tasks. Advanced L2 learners of English from various L1 backgrounds were required to

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

generate the first word that came into their mind (Experiment 1), or to generate as many verbs as possible in 1 minute (Experiment 2), to fill in the slot in the 40 VAC frames (e.g., *he ___ across the…*). In both experiments, the frequency of each type of verbs generated by participants was regressed on verb frequency in the VAC, VAC-verb contingency (the degree to which a verb is preferred by a construction, or vice versa, measured by ΔP), and verb semantic prototypicality. Statistical analyses revealed that L2 English participants in all language groups were sensitive to VAC-verb contingency.

Based on the contrasting patterns reviewed in the preceding text, it remains unclear whether L2 speakers can be sensitive to contingency information of L2 MWS. Using eye tracking, Yi et al. (2017) found significant effects for contingency in both L1 and L2 speakers, after controlling for the effects of phrasal frequency and other confounding variables. Specifically, contingency effects of lexical bundles were observed on first-pass reading time for L2 speakers, and on fixation counts for both L1 and L2 speakers. In addition, a significant interaction between contingency and phrasal frequency was found for L2 speakers. Such results suggest that language users may be sensitive to contingency information of larger-than-word units, regardless of whether they are native speakers.

COGNITIVE APTITUDES AND THEIR RELEVANCE TO SECOND LANGUAGE ACQUISITION

Aptitude is regarded as an endowment (Carroll, 1973; Skehan, 1998) distinct from other cognitive and affective factors (Li, 2016; Skehan, 2012). Aptitude constructs an important source of individual differences in second language acquisition (Dörnyei & Skehan, 2003; Granena & Long, 2013), which is widely recognized as a composite consisting of different abilities, including working memory and aptitude for implicit and explicit learning. Working memory is defined as a cognitive system consisting of a temporary storage and an attentional control component (Baddeley, 2007) responsible for temporary holding and manipulation of information (Williams, 2012). Implicit language aptitude is conceptualized as the ability to recognize and acquire patterns underlying the input through implicit induction (Granena, 2012, 2013b; Kaufman et al., 2010). In contrast, explicit language aptitude is generally relevant to explicit learning, with metalinguistic reasoning, explicit induction, or analysis involved.

Research has demonstrated that working memory plays a significant role in different domains of L2 acquisition and processing (Linck, Osthus, Koeth, & Bunting, 2014; for a review, see Williams, 2012). Meanwhile, implicit and explicit language aptitudes have also been reported to have an impact on various aspects of second language acquisition and processing. For example, Granena (2013b) found that implicit language aptitude is predictive of ultimate attainment in grammatical agreement relationships for both early and late L2 learners. Suzuki and DeKeyser (2015) also showed that implicit language aptitude correlated with performance in a L2 word-monitoring task that used Japanese particles as the target structure. Where aptitude for explicit language learning is concerned, Granena (2012) found that both pre- and post-critical-period L2 learners with high explicit language aptitude demonstrated advantages when performing linguistic tasks that allowed controlled use of language knowledge, while Suzuki and DeKeyser (2017) showed that aptitude for explicit learning significantly predicted acquisition of automatized explicit grammatical knowledge, as measured by timed grammaticality judgment.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

When it comes to MWS, current literature suggests that memory capacity seems to influence the acquisition and processing of such larger-than-word units. For example, Bolibaugh and Foster (2013) found that individual differences in L2 learners' phonological short-term memory affected both the rate of learning and the ultimate attainment of native selection of word combinations. Tremblay and Baayen (2010) showed that working memory capacity predicted L2 participants' immediate recall of visually presented sequences. However, the role of implicit and explicit language aptitude in the acquisition and processing of MWS is less clear. Studies (e.g., Webb & Kagimoto, 2009) have demonstrated that MWS such as collocations can be learned through explicit instruction. Nevertheless, other researchers (e.g., Bolibaugh & Foster, 2013) hold that advanced command of MWS is less likely to depend on deliberate and conscious processes, given their ubiquity in language. Instead, one's grasp of MWS may be achieved mainly through implicit, meaning-focused interaction in social contexts (Robinson & Ellis, 2008). Such views have also been supported empirically. For instance, Webb et al. (2013) found that collocations can be learned incidentally through reading after as few as two encounters. Additionally, Granena and Long (2013) found significant correlations between language aptitude, as measured by the LLAMA (Meara, 2005), and scores on tests of lexis and collocations for post-critical-period L2 learners. Moreover, it was scores on subtest D of the LLAMA, shown to measure aptitude for implicit language learning (Granena, 2013a), that were most strongly correlated with the lexis and collocations scores.

THE CURRENT STUDY

As reported in the previous section, L1 and advanced L2 speakers are sensitive to phrasal frequency and contingency when processing larger-than-word units. However, it remains unclear whether such sensitivity is robust across experimental tasks and different subcategories of MWS. Although the current literature suggests that cognitive aptitudes, including working memory and implicit/explicit language aptitudes, may influence the acquisition and processing of MWS, few studies have been carried out on this topic. Moreover, no research has been carried out to determine whether cognitive aptitudes can moderate the degree of language users' statistical sensitivity to phrasal frequency and contingency. The current study aimed to test the robustness of language users' statistical sensitivity to phrasal frequency and contingency of English adjective-noun collocations, adopting a PJT. Cognitive aptitudes, including working memory capacity and aptitudes for implicit and explicit learning were also measured, so that their effects on PJT performance could be assessed, along with their interaction with the two types of statistical information during the processing of collocations. The following research questions were addressed:

1. Are L1 and/or L2 speakers of English sensitive to the phrasal frequency of collocations?
2. Are L1 and/or L2 speakers of English sensitive to the contingency (measured by MI) of collocations?
3. Do cognitive aptitudes interact with phrasal frequency and contingency during the processing of English collocations? If so, will L1 and/or L2 English participants' statistical sensitivity to phrasal frequency and contingency be moderated by their aptitude profiles?

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

4. Do cognitive aptitudes (i.e., working memory capacity, implicit and explicit language aptitude) influence L1 and/or L2 participants' PJT performance?

METHODOLOGY

PARTICIPANTS

Two groups of participants were recruited from a mid-Atlantic state university in the United States. The L1 group consisted of 30 native speakers of English (7 males, 23 females), and the L2 group were 32 (9 males, 23 females) Chinese learners of English. The L1 (M = 23.1 years, SD = 8.5) and L2 English speakers (M = 23.6, SD = 2.7) were comparable in terms of their age (Welch two-sample t test: $t(50.68) = .98$, $p = .33$). To participate in this study, L2 English speakers needed (a) to have lived in the United States for at least one year by the time the study started, and (b) to have taken TOEFL or IELTS. L2 English speakers' average length of residence in the United States was 25.8 months (SD = 14.3), and their average age of onset (AO) for learning English was 9 (SD = 2.3). L2 participants all started learning English no later than age 13. Thus, they may be classified as early starters for the acquisition of lexis and collocations based on current SLA literature (e.g., DeKeyser, 2000; Granena & Long, 2013). Nevertheless, Pearson correlation analyses revealed statistically nonsignificant correlations between AO and TOEFL scores (r = -.06, t = -.34, df = 27, p = .74), as well as between years of instruction and TOEFL scores (r = .19, t = 1.03, df = 27, p = .31). After interviewing eight L2 participants who started learning English earliest, it was found that none of them learned English in immersive or naturalistic environments, and most teaching was explicit. Given the explicit nature of the EFL classes in China, such early experience would have been of very limited benefit, an inference supported by the statistically nonsignificant correlation between AO and TOEFL scores mentioned previously. Based on these preceding considerations, the Chinese participants were classified as late L2 English learners.

Regarding L2 proficiency, 29 L2 participants reported their most recent TOEFL iBT score (M = 100.0, SD = 8.3). TOEFL iBT scores range from 0 to 30 for each of the four test sections (Reading, Listening, Speaking, and Writing). According to ETS, scoring above 22 to 26 on any of the skills places the test taker at the highest levels (ETS, 2017). For the purposes of the study, 30 L2 participants also took a 30-item cloze test (Bachman, 1985). Taking the scores of TOEFL iBT and the cloze test together, the Chinese-English bilinguals were regarded as advanced L2 English learners. A summary of their demographic information is presented in Table 1.

MATERIALS

One hundred and eighty English adjective-noun collocations were extracted from the British National Corpus (BNC) through the Phrases in English (PIE) database (Fletcher, 2011) and used as the critical stimuli for the PJT. Adjective-noun collocations were targeted following the practice of Wolter and Gyllstad (2013) because variability in determiners present in verb-noun combinations (e.g., *make a mistake* vs. *make progress*) does not exist in such combinations, thus allowing for more control over item

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

TABLE 1. L2 speakers' demographic information

Characteristics	*M*	*Min*	*Max*	*SD*
Age	23.6	18	28	2.7
Age of onset	9.0	5	13	2.3
Years of L2 instruction	14.6	7	22	3.5
Length of residence	25.8	12	66	14.3
TOEFL iBT score	100	70	110	8.3
Cloze test score	21.2	10	28	4.7

Note: TOEFL iBT scores were collected from 29 participants. Cloze test scores were collected from 30 participants.

consistency. Adopting a corpus-based approach, collocations were operationalized as word sequences consisting of two or more words that co-occur more frequently than would be predicted by chance, given the frequency of their constituent words (Wolter & Gyllstad, 2011, 2013).

The following procedures were followed when constructing the materials. First, 8,340 adjective-noun sequences were extracted from the PIE database. The raw frequencies of the sequences and their constituent words were then normalized (converted to number of occurrences per million words) and transformed using natural log. MI values of the adjective-noun sequences were also computed. Second, thresholds were set for logged frequency and MI of the sequences at 0.1 and 3.0, respectively. A sequence in the pool qualified as a collocation if it appeared at least once per million words in the BNC, and if the statistical association between its constituent words as measured by MI was higher than 3.0. Third, a stratified sampling procedure was carried out, leading to 180 collocations being selected. Specifically, three discrete bins (Siyanova-Chanturia & Spina, 2015) were formed for collocation frequency (high: 2.0–4.0; medium: 1.1–1.6; low: 0.1–0.7) and MI values (high: 8.0–13.0; medium: 6.8–8.0; low: 3.0–6.0). Twenty collocations were then extracted from the pool to represent each of the nine levels formed by collocation frequency and MI. Familiarity ratings based on a five-point scale, from totally unknown to extremely familiar, were then collected from five advanced Chinese English L2 speakers who did not participate in this study. The average familiarity rating was 4.5 ($Min = 3.4$, $Max = 5$, $SD = 0.4$).

Filler items consisting of 180 implausible word combinations (e.g., *popular hour*, *religious morning*) were created by randomly matching a list of adjectives to another list of nouns. Words in the lists of adjectives and nouns were taken from those that appeared in the collocations. All filler items were then checked against the BNC to ensure that they did not appear in the corpus. The characteristics of the collocations are summarized in Table 2.

THE PHRASAL ACCEPTABILITY JUDGMENT TASK

A phrasal acceptability task was used to assess language users' online processing of English collocations. The items were presented one at a time in random order. Participants were required to judge whether a word combination is used in English by pressing the button as accurately and as fast as possible. The task began with a fixation

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

TABLE 2. Summary of characteristics of the collocations

Frequency bin	MI bin	Collocation frequency *M* (*SD*)	MI *M* (*SD*)	Word1 frequency *M* (*SD*)	Word2 frequency *M* (*SD*)	Length *M* (*SD*)
high	high	3.0 (0.4)	9.5 (1.2)	6.0 (0.8)	5.8 (0.9)	10.0 (2.7)
high	medium	2.4 (0.4)	7.1 (0.6)	5.6 (0.8)	5.8 (0.5)	12.1 (2.2)
high	low	2.1 (0.1)	5.2 (0.7)	5.6 (0.9)	5.5 (0.8)	11.2 (3.3)
medium	high	1.3 (0.1)	9.4 (1.1)	4.6 (1.0)	4.1 (0.7)	14.7 (3.2)
medium	medium	1.3 (0.1)	6.9 (0.6)	5.4 (0.5)	4.9 (0.7)	11.9 (2.8)
medium	low	1.3 (0.1)	4.6 (0.9)	6.1 (0.8)	5.8 (0.8)	10.2 (2.0)
low	high	0.4 (0.2)	9.6 (1.3)	3.7 (0.9)	3.9 (0.7)	13.6 (2.7)
low	medium	0.3 (0.2)	7.1 (0.6)	4.9 (0.8)	4.4 (0.8)	13.2 (2.6)
low	low	0.3 (0.2)	5.1 (0.8)	5.3 (0.8)	5.3 (1.0)	12.1 (3.1)

Note: Frequencies were normalized and logged (natural log). Length refers to the number of letters in the collocation.

cue consisting of a row of eight asterisks presented at the center of the monitor for 500 milliseconds. Following this was an item to which the participant had to respond, with a timeout of 4,000 milliseconds and no feedback. Reaction time and accuracy were automatically recorded by DMDX (Forster & Forster, 2003).

MEASURES OF COGNITIVE APTITUDES

The LLAMA Test

The LLAMA test (Meara, 2005) is a language-independent verbal aptitude test battery that has been reported to be relatively reliable (e.g., Granena, 2013a; Rogers, Meara, Barnett-Legh, Curry, & Davie, 2017). In addition, exploratory factor analyses reported by Granena (2013a) showed that the four LLAMA subtests (LLAMA B, LLAMA D, LLAMA E, LLAMA F) load on two dimensions of domain-general aptitude, with LLAMA D associated with implicit language aptitude, and the others with explicit language aptitude. LLAMA E is a test of the ability to associate sounds and their written form. Hence, it was not considered a measure primarily related to the processing of collocations. LLAMA B is a vocabulary-learning test in which participants are allowed 2 minutes to learn the names of as many of 20 images presented on the computer as they can. LLAMA D is a sound recognition test, which requires participants to first listen to 10 computer-synthesized sound sequences in a British-Columbian Indian language and then to judge whether certain sequences have been heard in the previous listening session. LLAMA F is a grammatical-inferencing test. In this task, participants are allowed 5 minutes to learn a set of 20 sentences, each associated with an image describing it. In the following testing session, their task is to choose the correct sentence out of two choices (grammatical vs. ungrammatical) for each image. Possible scores ranged from 0 to 100 for LLAMA B and LLAMA F, and from 0 to 75 for LLAMA D. The reliability of two of the three LLAMA subtests employed in the study was acceptable (LLAMA B: Cronbach's alpha = .79, k = 59; LLAMA F: Cronbach's alpha = .73, k = 60), but that of LLAMA D less so (Cronbach's alpha = .46, k = 59).

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Three-Term Contingency Learning (3-Term) Task

This task is designed to measure explicit associative learning ability (Kaufman et al., 2010). It consists of four blocks, each block containing a learning session and an immediate test session. Within each block, participants are presented with 10 unique stimulus words (e.g., "*FAR*"), each associated with three different outcome words (e.g., "*SAP*," "*COD*," "*PUG*") depending on which key (A, B, C) is pressed. Participants are required to learn the stimulus-outcome word associations. They are tested by typing the outcome words corresponding to each stimulus word under a given cue (A, B, or C). Feedback is provided to answers given by participants in the testing session. The duration of exposure to each association is self-paced, with a timeout of 2.5 seconds. Across the blocks, the stimulus and outcome words are the same, yet the trial order is randomized. The overall scores range from 0 to 120, and the reliability of this task was acceptable (split-half reliability measured by Spearman-Brown coefficient: .62).

Serial Reaction Time Task

A probabilistic version of the serial reaction time (SRT) task used by Kaufman et al. (2010) was adopted. This task is recognized as a measure of aptitude for implicit language learning (Granena, 2012, 2013b; Kaufman et al., 2010). A sequence of visual stimuli can appear at one of four positions (designated by "V," "B," "N," and "M" on the keyboard) arranged from left to right on the computer screen. Participants' task is to press the key corresponding to the position of the visual stimuli as accurately and as fast as possible. Unknown to participants, the position of the visual stimuli followed two possible sequence patterns, one probable and the other improbable. The probable sequence (1–2–1–4–3–2–4–1–3–4–2–3) occurred with a probability of .85, and the improbable sequence (3–2–3–4–1–2–4– 3–1–4–2–1) with a probability of .15. The two sequences produced no difference in terms of first-order transition probability (i.e., the probability of occurrence in each location); yet they differed exclusively in terms of second-order transition probability (the probability of the third position, given the two positions prior to it). The two sequences were intermixed such that the transition of the position of the visual stimuli switched between the two sequences. Generally, if a learning effect happens, reaction time for probable trials will be less than that for improbable trials. The current SRT task consisted of one practice block and eight learning blocks, with 120 trials in each block. Following the scoring method developed by Kaufman et al. (2010), the scores for each block were summed and used as the total score for each participant, ranging from 0 to 6. Split-half reliability measured using the Spearman-Brown correction was .59 (k = 55). Compared with previous studies (Granena, 2013b; Kaufman et al., 2010; Suzuki & DeKeyser, 2015), this reliability was considerably higher.

Operation Span Task

Operation Span Task (OSPAN) (Stone & Towse, 2015) is a verbal span test that measures working memory capacity. This dual-task test involves a repeated cycle of memory and processing components. Participants first saw a series of individually

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

presented integers (ranging from 10 to 99) to be stored and recalled in the correct order by the end of each trial. Immediately after the presentation of each integer, participants saw a mathematical operation (e.g., "7 + 11 = 36") and had to indicate whether they judged it correct or incorrect by pressing a key. The digits and operations were randomly generated for each trial, and the span size of the integers ranged from 2 to 7. Participants were required to respond as fast and as accurately as possible. A percentage scoring method was used. The reliability of the OSPAN task was acceptable (Spearman-Brown split-half coefficient: .80, $k = 55$).

PROCEDURES

Participants were tested individually or in groups of two at a time in a lab. On entering the lab, they signed the consent form and were given a brief manual containing instructions for each task. All tasks were administered on computers. The experiment started with the PJT, followed by six cognitive aptitude tests, the order counterbalanced for each participant. There was a pause to let participants rest after finishing the first half of the PJT task. Cognitive aptitude tasks were divided into blocks of two, and participants were given a 5-minute break after completing each block. When administering the LLAMA tests, note taking was not allowed. After completing all the tasks, participants were administered a brief survey to obtain demographic information. The whole experiment lasted for about 1 hour and 40 minutes. Participants received $20 for their participation.

STATISTICAL ANALYSES

PRINCIPAL COMPONENTS ANALYSIS

Principal Components Analysis (PCA) was conducted using SPSS 20 to determine the latent relationships between different aptitude tests, and to reduce the number of aptitude predictors in the following mixed-effects modeling. PCA investigates the correlations among variables, and such correlations are impacted by reliability. When reliability is low, correlations are reduced, hence diminishing the magnitude of the components (Kanyongo, 2005). A .70 cutoff is considered acceptable in social sciences for all scales of reliabilities (Lance, Butts, & Michels, 2006), and a more lenient .60 cutoff is sometimes used (Granena, 2013a). Participants in this study performed poorly in LLAMA D (average score: 29.5 out of 75), and LLAMA D only had a reliability coefficient of .46. Therefore, LLAMA D was excluded from the PCA.[1] Given that cognitive aptitudes are regarded as interrelated, the oblique rotation method (Direct-Oblimin) was used. Furthermore, Jolliffe's criterion of eigenvalue greater than .70 (Loewen & Gonulal, 2015), along with cumulative percentage of explained variance and the scree plot, were employed to determine the number of components extracted from the dataset. The Kaiser-Meyer-Olkin measure of sampling adequacy was .76, and the Barlett's test of sphericity was significant ($p < .001$).

The PCA resulted in three components with eigenvalues higher than .70. The total proportion of variance explained by the three components was 79.79%. The first component had an eigenvalue of 2.37 and accounted for 47.34% of the variance. The second component had an eigenvalue of 0.96 and accounted for an additional 19.28% of

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

the variance. The third component has an eigenvalue of .78 and accounted for another 13.17% of the variance. The rotated pattern matrix showed that three tests loaded on the first component with loadings greater than .3: LLAMA B (λ = .919), LLAMA F (λ = .769), and 3-Term (λ = .633). Regarding the second component, only the SRT (λ = .985) loaded strongly on it. Lastly, only the OSPAN (λ = .949) loaded strongly on the third component. Current research has shown that the SRT task may relate to implicit learning (Granena, 2012, 2013b; Kaufman et al., 2010), while LLAMA B and LLAMA F (Granena, 2013a), along with the 3-Term test (Kaufman et al., 2010), are likely to tap explicit learning. In contrast, the OSPAN task has been generally used as a measure of working memory capacity. Given this literature and the factor loading patterns revealed previously, the three components were named implicit language aptitude (SRT), explicit language aptitude (LLAMA B, LLAMA F, 3-Term), and working memory (OSPAN), respectively. Composite scores for the aptitude components were then computed using the regression-weighted method in SPSS and used for mixed-effects modeling in the following section. Descriptive statistics for all aptitude measures, as well as correlations between different aptitude measures, were provided in Appendix 2 and Appendix 3, respectively, in the supporting information online.

MIXED-EFFECTS MODELING

Mixed-effects models were built to analyze the PJT data, using the *lme4* package (version 1.1-12, Bates et al., 2015) in R (version 3.3.1, R Core Team, 2016). Specifically, linear mixed-effects models were constructed for reaction time, whereas mixed-effects logistic models were constructed for accuracy. For the linear mixed-effects models, reaction time was the dependent variable; for the mixed-effects logistic model, accuracy (1 = correct; 0 = wrong) was the dependent variable. For both groups of statistical models, independent variables include speaker (L1 vs. L2 English speakers), collocation frequency, MI, as well as explicit language aptitude, implicit language aptitude, and working memory capacity. Word1 frequency, Word2 frequency, and length were treated as covariates. Speaker was dummy coded with L2 speakers as the reference group. To achieve normality, reaction time, as well as the frequencies, were transformed using natural log. Medium correlations were found between working memory capacity and explicit language aptitude ($r = .432$, $p < .001$), as well as between Word2 frequency and length ($r = -.355$, $p < .001$). To reduce collinearity, each continuous variable was centered at its mean.

Statistical models were implemented using a maximum likelihood technique with forward-model-selection procedures (Cunnings, 2012; Yi et al., 2017). Under stepwise modeling, the order of entering the variables ultimately has no influence on the final model. However, entering variables of most interest first and adding them into the model by group makes the model-building procedure more sensible. Model selection started from random-intercept-only models, with independent variables entered first, then followed by covariates, and finally with random slopes tested. Model comparisons were conducted using the *anova* function. For the final model selected through model comparisons, a refined model was built upon it by removing the influential data points, detected based on the visual examination of the distribution of Cohen's *d* using the *influence.ME* package (version 0.9-8, Nieuwenhuis, te Grotenhuis, & Pelzer, 2012). The

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

alpha-level was set at .05, and *p*-values of the linear mixed-effects model were computed using the formula recommended by Baayen (2008).[2]

RESULTS

DATA TRIMMING

Reaction time data of the PJT were trimmed before running statistical analyses. First, erroneous responses in which a collocation received a response of "no" were removed. Second, responses that took longer than 3,000 milliseconds or shorter than 400 milliseconds were also excluded. Such cutoffs were adopted after visually examining exploratory graphics illustrating the distribution of subjects' RT and followed prior research practice by Yamashita and Jiang (2010). Third, reaction times that fell outside two standard deviations from the average for each participant was also removed. Data loss was 0.33% and 4.59% for steps two and three, respectively. The average RT of the PJT task was 848.8 ms (SD = 327.8 ms) for L1 speakers, and 1,180.6 ms (SD = 399.0 ms) for L2 speakers. For the analysis of accuracy rates, both correct and incorrect responses were retained. Participants' accuracy rates were all higher than 80% (M = 89.7%, SD = 3.7%). One letter was found missing from item 171. In addition, responses to item 1 were lost for most participants. Both items were excluded from the analyses. Correlations between the PJT performance and all aptitude measures were provided in Appendix 4 in the supporting information online.

REACTION TIME

The intercept-only model that included speaker, collocation frequency, MI, and their interactions, along with implicit language aptitude and its interaction with speaker, was found to be best-fitted. Model comparisons found that moderating effects between collocation frequency, MI, and any of the aptitude variables were redundant. Using the *influence.ME* package, item 91 and subject 37 were found to be influential. A refined model was then built upon the best-fitted model after removing data collected from item 91 and subject 37. Result patterns remained stable, although coefficient estimates increased. Using the *piecewiseSEM* package (version 1.2.1, Lefcheck, 2015), the proportion of variance explained by the fixed effects in the refined model was 40.23%, whereas the full model explained 59.74% of the variance. Model results are summarized in Table 3.

The effect of speaker ($Estimate$ = -.38, t = -.14.2, $p < .001$) indicated that L1 English speakers' average reaction time [exp(7.02 – 0.38) = 765.1 milliseconds] was significantly less than that of L2 speakers [exp(7.02) = 1,118.1 milliseconds]. The effect of collocation frequency ($Estimate$ = -.07, t = -.6.9, $p < .001$) implied that both L1 and L2 English speakers were sensitive to the collocation frequency. Moreover, the significant interaction between speaker and collocation frequency ($Estimate$ = .04, t = 8.0, $p <$ .001) indicated that the degree of the sensitivity to collocation frequency in L2 speakers was much greater than that in L1 speakers. Specifically, one unit of increase in collocation frequency (logged) led to a decrease in reaction time by 6.8% [1 – exp(-0.07)] and 3% [1 – exp(-0.07 + 0.04)] for L2 and L1 speakers, respectively. Similarly, the

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

TABLE 3. Linear mixed-effects model results for reaction time

	Fixed effects			Random effects			
				By Subject		By Item	
Parameters	*Estimate*	*SE*	*t*	*Variance*	*SD*	*Variance*	*SD*
Intercept	7.02	0.02	351.80***	0.01	0.10	0.01	0.10
Speaker	-0.38	0.03	-14.2***	—	—	—	—
Collocation frequency	-0.07	0.01	-6.9***	—	—	—	—
MI	0.02	0.00	6.2***	—	—	—	—
Speaker: Collocation frequency	0.04	0.01	8.0***	—	—	—	—
Speaker: MI	-0.01	0.00	-8.8***	—	—	—	—
Collocation frequency: MI	-0.01	0.00	-2.1*	—	—	—	—
Speaker: Implicit language aptitude	-0.06	0.02	-2.7**	—	—	—	—
Speaker: Collocation frequency: MI	0.01	0.00	4.1***	—	—	—	—

Note: There are 9,522 observations, where one observation is equal to one RT measurement for one colocation read by one participant. Model formula: RT (logged) ~ Speaker*Collocation frequency*MI + Speaker: Implicit language aptitude + (1|Subject) + (1|Item). Speaker was dummy coded, with L2 participants as the reference group.
*$p < .05$; ** $p < .01$; *** $p < .001$

significant effects of MI (*Estimate* = .02, $t = 6.2$, $p < .001$) suggested that both groups of participants were sensitive to the contingency of English collocations. Nevertheless, the significant interaction between MI and speaker (*Estimate* = -.01, t = -.8.8, $p < .001$) indicated that L2 speakers' sensitivity to contingency was also stronger than L1 speakers. One unit of increase in MI would result in an increase in reaction time by 2% [exp(0.02) – 1] and 1% [exp(0.02 – 0.01) – 1] for L2 and L1 speakers, respectively.

Collocation frequency was also found to interact with MI. The significant two-way interaction between collocation frequency and MI (*Estimate* = -.01, t = -2.1, p = .04), along with the three-way interaction between speaker, collocation frequency, and MI (*Estimate* = .01, $t = 4.1$, $p < .001$), suggested that collocation frequency interacted with MI, but only for L2 speakers. For L2 speakers, the higher the MI value, the larger the coefficient of collocation frequency, that is, the stronger the facilitative effect of collocation frequency on reaction time. In contrast, the higher the frequency, the weaker the negative influence of MI on the processing of English collocations. Such interactions are plotted in Figure 1.

Finally, the significant two-way interaction effect between speaker and implicit language aptitude (*Estimate* = -.06, t = -2.7, p = .007) indicated that implicit language aptitude was predictive of L1 English speakers' reaction time. For each unit of increase in implicit language aptitude, PJT reaction time was expected to decrease by 5.8% [1 – exp (-0.06)]. However, such an effect was not statistically significant for L2 speakers. The interaction effect is plotted in Figure 2.

ACCURACY

The intercept-only model, which included speaker, collocation frequency, MI, as well as the two-way interactions between speaker and MI and between speaker and explicit

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

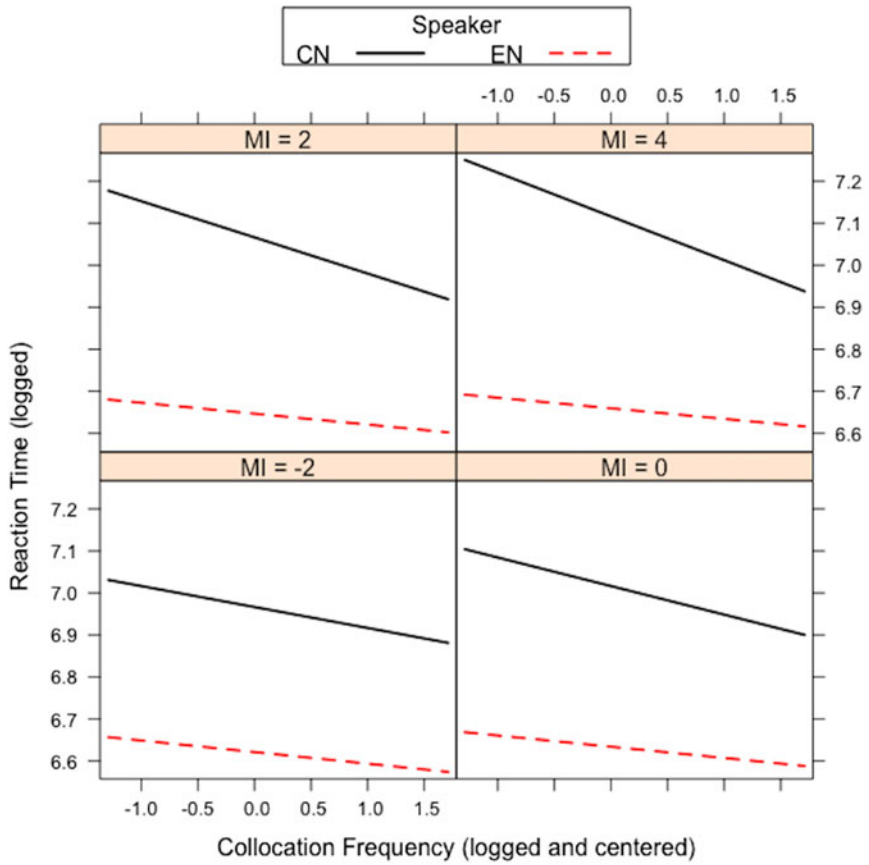


FIGURE 1. Three-way interaction effect between Speaker, Collaction frequency, and MI on reaction time (logged).

language aptitude was best fitted. Model comparisons found that moderating effects between collocation frequency, MI, and each of the aptitude predictors were redundant. Based on the best-fitted model, item 54 and subject 6 were found to be influential. A refined model was then built, based on the best-fitted model, with the data collected from item 54 and subject 6 removed. The full model explained about 31.3% of the total variance. The results are summarized in Table 4.

The effect of speaker (*Estimate* = .40, $z = 2.85$, $p = .004$) indicated that L1 English speakers were more likely to achieve correct responses than L2 speakers. Setting the values of collocation frequency, MI, and explicit language aptitude at their means, the average log-transformed odds of making a correct response was 2.93 and 3.33 (2.93 + 0.40) for L2 and L1 speakers, respectively. Transformed back to probabilities of achieving correct responses, this means the average accuracy rates for L2 and L1 speakers were 94.9% and 96.5%, respectively.[3] The significant effect of collocation frequency (*Estimate* = .44, $z = 4.21$, $p < .001$) indicated that both groups of participants

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

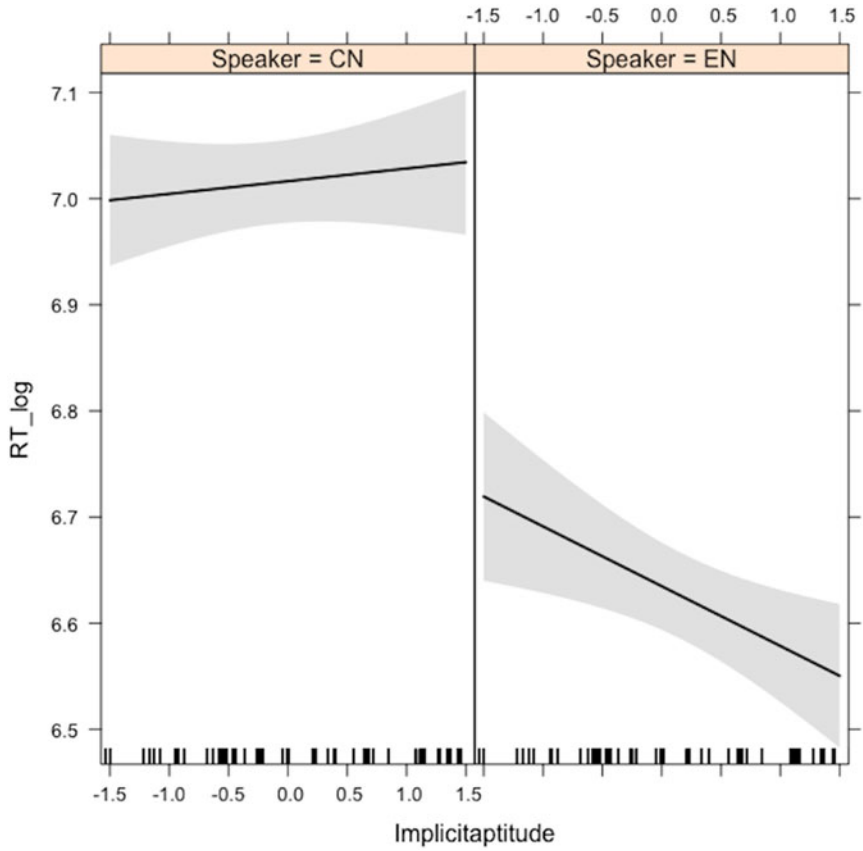


FIGURE 2. Two-way interaction between Speaker and Implicit aptitude on reaction time (logged).

were sensitive to collocation frequency. Specifically, one unit of increase in collocation frequency (logged) was predicted to enhance the odds of making a correct response by 1.6 times [exp(0.44)]. In addition, the significant effect of MI (*Estimate* = -.09, z = -1.92, p = .055) indicated that both groups of participants were also sensitive to MI. However, considering the significant two-way interaction between speaker and MI (*Estimate* = .16, z = 4.23, p < .001), it suggested that the influence of MI on the processing of collocations was the opposite in the two groups of participants. For L1 speakers, one unit of increase in MI was predicted to enhance the odds of making a correct response by 7.3% [exp(-0.09 + 0.16)]; for L2 speakers, one unit of such increase was predicted to decrease the odds of making a correct response by about 8.6% [1 – exp(-0.09)].

Lastly, the significant effect of explicit language aptitude (*Estimate* = .29, z = 2.47, p = .013) indicated that explicit language aptitude was predictive of the accuracy for both groups of participants. However, combining the significant

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

TABLE 4. Mixed-effects logistic model results for accuracy

	Fixed effects			Random effects			
				By Subject		By Item	
Parameters	*Estimate*	*SE*	*z*	*Variance*	*SD*	*Variance*	*SD*
Intercept	2.93	0.13	22.79***	0.20	0.45	1.06	1.03
Speaker	0.40	0.14	2.85**	—	—	—	—
Collocation frequency	0.44	0.10	4.21***	—	—	—	—
MI	-0.09	0.05	-1.92*	—	—	—	—
Speaker: MI	0.16	0.04	4.23***	—	—	—	—
Explicit language aptitude	0.29	0.12	2.47**	—	—	—	—
Speaker: Explicit language aptitude	-0.31	0.15	-2.11*	—	—	—	—

Note: There are 11,036 observations, where one observation is equal to one accuracy measurement for one collocation read by one participant. Model formula: Accuracy ~ Speaker + Collocation frequency + MI + Speaker: MI + Explicit language aptitude + Speaker: Explicit language aptitude + (1|Subject) + (1|Item). Speaker was dummy coded, with L2 participants as the reference group.
* $p < .05$; ** $p < .01$; *** $p < .001$

interaction between speaker and explicit language aptitude (*Estimate* = -.31, z = -2.11, $p = .004$) suggested that the role of explicit language aptitude differed between L1 and L2 speakers. Specifically, for L1 speakers, one unit of increase in explicit language aptitude would lead to a decrease in terms of the odds of making correct responses by 2% [1 - exp(0.29 – 0.31)]. In contrast, for L2 speakers, this would lead to an increase in terms of the odds of making correct responses by 34% [exp(0.29) - 1]. The two-way interaction between speaker and explicit language aptitude is illustrated in Figure 3.

DISCUSSION

The following findings were obtained through this study. First, both L1 and advanced L2 English speakers were sensitive to collocation frequency and contingency, with supporting evidence from reaction time and accuracy rates. Second, advanced L2 English speakers' statistical sensitivity to phrasal frequency and contingency was much stronger than that of L1 speakers. Third, for both L1 and advanced L2 English speakers, none of the cognitive aptitudes of interest (i.e., implicit/explicit language aptitude, working memory capacity) was found to moderate their sensitivity to either phrasal frequency or contingency of collocations. Finally, implicit and explicit language aptitude were found to influence participants' PJT performance, whereas working memory showed no impact on PJT performance.[4] For L1 speakers, implicit language aptitude played a facilitative role by reducing the reaction time, whereas explicit language aptitude played a negative role, lowering the probability of achieving correct responses. In contrast, for advanced L2 speakers, implicit language aptitude had no influence on the reaction time, while explicit language aptitude played a facilitative role, enhancing the probability of making correct responses.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

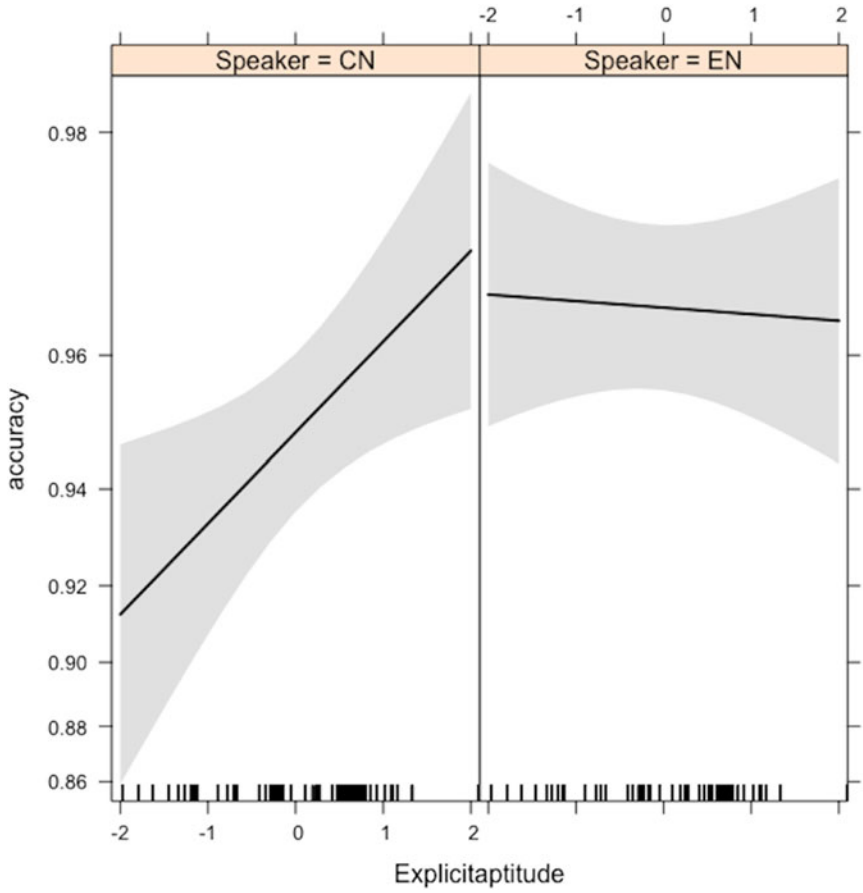


FIGURE 3. Two-way interaction effect between Speaker and Explicit aptitude on accuracy.

STATISTICAL SENSITIVITY TO PHRASAL FREQUENCY AND CONTINGENCY

Although current research has revealed the statistical sensitivity to phrasal frequency or contingency of MWS in L1 or L2 speakers, very few studies have found that such sensitivity exists in both domains and for both L1 and L2 speakers. This study replicated the findings obtained by Yi et al. (2017) that language users are sensitive to statistical regularities of MWS, including phrasal frequency and the contingency, regardless of their identity as L1 or L2 speakers. Such findings support the claim of Ellis (2006a, 2006b) that language acquisition is essentially statistical learning, and language users can capture the underlying distributional information through exposure to language input. Meanwhile, it also provided evidence for the idea that statistical learning is an ability retained by L2 speakers, with second language learners tuned to both phrasal frequency and the contingency of MWS. As noted previously, Ellis et al. (2008) first doubted

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

whether L2 speakers can be sensitive to contingency information of MWS. Yet findings in subsequent studies by Ellis and his colleagues (Ellis, 2016; Ellis et al., 2014) refuted their original argument, as they obtained robust contingency effects among advanced L2 learners. Moreover, as reported by Yi et al. (2017), the average length of formal L2 instruction received by their participants (who showed statistical sensitivity to both phrasal frequency and contingency) was less than four and a half years (M = 52.5 months, Max = 96 months, Min = 32 months, SD = 19.7 months). Hence, it is likely that L2 speakers' tuning of second language statistics may start from much earlier stages than previously thought, before reaching highly advanced levels.

This study also replicated earlier findings showing that L2 speakers tend to be more sensitive to statistical regularities than L1 speakers at the multiword level (Yi et al., 2017). Similar patterns have also been reported in the literature on bilingual word processing. For example, using lexical decision tasks, Duyck et al. (2008) examined the visual word recognition of Dutch-English bilinguals. They found that word frequency effects were much larger when participants performed the task in the second language than in L1. For another example, Cop et al. (2015) examined the natural reading of Dutch-English bilinguals and found that L2 word frequency had a larger influence on fixation durations than L1 word frequency. One explanation for this imbalance of word frequency effects in bilinguals may be rooted in differences in exposure. As argued by Cop et al. (2015), L2 words are learned much later than L1 words and learners receive less exposure to them on average than to L1 words. Consequently, the threshold for activating the L2 words may be lower than that for L1 words. This reasoning may extend to the imbalance in statistical sensitivities to phrasal frequency and contingency of MWS. However, alternative explanations also seem plausible. For example, the difference between the statistical sensitivities in L1 and L2 speakers is likely to result from the power law of practice (DeKeyser, 2007; Ellis, 2002). That is, the effects of practice are greatest at early stages of learning and become progressively smaller as exposure to language input accumulates. Given that L2 speakers are limited in their vocabulary size, and are more limited in their L2 experience than L1 speakers, each encounter with MWS is likely to generate larger practice effects.

THE MODERATING EFFECTS OF COGNITIVE APTITUDES ON STATISTICAL SENSITIVITY

Another question addressed by this research was whether cognitive aptitudes can moderate language users' statistical sensitivity to phrasal frequency and contingency of collocations, and whether such moderating effects differ between L1 and L2 speakers. This question was answered by examining potential three-way interactions between cognitive aptitudes, statistical regularities, and types of speakers. Nevertheless, mixed-effects modeling for both reaction time and accuracy rates failed to incorporate any of these interactions, due to redundancy. Thus, cognitive aptitudes were unlikely to moderate the statistical sensitivity to phrasal frequency and contingency of collocations, either for L1 or L2 speakers.

The independence of statistical sensitivity from cognitive aptitudes has significant theoretical implications. First, it supports the possibility that statistical learning is a mechanism that is not constrained by cognitive aptitudes. As reviewed in the opening

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

section, statistical sensitivity and statistical learning have been observed in infants, young children, and adults, as well as those with specific language impairments (Evans et al., 2009). If cognitive aptitudes such as implicit/explicit language aptitude and working memory capacity do moderate language users' statistical sensitivity, then one may observe significantly larger frequency or contingency effects for those with advantageous aptitude profiles. However, such patterns were not found in the current study. Statistical learning may interact with other language learning mechanisms (Saffran, 2003), yet the capacity to track sequential regularities, such as co-occurrence frequency and contingency, is an innate endowment that serves as an essential part of the cognitive architecture of human beings (Santolin & Saffran, 2017). Second, retention of the statistical sensitivity to phrasal frequency and contingency among adult L2 speakers also provides support for the argument that maturational constraints may not affect every aspect of language learning, at least not for statistical learning (see, also, Rastelli, 2014). As is well recognized in the field of second language acquisition, ultimate L2 attainment in various domains, including lexis and collocations (Granena & Long, 2013; Spadaro, 2013), is subject to maturational constraints. Consequently, post-sensitive-period L2 learners are unlikely to achieve nativelike proficiency (Granena & Long, 2013). Researchers (Abrahamsson & Hyltenstam, 2008; DeKeyser, 2000) have shown that cognitive aptitudes may partially compensate for such age effects, in that L2 learners with relatively high analytic aptitude may still be able to achieve a high level, albeit not nativelike, of second language proficiency. Based on results obtained from the current study, it seems that maturational constraints have differential effects on statistical sensitivity and statistical learning than on ultimate L2 attainment. Post-sensitive-period L2 learners showed stronger sensitivity to statistical regularities of collocations, and they probably share common statistical learning mechanisms with L1 speakers.

THE IMPLICIT/EXPLICIT NATURE OF COLLOCATIONAL PROCESSING/KNOWLEDGE

Finally, this study also raises interesting questions about the implicit versus explicit nature of language users' statistical knowledge. As reported earlier, cognitive aptitudes were found to influence language users' online PJT performance. Specifically, for L1 speakers, implicit language aptitude was found to facilitate the processing of English collocations by reducing reaction time, while explicit language aptitude was found to impede the processing of English collocations by lowering accuracy. In contrast, for L2 speakers, implicit language aptitude was not predictive of PJT reaction time, but explicit language aptitude considerably enhanced their PJT accuracy rates. Such disassociation patterns suggest that contrasting differences existed between L1 and advanced L2 speakers of English when dealing with the PJT.

Locating the source of such differences appears not that straightforward. Given the predictive patterns between implicit/explicit language aptitude and the PJT performance, one may think that the contrasting relational patterns might have resulted from the implicit/explicit nature of the task. Following the logic used by Suzuki and DeKeyser (2015, 2017), when examining the correlational relationships shown by the data, one would conclude that the PJT is an implicit task for the L1 participants, while it may be an explicit task for the L2 participants. Such a conclusion seems to make sense, given the presence/absence of a correlation between implicit language aptitude and PJT reaction

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

time and the negative/positive correlation between explicit language aptitude and PJT accuracy rates among the L1/L2 participants. However, if we accept that the PJT is an implicit task for L1 participants, and that this implicitness originates mostly from the online nature of the task (i.e., the time pressure), then labeling the PJT as an explicit task for the L2 participants will be problematic. Both L1 and L2 participants were under time pressure when conducting the PJT task. Given that the L2 participants took much longer than their L1 counterparts to process the lexical information, this task can be understood to be even more mentally taxing for the L2 participants. Therefore, the greater time pressure imposed to the L2 participants leads us to conclude that the PJT task should be implicit for L2 participants, taking the L1s as the baseline.

Another possible source leading to the contrasting patterns lies in the different ways L1 and L2 participants processed the collocations, and the different nature of their collocational knowledge. In a validation study of the construct of L2 proficiency using confirmatory factor analyses, Zhou and Ross (2017) reported that reaction time measures (based on phonetic/lexical/grammatical decision tasks) and accuracy measures (based on multiple-choice tests and imitation tasks) loaded onto different latent variables. Similar results were also obtained by Suzuki (2017) and Vafaee et al. (2017). Using confirmatory factor analyses, Suzuki (2017) and Vafaee et al. (2017) found that reaction time measures[5] (obtained from self-paced reading tasks and word-monitoring tasks) and accuracy measures (based on grammaticality judgment tasks) loaded onto different latent factors, and those factors were not meaningfully correlated (e.g., $r = .26$ in Vafaee et al., 2017; $r = .22$ in Suzuki, 2017). These research findings suggest that reaction time and accuracy rates are distinct measures and may reflect different latent factors: Reaction time measures directly tap into the processing of language, while accuracy rates (even collected from online tasks) reflect knowledge of language. Following this reasoning, the reported contrasting patterns should be examined separately for reaction time and accuracy rates, and from the perspectives of processing and knowledge, respectively. As a result, the presence/absence of a correlation between implicit language aptitude and PJT reaction time in L1/L2 speakers indicates that L1 English participants processed the collocations implicitly, while advanced L2 participants processed the same stimuli more explicitly. Similarly, the negative/positive influence of explicit language aptitude and accuracy rates in L1/L2 participants suggests that the collocational knowledge of L1 participants is likely to be implicit, while that of L2 participants may be explicit. Given the limited immersive L2 exposure and the explicit classroom instruction received by the Chinese-English bilinguals in this study (see the "Participants" section), such conclusions are not without basis. However, caution should be taken for such an interpretation, as direct empirical evidence supporting the distinction is still lacking.

CONCLUSION

The current study set out to investigate whether L1 and L2 speakers are sensitive to the phrasal frequency and contingency of collocations. Meanwhile, it also examined the moderating effects of cognitive aptitudes on such statistical sensitivity. Evidence obtained from this research supports that both L1 and advanced L2 speakers are tuned to statistical regularities underlying MWS. Moreover, the lack of moderating effects of cognitive aptitudes on language users' statistical sensitivity suggests that statistical

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

learning is an essential mechanism for both first and second language acquisition, which may not be constrained by sensitive-period effects. Lastly, taking an innovative approach to separately examining the disassociation patterns as found between the PJT performance and its relationship with implicit or explicit language aptitude, preliminary conclusions were made regarding the differences between L1 and L2 speakers in terms of the processing and knowledge of collocations. For L1 English speakers, they are likely to process collocations implicitly, and their collocational knowledge also seems to be implicit. However, for L2 English speakers (even at advanced levels), they seem to process collocations more explicitly, and their knowledge of collocations is likely to be explicit. These findings shed important light on the understanding of statistical learning, cognitive aptitudes, as well as the acquisition and processing of collocations. Future studies are needed to explore whether L1 and L2 speakers differ from each other in terms of the implicit/explicit nature of the knowledge of collocations and the way they process such larger-than-word units. In addition, more research should be carried out to investigate whether and how working memory influences the processing and acquisition of L2 collocations.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263118000141

NOTES

[1]When LLAMA D was kept for the PCA, SRT, and LLAMA D loaded onto a common factor, yet the two were not correlated.

[2]The formula is 2 * (1 - pt(abs(X), Y - Z)). X is the *t* value, Y is the number of observations, and Z is the number of fixed-effect parameters.

[3]In logistic regression, the dependent variable is log-transformed odds (in the current study, the odds of correct responses are the ratio of the probability of correct responses over the probability of incorrect responses). Log-transformed odds can be transformed back to a probability using the following formula: probability of correct responses = exp(logged odds)/(1 + exp(logged odds)).

[4]Working memory was dropped through the model selection procedure.

[5]Suzuki (2017) and Vafaee et al. (2017) used a measure of the difference in reaction time between grammatical and ungrammatical items.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481–509.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535–556.

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford, UK: Oxford University Press.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Longman.

Bolibaugh, C., & Foster, P. (2013). Memory-based aptitude for nativelike selection: The role of phonological short-term memory. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 205–230). Amsterdam, The Netherlands, and Philadelphia, PA: John Benjamins Publishing Company.

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society*, *34*, 421–435.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, *22*, 1216–1234.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). New York, NY: Routledge.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In D. Catherine & M. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Malden, MA: Blackwell Publishing.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*, 125–155.

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

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, N. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In D. Catherine & M. Long (Eds.), *The handbook of second language acquisition* (pp. 63–103). Oxford, UK: Blackwell.

Ellis, N. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*, 1–24.

Ellis, N. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*, 164–194.

Ellis, N. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372–405). New York, NY, and London, UK: Routledge.

Ellis, N. (2016). Online processing of verb–argument constructions: Lexical decision and meaningfulness. *Language and Cognition*, *8*, 391–420.

Ellis, N., O'Donnell, M., & Römer, U. (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*, 405–431.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–396.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*, 29–62.

ETS. (2017). Understanding your TOEFL iBT® test scores. Retrieved from https://www.ets.org/toefl/ibt/scores/understand/ (accessed December 20, 2017).

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*, 321–335.

Fletcher, W. H. (2011). Phrases in English (PIE). Retrieved from http://phrasesinenglish.org/ (accessed February 23, 2016).

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117–125.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: Chicago University Press.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*, 178–186.

Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment* (Unpublished doctoral dissertation). College Park, MD: University of Maryland.

Granena, G. (2013a). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). Amsterdam, The Netherlands, and Philadelphia, PA: John Benjamins Publishing Company.

Granena, G. (2013b). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*, 665–703.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society*, *35*, 151–166.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255.

Hamrick, P. (2014). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*, 247–278.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–96). Cambridge, MA: MIT Press.

Kanyongo, G. Y. (2005). The influence of reliability on four rules for determining the number of components to retain. *Journal of Modern Applied Statistical Methods*, *5*, 7.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220.

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford, CA: Stanford University Press.

Lefcheck, J. S. (2015). piecewiseSEM: Piecewise structural equation modeling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*, 573–579.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861–883.

Loewen, S. & Gonulal, T. (2015). Exploratory factor analysis and principal component analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182–212). New York, NY: Routledge.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*, 122–134.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751.

Meara, P. (2005). *LLAMA language aptitude tests: The manual*. Swansea, UK: Lognostics.

Nieuwenhuis, R., te Grotenhuis, H. F., & Pelzer, B. J. (2012). Influence.ME: tools for detecting influential data in mixed effects models. *R Journal*, *4*, 38–47.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). London, UK: Longman.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Rastelli, S. (2014). *Discontinuity in second language acquisition: The switch between statistical and grammatical learning*. Bristol, UK: Multilingual Matters.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Robinson, P., & Ellis, N. C. (2008). Conclusion: Cognitive linguistics, second language acquisition and L2 instruction–Issues for research. In P. Robinson & N. Ellis (Eds.), *The handbook of cognitive linguistics and second language acquisition* (pp. 489–545). New York, NY: Routledge.

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, *1*, 49–60.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114.

Saffran, J. R., Aslin, R. N, & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Santolin, C., & Saffran, J. (2017). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*, 52–63.

Schmidt, J. R. (2012). Human contingency learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1455–1456). New York: Springer Science & Business Media.

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, *65*, 533–562.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Skehan, P. (2012). Language aptitude. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 381–395). London, UK: Routledge.

Spadaro, K. (2013). Maturational constraints on lexical acquisition in a second language. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 43–68). Amsterdam: John Benjamins Publishing Company.

Stone, J., & Towse, J. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*, 5.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895.

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747–790.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1–42.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: acquisition and communication* (pp. 151–173). London, UK: Continuum.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95.

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, *43*, 55–77.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*, 91–120.

Williams, J. N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition*. (pp. 427–441). London, UK: Routledge.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*, 430–449.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, *35*, 451–482.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*, 647–668.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press

Yi, W., Lu, S., & Ma, G. (2017). Frequency, contingency and online processing of multiword sequences: An eye-tracking study. *Second Language Research*, *33*, 519–549.

Zhou, Q., & Ross, S. (2017). *An investigation of the construct validity of L2 Chinese language proficiency: A multitrait-multimethod approach*. Paper presented at the American Association of Applied Linguistics. March 18–21, Portland, OR.

Zuhurudeen, F. M., & Huang, Y. T. (2016). Effects of statistical learning on the acquisition of grammatical categories through Qur'anic memorization: A natural experiment. *Cognition*, *148*, 79–84.

https://doi.org/10.1017/S0272263118000141 Published online by Cambridge University Press