Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing*

SUHAD SONBUL
Umm Al-Qura University
University of Nottingham

This study explored whether native speakers of English and non-natives are sensitive to corpus-derived frequency of synonymous adjective-noun collocations (e.g., fatal mistake, awful mistake, and extreme mistake) and whether level of proficiency can influence this sensitivity. Both off-line (typicality rating task) and on-line (eye-movement) measures were employed. Off-line results showed that both natives and non-natives were sensitive to collocational frequency with clearer effects for non-natives as their proficiency increased. On-line, however, proficiency had no effect on sensitivity to frequency; both natives and non-natives showed early sensitivity to collocational frequency (first pass reading time). This on-line sensitivity disappeared later in processing for both groups (total reading time and fixation count). Results are discussed in light of usage-based theories of language acquisition and processing.

Introduction

Language users have been shown to be sensitive to the frequency of linguistic units along the continuum from the smallest units in language (i.e., phonemes) to the largest units (i.e., lexical constructions) (see Ellis, 2002, for an overview of evidence). This evidence is compatible with the emergentist theories of language acquisition. The most prominent of these are usage-based models which posit that linguistic knowledge is tightly related to language experience. Within these theories, language development is viewed as the associative learning (or *entrenchment*) of constructions (chunks) and their frequency-related aspects (Bybee, 2001, 2006; Langacker, 2000). Every exposure to a linguistic unit (from the smallest to the largest) acts as a memory trace which either reinforces or modifies existing knowledge. In contrast with this emergentist view, the words-and-rules approach (Pinker, 1999; Ullman, 2001) assumes that lexical knowledge is declarative (in the form of memorized forms) while grammatical knowledge is procedural (set of rules applied to combine these forms). According to this approach, frequency can affect the processing of memorized lexical forms but cannot influence the processing of longer constructions (which should be computed based on rules only).

Most of the available research on frequency effects on language processing has been concerned with the processing of individual words and has indisputably shown a clear effect of distributional properties (see, for example, Balota & Chumbley, 1984; Morton, 1969; Rayner & Duffy, 1986). More recently, a number of research studies have shown that language users are also sensitive to frequency effects on constructions larger than words. These constructions are collectively referred to as formulaic sequences (i.e., holistically processed sequences of words) and are categorized into various types: chunks (*by the way*), frozen metaphors (*fishing for compliments*), idioms (*kick the bucket*), and collocations (*fatal mistake*). However, there are reasons to believe that collocations (word combinations which typically co-occur) should be viewed differently. One characteristic of collocations which distinguishes them from other types of formulaic sequences is what Wray (2002) calls "fluidness" as opposed to the "fixedness" of idioms. Collocations have more to do with tendencies than exclusiveness. While the idiom *kick the bucket* is fixed in form and meaning and cannot be modified, the collocation *fatal mistake* is not as fixed. *Fatal* tends to occur frequently with *mistake*, but this does not exclude other (less frequent) combinations such as *awful mistake* which might sound acceptable as well.

Existing research exploring frequency effects on the processing of formulaic sequences, and more specifically

Address for correspondence:
Suhad Sonbul, P.O.Box: 10424 Makkah (zip code: 21955) Saudi Arabia
sssonbul@uqu.edu.sa

collocations, (e.g., Arnon & Snider, 2010; Bell, Jurafsky, Fosler-Lussier, Girand, Gregory & Gildea, 2003; Sosa & MacFarlane, 2002; Tremblay, Derwing, Libben & Westbury, 2011) is limited in a number of ways. First, most studies employed either off-line or on-line measures not allowing for the direct comparison of on-line/off-line processes.[1] Second, with the exception of two recent studies (i.e., Siyanova-Chanturia, Conklin & van Heuven, 2011b; Wolter & Gyllstad, 2013), most of the available research explored the issue with regard to native speakers only. Including both natives and non-natives in the same study would allow for testing the assumptions of usage-based models in both L1 (first language) and L2 (second language) learning contexts. Finally, semantic relatedness was only controlled for in one study (Siyanova-Chanturia et al., 2011b) which did not specifically look into the processing of collocations. Controlling for semantic relatedness is essential in on-line measures to minimize the potential influence of implausibility/unrelatedness on any facilitative frequency effects revealed.

The present study combines both off-line and on-line measures in order to assess native and non-native English language users' sensitivity to the corpus frequency of semantically-related (i.e., synonymous) adjective-noun collocations. Moreover, a rough estimate of proficiency (scores in the Vocabulary Levels Test) will be included to investigate whether sensitivity to collocational frequency is influenced by level of proficiency.

Frequency Effects on Formulaic Language Processing

Evidence for frequency effects on the processing of formulaic sequences has surfaced in the past few years. Arnon and Snider (2010) employed a timed off-line judgment task instructing their native-speaker participants to decide whether four-word sequences were possible in English or not (a YES/ NO phrasal decision task). In Experiment 1, the high cut-off bin (the condition with a big frequency difference) involved a 10 per million gap while the low cut-off bin (the condition with a small frequency difference) involved a 1 per million gap only.[2] Results showed sensitivity to frequency; a significant reaction time (RT) advantage was found for both cut-off bins. In Experiment 2 the effect was also established for a mid

cut-off bin (5 per million gap). Finally, a meta-analysis (regression analysis run on data from both experiments) confirmed the effect; raw sequence frequency predicted RT behaviour.

Another similar recent study (Wolter & Gyllstad, 2013) looked specifically at the effect of frequency on the off-line processing of collocations. The study employed a timed acceptability judgment task (similar to the one used by Arnon & Snider, 2010) and included adjective-noun collocations from various frequency levels. Both native speakers of English and advanced non-natives (L1 Swedish, L2 English) completed the task. Target collocations were extracted from the Corpus of Contemporary American English, COCA (Davies, 2008), to reflect both congruent (L1 = L2, e.g., *human rights*) and incongruent (L1 ≠ L2, e.g., *bottom line*) collocations. In addition to the target collocations, non-collocate pairs were constructed (through randomly matching adjectives with nouns, e.g., *angry use*). Natives showed sensitivity to collocational status (with a clear advantage for true collocations over non-collocate pairs) and to collocational frequency (with log collocational frequency predicting RT performance) regardless of congruence. As for the Swedish non-native learners, they showed a similar sensitivity both to collocational status and collocational frequency. However, unlike natives, their behaviour was clearly affected by congruence with shorter RTs for congruent items than incongruent items (leading to the conclusion that L1 exercises an effect on L2 processing even for advanced non-natives). Although these two studies provided clear evidence for frequency effects on native/non-native collocational processing, the task required an explicit judgment and, thus, did not assess on-line processing. On-line measures should tap into performance as it unfolds in real time (without explicit judgment) using RT methods such as priming/monitoring or the eye-tracking methodology.

Two studies have employed on-line RT measures in exploring frequency effects on natives' on-line collocational processing. Sosa and MacFarlane's (2002) native-speaker participants completed a monitoring task in which they had to detect the node *of* in two-word combinations with various parts of speech (e.g., *kind of*, *because of*). They divided their items (derived from the Switchboard Corpus, available online at http://catalog.ldc.upenn.edu/LDC98S75) into four frequency levels. Results showed that natives' reaction times were longer (i.e., slower processing) under the highest frequency level in comparison with the other three levels combined. It was concluded that more frequent

---

[1]   Marinis (2003, p. 144) distinguishes between two types of measures in the field of Second Language Acquisition: off-line and on-line. Traditional off-line measures are those which involve an explicit judgement such as grammaticality/typicality judgment tasks. On-line measures, on the other hand, tap into "the mental processes involved while reading or listening to words or sentences in real time" such as self-paced reading, priming, eye-tracking, and neurophysiological techniques.

[2]   The examples below represent the high and low cut off bins:

	High cut off bin		Low cut off bin
Don't have to worry	15.3 per million	*I want to sit*	3.6 per million
Don't have to wait	1.5 per million	*I want to say*	0.2 per million

collocations are stored holistically hindering the access to individual items. In a similar study employing the priming paradigm, Durrant and Doherty (2010) investigated both strategic and automatic collocational priming in two experiments.[3] Four frequency levels were included: low frequency combinations, moderate collocations, pure non-associated frequent collocations, and associated frequent collocations. Each level was paired with randomly constructed non-collocate pairs (e.g., *armed concept* as a control pair for *main concept*). Results of the strategic priming experiment showed facilitative effects for both types of highly frequent collocations over non-collocate pairs. On the other hand, automatic priming effects were only found for associated frequent collocations. The two studies presented above seem to suggest that native speakers are sensitive to collocational frequency on-line. It should be noted, however, that these studies suffer from an important limitation: they looked at arbitrarily divided levels of frequency but did not attempt to look at raw frequency effects across the continuum.

Reali and Christiansen (2007) explored raw frequency effects on collocational processing. They employed both off-line (complexity-rating task, Experiment 1) and on-line (self-paced reading task, Experiment 2) to test natives' processing of pronominal object-relative clauses comprising chunks of various frequency levels (high*, who I met disturbed*, versus low, *who I disturbed met*). Off-line, results showed that natives judged the sentences comprising the high-frequency chunks to be less complex than those containing the low-frequency ones. Similarly, on-line, sentences with highly frequent sequences were read faster (shorter RTs) than those with low-frequency ones: log raw clause frequency predicted RT performance.

The three studies reviewed above employed three RT measures (i.e., priming, monitoring and self-paced reading) to tap into the on-line processing of formulaic sequences. In comparison with these RT measures, an eye-tracking methodology allows for the separation of early (first pass reading time) and late (total reading time/number of fixations) comprehension processes during real-time reading (see Liversedge, Paterson & Pickering, 1998). Two of the earlier researchers who aimed at investigating whether collocational frequency exercises an influence on eye-movement behaviour were McDonald and Shillcock (2003). Two eye-tracking experiments explored natives' processing of highly predictable as opposed to low predictable verb-noun collocations (*avoid confusion* versus *avoid discovery*, respectively). Predictability was defined here as transitional probability

based on BNC (British National Corpus) frequency counts. Items were embedded in a short context in Experiment 1 but were inserted into a more natural reading passage in Experiment 2. Results of both experiments showed that natives' fixation durations were predicted by transitional probability (i.e., highly predictable collocations were fixated less often).

Up until now, all the studies reviewed did not control for an important item-related aspect: that is, semantic relatedness. Accordingly, it is not clear whether the facilitative frequency effects are real or false (i.e., emerging from the implausibility of control sequences or from variance in semantic relations among target sequences). One recent study by Siyanova-Chanturia et al. (2011b) overcame this limitation. The study employed the eye-tracking methodology in exploring natives' and advanced non-natives' (assumed to reflect variation in exposure to language) sensitivity to the frequency of one specific type of a formulaic sequence controlled for semantic relatedness (i.e., the binominal phrase). Participants' eye movements were monitored as they read sentences containing frequent binominal phrases (*safe and sound*) or their less frequent reversed forms (*sound and safe*). The analysis included three main variables and the interaction between them: raw BNC phrasal frequency, phrase type (binominal versus reversed), and proficiency (based on self-reported ability scores). Results demonstrated that: (1) raw BNC frequency had an overall facilitative effect on reading times (with fewer fixations and shorter reading times: first pass and total) but did not interact with proficiency and (2) binominals were read faster (early and late in processing) than their reversed forms with a clear effect of proficiency (i.e., sensitivity to binominals' configuration improved with proficiency). These results were taken as evidence for usage-based models of language acquisition (more exposure to language, reflected in higher self-reported proficiency, leads to clearer sensitivity). It should be noted, however, that the lack of interaction between raw BNC frequency and proficiency might leave us cautious about the above conclusion. It might be assumed that the effect of proficiency (which in itself is a subjective measure) observed in this study is more related to the special nature of binominals (with a specific, fixed configuration) per se than to the frequency of exposure to formulaic sequences in general.

In sum, research has generally shown that language users (both L1 and L2) are sensitive to the frequency of formulaic sequences both off-line and on-line providing clear evidence for usage-based theories over the words-and-rules approach. However, most of the available research has only looked either at the distinction between natives and non-natives or that between off-line/on-line processes but not both. Moreover, almost all studies used control items which were not semantically plausible (e.g.,

---

[3] Automatic priming is contrasted with strategic priming in that the former operates rapidly without the user's conscious intention (implicit knowledge) while the latter requires intention and reflects slower processes (explicit knowledge) (Dagenbach, Horst & Carr, 1990).

angry use, armed concept) and might, thus, have biased the results. Two of these issues (i.e., including both L1 and L2 users and controlling for semantic relatedness) were tackled in a recent study (Siyanova-Chanturia et al., 2011b). However, the study looked at a very special type of formulaic sequences, that is, binominals. It is not clear, thus, whether the facilitative frequency effects reported would apply to other, less fixed, categories such as collocations.

Nesselhauf (2005) distinguishes between three types of word pairs: free combinations (no sense restrictions on either word, e.g., *want a car*, where both words can be used in other combinations freely), collocations (sense of one word is restricted, e.g., *take a picture*, where the verb *take* in this sense cannot be used with nouns like *film*), and idioms (sense of both words is restricted, e.g., *sweeten the pill*). Hence, the question is whether Nesselhauf's second category follows the same pattern of fixed formulaic sequences (with corpus frequency predicting processing both for natives and non-natives).

Current Study

The current study deals with limitations of the previous research through employing various tasks in the assessment of natives' and non-natives' sensitivity to the frequency of synonymous adjective-noun collocations. The term *collocation* is defined here according to corpus-derived frequency as:

A non-idiomatic pair comprising two open class lemmas (adjective + noun) which occurs in a corpus (within a window of 4 characters to the right and left of the node, ±4) above chance (Mutual Information, MI > 1).[4]

The first aim of the present experiment is to establish whether native and non-native speakers of English are sensitive to corpus-derived collocational frequency both off-line in a rating task and on-line in an eye-tracking experiment. Another aim of the study is to investigate whether sensitivity to frequency (both off-line and on-line) is influenced by level of proficiency (as indicated by objective scores in the Vocabulary Levels Test, VLT).[5] The following research questions will be addressed:

(1) Are native speakers of English and non-natives sensitive to corpus-derived collocational frequency off-line (typicality rating scores)? What effect does

proficiency have on sensitivity to collocational frequency off-line?

(2) Are native speakers of English and non-natives sensitive to corpus-derived collocational frequency on-line (real-time eye-movement behaviour)? What effect does proficiency have on sensitivity to collocational frequency on-line?

A list of synonymous collocations belonging to a variety of frequency levels were extracted from the BNC along with non-collocations (two-word combinations with zero occurrences or negative MI scores). The research questions are addressed through exploring effects of raw corpus-derived frequency (irrespective of the arbitrary division of frequency levels) on natives' and non-natives' off-line/on-line collocational performance (through a number of Linear Mixed Effects, LME, models). If frequency exerts an influence on off-line/on-line processing, raw BNC frequency should significantly predict collocational performance. Also, if level of proficiency plays a role on how sensitive language users are to collocational frequency, then VLT score should significantly interact with collocational frequency, supporting the usage-based view that linguistic units (of various sizes) are modified based on the cumulative experience with language.

Methods

Participants

Thirty natives and 30 non-native participants (12 males and 48 females) took part in the present study. They all had normal or corrected to normal vision and ranged in age between 19 and 43 ($M = 24.27$, $SD = 6.27$). The native participants' average age was 19.40 ($SD = 0.50$). Non-natives, on the other hand, had an average age of 29.13 ($SD = 5.54$). An independent samples t-test revealed a significant difference between the two groups' average age: ($t(58) = -9.59$, $p < .001$, *eta squared* $= 0.61$). Thus, age was added as a variable in the LME models to control for its effect (see *Data analysis* section below).

The natives were undergraduate students at Nottingham University who participated for course credit. The non-natives, on the other hand, were postgraduates (master's $= 12$, Ph.D $= 18$) who had all met the university entry requirement (minimum IELTS score of 6.0 or TOEFL score of 550) and were offered a payment of £6 for their participation. These non-native participants came from a variety of L1 backgrounds (Arabic $= 5$, Chinese $= 4$, Farsi $= 1$, Finnish $= 1$, French $= 1$, German $= 1$, Hungarian $= 1$, Icelandic $= 1$, Indonesian $= 2$, Malay $= 1$, Russian $= 2$, Spanish $= 3$, Tamil $= 3$, Thai $= 3$, Vietnamese $= 1$) and had spent a mean of 26.07 months in English-speaking countries ($SD = 27.63$, *Min* $= 6$, *Max* $= 132$). They were

---

[4] Although the established MI threshold value for 'significant' collocations is 3 (Hunston, 2002), Evert (2008, p. 6) distinguishes between the threshold and the ranking approaches to operationalizing collocations. When collocations are treated as forming a cline (ranked from weak to strong collocations), then the threshold value can go down and pairs should be viewed as "more collocational" or "less collocational" according to their MI scores.

[5] See Meara and Jones (1988) for evidence of a linear relationship between vocabulary size and proficiency.

first exposed to English at an average age of 8.92 years (*SD* = 3.86, *Min* = 3, *Max* = 17). Their self-rated proficiency scores (on a scale from 1 = very poor to 5 = excellent) were: reading *M* = 4.23, *SD* = 0.82; writing *M* = 3.80, *SD* = 0.96; speaking *M* = 4.13, *SD* = 0.73, and listening *M* = 4.23, *SD* = 0.73. Their overall self-rated proficiency score (averaged across skills) was 4.55, *SD* = 0.67.

All participants were administered the 3K and 5K levels of the VLT (Schmitt, Schmitt & Clapham, 2001) as a rough objective estimate of their proficiency. It was decided that these two levels were the most relevant to the non-native participants in the present study (the 2K level might be redundant while the 10K level might be too difficult). The native participants scored almost a perfect score (*Mean* = 59.77 out of a maximum 60, *SD* = 0.43). Non-natives, on the other hand, scored an average of 52.73 (*SD* = 6.05) out of 60 with a mean score of 28.17 out of 30 on the 3K level (*SD* = 2.00) and a mean score of 24.57 out of 30 on the 5K level (*SD* = 4.45). An independent samples t-test revealed a significant difference between the two groups' overall VLT scores: (t (58) = 6.35, $p <$ .001, *eta squared* = 0.41).

Stimuli

Initially, the aim was to find three different synonymous adjective collocates for a number of noun nodes in English representing three levels of collocational frequency (lower: 5–15 occurrences in the whole BNC; mid: 25–45 occurrences; and higher: over 55 occurrences) along with a fourth non-collocate adjective. However, since an important aim was to control for semantic relatedness, I ended up with three different sets reflecting the different combination of collocational levels where each noun node is matched with 2 collocates and one non-collocate (Set 1: non, lower and mid; Set 2: non, mid and higher; Set 3: non, lower and higher). The division of levels was arbitrary (based on frequency range), thus these levels were not employed in the analysis. Instead, raw BNC frequency (dealt with in the form of quantiles) was included as a more reliable predictor in the LME models (see *Data analysis* section for details).

In order to arrive at a set of candidate collocational arrays, a number of steps were followed:

(1) Davies' BNC interface (Davies, 2004) was consulted to find adjective collocates for the most frequent 2,000 lemmas in the BNC (Leech, Rayson & Wilson, 2001) in a window of ±4.

(2) The extracted adjective collocates for each noun were examined to find two candidate synonymous adjectives which fit the following criteria:
   a. The collocations belong to two of the three specified levels (lower, mid and higher).
   b Each collocation has an MI-score of 1 or above.

(3) Concordance lines for each adjective-noun collocation were checked to make sure that the adjective modifies the noun. Any occurrences where the adjective is not describing the noun (e.g., where the adjective is occurring outside sentence boundary) were excluded from the frequency count of the collocation.

(4) For each candidate item, a non-collocate synonymous adjective was searched for. This adjective returned 0–2 occurrences in the BNC within the ±4 span and a negative MI score.[6]

(5) Finally, two association databases were checked for the forward and backward association strength of the target collocates and non-collocates (Kiss, Armstrong, Milroy & Piper, 1973; Nelson, McEvoy & Schreiber, 1998). This step was intended to control for association strength since previous research (e.g., Durrant & Doherty, 2010) revealed its significant role in collocational processing. Only adjectives that were not produced in response to (and did not elicit) the target nouns were included in the item pool.[7]

It should be noted here that I did not experimentally control for individual words' length or frequency. This is due to the fact that LME modelling is employed in analysing the results. This type of analysis allows for the statistical control of item-related variables and thus eliminates the need for their experimental control.

In the end, 30 adjective-noun arrays were selected (10 under each set, see the appendix). Sentence contexts were adapted from the BNC for the off-line/on-line tasks. For each noun node, three versions of the same sentence were created with the difference being in the adjective collocating with each noun. Here is an example:

The engineer made one *fatal mistake* which weakened the bridge.   *Frequency=26*
The engineer made one *awful mistake* which weakened the bridge.   *Frequency=4*
The engineer made one *extreme mistake* which weakened the bridge.   *Frequency=0*

---

[6] Candidate adjective-noun collocations and non-collocations were also checked in the same interface for the COCA frequency (normalized to 100 million). This step was intended to ensure similar frequency breakdowns for collocations and no possibility for non-collocations in another corpus.

[7] Items with association strength of 0.00 or 0.01 (i.e., where only one participant out of a thousand or 1 out of a hundred report an association) were also treated as non-associates since they only reflect idiosyncratic associational behaviour.

The final sentences were piloted with 10 native speakers of English who did not participate in the main study to ensure naturalness. Results showed that the sentences were rated very high on a scale from 1 (not natural) to 6 (very natural): *Mean* = 5.79, *SD* = 0.20, *Min* = 5.37, *Max* = 6.00.

Measures

On-line eye-tracking experiment
To assess on-line processing of collocations, an eye-tracking experiment was designed on Experiment Builder and run using the SR head-mounted eye-tracker (see *procedures* below for more details). In this experiment, participants were presented with the target sentence contexts and were instructed to read as naturally as possible for comprehension.

Critical stimuli for this experiment included the 60 adjective-noun collocations and the 30 non-collocate pairs. The collocations and non-collocations were divided into three counterbalanced lists such that one adjective from each one of the 30 arrays is combined with the noun node. A noun which was matched with an adjective from a specific level (non-collocate level, or one of the two collocational levels) in one list was matched with an adjective from a different level in the other two lists. No noun or adjective was used more than once in any of the lists. In addition, four practice trials and 62 filler trials were included in which non-target combinations were inserted. Some filler sentences included non-collocations ($K = 20$) while others included collocations from the three frequency levels (low, $K = 14$; mid, $K = 14$; high, $K = 14$). Fillers were intended to include collocations (and non-collocations) so that targets are not particularly marked (and are not noticeable by participants). Thus, in each stimuli list, participants were presented with 96 trials (30 targets, 62 fillers, and 4 practice trials).

Off-line rating task
The off-line sensitivity measure was a rating task with three counterbalanced lists of sentences (the same as the eye-tracking list for each participant). In this task, participants were presented only with the 30 experimental sentences under each list (with no fillers) with the target combinations underlined. They were instructed to rate each underlined adjective-noun combination on how typical it is in English on a scale from 1 (not typical) to 6 (very typical).

Procedures

The study was performed in three stages. First, upon arrival at the lab, the participant signed the consent form in which he/she was only briefed on the purpose of the study (without a detailed account of its various stages). Then, the on-line eye-tracking experiment started. Eye movements were recorded using SMI EyeLink I (SR Research Ltd., Mississauga, Ontario, Canada). A 9-point grid calibration procedure was done prior to the experiment. The first four trials were always practice trials. The eye-tracker was calibrated at least four times during the experiment. Each trial started with a fixation point that appeared in the middle of the screen. After participants fixated it and a calibration check was conducted, a sentence appeared across one line in the middle of the screen (in Courier New, 14 point, font). The task was to read the sentences as quickly as possible for comprehension. One quarter of the sentences were followed by a comprehension question. The rest were followed by "Ready?" Both groups of participants had no difficulty answering the comprehension questions (Natives: 94.80%; Non-natives: 91.33%). The order of trials was randomized across subjects to avoid potential order effects. This task took around 20 minutes.

Immediately after that, participants performed a distractor task which was mainly intended to minimize any effect of the on-line measure on the subsequent off-line task. This task consisted of only two levels of the VLT: the 3K and 5K levels plus the language background questionnaire. It was checked very carefully that none of the words constituting the target collocate and non-collocate pairs were used in the VLT. Another purpose of this test was to arrive at a rough estimate of non-natives' proficiency.

Finally, participants moved on to the final stage of the study: the off-line typicality rating task (conducted on E-Prime for randomization of trials) which took no more than 10 minutes. There were four practice trials in the beginning. This was followed by the 30 target items under each list.

Data analysis

I conducted the analysis with R version 2.15.0 (R Development Core Team, 2010) using LME models. Four LME models were fit: one model was fit on the off-line rating scores and three models were fit on three eye-tracking measures (first pass reading time, total reading time, and fixation count of the target combinations). The analysis was conducted both in the forward and backward step-wise model selection procedures.[8] The resulting

---

[8] The forward method starts with the simplest (null) model which only includes the dependent measure and the random variables (participants and items). Fixed effects are then added incrementally and X^2 (likelihood ratio)-tests are used to check whether the inclusion of additional predictors contributes significantly ($p < .05$) to the model. The backward method, on the other hand, starts with all predictors tested in the forward method. Predictors are then excluded stepwise and X^2-tests are used to check whether the exclusion of a predictor has a significant effect on the model. For more details on

best-fit models in both directions for all dependent measures were identical. Positive *t* values in the best-fit LME model indicate a direct relation (i.e., the higher the value of the predictor, the higher the value of the dependent measure) while negative *t* values indicate an inverse relation (i.e., the higher the value of the predictor, the lower the value of the dependent measure).

The above-mentioned eye-movement measures were chosen as they are the most commonly used in multi-word sequences research (e.g., Siyanova-Chanturia, Conklin & Schmitt, 2011a; Siyanova-Chanturia et al., 2011b). *First pass reading time* is the sum of fixation durations within the area of interest before it was left (to the left or to the right). It is the primary measure of interest when target items are longer than single words reflecting early/immediate comprehension processes. *Total reading time* is the sum of all fixation durations made within the area of interest (including re-reading) and is assumed to reflect later comprehension processes (e.g., textual integration). *Fixation count* is the number of all fixations made within the area of interest and is also assumed to reflect later comprehension processes (see Liversedge et al., 1998; Rayner, 1998, for an overview).

The following predictors (fixed independent factors) were tested for each dependent measure: age, trial number, VLT score, pair length (number of characters), log Word1 frequency, log Word2 frequency, and log collocational frequency (all frequency counts are based on occurrences per 100 million words).[9] As I did not experimentally control for pair length, individual word frequency, or participants' age, I statistically controlled for these variables prior to testing for the effect of collocational frequency. Finally, I tested the interaction between the two major predictors (log collocational frequency and VLT score) and each one of the other predictors (for each dependent measure). All dependent measures (rating scores, first pass reading times, total reading times, and fixation counts) were log-transformed to reduce skewness in the distribution.

As for the on-line eye-movement measures, data cleaning was essential prior to fitting the models. Single fixation durations shorter than 100 ms and longer than 800 ms were excluded. The missing data accounted for 4.31% of the total data for native speakers and 5.27% for non-natives. I also excluded cumulative fixation durations shorter than 200 ms and longer than 2,000 ms (per combination) to further reduce skewness in the data distribution. This resulted in the loss of 10.71% of the data for total reading time and 10.15% of the data for first-pass reading time. For the fixation count measure, I

fitting LME models and interpretation of their results, see Sonbul and Schmitt (2013).
[9] Collocational frequency was log transformed after adding 1 to all values.

also excluded fixation counts of 12 or more (only 2.51% of the data).[10]

When fitting mixed-effects models, variables which correlate strongly can cause serious problems, so residuals should be calculated for strongly correlated variables (see Baayen, 2008). Collinearity among item-related variables was checked prior to fitting the LME models, and residuals were calculated. I orthogonalized pair length by fitting a linear model in which pair length was predicted by log W2 frequency. Residuals of this model (ResidPairLength) correlated with the original (pair length) predictor (*r* = 0.97, p < .001). I also checked for collinearity between the two participant-related factors (VLT and age). They were found to correlate, so residuals were calculated by fitting a linear model in which VLT score was predicted by age. The residuals (ResidVLT) correlated with the original VLT predictor (*r* = 0.91, p < .001). Finally, all continous predictors were centred. A summary of the continuous variables is presented in Table 1.

Results

On-line eye-tracking

Table 2 presents the mean reading times and fixation counts for both natives and non-natives under various collocational frequency quantiles. It can be clearly seen that there is not a clear trend for any of the eye movement measures across frequency quantiles. This might be related to the fact that length and frequency of individual words were not experimentally controlled for, and might, thus, have concealed frequency effects. As stated above, the analysis started with controlling for these lower-level variables and then tested for collocational frequency effects.

First pass reading time

The best-fit model for variables predicting first pass reading time is presented in Table 3. There are a number of significant main effects. First, the VLT score had a significant effect on reading times: the higher the VLT score, the shorter the reading times (i.e., the more proficient the participant, the faster he/she read the combinations). Second, pair length influenced reading times in that the longer the pair, the longer the first pass

[10] I also conducted the analysis with the full (untrimmed) data to explore the effect of excluding a large percentage of outliers. Results were identical for the fixation count measure. However, two differences emerged for the first pass reading time and the total reading measures (one for each). First the (VLT score x trial number) interaction, which was significant in the final LME first pass reading time model, was not significant with the full data set. Second, the (VLT score x pair length) interaction, which was not significant in the final total reading time model, was significant with the full data. Other than that, all significant/insignificant predictors remained the same.

Table 1. *Summary of the Continuous Variables.*

Variable	Range (adjusted range)	SD	Mdn
Trial Num (off-line)	5–34 (−14.49–14.51)	8.66	−0.49
Trial Num (on-line)	5–96 (−45.04–45.96)	26.46	−0.04
Age	19–43 (−5.26–18.74 years)	6.22	−4.26
Resid VLT Score	39–60 (−15.90–9.52 points)	5.02	1.84
Resid Pair length	8–23 (−6.56–8.12 characters)	4.04	−0.47
Log Word1 Frequency	6.27–9.27 (−1.96–1.03 log units)	0.57	0.17
Log Word2 Frequency	7.93–9.18 (−0.80–0.45 log units)	0.24	−0.01
Log Collocational Frequency	0–5.09 (−2.25–2.85 log units)	1.77	0.15

Note: The second column shows the range of the variables. The adjusted range after transformation, partialing out correlated predictors and/or centering, is presented in parentheses. Standard deviations and medians refer to the predictor values in the models. All variables are centred, and their means are zero.

Table 2. *Mean (SE) on-line reading times and fixations counts for both natives and non-natives under various collocational frequency quantiles.*

Measure	Native Speakers				Non-native Speakers			
	Collocational Frequency Quantiles				Collocational Frequency Quantiles			
	1st / 2nd Q [a]	3rd Q	4th Q	5th Q	1st / 2nd Q [a]	3rd Q	4th Q	5th Q
First Pass Reading Time	470.63	483.43	423.28	440.61	703.68	690.31	662.12	661.96
	(11.07)	(16.52)	(10.79)	(12.64)	(19.68)	(26.59)	(21.92)	(24.52)
Total Reading Time	708.3	675.08	622.13	707.43	944.73	917.1	871.81	926.92
	(21.39)	(27.15)	(23.61)	(27.56)	(28.98)	(36.47)	(30.82)	(35.39)
Fixation Count	3.43	3.11	2.83	3.26	4.32	4.10	3.82	4.45
	(0.11)	(0.12)	(0.10)	(0.12)	(0.14)	(0.17)	(0.14)	(0.18)

Note: [a] The first and second quantiles are identical due to the presence of zero frequency values (comprising non-collocates).

reading time was. Third, the more frequent the second word of the pair (noun node), the shorter the reading time was. Fourth, collocational frequency had a significant effect on the first pass reading time measure: the higher the frequency of the pair, the shorter the reading time. Fifth, participants' age had a significant main effect on reading times (the older the participant, the longer the reading time). Finally, it is interesting to note that trial number had no significant direct effect on reading times, though it interacted with the VLT score. The interaction is depicted in Figure 1. It can be clearly seen that the usual direction of trial number effects (shorter reading times with increase in trial number) is actually reversed for participant with low VLT scores. This result might be related to the fact that the non-native participants with lower proficiency levels paid extra attention to answering the comprehension questions properly. This might have led them to spend more time in the initial reading of sentences as the experiment proceeded in order to make sure that they do not lose any detail.

Above all, results of the first pass reading time measure show that both natives and non-natives are clearly sensitive to collocational frequency early in processing. Once pair length and frequency of individual words were controlled for, collocational frequency had a clear effect on collocational processing. Interestingly, this sensitivity was not affected by level of proficiency (no interaction between VLT scores and collocational frequency). No difference was observed between the lower-level non-natives and the higher-level participants (i.e., natives and non-natives achieving native-like scores in the VLT).

Total reading time
Table 4 presents the best-fit LME model for the total reading time measure. Only four main predictors were found to be significant: trial number (the bigger the number, the shorter the total reading time), VLT score (the larger the score, the shorter the total reading time), pair length (the longer the pair, the longer the total reading time), and age (the older the participant, the longer the total reading time). Collocational frequency did not surface as a significant (main or interacting) predictor.

Table 3. *Summary of the Best Fit LME Model for Variables Predicting Log First Pass Reading Time (N = 1611, R² = 0.47).*

Predictor	Estimate	MCMC	SE	t	pMCMC	pr(>|t|)
(Intercept)	6.22	6.22	0.03	248.04	0.001	< .001
Trial Num	−0.00	−0.00	0.00	−1.28	0.18	0.20
Resid VLT	−0.04	−0.04	0.00	−7.76	0.001	< .001
Resid pair length	0.03	0.03	0.00	10.17	0.001	< .001
Log W2 frequency	−0.11	−0.11	0.05	−2.29	0.02	0.02
Log collocational frequency	−0.02	−0.02	0.01	−2.47	0.01	0.01
Age	0.02	0.01	0.00	3.91	0.001	< .001
Resid VLT x Trial Num	−0.00	−0.00	0.00	−2.22	0.03	0.03

Note: The model has random intercepts for participants and items; MCMC = Monte Carlo Markov chain; pMCMC = p values estimated by the MCMC chain method using 10,000 simulations; *Pr(>|t|)* = p values obtained with the t test using the difference between the number of observations and the number of fixed effects as the upper bound for the degrees of freedom.
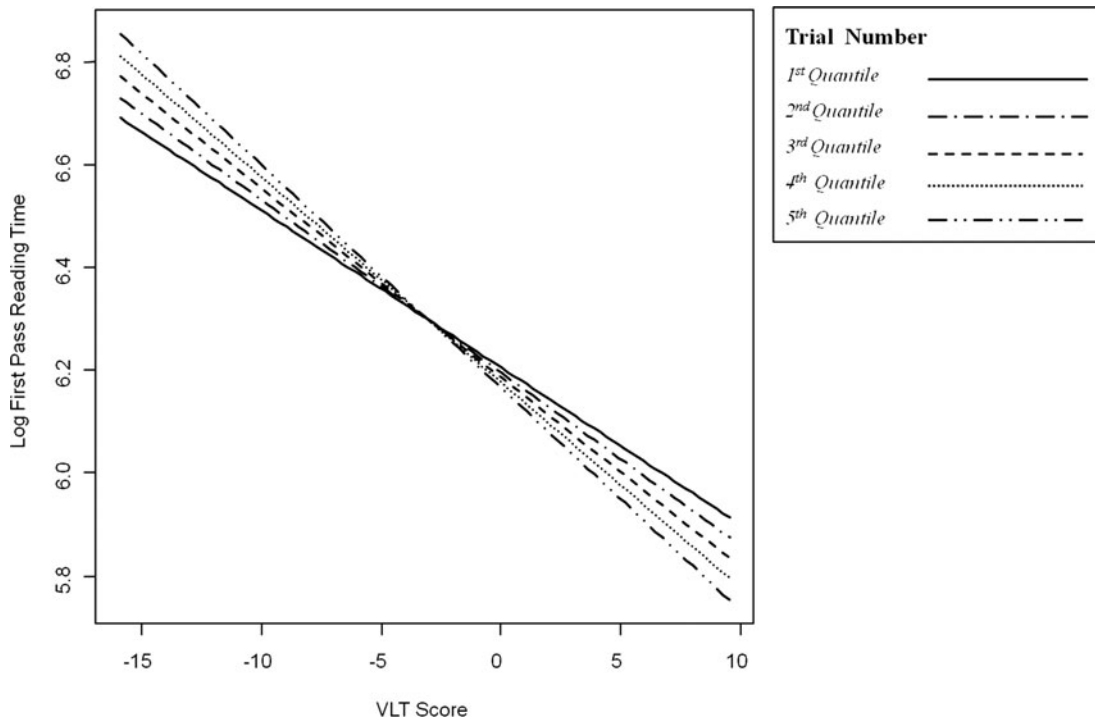

Figure 1. Interaction between VLT score and Tiral Number for the on-line first pass reading time measure.

Thus, results of the total reading time measure suggest that collocational frequency has no effect on later integrative processes during reading. This lack of sensitivity was not modulated by proficiency. Natives and non-natives (across proficiency levels) did not experience any late processing difficulty when reading less frequent collocations/non-collocate pairs than more frequent ones.

Fixation count

The best-fit model for the on-line fixation-count measure is presented in Table 5. As both measures (total reading time and fixation count) are late measures indicative of

textual integration (see *Data analysis* section above), their best-fit models were identical. Only trial number, VLT score, pair length, and age significantly predicted fixation counts (the first two with a negative and the last two with a positive effect). Thus, similar to the total reading time measure, results of the fixation count show that collocational frequency had no effect on later integration collocational processes.

To recap on-line results, natives and advanced non-natives showed similar effects. Collocational frequency had a significant main effect on initial reading times (not modulated by proficiency) but did not influence later

Table 4. *Summary of the Best Fit LME Model for Variables Predicting Log Total Reading Time (N = 1601, R^2 = 0.58).*

Predictor	Estimate	MCMC	SE	t	pMCMC	pr(>|t|)
(Intercept)	6.59	6.58	0.04	160.25	0.001	< .001
Trial Num	−0.00	−00.00	0.00	−3.37	0.002	0.001
Resid VLT	−0.04	−0.04	0.01	−7.89	0.001	< .001
Resid pair length	0.03	0.03	0.01	3.43	0.001	0.001
Age	0.02	0.01	0.00	4.40	0.001	< .001

Note: The model has random intercepts for participants and items; MCMC = Monte Carlo Markov chain; pMCMC = p values estimated by the MCMC chain method using 10,000 simulations; $Pr(> |t|)$ = p values obtained with the t test using the difference between the number of observations and the number of fixed effects as the upper bound for the degrees of freedom.

Table 5. *Summary of the Best Fit LME Model for Variables Predicting Log Fixation Count (N = 1748, R^2 = 0.58).*

Predictor	Estimate	MCMC	SE	t	pMCMC	pr(>|t|)
(Intercept)	1.18	1.17	0.04	28.18	0.001	< .001
Trial Num	−0.00	−0.00	0.00	−4.10	0.001	< .001
Resid VLT	−0.03	−0.03	0.00	−6.46	0.001	< .001
Resid pair length	0.03	0.03	0.01	3.73	0.001	< .001
Age	0.02	0.01	0.00	4.11	0.001	< .001

Note: The model has random intercepts for participants and items; MCMC = Monte Carlo Markov chain; pMCMC = p values estimated by the MCMC chain method using 10,000 simulations; $Pr(> |t|)$ = p values obtained with the t test using the difference between the number of observations and the number of fixed effects as the upper bound for the degrees of freedom.

reading processes for either group of participants. It is also worthy of notice that participants' age had a significant positive main effect on all eye-tracking measures but did not interact with the main factors of collocational frequency or proficiency.

Off-line rating

Mean typicality rating scores both for natives and non-natives under various collocational frequency quantiles are presented in Table 6. It can be clearly seen that the mean rating score increased with frequency for both groups. The difference between the 3rd and 4th quantiles was not very clear for the natives, though. It is also notable that the mean scores for the 1st/2nd quantiles (representing non-collocates) were surprisingly moderate (around 3.50 out of 6 for both groups). This might be due to the fact that non-collocations in the present study were semantically controlled and, thus, did not look totally implausible to participants.

The final best-fit model is presented in Table 7. Significant main effects can be summarized as follows: (1) the higher the frequency of the second content word in the pair, the lower the rating score, (2) the higher the frequency of the collocation, the higher the typicality rating score. The VLT score had no significant main effects on rating scores, but it interacted with first word frequency and with collocational frequency modulating some of the main effects. First, the interaction between the VLT score and the log frequency of the first content word is depicted in Figure 2. The higher the frequency of the first word (i.e., the adjective), the higher the rating score was for the lower-level non-natives. This interaction went in the opposite direction for the higher-level non-natives and natives. Also, this VLT effect was clearer for the first frequency quantile. In order to explain this interaction, it should be noted that while the effect of the second word frequency on rating scores was significant regardless of proficiency (with lower rating scores as the frequency increased, see above), the effect of the first word frequency on rating scores was modulated by proficiency (i.e., highly proficient participants showed a similar effect to that of the second word frequency while the lower-level participants showed the effect in the opposite direction). It seems that for the lower-level non-natives, collocations comprising very non-frequent adjectives are more likely to form part of a less typical collocation.

Second, and more importantly, the VLT score interacted with log collocational frequency (see Figure 3).

Table 6. *Mean (SD) off-line rating scores for both natives and non-natives under various collocational frequency quantiles.*

	Native Speakers				Non-native Speakers			
	Collocational Frequency Quantiles				Collocational Frequency Quantiles			
Measure	1st / 2nd Q [a]	3rd Q	4th Q	5th Q	1st / 2nd Q [a]	3rd Q	4th Q	5th Q
Rating Score	3.36	5.07	5.58	5.57	3.44	4.82	5.03	5.32
	(1.59)	(1.22)	(0.86)	(0.82)	(1.84)	(1.50)	(1.43)	(1.27)

Note: [a] The first and second quantiles are identical due to the presence of zero frequency values (comprising non-collocates).

Table 7. *Summary of the Best Fit LME Model for Variables Predicting Log Rating Score (N = 1793, $R^2 = 0.45$).*

Predictor	Estimate	MCMC	SE	t	pMCMC	pr(>|t|)
(Intercept)	1.43	1.43	0.03	53.96	< .001	< .001
Resid VLT	0.00	0.00	0.00	0.77	0.40	0.44
Log W1 frequency	−0.03	−0.03	0.03	−0.93	0.33	0.35
Log W2 frequency	−0.15	−0.15	0.07	−2.04	0.03	0.04
Log collocational frequency	0.15	0.15	0.01	14.51	< .001	< .001
Resid VLT x Log W1 frequency	−0.01	−0.01	0.00	−2.93	0.02	0.02
Resid VLT x Log collocational frequency	0.01	0.01	0.00	7.15	< .001	< .001

Note: The model has random intercepts for participants and items; MCMC = Monte Carlo Markov chain; pMCMC = p values estimated by the MCMC chain method using 10,000 simulations; *Pr(> |t|)* = p values obtained with the t test using the difference between the number of observations and the number of fixed effects as the upper bound for the degrees of freedom.
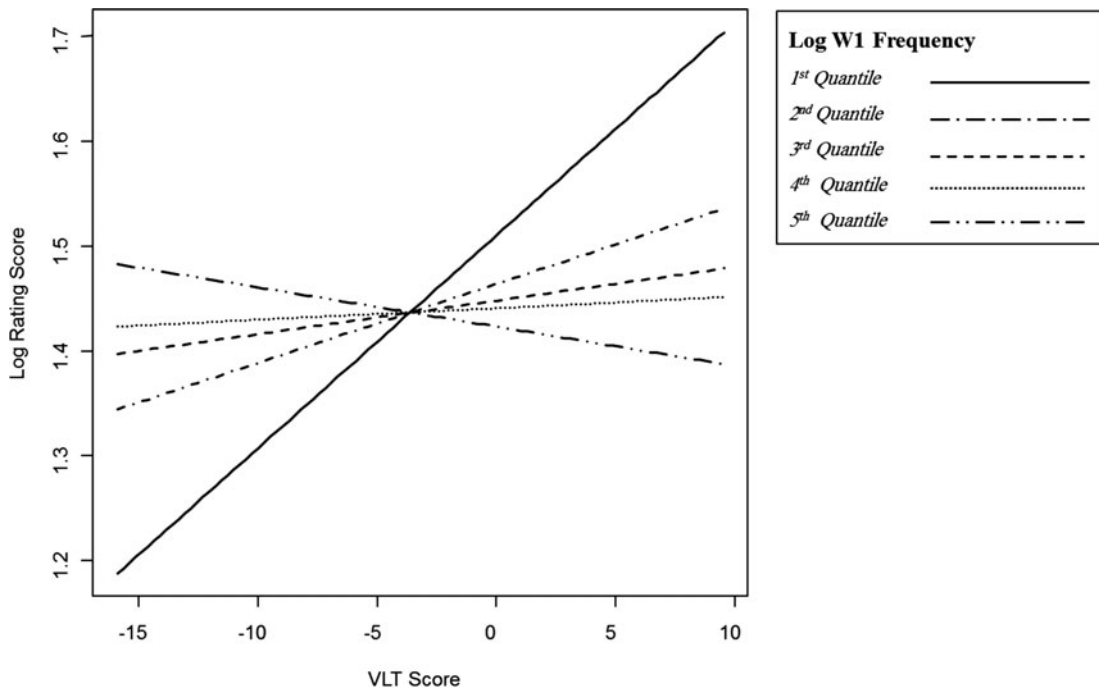

Figure 2. Interaction between VLT score and log W1 frequency for the off-line measure.
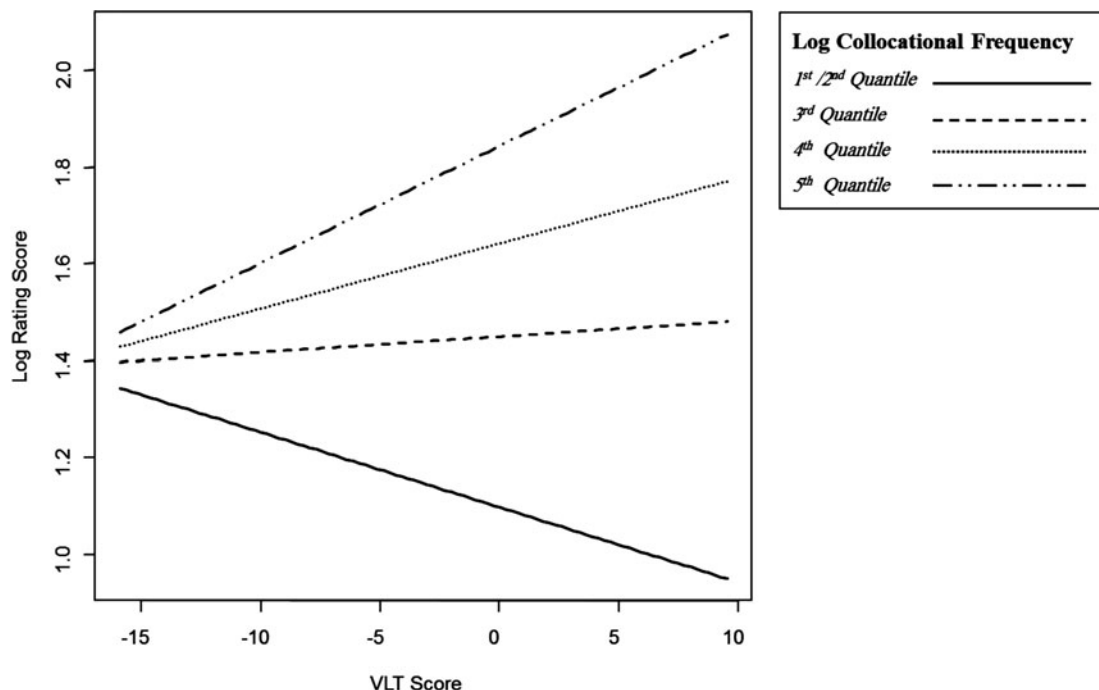
Figure 3. Interaction between VLT score and log collocational frequency for the off-line measure (Note: the first and second log collocational frequency quantiles are identical).

As can be clearly seen, the difference between the collocational frequency quantiles was present for participants across all proficiency levels. However, sensitivity to frequency increased with proficiency: participants with higher VLT scores showed bigger differences across frequency quantiles.

In summary, results of the off-line task show a clear effect of collocational frequency on the typicality rating behaviour of both natives and non-natives. This effect was modulated by proficiency in that the natives and highly proficient non-natives (those achieving native-like scores in the VLT) showed stronger sensitivity to collocational frequency. Age did not have any main or interacting effect on off-line results.

Discussion

Previous research addressing frequency effects on the processing of formulaic sequences focused on fixed sequences (e.g., binominals). The present study looked at a different type of formulaic sequences (i.e., collocations) which are far less fixed. More importantly, the present study (1) employed both off-line (rating) and on-line (the highly sensitive eye-movement) measures, (2) tested both natives and non-natives, and (3) controlled for the semantic relatedness of word pairs across frequency levels.

Results of the off-line typicality rating task showed very clear effects of frequency with higher rating scores as the frequency of the collocation increased. More

interestingly, level of proficiency influenced rating scores with bigger differences across frequency levels as the proficiency increased. This effect was present over and above all lower-level item-related variables. Thus, to answer the first research question, both natives and non-natives show off-line sensitivity to corpus-derived collocational frequency with higher levels of sensitivity as proficiency increases. This result replicates Arnon and Snider's (2010) finding that natives are sensitive to the frequency of formulaic sequences off-line. Moreover, similar to Wolter and Gyllstad (2013), current findings show that both natives and non-native language users are similarly sensitive (during off-line performance) to the frequency of a specific type of formulaic sequences, i.e., collocations. These findings surfaced in the present study even when semantic plausibility was controlled for (an aspect which was not considered in Wolter and Gyllstad's (2013) study). Additionally, the present study shows that non-natives' off-line sensitivity is affected by their level of proficiency.

This proficiency effect was not present during on-line processing. Both groups of participants (natives and non-natives) manifested only early sensitivity to collocational frequency during reading, which surfaced over and above lower-level factors and which was not modulated by proficiency. This early sensitivity to collocational frequency disappeared later in processing; collocational frequency was not a significant main or interacting predictor of total reading time or fixation count. These findings provide a clear answer to the second research

question: both natives and non-natives show early on-line sensitivity to the corpus-derived frequency of collocations with no clear effect of proficiency.

Collocations versus other formulaic sequences

Eye movement results of the present study stand in contrast with Siyanova-Chanturia et al.'s (2011b) finding of frequency effects on both early and late processing of fixed binominals. This discrepancy can be attributed to (and essentially reinforces) the inherent difference between collocations and other fixed categories of formulaic sequences. Natives and non-natives are able to recover quickly from difficulties when reading an infrequent collocation (or even unattested one) but cannot that easily cope with a very marked binominal in the reversed form. It seems that, while reading naturally, language users are more likely to accept deviations from typical collocations than those invovling the fixed configuration of binominals. Altering the adjective most typically used with a given noun is likely to cause only initial reading difficulty. However, a change in the most distinctive word order in binominals can cause initial processing disadvantage that persists. Collocations are not as fixed as other types of formulaic sequences, and thus language users might be more tolerant of alterations in their structure.

This conclusion is further supported by the off-line results of the present study where the average rating scores for non-collocations (1st and 2nd quantiles) by both natives and non-natives were above 3.00 on a 6-point scale (3.36 and 3.44, respectively: see Table 6). These relatively high ratings for non-collocations seem to emphasize the "fluidness" of collocations in comparison with other types of formaulic sequences (see *Introduction* section). Wolter and Gyllstad (2013) also touched upon this point given the finding that error rates in their off-line collocational decision task were significantly higher for non-collocates than for real collocations (i.e., more YES responses to non-collocates than NO responses to real collocations) both for natives and non-natives. They explained this finding in light of the usage-based notion of "schemata" (recurrent, generalized patterns abstracted from frequent experience) leading to various adjective + noun pairs (even the unattested ones) being judged as common (this point will be revisited below upon discussing implications of the present study to usage-based models).

L2 proficiency and on-line/off-line performance

As for the effect of proficiency, it is interesering to note that natives and non-natives do not differ in the way they respond to variation in collocational frequency during real time reading. Both groups showed an on-line processing advantage for highly frequent collocations over less frequent ones initially but recovered later. Off-line, however, proficiency had a positive modulating effect on the reaction to collocational frequency; participants manifested bigger differences across frequency levels as proficiency increased.

Based on this distinction between off-line and on-line results, it might be concluded that non-natives develop implicit native-like sensitivity to collocational frequency through exposure but cannot as easily develop explicit native-like sensitivity as evident in rating scores. This result seems to support McLaughlin, Osterhout and Kim's (2004) finding that non-native learners develop implicit sensitivity to lexical aspects of language prior to their ability to explicitly judge them. Thus, researchers have to be careful when employing off-line behavioural measures of lexical (in particular, collocational) knowledge as these might underestimate non-natives' actual knowledge.

Another important point to note here is related to the false discrepancy between the current on-line, eye-tracking, results and those reported in Siyanova-Chanturia et al. (2011b) when it comes to effects of proficiency. One might misleadingly argue that results of that study manifest a modulating effect of proficiency on non-natives' on-line sensitivity to frequency while the present study does not. It should be noted, however, (as pointed out earlier) that the modulating effect of proficiency reported in that study is more related to sensitivity to binominals' word order (one of their special features) than to the raw corpus-derived frequency. Accordingly, it might be argued that proficiency had no modulating effect on sensitivity to frequency in either study. Although this finding might be viewed as going against predictions of usage-based theories (i.e., more exposure should lead to entrenchment of connections), it should be noted that both studies tested non-native participants of a fairly advanced level (passing the language requirements of a master's/Ph.D course in an English-medium university) and with little variation. It might be claimed that, at this advanced level, natives and non-natives are equally sensitive to the distibutional properties of formulaic sequences (at least on-line while reading).

Speaking about non-native participants in the present study and in Siyanova-Chanturia et al. (2011b), an important point to note is that L2 users in both studies came from a wide variety of L1 backgrounds. Recent research on collocations has revealed L1 congruence as a significant factor during both off-line (Yamashita & Jiang, 2010) and on-line (Wolter & Gyllstad, 2011) processing. Wolter and Gyllstad (2013) did test the relationship between L1 congruence and off-line sensitivity to collocational frequency and did not find any interaction. However, a totally different picture might emerge for on-line measures and/or for lower-level non-natives (shown to be more sensitive to L1 congurence than higher-level non-natives in Yamashita and Jiang (2010)).

Usage-based theories of language acquisition

Results of the present study are incompatible with the words-and-rules approach to language processing/acquisition (Pinker, 1999). Natives and advanced non-natives are not only sensitive to the frequency of memorized individual lexical items in the language (as established by earlier research), but are also sensitive to the distributional properties of longer constructions (here collocations).

In contrast with the words-and-rules approach, usage-based theories (Bybee, 2001, 2006; Langacker, 2000) assume that the processing of any linguistic unit (from the smallest to the largest) is influenced by the frequency of exposure to/experience with that unit in the language. These assumptions are supported by results of the present study. Corpus-derived frequency plays an important role in natives' and non-natives' processing of collocations both off-line (while making explicit judgements) and on-line (while reading for comprehension). This effect was clearer off-line than on-line, however, where the effect was found only early during processing. The fact that frequency effects disappeared later might be related, as pointed out above, to the non-exclusive nature of collocations. Usage-based theories are capable of accounting for this lack of effect in late reading times through the notion of "schemata" (see Tomasello, 2000). It might be argued that, upon encountering an unattested (non-collocate) pair like *extreme mistake* during reading, the language user (whether native or non-native) will intuitively respond by spending some time to try to tackle the unnaturalness. However, given the fact that collocations are composed of two open-class lemmas filling in slots in a common abstract pattern/"schema" (adjective+noun), language users will quickly cope with the abnormality. They might be able to quickly generalize the common "schema" to the novel (unattested) pair.

Although the present research has provided interesting insights into usage-based models of language processing both off-line and on-line, the study is limited in a number of ways. First, an important assumption of usage-based models is that more exposure to the language should lead to deeper entrenchment of a unit's memory. It should be noted, however, that all participants in the present study (natives and non-natives) had high levels of academic achievement (B.A and master's/Ph.D students, respectively) and should, thus, have been exposed to a large repertoire of texts in English. Second, as pointed out above, the non-natives in the present study had little variation in their language ability and came from a variety of L1 backgrounds. Finally, the present study did not consider the influence of working memory on the retention of frequency-related information from L1/L2 exposure (see Martin & Ellis, 2012). All of these factors might have affected sensitivity to collocational frequency both off-line and on-line. In order to fully test assumptions of usage-based theories, future research will need to consider a wider range of variation not only in language users' academic attainment and working memory resources but also in their language proficiency (for non-natives). It should also consider the effect of resemblance/difference between L1 and L2 as a potential factor influencing sensitivity to the distributional properties of collocations.

Conclusion

The present study was conducted in an attempt to gain greater understanding of the emerging issue of frequency effects on the processing of formulaic sequences. The study dealt with limitations of the previous research through combining off-line and on-line measures in testing both natives and non-natives. Results showed clear off-line and early on-line sensitivity to frequency for both natives and non-natives with an effect of proficiency on the former measure only. Future research in this area will need to test a homogenous group of bilinguals in order to detect any congruence effects (L1 = L2 or L1 ≠ L2) on sensitivity to collocational frequency. Moreover, the potential contribution of factors such as level of proficiency, working memory resources, and level of academic attainment should also be considered (along with the interaction among them) in order to arrive at a deeper understanding of usage-based models of language acquisition/processing both in L1 and L2.

Appendix: Target Items

#	Noun Node	Non-collocate				Lower (5–15)				Mid (25–45)			
		Adjective	Coll Freq [a]	MI	Final Coll Freq [b]	Adjective	Coll Freq [a]	MI	Final Coll Freq [b]	Adjective	Coll Freq [a]	MI	Final Coll Freq [b]
1	COMPONENT	tremendous	0	_	0	substantial	9	1.60	6	fundamental	42	4.28	40
2	LINK	grand	1	−1.56	0	crucial	18	2.72	16	vital	50	4.00	44
3	COMBINATION	genuine	0	_	0	ideal	10	2.02	9	perfect	24	3.27	24
4	MISTAKE	extreme	0	_	0	awful	5	1.97	4	fatal	26	5.49	26
5	ALTERNATIVE	charming	0	_	0	pleasant	5	2.17	4	acceptable	39	4.65	37
6	CRITERION	heavy	2	−1.99	0	precise	10	3.01	6	strict	29	4.82	26
7	OBSERVATION	narrow	0	_	0	accurate	10	3.90	9	careful	39	5.06	36
8	WINTER	shocking	0	_	0	cruel	7	3.10	5	harsh	31	4.92	28
9	CODE	exact	0	_	0	rigid	11	3.88	10	strict	36	4.83	32
10	MAP	perfect	1	−1.36	0	complete	17	1.56	5	detailed	43	3.90	38

The table above has the following spanning header structure:

Set 1

Frequency Level

Appendix: Continued

		Set 2											
		Frequency Level											
		Non-Collocate				Mid (25–45)				Higher (55-)			
#	Noun Node	Adjective	Coll Freq [a]	MI	Final Coll Freq [b]	Adjective	Coll Freq [a]	MI	Final Coll Freq [b]	Adjective	Coll Freq [a]	MI	Final Coll Freq [b]
1	AIM	outstanding	1	−0.90	0	basic	30	2.14	27	primary	99	4.09	89
2	CAUSE	outstanding	1	−1.28	0	principal	28	2.84	25	underlying	109	5.56	109
3	REDUCTION	valuable	1	−0.97	0	considerable	35	2.86	30	substantial	76	4.59	76
4	SCALE	rich	2	−2.04	0	vast	31	3.47	29	grand	138	5.31	131
5	PAIN	supreme	0	_	0	intense	44	4.84	39	severe	92	4.92	86
6	EXPLANATION	impressive	0	_	0	convincing	38	5.74	37	satisfactory	63	5.81	61
7	LOCATION	strict	0	_	0	precise	38	4.88	36	exact	67	6.05	67
8	CONSIDERATION	solid	1	−1.21	0	proper	43	3.39	38	careful	164	5.65	162
9	CROWD	grand	1	−1.07	0	vast	25	3.62	24	huge	64	4.24	62
10	FLOW	definite	0	_	0	constant	43	4.27	42	steady	55	5.49	55

Appendix: Continued

		Set 3											
		Frequency Level											
		Non-collocate				**Lower (5–15)**				**Higher (55-)**			
#	**Noun Node**	**Adjective**	**Coll Freq [a]**	**MI**	**Final Coll Freq [b]**	**Adjective**	**Coll Freq [a]**	**MI**	**Final Coll Freq [b]**	**Adjective**	**Coll Freq [a]**	**MI**	**Final Coll Freq [b]**
1	**RECOGNITION**	standard	1	−2.60	0	proper	12	1.98	10	official	71	3.95	69
2	**WEATHER**	acute	0	_	0	harsh	11	3.69	11	severe	59	4.76	57
3	**CHARACTERISTIC**	absolute	1	−0.53	0	primary	13	1.72	8	essential	66	4.18	63
4	**SOLUTION**	supreme	0	_	0	attractive	10	1.35	9	ideal	72	4.06	68
5	**DAMAGE**	supreme	1	−1.34	0	immense	15	3.82	14	extensive	74	4.57	71
6	**MANNER**	average	1	−2.45	0	typical	7	1.39	6	usual	71	4.37	65
7	**INVESTIGATION**	correct	1	−1.90	0	lengthy	9	3.82	7	detailed	127	5.14	123
8	**INTERPRETATION**	perfect	0	_	0	precise	9	2.83	6	correct	62	4.38	59
9	**OPENING**	proper	1	−1.49	0	formal	18	2.68	14	official	103	4.60	100
10	**CONTRAST**	severe	1	−1.30	0	extreme	16	3.03	10	sharp	152	5.90	151

[a]Raw BNC frequency
[b]Raw BNC frequency after excluding occurrences where the adjective was not modifying the noun

References

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.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340–357.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America, 113,* 1001–1024.

Bybee, J. (2001). *Phonology and language use.* Cambridge: Cambridge University Press.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language, 82(4)*, 711–733.

Dagenbach, D., Horst, S., & Carr, T. H. (1990). Adding new information to semantic memory: How much learning is enough to produce automatic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 581–591.

Davies, M. (2004). BYU-BNC: The British National Corpus. Available online at http://corpus.byu.edu/bnc.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. Available online at http://www.americancorpus.org.

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.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24,* 143–188.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics. An international handbook (extended manuscript of Chapter 58, available online at http://cogsci.uni-osnabrueck.de/~severt/PUB/Evert2007HSK_extended_manuscript.pdf)*. Berlin: Mouton de Gruyter.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey & N. Hamilton-Smith (eds.), *The Computer and Literary Studies*. Edinburgh: University Press. Available online at http://www.eat.rl.ac.uk/.

Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (eds.), *Usage-Based Models of Language* (pp. 1–63). Stanford: CSLI.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwood (ed.), *Eye guidance in reading and scene perception* (pp. 55–75).

Marinis, T. (2003). Psycholinguistic techniques in second language acquisition research. *Second Language Research, 19,* 144–161.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34,* 379–413.

McDonald, S. A., & Shillcock, R. C. (2003). Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science, 14,* 648–652.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second language word learning: Minimal instruction produces rapid change. *Nature Neuroscience, 7,* 703–704.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (ed.), *Applied linguistics in society: British studies in applied linguistics 3* (pp. 80–87). London: CILT.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available online at http://w3.usf.edu/FreeAssociation/.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: HarperCollins.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124,* 372–422.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191–201.

Reali, F., & Christiansen, M. H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161–170.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18,* 55–88.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011a). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27,* 251–272.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011b). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 776–784.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning, 63,* 121–159.

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language, 83,* 227–236.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive linguistics, 11,* 61–82.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks. *Language Learning, 61,* 569–613.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4,* 105–122.

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: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition, 35,* 451–482.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

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.