

RESEARCH ARTICLE 

Quantifying proper nouns' influence on L2 English learners' reading fluency

Christopher Nicklin^{1*} , Allie Patterson²  and Stuart McLean³ ¹Rikkyo University, Tokyo, Japan; ²Tokyo Kasei University, Tokyo, Japan; ³Momoyama Gakuin University, Osaka, Japan*Corresponding author. E-mail: c.nicklin@rikkyo.ac.jp

(Received 11 February 2022; Revised 16 November 2022; Accepted 21 November 2022)

Abstract

Proper nouns constitute a lexical class with special properties and are thus treated differently from other words by second language acquisition researchers. An assumption exists that even low-proficiency learners will find them unproblematic, yet research suggests this assumption might be misplaced. The present study involved two self-paced reading experiments designed to investigate proper nouns' influence on Japanese university students' reading fluency. In Experiment 1, participants were presented with 60 decontextualized sentences containing 30 proper nouns and 30 common nouns to determine whether they are processed in a similar manner. In Experiment 2, participants read another 60 sentences comprising a book chapter to explore the effects of repeated exposure to a set of proper nouns. The results indicated that proper nouns are processed in a similar manner to common nouns in terms of disrupting reading fluency. The implications for language learning pedagogy, in particular extensive reading, are discussed.

Introduction

Proper nouns are a lexical item class that appear simple on the surface but comprise a set of qualities that belies their façade. At their very core, *proper nouns* (PNs) constitute names, specifically single names that might be considered unique, such as people, locations, objects, institutions, or artworks (Valentine et al., 1996). In English, proper noun status is typically indicated by a capitalized initial letter, such as with *Tokyo* or *Disneyland*. According to Quirk et al.'s (1985) hierarchical taxonomy of word class, PNs can be considered subsidiary to *nouns* and are sometimes distinguished in this subclass from *proper names*, which constitute multiword, single-unit items, such as *Tokyo Disneyland*.¹ Grammatically, PNs are distinguished from common nouns (CNs; e.g., *handle*, *liquid*, or *puzzle*) by their freedom from articles and determiners. In instances in which such syntax precedes a PN, as with *the Kennedys*, the PN tends to lose its status and is recast as a common noun phrase (Allerton, 1987). Their special status is fortified by research indicating that they are the lexical class with the highest

¹Henceforth, *PN* serves as an umbrella term denoting both proper nouns and proper names.

probability of inducing retrieval issues and are the only lexical class to incur difficulties for certain aphasiacs (Valentine et al., 1996).

Another aspect of PNs that distinguishes them from CNs is the argument over whether they constitute encyclopedic knowledge or an analyzable lexical class in their own right, which was summarized by Klassen (2021). Causal-historical theory (Kripke, 1980), which stems from philosophical enquiry, argues that PNs constitute encyclopedic knowledge. For instance, they should not be considered a part of language because they do not address qualities of entities that they designate. Instead, proper nouns are defined by their causal history, which relates to the chain of events beginning from the birth of the PNs bearer. For instance, the name *Albert Einstein* relates to actions performed by the bearer of the name throughout their lifetime. It does not contain any inherent meaning. However, in a meta-analysis conducted on proper name reference research, van Dongen et al. (2021) concluded that philosophical, semantic intuition research is not as reliable as philosophical practice assumes. Descriptivist theory provides the counterargument to causal-historical theory, suggesting that PNs are a part of language using connoted, presuppositional meaning (Van Langendonck, 2007). For instance, upon reading the name *Rupert*, a British reader is likely to imagine the bearer as being an upper-class male, while *Paris* suggests the categorical information of *city*, with connotations of *fashion* and *romance*, unlike *Pyongyang*, which perhaps invokes contrasting concepts. Furthermore, certain PNs unarguably contain meaning, such as *Einstein* in the expression *he's no Einstein* (Nation & Kobeleva, 2016).

The aim of the present study is to answer calls for empirical research to investigate the extent to which second language learners are affected by PNs (Brown, 2010; Klassen, 2021; Nation & Kobeleva, 2016; Webb, 2021). To do so, the effect of PNs on Japanese English as a foreign language (EFL) students' reading fluency was assessed with two self-paced reading experiments. The first involved presenting participants with PNs embedded in decontextualized sentences, while the second involved repetitions of PNs embedded in a short text to emulate a more naturalistic reading process. The experiments were conducted to investigate the extent to which PNs disrupt second-language learners' reading fluency, and to determine whether the frequency effect found in vocabulary research holds true for PNs.

Literature review

Proper nouns in L2 vocabulary research

Second language (L2) vocabulary researchers have overwhelmingly subscribed to a causal-historical approach, whereby PNs are regarded as unproblematic for L2 learners, and this is manifested in three ways. The first and most common way that PNs are regarded as unproblematic is to assume that they are *all known* to learners or can be easily identified as PNs with recourse to context (Nation, 2006; Nation & Webb, 2011; Webb & Chang, 2015). However, research suggests that the assumption might be misplaced. Japanese university students' speed-reading times were investigated by Kramer and McLean (2019), who observed slower than predicted times for a hypothetical student reading at 600 characters per minute. This discrepancy was partly attributable to the large number of PNs contained in Nation and Malarcher's (2007) reading fluency development texts, such as *Rabindranath* and *K'ung Fu-tzu*, which are orthographically, phonologically, and morphologically contrasting to English words and thus potentially impede processing (Elgort & Warren, 2014).

Skepticism regarding the assumption that PNs are known or can easily be understood through context is also raised by research demonstrating that unfamiliar PNs hinder learners' comprehension. While listening to news stories, Kobeleva's (2012) intermediate and advanced ESL learners' scores on comprehension and referent derivation questions suffered when PNs were unfamiliar in contrast to when they were familiar. Furthermore, both groups of learners reported increased difficulty in tasks involving unfamiliar PNs. In Erten and Razi's study (2009), trainee English teachers at a Turkish university exhibited improved comprehension when person and place names were nativized from English to Turkish to render them more familiar. However, in a study with intermediate proficiency Japanese university students, Klassen (2020) failed to replicate those results. The failure was attributed to two factors. First, the proficiency difference between the Turkish trainee English teachers and Klassen's intermediate-level learners, whereby vocabulary issues hindered the lower-proficiency sample more so than the familiarity of the PNs. Second, due to the closer linguistic distance between Turkish and English orthography in contrast to Japanese and English orthography, the Turkish participants held a processing advantage over the Japanese participants for the reading activities. When synthesized, these results indicate that a lack of familiarity with PNs inhibits the comprehensibility of written and spoken text for L2 English learners, and potentially impedes reading speed, which in itself can diminish comprehension (Beglar *et al.*, 2012).

The second way that PNs are regarded as unproblematic is by the fact that they are signaled through initial letter capitalization (e.g., Nation, 2006; Nation & Webb, 2011; Webb 2021; Webb & Macalister, 2013). However, this belief is also based upon assumption and has seldom been explicitly researched. One such study was conducted by Opitz and Bordag (2021), who investigated the effect of capitalization with L1 German speakers and L1 Czech advanced learners of German. The Czech speakers utilized capitalization in a similar manner to the L1 German speakers, whereby capitalization was employed to clarify word sense. All nouns are capitalized in written German, rendering the language ideal for the study. However, both German and Czech orthographies are rendered with the Roman alphabet, and the authors concluded that further research is warranted with other L2 populations. It is feasible that readers of L1s with orthographies that are linguistically distant from Roman script, such as Japanese, are not as attuned to capitalization as the Czech readers in Opitz and Bordag's study.

The third way that PNs are regarded as unproblematic is through their inclusion in word coverage figures as known words. L2 researchers customarily endorse 95% or 98% lexical coverage benchmarks as sufficient for comprehension, which ordinarily includes PNs (Laufer & Ravenhorst-Kalovski, 2010). In practice, when reporting the composition of reading materials in terms of lexical coverage, L2 researchers often report the percentage of PNs included, which is commendable (e.g., Webb & Chang, 2015; Webb & Macalister, 2013). However, research also indicates that PNs account for between 4–5% of a typical written text (e.g., Francis & Kučera, 1982). Therefore, if a text contains PNs that are not known by an English learner, it is possible that learners would not recognize 1 in 20 of the words, which would hinder comprehension according to the customary 95–98% threshold.

Proper nouns and extensive reading

The assumption that PNs are unproblematic and do not inhibit reading comprehension is pertinent for extensive reading (ER), which involves language learners reading and

comprehending large amounts of text with both speed and fluency (Waring & McLean, 2015). To achieve this, students are provided with graded readers, in which vocabulary and syntax are controlled for varying proficiency levels. The large amounts of lexically and semantically simplified text provided in graded readers lower the cognitive burden of reading and develop learners' reading fluency. Reading fluency is the "ability to read rapidly with ease and accuracy" (Grabe, 2009, p. 291) and is the product of three mutually inclusive subprocesses: automaticity, rate, and accuracy (Kuhn & Stahl, 2003), although other variables, such as linguistic distance between the learners L1 and the target language, also play a role (Nisbet et al., 2021). *Automaticity* relates to the effortless conduct of a skill developed through countless hours of repetitive practice. Graded readers facilitate the automatization of lower-level psycholinguistic processes that comprise reading, which include word recognition, syntactic parsing, and semantic-proposition encoding (Grabe, 2009). The more automatic these processes become, the faster the *rate* of reading becomes, which enables even larger amounts of input and skill development. Finally, rapid automatized skills are rendered pointless if unaccompanied by *accuracy*. For instance, an L2 Chinese reader with rapid, automatized character recognition skills faces inhibited comprehension unless the characters are semantically processed with accuracy.

Research investigating ER's value predominantly indicates that the activity is beneficial for reading fluency development. Significant reading rate gains have been observed for Japanese learners of English when 200,000 words of simplified text are read annually, and 95 to 98% of the word families are familiar (e.g., Beglar et al., 2012; Beglar & Hunt, 2014). Increases of 20% in reading speed were also observed by Bui and Macalister (2021) for Vietnamese learners during a 10-week ER program. However, such findings are inevitably based upon text coverage counts that assume PNs comprehensibility. If unfamiliar PNs disrupt reading fluency, whether in terms of form-recognition, processing rate, or any other component skill related to reading fluency, it is possible that PN-dense texts are less suitable for ER.

Incidental vocabulary acquisition through reading

Although research suggests that ER is rewarding in terms of reading fluency, it also demonstrates that reading in general is an inefficient method for incidental vocabulary acquisition, with words requiring numerous exposures to develop form and meaning recognition. In an eye-tracking study, Pellicer-Sánchez (2016) exposed participants to six nonwords embedded eight times each in a 2,300-word short story. Immediate posttest revealed the form of 85.50% of the nonwords was recognized, the meaning of 78.26% was recognized, and the meaning of 60.87% was recalled. The author also reported significant increases in nonword reading times (RTs) following three to four exposures and reading rates in sync with typical known words by the eighth occurrence. Similar results were observed by Webb (2007), who found significantly improved form recognition scores for nonwords following three exposures, but that 10 or more exposures were required to develop deeper knowledge. Elgort and Warren's (2014) investigation into contextual word learning while reading under naturalistic conditions revealed that more than 12 encounters with pseudowords were required to develop explicit knowledge regardless of proficiency. Furthermore, implicit knowledge of form and meaning was triggered but was not robust even when more than 12 exposures were provided.

The research reviewed in the preceding text focused upon non-PN vocabulary, and as mentioned previously PNs possess characteristics that distinguish them from CNs,

such as the fact that they are the word class most likely to induce retrieval issues (Valentine *et al.*, 1996). Furthermore, L1 listening research on incidental PN learning with *earwitness* testimony indicated the “pronounced difficulty of proper name learning” (Swanson *et al.*, 2021, p. 1), which in some instances resulted in participants implicating innocent parties as guilty. Thus, although L2 research suggests that approximately four encounters are sufficient for the development of form knowledge while approximately 8 to 10, or perhaps more, encounters are requisite for semantic knowledge, it is incautious to generalize the findings to PNs without specific research.

To summarize, L2 researchers have adopted a casual-historic approach to PNs and assume that they do not inhibit reading comprehension. However, this assumption has seldom been explicitly researched. The aim of simplified texts such as graded readers for ER is to lower the cognitive burden of reading for learners, enabling them to increase reading fluency through the automaticity of reading processes. To this end, low-frequency nouns are avoided by graded reader authors to simplify the texts. This decision is supported by reading research indicating that new words disrupt reading fluency and require between 3 to 10 exposures before they are processed with comparable RTs to known words. If it is the case that PNs are as detrimental to the reading process as low-frequency lexis, then the inclusion of PNs in graded reader lexical coverage counts warrants consideration.

The present study

In the present study, two psycholinguistic experiments were conducted with Japanese university students to quantify the effect of PNs on L2 reading fluency. In Experiment 1, participants read 60 sentences, which were extracted from a small corpus of graded readers. Half the sentences included PNs, while the other half contained CNs that were matched with the PNs in frequency. The aim was to determine the extent to which the PNs disrupted L2 reading fluency in comparison with CNs, and the following hypotheses were addressed:

1. Proper nouns significantly disrupt second-language learners’ reading fluency.
2. Less frequent proper nouns will be more detrimental to reading fluency than high-frequency ones.
3. The reading times elicited by proper nouns will be comparable to the reading times elicited by common nouns matched in frequency to the proper nouns

In Experiment 2, participants read another 60 sentences comprising a chapter of a low-level graded reader that contained numerous repetitions of a set of PNs. The aim was to explore the extent to which reading fluency disruptions rendered by PNs decrease with repeated exposure and the following research hypothesis was addressed.

4. Repeated exposures during naturalistic reading will reduce a proper nouns’ disruptiveness to reading fluency.

Experiment 1

Experiment 1 was designed to investigate the extent to which PNs result in reading disfluencies and involved a comparison between PNs and CNs that were matched in terms of frequency and embedded in decontextualized sentences extracted from graded readers.

Method

Participants

Forty-four participants were recruited from a women's university in Tokyo, Japan. A sample size of 30 was determined using power simulation with the *simr* (Green & MacLeod, 2016) package for R (R Core Team, 2020) and pilot study data (see Supplementary Materials A for details). However, considering the potential loss of data, 44 participants were recruited. The sample comprised Japanese female students aged between 18 and 19 from six intermediate English classes. Although access to standardized placement test results was unavailable, student proficiency in the intermediate classes approximated B1 on the CEFR scale. The participants had completed at least six years of formal English education prior to university. All participants signed a consent form before the experiment and were compensated with a 1,000-yen Amazon voucher. The participants were randomly assigned to one of two groups based upon experiment order. One group completed Experiment 1 first, the other group completed Experiment 2 first. The results of 11 participants were excluded from the Experiment 1 analyses for responding correctly to less than 70.00% of the self-paced reading task comprehension questions. Participants completed a 60-item meaning-recall vocabulary test at the beginning of the semester and the scores were used to assess whether the two groups in each experiment were comparable in terms of vocabulary knowledge. Descriptive statistics for the eligible participants are presented in Table 1. The confidence interval (CI) overlaps between the two groups with regard to self-paced reading task comprehension scores, average reading time across both experiments, and meaning recall scores indicated that there was no significant difference between the two groups in terms of these measures (Cumming, 2012).

Instruments

Self-paced reading task

To address the RQs, self-paced reading tasks for each experiment were built with *PsychoPy* (Peirce et al., 2019). Self-paced reading is a computer-based research method in which sentences are broken into words or segments and visually presented to participants. Participants control the speed at which the words are presented using keystrokes, and the participants' RTs are recorded and analyzed.

Table 1. Experiment 1 participant descriptive statistics

Order	Parameter	<i>n</i>	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurt	95% CIs ^a	
									<i>LL</i>	<i>UL</i>
1-2	Comprehension (%)	18	77.41	4.79	70.00	86.67	0.19	-1.07	75.03	79.79
	Mean RT	18	565.16	72.50	442.11	698.45	0.07	-0.98	529.10	601.22
	Meaning recall	16	43.06	8.73	28.00	53.00	-0.44	-1.52	38.41	47.71
2-1	Comprehension (%)	15	81.11	6.51	70.00	93.33	0.21	-1.07	77.50	84.72
	Mean RT	15	549.85	93.46	392.78	800.37	0.91	1.13	502.48	597.21
	Meaning recall	14	44.50	10.45	31.00	60.00	-0.06	-1.82	38.47	50.53

Note: Three participants did not complete the meaning recall test.

^aCIs calculated with values from *t* distributions, *t*(17) = 2.11, *t*(15) = 2.13, *t*(14) = 2.15, and *t*(13) = 2.16 to account for small samples.

Table 2. Distribution of proper nouns in the graded reader corpus

Stage	Title	Pages	Word count	PN count	PNs per Page	PN%
2	<i>Dead Man's Island</i>	40	5,194	334	8.35	6.43
	<i>The Jungle Book</i>	41	6,523	595	14.51	9.12
	<i>The Year of Sharing</i>	39	6,401	170	4.36	2.66
3	<i>Goldfish</i>	56	10,670	374	6.68	3.51
	<i>The Secret Garden</i>	57	10,724	568	9.96	5.30
	<i>Through the Looking Glass</i>	57	10,503	770	13.51	7.33
4	<i>Death of an Englishman</i>	71	13,507	1,063	14.97	7.87
	<i>The Moonspinners</i>	72	13,923	858	11.92	6.16
	<i>The Whispering Knights</i>	73	15,537	921	12.62	5.93
5	<i>Deadlock</i>	85	19,331	1,630	19.18	8.43
	<i>Do Androids Dream...</i>	103	24,607	1,922	18.66	7.81
	<i>Great Expectations</i>	86	24,096	1,503	17.48	6.24
6	<i>A Passage to India</i>	120	30,896	1,930	16.08	6.25
	<i>The Woman in White</i>	104	31,823	2,484	23.88	7.81
	<i>Vanity Fair</i>	120	33,221	3,002	25.02	9.04

Stimuli sentences. For Experiment 1, 30 sentences containing centrally located, two-syllable PNs and 30 sentences containing centrally located, two-syllable CNs were extracted from a small, graded reader corpus. The corpus comprised 15 graded readers from a series that spans seven proficiency levels, beginning with *Starter*, and then progressing through stages 1 to 6, with 6 being the most challenging and containing the most frequent 2,500 headwords from an unspecified corpus. The self-paced reading task required sentences with centrally located PNs to avoid wrap-up effects, whereby sentence final words elicit longer RTs (Jiang, 2012). This requirement rendered the short sentences that are a feature of *Starter* and *Stage 1* books ineligible, and so three books were randomly selected from each of stages 2 through 6, resulting in 15 books, which are summarized in Table 2. The 15 books were scanned, converted into Word documents, and tagged with TagAnt (Anthony, 2015), resulting in a 256,956-word corpus.

In total, 127 unique two-syllable PNs were extracted from the corpus. Two-syllable PNs were selected to control for the effect of syllable number. The 100 most frequent were presented to 216 participants from the target population in a norming study to determine the familiarity of the PNs. The participants read the 100 PNs along with 65 distractor items in a randomized order on a Google Form and were instructed to rate the familiarity of each of the 165 items on a scale of 1 to 4. Rasch analysis was conducted on the norming study data and the resulting logits were utilized to select 30 target PNs with a spread of familiarity according to the target population (see Nicklin [2021] for a detailed report of the norming study).

For each target PN, sentences involving the PN located centrally, at least three words from the end of the sentence, were extracted from the corpus. To ensure the sentences were comparable and able to fit across the computer screen on a single line, all sentences were between 49- and 87-character spaces and 10 to 18 words long. Nontarget PNs were changed to pronouns or “*The wo/man*” and nontarget PNs were simplified. For instance, *The Mansion Hotel* was changed to *the hotel*. Sentences containing two (or more) PNs in addition to the target PN were not included because of potential confusion when replaced with pronouns. In sentences containing two consecutive PNs, such as *Rick Deckard*, only the target PN was kept. Slight alterations were made to the majority of sentences to meet the required criteria (see Supplementary Materials B for

original and altered sentences). This process resulted in 30 sentences for the self-paced reading task that were representative of sentences found in graded readers.

In addition, 30 sentences containing two-syllable CNs in the central position were created with the same method utilized for the PN sentences. All two-syllable CNs that appeared once ($N = 289$) were extracted from the corpus. Words beginning with capital letters (e.g., *Friendship*) were removed because this indicated that the single appearance was at the beginning of a sentence, and words with clear multiple POSs, such as gerunds (e.g., *meeting*), were also omitted. For the remaining 146 CNs, frequency values from the Corpus of Contemporary American English (COCA; Davies 2008–) were obtained. The 30 target PNs were matched with CNs from the list of 146 according to log-transformed COCA frequency. For instance, *Frances* (log frequency = 3.75) was matched with *betrayal* (log frequency = 3.75). Log COCA frequency was utilized instead of raw frequency because the final analysis involved transformed values to avoid the effects of the Zipfian distribution (Zipf, 1935) to which raw frequency values generally adhere. When exact log frequency matches were not possible, a CN with as close a value as possible was selected. The difference between the PN and CN in terms of log frequency never exceeded 0.26 (e.g., *Jaggers* [log frequency = 2.22] paired with *signpost* [log frequency = 2.48]), with the mean difference being 0.04 ($SD = 0.06$).

Descriptive statistics for the CNs and PNs, and the sentences containing them are presented in Table 3. The overlap between the CIs for the CNs' and PNs' log frequency values provided additional support for the claim that items were well matched in terms of frequency. Despite this, the lack of overlap between the CIs for word length and sentence length indicated a significant difference between the values. The CNs were approximately one letter longer in general, and the sentences containing the PNs were approximately one word longer in general. However, in practical terms this difference is negligible.

Comprehension questions

Once the target sentences were finalized, a set of comprehension questions were developed, which functioned as attention checks to ensure that participants were engaged in the self-paced reading task. For each sentence, a true-or-false statement related to the sentence was written. For instance, the Experiment 1 sentence containing the PN *Jeffreys* (*They passed another police car and Jeffreys was surprised when the*

Table 3. Descriptive statistics for words and sentences in Experiment 1

Parameter	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurt	95% CI		
							<i>LL</i>	<i>UL</i>	
CNs	Length (letters)	6.93	1.34	5.00	10.00	0.41	-0.50	6.85	7.02
	COCA freq	14,838.83	20,451.65	32.00	72,410.00	1.82	2.44	13,502.66	16,175.01
	Log COCA freq	3.62	0.95	1.51	4.86	-0.81	-0.17	3.56	3.68
	Sent. length (letters)	67.57	8.88	49.00	84.00	-0.09	-0.56	66.99	68.15
PNs	Sent. length (words)	13.40	1.96	10.00	16.00	-0.11	-1.24	13.27	13.53
	Length (letters)	5.83	1.23	4.00	9.00	0.57	-0.10	5.75	5.91
	COCA freq	15,428.30	20,548.79	7.00	69,891.00	1.66	1.84	14,085.78	16,770.82
	Log COCA freq	3.59	1.00	0.85	4.84	-0.99	0.50	3.52	3.65
	Sent. length (letters)	72.17	8.96	54.00	87.00	-0.18	-0.88	71.58	72.75
	Sent. length (words)	14.27	1.86	10.00	18.00	0.00	-0.29	14.15	14.39

drivers did not wave) was followed by the comprehension statement *The sentence suggested that they passed a bus*. In psycholinguistic experiments involving decontextualized sentences, only 30% of sentences are required to be followed by comprehension questions (e.g., eye-tracking [Godfroid, 2020]), however 30 (50.00%) were utilized in Experiment 1. The comprehension questions were piloted with seven advanced users of English, who reported no issues. Supplementary Materials C contains the comprehension questions.

Procedure

Before the experiment, participants completed consent forms and read Japanese self-paced reading task instructions, which are contained in Supplementary Materials D. Figure 1 illustrates the experiment procedure. For Experiment 1, sentence order was randomized and participants saw each PN only once. Before each sentence, participants were shown a series of dashes that indicated the position of each letter within each word within the sentences. The position of the first word was indicated by a fixation asterisk. After pressing the response button, the first word of the sentence appeared. After each keystroke, the participants were shown the next word in the

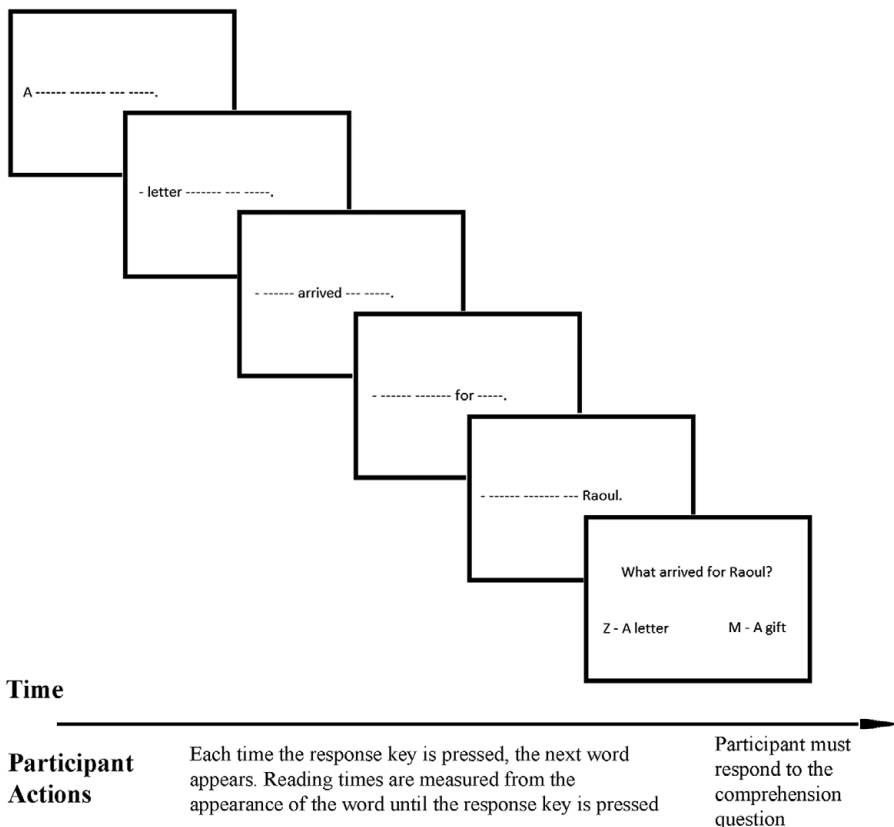


Figure 1. Self-paced reading task procedure.

sentence and the time between keystrokes was recorded as a RT in milliseconds (ms) for analysis. Half the participants were randomly assigned to complete Experiment 1 before Experiment 2, while the other half completed Experiment 2 first. This decision was made to account for (a) fatigue and (b) the belief that RTs might become faster toward the end of the test as a result of familiarity with the self-paced reading paradigm.

Analysis

The RT data was analyzed with linear mixed-effects models (LMMs) using the *lme4* (Bates et al., 2015) R package. For Experiment 1, two separate LMMs were built to model the RTs—one for PNs and another for CNs (see Supplementary Materials E for justification). Before the analysis, the RTs were logarithmically transformed (\log_e) to control for skewness (e.g., Bultena et al., 2014; Frank et al., 2016; Stewart et al., 2021). Additionally, RTs shorter than 150 ms and greater than 2,000 ms were winsorized before being log-transformed in the models. The 150 ms lower boundary was selected based upon Hsu et al. (2011), who demonstrated that L1 speakers require between 150–200 ms to recognize a word. Although the participants in the present study were L2 speakers, the boundary was adhered to for data preservation. An upper boundary of 2,000 ms was utilized because longer RTs are plausibly the result of processes unassociated with those of primary interest (Baayen, 2008), and boundaries larger than 2,000 ms resulted in a negligible amount of data being saved. Winsorizing was preferred to trimming because it has been shown to produce similar results while preserving potentially valid data points (Nicklin & Plonsky, 2020). In total, 5.61% of the data lay outside of the boundaries, with 0.27% outside of the lower boundary, and 5.34% outside of the upper one.

Both the PN and CN LMMs were constructed with theoretically justifiable variables, terms, and interactions, and included by-participant and by-item random intercepts to resolve the nonindependence stemming from recording numerous responses for each participant and item (Winter, 2013). By-subject slopes were modeled for length in characters and log COCA frequency (participants) to account for the fact that these variables were expected to uniquely influence the RTs for each participant. Details of the random-effect selection process are contained in Supplementary Materials E. Experiment presentation order, word length in characters, and log-transformed COCA frequency were the independent variables, which were modeled as fixed effects. All continuous variables were centered to aid regression coefficient interpretation (Winter, 2020).

The final variable included was a five-level categorical variable, *nLocation*, which was constructed to investigate the extent to which the target words and the words in the spillover region following the target words affected the RTs. The RTs in this area were the only ones analyzed because this is the area where the research hypotheses could be addressed. In self-paced reading tasks, spillover effects describe the phenomenon whereby the expected effect occurs on the word, or words, following the target word (Jiang, 2012). To account for this, the 5-level *nLocation* variable comprised the RTs for the target word (t_0), and the three words following (t_1 , t_2 , and t_3). The fifth level was the word preceding the target word (t_{-1}), which was the reference level and allowed a comparison with a word that was unaffected by the target (see Figure 2). Only these five

The man would like to

help	Sofia	but	he	does
------	-------	-----	----	------

 not like her husband
t-1 t0 t1 t2 t3

Figure 2. The *nLocation* variable's five levels.

RTs were modeled. The nLocation variable was also modeled as an interaction term with target-word frequency, which was a variable designed to investigate the extent to which the frequency of the target word affected the RTs in the spillover region. Following Meteyard and Davies's (2020) best practice guidelines, LMM assumptions (i.e., linearity, random distribution of residuals, and homoscedasticity) were assessed (see Supplementary Materials E) and collinearity was assessed by calculating variance inflation factors (VIFs) with Frank's (2014) R function.

Results

To address the first hypothesis, stating that PNs were expected to disrupt the participants' reading fluency, the PN RTs were plotted and the results of the PN LMM were analyzed. Figure 3 and the descriptive statistics in Table 4, which are reported with Median and interquartile range (IQR) to account for the skewed distributions of the raw RTs, indicated that the PN RTs (t_0 ; Median = 830.40 ms, IQR = 938.82 ms), were longer and more widely distributed than the preceding words' RTs (t_{-1} ; Median = 477.04 ms, IQR = 279.12 ms). The boxplots also suggested a weak spillover effect, whereby the RTs for words directly following the target region (t_1 ; Median = 549.68 ms, IQR = 410.20 ms) were more similar to the pattern in the t_{-1} region than in the target region.

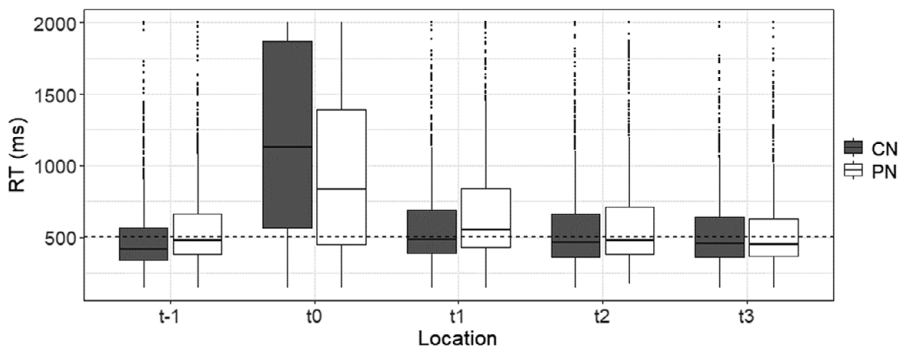


Figure 3. Boxplots of raw reading times for Experiment 1.
Note: Dotted line represents Median = 502.36 ms, IQR = 443.80.

Table 4. Descriptive statistics for reading times (ms) by noun type and location

Noun type	Location	<i>n</i>	Median	<i>IQR</i>	Min	Max	Skew	Kurt
Proper noun	t-1	990	477.04	279.12	150.00	2,000.00	2.19	4.72
	t0	990	830.40	938.82	150.00	2,000.00	0.57	-0.96
	t1	990	549.68	410.20	150.00	2,000.00	1.62	2.12
	t2	990	477.93	333.17	172.85	2,000.00	1.98	3.67
	t3	990	450.81	260.92	150.00	2,000.00	2.25	5.11
Common noun	t-1	990	417.65	226.00	150.00	2,000.00	2.50	6.71
	t0	990	1,131.10	1,307.92	150.00	2,000.00	0.05	-1.42
	t1	990	481.95	303.92	150.00	2,000.00	2.14	4.82
	t2	990	459.34	298.84	150.00	2,000.00	2.08	4.49
	t3	990	457.38	278.40	150.00	2,000.00	2.36	6.38

Table 5. Experiment 1 proper nouns LMM

Parameters	Fixed effects						Random effects	
	β	SE	df	t	p	VIF	Person	Item
							SD	SD
(Intercept)	-0.50	0.04	52.77	-11.11	.000		0.20	0.06
t-1 vs. t0	0.11	0.03	4,155.41	3.38	.000	2.39		
t-1 vs. t1	0.10	0.02	4,841.18	4.94	.000	1.73		
t-1 vs. t2	-0.04	0.02	4,839.81	-1.92	.055	1.70		
t-1 vs. t3	-0.06	0.02	4,839.99	-2.89	.004	1.65		
Target log frequency	-0.04	0.02	108.74	-2.09	.039	1.98		
Experiment order	-0.15	0.05	31.00	-3.38	.002	1.00		
Length	0.17	0.01	51.40	13.89	.000	1.16	0.04	
Log frequency	-0.07	0.02	70.35	-3.11	.003	1.46	0.09	
t-1 vs. t0: Target log freq	0.01	0.02	4,828.81	0.69	.493	1.91		
t-1 vs. t1: Target log freq	0.04	0.02	4,816.57	2.37	.018	1.69		
t-1 vs. t2: Target log freq	0.05	0.02	4,821.48	2.65	.008	1.72		
t-1 vs. t3: Target log freq	0.06	0.02	4,816.69	2.99	.003	1.70		

Note: Model formula = $\log(\text{rt}) \sim \text{pnLocation} * \text{t.log_freq} + (1 + \text{length} + \text{log_freq} | \text{id}) + (1 | \text{item}) + \text{exptOrder} + \text{length} + \text{log_freq}$.

The LMM results in Table 5 provided initial evidence of the PNs' tendency to disrupt reading fluency, and revealed the existence of a spillover effect once the random and fixed effects were accounted for. The fixed-effect coefficients reported in Table 5 showed that the RTs in the region preceding the target word were significantly differently from the target region (t-1 vs. t0), $\beta = 0.11$, $t(4,155.41) = 3.38$, with a slightly stronger effect displayed in the spillover region (t-1 vs. t1), $\beta = 0.10$, $t(4,841.18) = 4.94$. Despite no significant effect being observed in the second spillover region, there was a small effect in the third (t-1 vs. t3), $\beta = -0.06$, $t(4,839.99) = -2.89$. The negative coefficients in these latter two regions suggested that the participants' RTs increased in comparison with the t-1 region. Besides region length, $\beta = 0.17$, $t(51.40) = 13.89$ and log frequency, $\beta = -0.07$, $t(70.35) = -3.11$, were significant predictors of RTs, and the significant effect of experiment order, $\beta = -0.15$, $t(31.00) = -3.38$, implied that the participants who completed Experiment 2 first responded faster to the PNs in Experiment 1 than those who completed Experiment 1 first. This latter result justified our decision to randomly assign half the participants to complete the experiments in reverse order to account for fatigue and familiarity effects. With regard to the random effects, the by-person intercepts, $SD = 0.20$, displayed greater variance around the model intercept than the by-item intercepts, $SD = 0.06$, suggesting that person variance was more influential on RTs than item variance. The variance of the random slopes was similarly small, with the largest, $SD = 0.09$, registered for frequency. Effect sizes calculated with the *MuMin* package (Barton, 2020) for R revealed that the model explained approximately 42% of the variance in the data, conditional $R^2 = .42$, with just more than half accounted for by the fixed effects, marginal $R^2 = .23$.

In response to the second hypothesis, the LMM results suggested that PN frequency was a weak predictor of the RTs. In general, target-area frequency proved to be a weak predictor of the RTs across the t-1 to t3 regions, $\beta = -0.04$, $t(108.74) = -2.09$. When modeled as an interaction term with location, a significant effect was observed between the t-1 and t1 regions, $\beta = 0.04$, $t(4,817.57) = 2.37$, but not t-1 and t0, $\beta = 0.01$, $t(4,828.81) = 0.69$. Curiously, the strongest effect was observed in the third spillover

Table 6. Experiment 1 common nouns LMM

Parameters	Fixed effects						Random effects	
	β	SE	df	t	p	VIF	Person	Item
							SD	SD
(Intercept)	-0.53	0.05	54.58	-11.59	.000		0.21	0.07
t-1 vs. t0	0.25	0.03	4,470.74	8.22	.000	2.04		
t-1 vs. t1	0.11	0.02	4,822.48	6.07	.000	1.47		
t-1 vs. t2	0.00	0.02	4,829.96	-0.11	.910	1.66		
t-1 vs. t3	-0.06	0.02	4,841.60	-3.09	.002	1.81		
Target log frequency	0.00	0.02	86.59	-0.18	.860	1.77		
Experiment order	-0.17	0.05	31.00	-3.60	.001	1.00		
Length	0.20	0.01	53.14	16.08	.000	1.20	0.02	
Log frequency	-0.05	0.02	71.51	-2.08	.042	1.51	0.09	
t-1 vs. t0: Target log freq	-0.09	0.02	4,831.37	-4.10	.000	2.04		
t-1 vs. t1: Target log freq	-0.02	0.02	4,814.71	-1.25	.210	1.68		
t-1 vs. t2: Target log freq	-0.01	0.02	4,816.95	-0.79	.428	1.69		
t-1 vs. t3: Target log freq	0.00	0.02	4,816.89	-0.02	.981	1.68		

Note: Model formula = $\log(\text{rt}) \sim \text{nLocation} * \text{t.log_freq} + (1 + \text{length} + \text{log_freq} | \text{id}) + (1 | \text{item}) + \text{exptOrder} + \text{length} + \text{log_freq}$.

region (t-1 vs. t3), $\beta = 0.06$, $t(4,816.69) = 2.99$, suggesting that the higher-frequency PNs were associated with longer RT latencies later in the sentences.

To address the third hypothesis, which stated that the disruptions exerted by PNs would be comparable to CNs, the RT region comparisons in the fixed effects (i.e., t-1 vs. t0; t-1 vs. t1; t-1 vs. t2; and t-1 vs. t3) for the CN LMM displayed in Table 6 were compared with Table 5. The results show that these comparisons followed a similar pattern in both models, with t-1 vs. t0 displaying the largest effect, followed by t-1 vs. t1, with t-1 vs. t2 being nonsignificant and the RTs in t-1 being significantly longer than those in t3. The similarity of these patterns in both models lends support to the hypothesis that the disruptions exerted by the PNs and CNs were comparable. As with the PN LMM, significant effects for the t-1 vs. t0 comparison, $\beta = 0.25$, $t(4,470.74) = 8.22$, and the t-1 vs. t1 comparison, $\beta = 0.11$, $t(4,822.48) = 6.07$, were observed with the CN LMM. These results were the reverse of the PN LMM, in that CNs exerted longer RTs in the target region than in the spillover region. Also, in tandem with the PN LMM, the RTs were significantly faster in the third spillover region than in the region before the target word (t-1 vs. t3), $\beta = -0.06$, $t(4,841.60) = -3.09$. Once more, length, $\beta = 0.20$, $t(53.14) = 16.08$, frequency, $\beta = -0.05$, $t(71.51) = -2.08$, and experiment order, $\beta = -0.17$, $t(31.00) = -3.60$, were significant predictors, while target word frequency influenced the RTs in the t-1 vs. t0 comparison only, $\beta = -0.09$, $t(4,831.37) = -4.10$. The CN LMM random effects revealed practically identical results to the PN LMM. The model explained approximately 50% of the variance, conditional $R^2 = .50$, which was 8% more than the PN LMM. Approximately 33% was accounted for by the fixed effects, marginal $R^2 = .33$, which was double the approximately 16% explained by the random effects.

Further support for the third hypothesis was collected with a LMM containing the random-effects structure along with *part-of-speech* as the sole fixed effect predicting the RTs. Part-of-speech comprised the words from the five target regions dummy coded as either CN or PN. The resulting regression coefficient presented in Table 7, $\beta = -0.03$, $t(58.00) = 1.09$, suggested that there was no significant difference between the influence of PNs and CNs on the RTs (see Supplementary Materials E for model

Table 7. Experiment 1 part of speech LMM

Parameters	Fixed effects						Random effects	
	β	SE	df	t	p	VIF	Person	Item
							SD	SD
(Intercept)	-0.56	0.06	45.30	-12.33	.000		0.24	0.10
PN vs. CN	0.03	0.03	58.00	1.09	.281	2.04		

Note: Model formula = $\log(rt) \sim (1|id) + (1|item) + pos$.

details and assumption check). The effect size, marginal $R^2 = .00$, indicated that the fixed effect explained practically none of the variance in the RTs, while approximately 20%, conditional, $R^2 = .20$, was explained by the random effects, which was comparable to the other LMMs.

Experiment 2

Experiment 1 demonstrated that PNs were processed in a comparable manner to CNs when the items were matched for frequency and presented in decontextualized sentences. However, in naturalistic reading, texts usually contain continuing narratives that provide contextual explanations for PNs. Locations and characters are introduced and contextualized within the narrative, while multiple instances of the same PNs will be found throughout, and frequently on the same page. Repeated exposure may lead to habituation, which may diminish disfluencies caused by PNs in a similar manner to how repeated exposure affects CN processing (e.g., Elgort & Warren, 2014; Pellicer-Sánchez, 2016; Webb, 2007). Whereas Experiment 1 involved PNs in decontextualized sentences, Experiment 2 was designed to investigate the effect of repeated exposures to PNs on RTs while reading an authentic text. Participants read a graded reader extract and the effect of repeated exposures to PNs within a continuing narrative was quantified with a LMM.

Method

Participants

Experiment 2 was conducted to address the fourth hypothesis, which concerned the effects of repeated exposure to PNs on RTs. The experiment was administered to the same participants as Experiment 1 during the same session. Six participants' results were omitted from analysis because they responded correctly to less than 70.00% of the comprehension questions, leaving 38 participants' results for analysis. The CI overlap between the self-paced reading task comprehension question scores and mean reading times in Table 8 indicated no significant difference between the groups.

Instruments

Self-paced reading task

The self-paced reading task architecture from Experiment 1 was utilized in Experiment 2, with the only difference being the stimuli sentences (see Supplementary Materials F) and the nonrandom presentation.

Table 8. Experiment 2 participant descriptive statistics

Order	Parameter	<i>n</i>	<i>M</i>	<i>SD</i>	Min	Max	Skew	Kurt	95% CIs ^a	
									<i>LL</i>	<i>UL</i>
1-2	Comprehension (%)	19	80.00	7.82	70.00	95.00	0.08	-1.24	76.23	83.77
	Mean RT	19	567.94	73.85	442.10	698.45	-0.06	-1.05	532.36	603.52
	Meaning recall	17	44.24	9.41	28.00	58.00	-0.39	-1.42	39.40	49.08
2-1	Comprehension (%)	19	85.53	8.96	70.00	100.00	-0.26	-1.19	81.21	89.95
	Mean RT	19	560.41	126.06	281.94	800.37	0.08	-0.09	499.68	621.14
	Meaning recall	18	43.17	10.61	24.00	60.00	-0.10	-1.48	37.89	48.45

Note: Three participants did not complete the meaning recall test.

^aCIs calculated with values from a *t* distribution, $t(18) = 2.10$, $t(17) = 2.11$, and $t(16) = 2.12$ to account for small samples.

Stimuli sentences. In total, the self-paced reading task for Experiment 2 involved 68 sentences presented to participants in a nonrandomized order. Sixty sentences were extracted from a chapter of the graded reader version of *The Jungle Book*, which was chosen because a relatively high 9.12% of the text consisted of PNs (see Table 2), despite it being targeted toward low-proficiency learners. Eight practice sentences were created to contextualize the narrative and familiarize participants with the self-paced reading task. Unlike Experiment 1, Experiment 2 included reported speech because it was constructed to mirror the natural reading process as closely as a self-paced reading task can. However, some sentences were altered to ensure that (a) the target PN did not occur toward the end of the sentence, (b) only one PN occurred in each sentence, and (c) the sentences were no longer than 90 characters long to fit on a single line. The stimuli sentences contained four unique PNs that were compared; *Bagheera* (9 occurrences), *Kaa* (10 occurrences), *Mowgli* (8 occurrences), and *Baloo* (8 occurrences).

To ensure that the participants would not be familiar with the target PNs contained within the stimuli sentences, a short Google Form questionnaire was administered to 80 Japanese students at a different university from the self-paced reading experiment sample. When shown pictures of *Mowgli* and *Baloo* from the 1967 animated movie and the 2016 live-action movie, only four (5.00%) students correctly identified *Mowgli* and only two (2.50%) correctly identified *Baloo*. This approach was preferred over more traditional familiarity rating, such as Likert scales, to avoid participants overestimating their knowledge of the PNs. These results suggested that the target population of Japanese university students were not familiar with the target PNs.

Procedure

The procedure of Experiment 2 mirrored that of Experiment 1. Each stimulus sentence was presented in the same order as in the graded reader and each of the four target PNs occurred between 8 and 10 times. Twenty (33.33%) of the 60 target sentences were followed by comprehension questions (see Supplementary Materials C).

Analysis

As with Experiment 1, the self-paced reading data were analyzed with boxplots and LMMs. RTs below 150 ms and above 2,000 ms were winsorized, affecting 4.29% of the data (0.67% and 3.62% at the lower and upper boundaries, respectively). The random-effects structure was the same as the LMMs in Experiment 1, except random intercepts

were also modeled for each of the target PNs. Log COCA frequency was omitted from the fixed effects because it proved to be a weak predictor and resulted in the VIF for the $t-1$ vs. t_0 contrast to exceed 2.5. An additional eight-level categorical variable, *pnOccurrence*, was included in the model that represented the chronological order of appearance of each PN, whereby "1" represented the RT of the first occurrence of each PN and also acted as the reference level. This allowed all of the RTs for the following PN occurrences to be compared with the first occurrence, enabling the effect of PN repetitions to be quantified.

Results

The RTs observed in the target region (t_0) were plotted to analyze the difference in median and RT distribution at each occurrence (see Figure 4) and to observe the mean RT fluctuation for each PN occurrence across the 60 sentences (Figure 5). The target region was focused upon because this is where the RTs were longest in Experiment 1 (see Figure 3). In both figures, the dotted line represents the median RT (466.44 ms; IQR = 342.09) for all words and participants. The boxplots in Figure 4 revealed that for three of the target PNs (*Bagheera*, *Baloo*, and *Mowgli*), the RTs at the first occurrence were larger and more dispersed but regressed toward the mean RT as occurrences increased. The exception was the single-syllable PN, *Kaa*, which displayed a relatively small decrease across the eight occurrences. The boxplots for *Mowgli* illustrated a disruption to this pattern between the fourth and fifth occurrence. A comparison with the line charts in Figure 5 showed that there was a long gap of more than 20 sentences

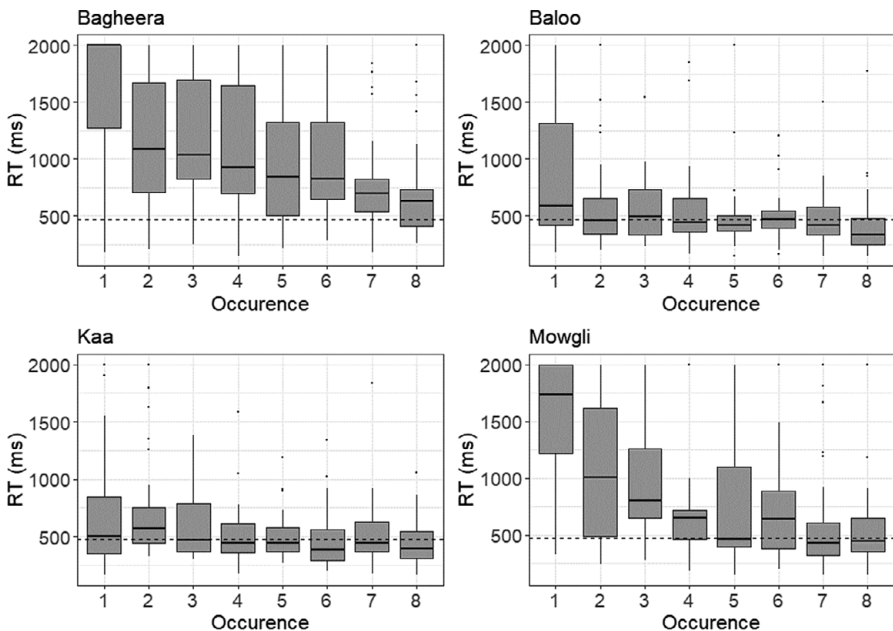


Figure 4. Boxplots of proper noun raw RTs by occurrence.
Note: Dotted line represents Median = 466.44 ms, IQR = 342.09.

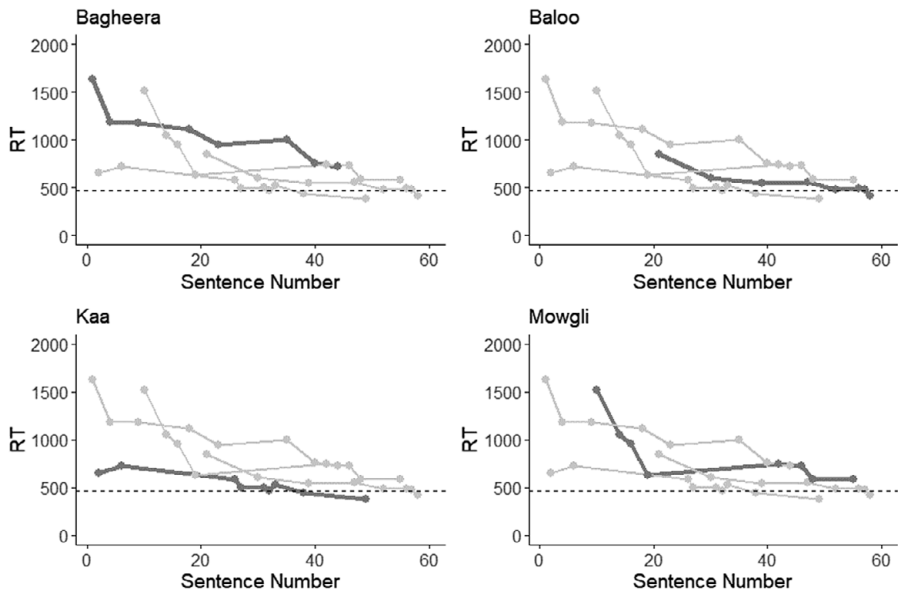


Figure 5. Line charts mapping proper noun mean RTs by appearance (sentence number).
 Note: Dotted line represents Median = 466.44 ms, IQR = 342.09.

(more than a third of the passage's length) between these two occurrences, which most likely influenced the larger RTs visible at the fifth occurrence.

The fixed effects for the LMM displayed in Table 9 also supported the hypothesis that PN processing speeds increased with a relatively small amount of repetition. Although the difference between the first two occurrences across all four PNs (Occurrence 1 vs. 2) was nonsignificant, $\beta = -0.07$, $t(21.36) = -1.52$, the difference by the third occurrence was, $\beta = -0.16$, $t(20.76) = -3.70$. The effect grew across the occurrences and was at its strongest by the eighth appearance, $\beta = -0.42$, $t(20.62) = -9.54$. When the variance resulting from the occurrences was taken into consideration, the difference between the t-1 and t0 regions was not significant, $\beta = 0.03$, $t(179.34) = 1.33$, nor was the t-1 vs. t1 comparison, $\beta = -0.01$, $t(5,678.71) = 0.72$. However, the t-1 vs. t2 and t-1 vs. t3 comparisons were significant with negative coefficients, indicating that the RTs observed in the second and third spillover regions were statistically faster than the pretarget region. In line with the Experiment 1 results, length, $\beta = 0.16$, $t(45.87) = 13.74$, and experiment order, $\beta = 0.17$, $t(36.00) = 2.64$, were significant predictors of RT latencies. Again, this latter result justified our decision to randomly assign half the participants to complete the experiments in reverse order. The random effects followed a similar pattern to those observed in Experiment 1. The by-participant intercepts displayed greater variance than the by-item and by-PN intercepts. The model explained 42% of the variance, conditional $R^2 = .42$, with little less half accounted for by the fixed effects, marginal $R^2 = .20$.

Discussion

In the present study, two self-paced reading experiments were conducted to investigate four hypotheses relating to the effect of PNs on L2 English learners' reading fluency.

Table 9. Experiment 2 proper nouns LMM

Parameters	Fixed effects						Random effects		
	β	SE	df	t	p	VIF	Person	Item	PN
							SD	SD	SD
(Intercept)	-0.48	0.07	14.73	-6.52	.000		0.22	0.05	0.09
Occurrence 1 vs. 2	-0.07	0.04	21.36	-1.52	.142	1.73			
Occurrence 1 vs. 3	-0.16	0.04	20.76	-3.70	.001	1.75			
Occurrence 1 vs. 4	-0.30	0.04	20.90	-6.94	.000	1.74			
Occurrence 1 vs. 5	-0.30	0.04	21.48	-6.84	.000	1.73			
Occurrence 1 vs. 6	-0.34	0.04	20.69	-7.72	.000	1.75			
Occurrence 1 vs. 7	-0.36	0.04	21.38	-8.18	.000	1.73			
Occurrence 1 vs. 8	-0.42	0.04	20.62	-9.54	.000	1.74			
t-1 vs. t0	0.03	0.02	179.34	1.33	.184	1.47			
t-1 vs. t1	0.01	0.02	5,678.71	0.72	.473	1.54			
t-1 vs. t2	-0.05	0.02	5,686.01	-2.71	.007	1.53			
t-1 vs. t3	-0.11	0.02	5,709.42	-6.30	.000	1.52			
Experiment order	0.17	0.06	36.00	2.64	.012	1.00			
Length	0.16	0.01	45.87	13.74	.000	1.05	0.06		
Log freq	-	-	-	-	-	-	0.03		

Note: Model formula = log(rt) ~ (1 + length + freq|id) + (1|pn) + (1|item) + pnOccurrence + pnLocation + exptOrder + length.

The first hypothesis, which stated that PNs would significantly disrupt second-language learners' reading fluency, was supported by the self-paced reading task results. When compared with the RTs elicited by words preceding and following the PNs, the RTs elicited by the PNs were significantly longer. The second hypothesis, which stated that less frequent PNs would disrupt reading fluency more than high-frequency ones, was not confirmed. The third hypothesis, which stated that disruptions to reading fluency exerted by PNs would be comparable to those exerted by frequency-matched CNs, was supported. The boxplots and the regression coefficients showed that although CNs seemed to have a slightly greater effect on RTs, the patterns observed for PNs and CNs were comparable. The final hypothesis, which stated that repeated exposures to PNs would reduce RTs, was supported by the Experiment 2 results. The difference between the first occurrence RTs and the following occurrences gradually increased, as attested to by the regression coefficients and line charts.

When synthesized, the results indicated that PNs were processed by the L2 learners in a manner comparable to CNs that were matched for frequency. This suggests that L2 vocabulary researchers' causal-historical approach, whereby PNs have been treated as encyclopedic knowledge and assumed as known, deserves reassessment, and a descriptivist approach acknowledging PNs as a part of language might be more apt. The implications of both experiments' results will be discussed in relation to three points of interest: ER pedagogy, the qualities of the PNs, and an observed dispersion effect on the PN RTs.

The first point of interest relates to the implication for ER pedagogy. ER necessitates learners comprehending vast amounts of text with speed and fluency (Waring & McLean, 2015), and to this end low-frequency vocabulary is controlled by the implementation of lexical coverage thresholds. This information is provided by publishers to inform learners and teachers of GR suitability. The RTs observed in this study's two experiments indicated that PNs disrupt fluency in a similar manner to CNs of comparable frequency. Furthermore, Table 2 illustrates that the distribution of PNs

in GRs across proficiency stages is inconsistent. In *The Jungle Book*, which is intended for low-proficiency learners, 9.12% of the words are PNs, equating to more than 14 PNs per page on average. This percentage is greater than any other GR in the sample, including those intended for the most proficient readers. When considering that knowledge of 95–98 percent of vocabulary is required for comprehension (e.g., Laufer & Ravenhorst-Kalovski, 2010), 9.12% of potentially unknown words will undoubtedly disrupt fluency, which in turn may hinder comprehension (Beglar *et al.*, 2012).

The present study's results suggest that PNs warrant consideration by GR publishers in the same manner as CNs do. We are not suggesting that PN occurrences should be reduced by GR authors to meet arbitrary thresholds, although it might be the case that PNs should be treated as off-list words in the same manner that low-frequency CNs are. We do, however, believe that it is necessary for publishers to take PN coverage percentages into consideration when assigning GRs to proficiency levels. Additionally, because PNs affect reading fluency, and the main benefit of ER is arguably the development of reading fluency, it would be worthwhile for both language learners and teachers if publishers provided information regarding PN coverage percentages to raise awareness of which GRs are potentially more challenging.

The second point of interest relates to the qualities of PNs, specifically length and frequency. In both experiments, length proved to be the strongest predictor of reading times. The effect was slightly stronger for CNs, which were also one letter longer on average (see Table 3). The fact that longer nouns elicited longer reading times was expected, however the extent to which it dominated the model over frequency was not. In previous L2 SPR research, frequency was a stronger predictor of RTs than length, which was a nonsignificant predictor (Shantz, 2017). However, Shantz investigated the effect of grammaticality on four-word sequences, not single words like the present study. Furthermore, when length has been shown to be a significant RT predictor, such as in Tamura *et al.* (2019), the effect has not been as strong as in the present study. Although syllable length was controlled for in the present study, our results highlight the importance of controlling for letter length in L2 behavioral research, either at the item design stage or as a fixed effect in regression models.

Frequency was a significant RT predictor across the regions of interest for both PNs and CNs in Experiment 1, while the location by target word frequency interaction term revealed a relatively small target word frequency effect in the t-1 vs. t1 comparison for PNs, and a larger effect in the t-1 vs. t0 for the CNs. Although it should be acknowledged that t-1 vs. t1 frequency effect for PNs in Experiment 1 was a small effect, the results do indicate that corpus-based frequency is a weaker predictor of PNs than CNs that are matched for frequency. This is understandable because although the frequency effect is one of the most robust findings in psycholinguistic word recognition research (Cortese & Balota, 2012), frequency values are a mere proxy for an idiosyncratic concept: the number of encounters with a word. The learners in the present study might have previously encountered target CNs such as *trumpet*, or might be expected to parse the component words of compounds, such as *backache*, fairly easily. However, it is less likely that they would have encountered the target PNs matched with these CNs for frequency, which were *Garland* and *Deckard*, respectively, unless they had read the relevant graded readers. Furthermore, it is debatable whether the role of frequency is as relevant for L2 acquisition as for L1 (e.g., Von Stutterheim *et al.*, 2021), hence the relatively small role for frequency in the present study. For instance, the frequency values in the present study were extracted from COCA because it was the only reference corpus that contained all the PNs. However, it is questionable how pertinent the materials gathered in COCA are to the linguistic experience of Japanese university

students. To summarize, although a frequency effect was hypothesized for PNs, the size of the effect was relatively small in comparison with the CNs and was nonexistent in some of the expected regions.

The final point of interest relates to a dispersion effect observed between repetition and PN RTs. In accordance with Pellicer-Sánchez's (2016) findings, the results indicated that form recognition of the four PNs in Experiment 2 seemed to have been achieved within eight encounters. In fact, Figures 4 and 5 indicated that the RTs for *Bagheera* had stabilized by seven encounters, *Mowgli* by four encounters, *Kaa* by three, and *Baloo* by as few as two. This result aligns with Webb (2007), who found that form recognition scores for nonwords significantly improved after three exposures. However, the line charts in Figure 5 indicated that RTs were not merely influenced by the number of occurrences, but also the distance between, or the *dispersion* of those occurrences. If an unfamiliar PN occurs four times within the space of 10 sentences, it is likely to be processed with similar speed as any other word upon the final occurrence. For instance, the first four occurrences of *Mowgli*. However, after a gap of more than 15 sentences the RTs increased again (see Figure 5; *Bagheera* and *Mowgli*). It is not illogical to suggest that a longer gap of perhaps 100 sentences might even result in RTs closer to those of the first occurrence than to the mean RT. More precise characterization of dispersion effects deserves attention in future research.

The dispersion effect might also have been moderated by two other variables: length and orthotactic probability. The shortest PN, *Kaa*, comprised three letters and one syllable. In contrast with the other PNs, Figure 5 illustrates how the RTs for *Kaa* were similar to the mean RT from the first occurrence and decreased at a slower rate. Furthermore, the third appearance of *Kaa* occurred approximately 20 sentences after the second appearance, but unlike with *Bagheera* and *Baloo*, there was no dispersion effect displayed for *Kaa*. High orthotactic probability, which relates to the sequential letter probability, has been shown to facilitate intentional L2 vocabulary acquisition (Bordag et al., 2017). Figure 5 indicates that *Bagheera* and *Mowgli* resulted in larger RT latencies than *Kaa* and *Baloo*. Both *Bagheera* and *Mowgli* contain character-level *n*-grams that might be considered rare in English, such as *ghe* in *Bagheera* or *wgl* in *Mowgli*. It might be the case that such *n*-grams influence processing latencies, and are more likely to occur in non-English names. An investigation of the influence of character-level *n*-gram probability on PN RTs was beyond the scope of the present study, but future research is warranted.

Limitations

Despite our best intentions, the present study is not without limitations. First, self-paced reading forbade participants from regressing to previously read words, which renders the reading process somewhat unnatural and is particularly pertinent to the passage reading investigated in Experiment 2. Alternative methodology, such as eye-tracking, would allow regressions to be quantified and potentially sheds more light on our findings. Although eye-tracking was not possible for the present study, our materials are available on the Open Science Framework (OSF; <https://osf.io/xn7ya/>) for replication purposes, along with the Supplementary Material files, data, and R scripts. Second, in recent surveys of L2 instructed vocabulary acquisition studies, researchers have called for improved experimental designs using power analysis, randomization, and multisite samples (Nickin & Vitta, 2021; Vitta et al., 2022). Despite two of the three requirements being incorporated in the present

study's design, a multisite sample was not recruited for the final experiment, which potentially harms the generalizability of the results. This was somewhat mitigated by the fact that the pilot study and main study participants were recruited from different universities and the findings were congruent with one another (see Supplementary Materials A). However, both universities were located in Japan, which constitutes a third limitation to this study. We cannot presume that the proper noun reading behavior displayed by the Japanese learners in the present sample is the same as would be displayed by English users from Germany, for example, whose L1 orthography is closer to English than Japanese. Future research should attempt to replicate these results with learners of non-Japanese L1 orthographies. Additionally, the sample comprised learners from CEFR B1 classes with similar L2 history. It is likely that less PN-induced disfluency is incurred by more advanced learners, thus research is required with a wide range of proficiency levels to certify this hypothesis. Future research involving graded readers and learners of non-English languages would also be worthwhile, as would work quantifying the relationship between PN-induced dysfluencies and comprehension, which was beyond the scope of the present study. Fourth, as mentioned previously, corpus-based frequency counts constitute a mere proxy for which the language that learners are exposed. It is questionable how relevant the COCA-based frequency counts obtained for the proper nouns in this study were with respect to the language experienced by the participants, hence the relatively weak frequency effect that was observed. Finally, due to the low frequency of many of the target PNs and CNs, the number of available sentences were limited, and thus a number of potentially influential variables were not controlled for, such as whether the PNs and CNs were in a subject or object position.

Conclusion

The present study involved two self-paced reading experiments to investigate the effect of PNs on English learners' reading fluency. Until now, L2 vocabulary researchers have worked under the assumption that PNs are known words or can be easily identified as PNs because of capitalization. In Experiment 1, a set of 30 PNs that were presented to participants in decontextualized sentences were processed in a similar manner to 30 CNs matched in terms of syllable length and COCA frequency, although the magnitude of the frequency effect was smaller for PNs than CNs. In Experiment 2, when the same participants were presented with PNs embedded in a narrative context, a reduction in reading times indicated that form recognition had been achieved within eight occurrences, which was synchronous with previous research involving non-PN vocabulary. However, the results also indicated the presence of a dispersion effect, whereby long gaps between the occurrences of certain PNs increased the reading times to previously observed lengths, and this dispersion effect warrants further research. These results have implications for graded reader publishers in that PN coverage should be considered when assigning books to proficiency levels, and that PN coverage percentages should be provided to teachers and students to help them determine the appropriacy of a book for extensive reading. It is also arguable that because PNs are processed in a similar manner to CNs of comparable frequency, they should not be counted as off-list for lexical coverage counts. In conclusion, the results of the present study imply that PNs should not be assumed to be known by second language learners, but should be assumed to disrupt reading fluency, and thus potentially inhibit reading comprehension as much as equally frequent or infrequent CNs.

Acknowledgments. We would like to thank Joseph P. Vitta and Paul Nation for their feedback on earlier drafts on this study.

Data Availability Statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials and data are available at <https://osf.io/xn7ya/>.

Declaration of Competing Interests. The authors hereby declare that there are no competing interests in relation to this project and submitted manuscript.

References

- Allerton, D. J. (1987). The linguistic and sociolinguistic status of proper names: What are they, and who do they belong to? *Journal of Pragmatics*, 11, 61–92. [https://doi.org/10.1016/0378-2166\(87\)90153-6](https://doi.org/10.1016/0378-2166(87)90153-6)
- Anthony, L. (2015). TagAnt (Version 1.2.0) [Computer Software]. Waseda University. <http://www.antlab.sci.waseda.ac.jp/>
- Baayen, R. H. (2008). *Analyzing linguistic data*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511801686>
- Barton, K. (2020). *MuMIn: Multi-model inference*. R package version 1.43.17. <https://CRAN.R-project.org/package=MuMIn>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beglar, D., & Hunt, A. (2014). Pleasure reading and reading rate gains. *Reading in a Foreign Language*, 26, 29–48. <https://doi.org/10.1257/66684>
- Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning*, 62, 665–703. <https://doi.org/10.1111/j.1467-9922.2011.00651.x>
- Bordag, D., Kirschenbaum, A., Rogahn, M., & Tschirner, E. (2017). The role of orthotactic probability in incidental and intentional vocabulary acquisition. *Second Language Research*, 33, 147–178. <https://doi.org/10.1177/0267658316665879>
- Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language*, 22, 355–361. <https://doi.org/10.1257/66842>
- Bui, T. B., & Macalister, J. (2021). Online extensive reading in an EFL context: Investigating reading fluency and perceptions. *Reading in a Foreign Language*, 33, 1–29. <https://doi.org/10.1257/67391>
- Bultena, S., Dijkstra, T., & van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, 67, 1214–1241. <https://doi.org/10.1080/17470218.2013.853090>
- Cortese, M. J., & Balota, D. A. (2012). Visual word recognition in skilled adult readers. In M. J. Spivey, K. McRae, & M. F. Joannis (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 159–185). Cambridge University Press. <https://doi.org/10.1017/CBO9781139029377>
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. <https://doi.org/10.4324/9780203807002>
- Davies, M. (2008–). *The Corpus of Contemporary American English (COCA)*. <https://www.english-corpora.org/coca/>
- Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, 64, 365–414. <https://doi.org/10.1111/lang.12052>
- Erten, I. H., & Razi, S. (2009). The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language*, 21, 60–77.
- Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage*. Houghton Mifflin Company.
- Frank, A. (2014). *R-hacks*. <https://github.com/auf frank/R-hacks/blob/master/mer-utils.R>
- Frank, S. L., Trompenaars, T., & Vasisht, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, 40, 554–578. <https://doi.org/10.1111/cogs.12247>
- Godfroid, A. (2020). *Eye-tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge. <https://doi.org/10.4324/9781315775616>
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

- Green, P., & MacLeod, C. J. (2016). *SIMR*: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7, 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Hsu, C.-H., Lee, C.-Y., & Marantz, A. (2011). Effects of visual complexity and sublexical information in the occipitotemporal cortex in the reading of Chinese phonograms: A single-trial analysis with MEG. *Brain and Language*, 117, 1–11. <https://doi.org/10.1016/j.bandl.2010.10.002>
- Jiang, N. (2012). *Conducting reaction time research in second language studies*. Routledge. <https://doi.org/10.4324/9780203146255>
- Klassen, K. (2020). The effect of cultural familiarity on reading comprehension: An approximate replication of Erten and Razi (2009). *Reading Psychology*, 41, 205–228. <https://doi.org/10.1080/02702711.2020.1768976>
- Klassen, K. (2021). Proper name theory and implications for second language reading. *Language Teaching*, 55, 149–155. <https://doi.org/10.1017/S026144482100015X>
- Kobeleva, P. P. (2012). Second language listening and unfamiliar proper names: comprehension barrier? *RELC Journal*, 43, 83–98. <https://doi.org/10.1177/0033688212440637>
- Kramer, B., & McLean, S. (2019). L2 reading rate and word length: The necessity of character-based measurement. *Reading in a Foreign Language*, 31, 201–225. <https://doi.org/10.1025/66930>
- Kripke, S. (1980). *Naming and necessity*. Blackwell.
- Kuhn, M., & Stahl, S. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95, 3–21. <https://doi.org/10.1037/0022-0663.95.1.3>
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22, 15–30. <https://doi.org/10.125/66648>
- Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 1–22. <https://doi.org/10.1016/j.jml.2020.104092>
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63, 59–82.
- Nation, I. S. P., & Kobeleva, P. P. (2016). Proper nouns. In I. S. P. Nation (Ed.), *Making and using wordlists for language learning and testing* (pp. 55–64). Benjamins. <https://doi.org/10.1075/z.208>
- Nation, I. S. P., & Malarcher, C. (2007). *Reading for speed and fluency*. Compass.
- Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Cengage Learning.
- Nicklin, C. (2021). Developing a measure of proper name familiarity for Japanese university students. *Vocabulary Learning and Instruction*, 10, 91–100. <https://doi.org/10.7820/vli.v10.2.nicklin>
- Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research: A synthesis and data re-analysis from self-paced reading. *Annual Review of Applied Linguistics*, 40, 26–55. <https://doi.org/10.1017/S0267190520000057>
- Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*, 105, 218–236. <https://doi.org/10.1111/modl.12692>
- Nisbet, K., Bertram, R., Erlinghagen, C., Pieczykolan, A., & Kuperman, V. (2021). Quantifying the difference in reading fluency between L1 and L2 readers of English. *Studies in Second Language Acquisition*, 44, 407–434. <https://doi.org/10.1017/S0272263121000279>
- Opitz, A., & Bordag, D. (2021). The impact of orthography on lexical access: The case of capitalization and word category information in L1 and L2 German. *Studies in Second Language Acquisition*, 44, 1194–1209. <https://doi.org/10.1017/S0272263121000711>
- Pearce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition *from and while* reading: An eye-tracking study. *Studies in Second Language Acquisition*, 38, 97–130. <https://doi.org/10.1017/S0272263115000224>
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.
- R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>
- Shantz, K. (2017). Phrase frequency, proficiency and grammaticality interact in non-native processing: Implications for theories of SLA. *Second Language Research*, 33, 91–118. <https://doi.org/10.1177/0267658316673403>

- Stewart, J., Vitta, J. P., Nicklin, C., McLean, S., Pinchbeck, G. G., & Kramer, B. (2021). The relationship between word difficulty and frequency: A response to Hashimoto (2021). *Language Assessment Quarterly*, 19, 90–101. <https://doi.org/10.1080/15434303.2021.1992629>
- Swanson, C. L., James, L. E., & Ingram, R. E. (2021). Incidental learning of proper names and “earwitness” recall. *Memory*, 29, 637–644. <https://doi.org/10.1080/09658211.2021.1925301>
- Tamura, Y., Fukuta, J., Nishimura, Y., Harada, Y., Hara, K., & Kato, D. (2019). Japanese EFL learners' sentence processing of conceptual plurality: An analysis focusing on reciprocal verbs. *Applied Psycholinguistics*, 40, 59–91. <https://doi.org/10.1017/S0142716418000450>
- Valentine, T., Brennen, T., & Brédart, S. (1996). *The cognitive psychology of proper names: On the importance of being Ernest*. Routledge. <https://doi.org/10.4324/9780203285763>
- van Dongen, N., Colombo, M., Romero, F., & Sprenger, J. (2021). Intuitions about the referent of proper nouns: A meta-analysis. *Review of Philosophy and Psychology*, 12, 745–774. <https://doi.org/10.1007/s13164-020-00503-8>
- Van Langendonck, W. (2007). *Theory and typology of proper names*. De Gruyter Mouton. <https://doi.org/10.1515/9783110197853>
- Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size-driven sample size planning, randomization, and multi-site use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, 44, 1424–1448. <https://doi.org/10.1017/s0272263121000541>
- Von Stutterheim, C., Lambert, M., & Gerwin, J. (2021). Limitations on the role of frequency in L2 acquisition. *Language and Cognition*, 13, 291–321. <https://doi.org/10.1017/langcog.2021.5>
- Waring, R., & McLean, S. (2015). Exploration of the core and variable dimensions of extensive reading research and pedagogy. *Reading in a Foreign Language*, 27, 160–167. <https://doi.org/10.125/66708>
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28, 46–65. <https://doi.org/10.1093/applin/aml048>
- Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33, 278–293. <https://doi.org/10.125/67407>
- Webb, S., & Chang, A. C-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, 19, 667–686. <https://doi.org/10.1177/1362168814559800>
- Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly*, 47, 300–322. <https://doi.org/10.1002.tesq.70>
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. <https://doi.org/10.48550/arXiv.1308.5499>
- Winter, B. (2020). *Statistics for linguists: An introduction using R*. Routledge. <https://doi.org/10.4324/9781315165547>
- Zipf, G. K. (1935). *The psycho-biology of language*. Houghton Mifflin.

Cite this article: Nicklin, C., Patterson, A. and McLean, S. (2023). Quantifying proper nouns' influence on L2 English learners' reading fluency. *Studies in Second Language Acquisition*, 45: 906–929. <https://doi.org/10.1017/S027226312200050X>