LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

Investigation of Native Speaker and Second Language Learner Intuition of Collocation Frequency

Anna Siyanova-Chanturia[a] and Stefania Spina[b]

[a]Victoria University of Wellington and [b]University for Foreigners Perugia

Research into frequency intuition has focused primarily on native (L1) and, to a lesser degree, nonnative (L2) speaker intuitions about single word frequency. What remains a largely unexplored area is L1 and L2 intuitions about collocation (i.e., phrasal) frequency. To bridge this gap, the present study aimed to answer the following question: How do L2 learners and native speakers compare against each other and corpora in their subjective judgments of collocation frequency? Native speakers and learners of Italian were asked to judge 80 noun-adjective pairings as one of the following: high frequency, medium frequency, low frequency, very low frequency. Both L1 and L2 intuitions of high frequency collocations correlated strongly with corpus frequency. Neither of the two groups of participants exhibited accurate intuitions of medium and low frequency collocations. With regard to very low frequency pairings, L1 but not L2 intuitions were found to correlate with corpora for the majority of the items. Further, mixed-effects modeling revealed that L2 learners were comparable to native speakers in their judgments of the four frequency bands, although some differences did emerge. Taken together, the study provides new insights into the nature of L1 and L2 intuitions about phrasal frequency.

Keywords collocation frequency; intuition; corpus; native speakers; L2 learners; Italian

Introduction

As Alderson (2007) noted, for many languages, there are no reliable corpora available. And even when available, existing corpora are not always used for

This research was supported by the Research Establishment Grant (Victoria University of Wellington) to the first author. We wish to thank Phil Durrant and the three anonymous reviewers for their very helpful comments on the earlier draft of this paper, as well as Paul Warren, Harald Baayen, and Stefan Gries for their advice with regard to some of the statistical analyses. Any inaccuracies are, of course, our own. The elicitation instruments used for this study can be accessed by readers in the IRIS digital repository (http://www.iris-database.org).

Correspondence concerning this article should be addressed to Anna Siyanova-Chanturia, School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington 6140, New Zealand. E-mail: anna.siyanova@vuw.ac.nz

DOI: 10.1111/lang.12125

language teaching purposes. Thus, it would be of theoretical and practical interest to know if subjective judgments of the relative frequency of lexical items in a language were comparable with objective frequency counts as attested in corpora (Alderson, 2007). A broader theoretical question is whether language users (i.e., not language teachers or linguists), with many years of experience using language on a daily basis, can demonstrate accurate intuitions about a lexical item's relative frequency. Overall, corpus linguists have cast doubts on the ability of native speakers to accurately judge language frequencies (e.g., Biber, Conrad, & Reppen, 1996; Hunston, 2002; Stubbs, 1995). According to Hunston, "it is almost impossible to be conscious of the relative frequency of words, phrases and structures except in very general terms" (p. 21). Similarly, Stubbs argues that native speakers cannot give accurate estimates of the frequency and distribution of different lexical items in a language. Ironically, both of these (and other similar) propositions were made on the basis of the researchers' own intuitions, rather than drawing on empirical investigations.

However, it is not just language teacher or native speaker intuitions that have interested researchers and language educators. Some consider the knowledge of a word's relative frequency to be an essential part of word knowledge (e.g., Nation, 1990; Richards, 1976), on par with knowing the word's meaning, pronunciation, spelling, and so on. And so, the question of whether or not second language (L2) learners can reliably use their frequency intuitions is an equally important and valid one. According to Richards and Nation, part of L2 competence is the ability to judge how frequently words (and, arguably, other lexical items) are used in a language.

It is widely accepted that frequency plays a key role in natural language processing. According to Ellis (2002), the language processor is tuned to input frequency because language users are sensitive to the frequencies of linguistic events in their daily experiences. Indeed, lexical frequency effects are some of the most robust in psycholinguistic research (e.g., Balota & Chumbley, 1984; Monsell, Doyle, & Haggard, 1989; Rayner & Duffy, 1986). As proposed by Bod, Hay, and Jannedy (2003, p. 10), "frequency effects are everywhere." Some researchers have even suggested that frequency may be the main factor responsible for the organization of the mental lexicon (Forster, 1976).

In recognition of the important role of frequency in language acquisition, processing, and use, Richards (1976) and Nation (1990) included knowledge of word frequency in their lists of what a learner must know to have full mastery of a word. According to Richards, "knowing a word means knowing the degree of probability of encountering that word in speech or print" (p. 83). Knowledge

of the relative frequency of a word is also important because it is interconnected with other aspects of knowing a word, such as style and register (e.g., Alexander, 1982; Laufer, 1990; Nation, 1990; Richards, 1976). Despite the practical and theoretical importance of researching frequency intuitions in native speakers and L2 learners of various proficiency levels, few studies have thus far endeavoured to investigate frequency intuitions empirically. Indeed, linguists have long called for a more thorough examination of subjective frequency estimates in speakers' first languages (L1s) and their L2s (e.g., Alderson, 2007; McGee, 2006; Stubbs, 2001, 2002). As McGee argues, there is "a real need" for empirical research to investigate intuition-corpus differences in the areas of collocation and frequency (p. 30).

Word Frequency Intuition Research

It has been four decades since Richards (1974, 1976) argued that native English speaker intuitions of word frequency can, by and large, be deemed accurate and reliable, one exception being "concrete nouns" (1976, p. 79). Although Richards (1976) made no suggestions as to what L2 learner intuitions might be like (compared to the L1 baseline or a reference corpus), a number of researchers have since attempted to measure the accuracy of native and, to a lesser extent, nonnative speaker intuitions about (single) word frequency.

Research into subjective frequency estimates dates back to the 1960s and 1970s. In one of the earliest studies (Tryk, 1968), 50 American university students ranked 100 nouns of different frequencies for their estimates of public and personal use of these words. The estimates for private and public use were found to be almost identical when correlated with the Thorndike and Lorge (1944) objective count data. In a similar study, Shapiro (1969) examined subjective frequency estimates of sixth-graders, ninth-graders, college sophomores, chemists, schoolteachers, and newspaper reporters (20 respondents in each group) and found no differences in their subjective judgments of a mixture of words (mostly nouns) and Thorndike and Lorge's (1944) and Kučera and Francis's (1967) objective (corpus) data. Two methods were used—multiple rank order and subjective magnitude estimation—both resulting in similarly high correlations. In Carroll (1971), 15 lexicographers and 13 nonspecialists were asked to provide subjective frequency estimates for 60 words (same as in Shapiro, 1969), using the subjective magnitude estimation method. Although lexicographers performed better than nonspecialists, both groups produced high correlations with the objective count data (which were Thorndike & Lorge,

1944, and Kučera & Francis, 1967). Overall, the studies that compared subjective frequency estimates with objective frequency count data in a L1 found generally high correlations (e.g., .92–.94 in Backman, 1976; .92–.97 in Carroll, 1971; .91–.95 in Frey, 1981; .92–.98 in Shapiro, 1969; and .75–.78 in Tryk, 1968).

While the above studies were some of the first ones to address the question of the accuracy of word frequency intuitions in a L1, Ringeling (1984) was, perhaps, the first to investigate such intuitions in a L2. Ringeling compared five advanced learners of English (L1 Dutch) with five native speakers of English (staff at a Dutch university) in their judgments of word frequency. Participants were required to rank a list of 24 words according to their (a) frequencies in the English language (public use) and (b) individual experiences with these words (personal use). L1 and L2 speaker ratings were found to correlate with the objective count data obtained from Carroll (1971): .68–.85 for L1 Dutch speakers and .79–.82 for L1 English speakers. This led Ringeling to conclude that advanced L2 learners were able to develop word frequency intuitions comparable to those of their native speaker counterparts. This study is important in that it was one of the first to address the question of word frequency intuitions in a L2. However, a number of methodological shortcomings—such as a very small participant pool and the use of either very high or very low frequency words (and the absence of medium frequency items)—bring into question the validity of the findings reported.

In a more recent study, Schmitt and Dunham (1999) asked a group of native and intermediate and advanced nonnative speakers of English to judge 12 sets of near synonyms against the corresponding anchor words; then, these judgments were compared with corpus data (British National Corpus [BNC] and COBUILD). Native speaker accuracy was 77% and 85% (depending on whether the core word was the reference against which the other words were judged), while L2 speaker accuracy was 71% and 79%. The correlation between L1 ratings and corpus data was found to be relatively low ($r = .53$); interestingly, L2 correlation appeared to be stronger than that for L1 speakers ($r = .58$). It was found that L1 speakers were rather heterogeneous in their intuitions, with the education level playing a role. Overall, educated L2 learners appeared to have intuitions comparable to or better than those of native speakers with less education, while educated L1 speakers seemed to have better intuitions than their educated nonnative counterparts. This study is significant in that it highlighted the relationship between education level (both in a L1 and L2) and the accuracy of word intuition. On the downside, the design and the task were rather complex, not least because participants had to deal with fractions (noted

by McGee, 2008). For each set of synonymous words, one anchor word was assigned, for example: *glisten*, *shimmer*, *sparkle*, *shine* (anchor), and *twinkle*. Participants rated the nonanchor words against the anchor word, such that, if they thought a word to be 10 times more frequent than the anchor word, they indicated "10"; if they thought it to be one-third as frequent, they indicated "1/3" or ".33," and so on. Second, the variability between the participants was very high. Third, many words were skipped entirely and thus were not judged by some of the participants, resulting in a large amount of missing data.

In a more recent study, Alderson (2007) reports on the results of three investigations of frequency judgments. In the first experiment, participants were required to indicate how often the 100 target verbs occurred in every million words of English. In the second experiment, participants ranked the 50 target verbs according to their frequency. The third experiment investigated participants' abilities to rank the 25 target verbs according to their relative frequencies. Overall, Alderson's findings suggest that the judgments of professional linguists (who acted as participants in all three studies) did not correlate highly with corpus-derived frequency estimates. As in Schmitt and Dunham (1999), considerable individual variation in frequency judgments was found. Alderson concluded that subjective frequency judgments could not serve as accurate enough measures of objective (i.e., corpus) frequencies and called for more research into the nature of intuitions about word frequency. It is noteworthy that Alderson's study (2007) and his position have been criticized by McGee (2006, 2008). Specifically, McGee argues, and empirically demonstrates, that the (small) size of the corpora used in the early studies (e.g., Ringeling, 1984) is not problematic from the methodological standpoint, which is in contrast with Alderson's concern regarding the use of small corpora. Further, McGee is critical of the methodology used by Alderson because of "the 'uncontrolled' variety between the relative frequencies of words in his word sets" (2008, p. 511), which is also an issue with Schmitt and Dunham (1999).

Finally, McCrostie (2007) investigated English teachers' word frequency intuitions versus those of university students. In the study, the two groups of native English speakers (21 teachers and 20 university students) were asked to rank two lists of 24 words in order of their frequency. The first of the two lists covered a range of frequency bands, while the second list contained words in the middle frequency range. McCrostie's study led to two findings. First, experienced English teachers' frequency judgments were very similar to those of undergraduate university students, implying that trained teachers possessed no special skills in judging word frequency. Second, both participant

groups seemed to have difficulties judging the frequency of words in the middle frequency range; at the same time, both groups of native speakers were found to be able to accurately judge the frequency of items with very high or very low frequencies (which appears to be in line with Ringeling, 1984). The fact that native speakers were more or less successful depending on the word's frequency range (high, middle, low) is an interesting finding that will be addressed further in the discussion section.

All of the studies reviewed above have focussed on word frequency intuitions. However, the lexicon is not made only of single words. Units larger than a word (e.g., collocations, idioms, multiword verbs, binomials, lexical bundles, speech routines) have been found to constitute 20% to 50% of spoken and written native-speaker discourse (Biber, Johansson, Leech, Conrad, & Finegan,1999; Erman & Warren, 2000; Foster, 2001; Howarth, 1998; Sorhus, 1977). According to Pollio, Barlow, Fine, and Pollio (1977), four such units are produced by a native speaker in every minute of spoken discourse. Clearly, such estimates suggest that knowing and using a wide range of units above the word level is an essential characteristic of mature linguistic competence. It is natural then that our research should focus not only on intuitions about word frequency but also on intuitions pertinent to phrasal frequency.

Phrasal Frequency Intuition Research

To the best of our knowledge, only two published studies have investigated (directly or indirectly) intuitions about phrasal frequency in both L1 and L2. In one study, Hoffman and Lehmann (2000) analyzed native and nonnative speaker intuitions about 55 word pairings whose two constituent words were strongly associated in the BNC. Each word pairing consisted of a low frequency node (50 to 100 occurrences in the BNC) and a collocate word found in the ±3 word span. Although the part of speech of the target pairings varied, the majority of the target items were adjective-noun (e.g., *connective – tissue*) or noun-noun combinations (e.g., *hustle – bustle*). A questionnaire was administered to 16 native and 16 nonnative speakers whose task was to provide collocates for the node presented in isolation. Thus, the task used in this study was not a frequency intuition judgment task per se (as, e.g., the tasks employed in the aforementioned word frequency intuition studies); rather, it was a collocate-recall task. What Hoffman and Lehmann (2000) required their participants to do was more akin to a word association task (i.e., coming up with a collocate having first been presented with a node), rather than judging frequency of a particular collocation. Hoffman and Lehmann found that native speakers

supplied appropriate collocates in 70% of the cases (all except one native speaker supplied 50% or more of the node-collocate pairings), which was described as "an astonishing feat" (p. 31); nonnative speaker performance was found to be only half as accurate (34%) as that of native respondents. However, a word of caution is warranted when interpreting these results. While the authors maintain that all nonnative participants spoke "fluent English" (p. 21), they also acknowledge that their informants represented a wide range of competences, including nonnative speakers who learned English at school and only used English while on holiday (no other details were provided about L2 level). Thus, a low percentage of appropriate collocations supplied (34%) might have been due to relatively low English language proficiency among some nonnative respondents.

The second study that investigated intuitions about collocation frequency in both L1 and L2 is Siyanova and Schmitt (2008, Study 2). In this study, the researchers looked at learner intuitions by asking a group of native and nonnative speakers to rate the commonness (a more colloquial term for frequency) of adjective-noun pairs. Two groups of collocations were selected from a corpus of L2 learner writings: frequent (native-like) and infrequent (learner) collocations. In addition, the frequent collocation group was subdivided into high (over 100 occurrences in the BNC) and medium (21–100 occurrences in the BNC) frequency bands. The target items were inserted in a questionnaire, in which participants (60 native and 60 nonnative English speakers) were asked to rate each collocation on a 6-point Likert scale (1 = *very uncommon*, 6 = *very common*). Analysis revealed that native speaker intuitions mirrored the BNC frequency data more accurately than did nonnative speaker ones. In addition, while L2 learner ratings did reliably distinguish frequent from infrequent collocations, they were, nevertheless, found to be similar for the high and medium frequency bands. On the contrary, native speakers reliably distinguished frequent from infrequent, as well as high from medium frequency collocations. Further, native speakers appeared to be more decisive in their ratings, drawing on a wider range of scores, such that (on average) they gave frequent collocations higher, and infrequent collocations lower, scores than did nonnative speakers. Perhaps, as a result, native speaker ratings also correlated more strongly with the BNC frequency data ($r = .58$) than did nonnative ratings ($r = .44$).

Two more studies that looked at L1 (but not L2) phrasal intuitions are of interest to the present investigation. Backman (1978) had a group of L1 Swedish university students rank the frequency of 18 Swedish three-word combinations, such as *kanske är det* "it may be." Backman used the magnitude estimation

technique that required the participants to rank three-word combinations against an anchor. The correlation between the participants' judgments and the objective data was found to be around .56. Backman concluded that collocations "can be supposed to have psychological counterparts" (p. 2), implying that language users have relatively accurate intuitions about the frequency of phrases in language. It needs to be noted, however, that a mixture of three-word phrases was used, such as literal and idiomatic, collocations and lexical bundles (e.g., English translations from Swedish, including *it may be*, *at heart*, *to devote oneself to*, *in the course of time*, *a great deal*, *of various kinds*). In addition, the correlation of .56 is a much lower figure than what was found in the earlier research into L1 word frequency intuition (e.g., Carroll, 1971; Shapiro, 1969; Tryk, 1968).

Finally, McGee (2009) reports on an experiment that was designed to compare BNC data and English language teacher intuitions (all being native speakers) about the most frequent collocates of 20 common English adjectives. Participants were presented with a list of target adjectives (e.g., *different*, *difficult*, *full*, *good*, *great*, *important*, *large*, *main*, *old*, *particular*, *personal*, *possible*, *real*, *recent*, *similar*, *small*, *special*, *strong*, *various*, and *young*) and were required to produce one most frequent (according to their intuitions) collocate for each adjective. Their responses were then compared to BNC frequency data for the same adjectives. The results suggested that, with the exception of three adjectives (*difficult*, *real*, *young*), the teachers' intuitions about the most frequent collocates of the target adjectives differed significantly from the BNC data, implying clear dissociation between the elicited (intuition) and the corpus data.

The Current Study

The above review suggests that, arguably, the biggest gap in frequency intuition research pertains to L2 learners and their subjective perceptions of phrasal frequencies. With this in mind, the present study addressed the following question: How do L2 learners and native speakers compare against each other and corpora in their subjective judgments of collocation frequency? It was hoped that the study would provide new insights into the poorly understood area of native and nonnative speaker perceptions of collocation frequency. In addition, given that most of the above studies were conducted in English, targeting English word and phrasal frequencies, it was decided to focus on a different, currently underresearched, language—Italian. Although we do not believe that learning and using Italian collocations (e.g., Spina, 2010) is fundamentally different from

learning and using English (or French, German, Spanish, etc.) collocations, we do believe that it is important for a wider range of L2s to be represented alongside English (see also Ortega, 2009, for a similar stance).

Method

Participants

Native speaker participants were 42 native speakers of Italian (25 females), with a mean age of 31 years (18–55, $SD = 9.6$), who came from various parts of Italy. Nonnative participants were 42 nonnative speakers of Italian (35 females), with a mean age of 23.9 years (18–29, $SD = 3.0$). Nonnative speakers came from a range of L1 backgrounds, including Armenian (8), Chinese, Spanish (5), English, German, Polish (3), Farsi, French, Russian (2), Catalan, Croatian, Czech, Danish, Greek, Lithuanian, Norwegian, Swedish, and Turkish (1). Both groups of participants were comparable in terms of their education and socioeconomic background; it was important to ensure a comparable educational background, as this factor has previously been found to affect intuitions (e.g., Alderson, 2007; McGee, 2009; Schmitt & Dunham, 1999). The participants were Ph.D. students, researchers, teachers, or young professionals, that is, individuals with a completed bachelor's degree. All participants were unpaid volunteers.

Nonnative speakers were required to complete a language background questionnaire and to rate their speaking, reading, writing, and comprehension on a 5-point Likert scale (1 = very poor, 2 = weak, 3 = ok, 4 = good, 5 = excellent; for a comparable procedure, see Siyanova-Chanturia, Conklin, & van Heuven, 2011). Participants also provided their Italian language qualification and level (e.g., A1, A2, B1, etc. using the Common European Framework of Reference for Languages, http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp). On the basis of the ratings and the reported level of the Italian language, nonnative speakers were divided into two proficiency groups: advanced and intermediate learners of Italian, with *t* tests showing that the two learner groups differed significantly in their self-reported speaking, reading, writing, and comprehension scores (all ps $< .05$) and in the amount of time spent in Italy ($p < .05$). The two groups' demographics and experience with Italian are summarized in Table 1.

Stimuli and Instrument

The purpose of the study was to investigate native and nonnative speaker intuition of Italian word pairings of various frequencies. Specifically, we wanted to focus on a range of frequencies (high, medium, low, very low) to allow for a more nuanced analysis of phrasal frequency intuition in both L1 and L2.

Table 1 Nonnative speaker (NNS) demographics and self-reported L2 experience

	NNS advanced ($n = 21$)			NNS intermediate ($n = 21$)		
Variable	*M*	*SD*	Range	*M*	*SD*	Range
Age[a]	25.7	2.2	21–29	22.1	2.6	18–29
First contact with Italian[a]	18.4	4.0	8–25	17.9	2.7	15–23
Time in Italy[b]	24.8	23.0	2–78	6.8	6.4	0–24
Speaking[c]	4.4	0.6	3–5	3.5	1.0	2–5
Reading[c]	4.6	0.5	4–5	4.0	0.8	2–5
Writing[c]	4.3	0.6	3–5	3.3	0.6	2–4
Comprehension[c]	4.6	0.5	4–5	4.0	0.9	2–5

Note. [a]In years. [b]In months; the data from three nonnative speakers were missing, so the values were calculated for 20 intermediate and 19 advanced speakers. [c]Based on a 5-point scale (1 = very poor, 2 = weak, 3 = ok, 4 = good, 5 = excellent).

Using the Perugia corpus,[1] a preliminary (random) pool of 265 noun-adjective pairings of various frequencies was selected. No criteria were applied other than part of speech of Word 1 being a noun and Word 2 an adjective; in addition, we did not consider word combinations with proper nouns or adjectives that denote nationalities, or anything that is normally capitalized in English or Italian. The selected items were then ranked according to their raw frequency. From the obtained list, we selected (using a stratified random procedure) three groups of collocations that, on the basis of their Perugia phrasal frequencies, were assigned to one of the following frequency groups: high frequency (e.g., *tempo libero* "free time," $n = 20$), medium frequency (e.g., *guida turistica* "tourist guide," $n = 20$), and low frequency (*libro interessante* "interesting book," $n = 20$). Although the collocations were extracted from one reference corpus (Perugia), their relative frequencies were compared against another Italian corpus—la Repubblica corpus.[2] The relative Perugia frequencies of the selected items were comparable with la Repubblica frequencies; the two were found to correlate significantly, as shown by a Spearman correlation test ($r = .96, p < .001$).

When selecting and grouping our collocations, it was deemed necessary to avoid using borderline items; that is, there were no collocations with similar frequencies but belonging to different frequency bands (see Table 2). In addition, another 20 noun-adjective phrases were created (rather than extracted

Table 2 Frequency profiles for target materials

	Perugia corpus			La Repubblica corpus		
Frequency group	*M*	*SD*	Range	*M*	*SD*	Range
High ($n = 20$)	2.1	0.2	1.9–2.6	3.4	0.4	2.9–4.5
Medium ($n = 20$)	1.4	0.1	1.3–1.6	2.4	0.2	1.9–2.7
Low ($n = 20$)	0.6	0.1	0.5–0.8	1.4	0.3	1.0–1.7
Very low ($n = 20$)	0	0		0	0	
Word 1 ($n = 80$)	3.1	0.7	1.2–4.3	4.2	0.8	2.1–5.7
Word 2 ($n = 80$)	2.7	0.6	1.3–4.1	3.8	0.8	1.7–5.3

Note. Frequency values are based on logarithmically transformed raw frequency estimates (e.g., Baayen, 2005).

from a corpus) in order to be assigned to a fourth frequency group, namely, very low frequency. These 20 items were meaningful and grammatically correct (as judged by the authors of this study, who are proficient nonnative and native speakers of Italian, respectively), however, they did not occur either in the Perugia or la Repubblica corpora (e.g., *nonni ospitali* "hospitable grandparents"). Further, we obtained individual frequencies for Word 1 and Word 2 of the selected 80 collocations. This was done to ensure that no rare words were present, which may have been unknown to our nonnative speakers. The pairings within the four frequency groups differed significantly in corpus frequency (Perugia and la Repubblica), as suggested by *t* tests (all $ps < .001$). Individual Word 1 and Word 2 frequencies are summarized in Table 2, and the Italian combinations used in the study are listed in Appendix S1 in the Supporting Information online.

As was mentioned above, two corpora were used during the norming stage; this was done to ensure that the frequencies extracted were not specific to one corpus (or one genre) and that they were representative of the Italian language in general. As an additional step, however, Google[3] frequencies for the 80 target items were obtained and then correlated with those from the Perugia and la Repubblica corpora. Both sets of corpus frequencies correlated very strongly with Google frequencies, as shown by Spearman correlation tests (Perugia vs. Google: $r = .91$, $p < .001$; la Repubblica vs. Google: $r = .89$, $p < .001$).

Once selected, the target items were incorporated into a questionnaire. Participants were asked to judge the target collocations on a 4-point scale: high frequency, medium frequency, low frequency, and very low frequency. Because our participants were nonnative speakers of Italian (as well as native speakers),

it was deemed necessary to include an "I don't know" option, in case some of the words were not familiar to nonnative speakers. This option (always appearing last) was not part of the 4-point scale. The questionnaire contained the following parts: instructions, the 80 collocations, and demographic and language background questionnaire (for L2 speakers only). Finally, two versions of the same questionnaire were created and administered, such that the order of the items in the two questionnaires was different (otherwise, the two versions were identical). Half of the participants completed Version 1 of the questionnaire, and the other half Version 2. In sum, the following design features set the present investigation apart from previous research (e.g., Ringeling, 1984; Schmitt & Dunham, 1999): a simple task employed that was unlikely to confuse the respondents, a relatively large pool of participants, a range of collocation frequencies used, and the use of three large reference corpora, which ensured the representativeness of the relative frequencies in the Italian language.

Procedure

All data collection was conducted by means of a questionnaire administered online; for this purpose, each participant was provided with a Web link. Although there was no time pressure to complete the questionnaire, participants were urged not to consult anything or anyone and to complete the questionnaire in one go. It is estimated that, on average, participants took around 15 minutes to complete the questionnaire. They were advised that the questionnaire was not a language test and that there were no right or wrong answers. Detailed instructions were provided both in Italian and English (the task, however, was entirely in Italian). The English version of the instructions is provided in Appendix S2 in the Supporting Information online.

Analysis and Predictions

As was mentioned above, participants were required to rate target collocations on a 4-point scale: high frequency, medium frequency, low frequency, and very low frequency. During the coding process, the four types of ratings were coded as follows: 4 = high frequency, 3 = medium frequency, 2 = low frequency, and 1 = very low frequency. Some items received the "I don't know" rating; these data (natives = 1.9%, intermediate = 8.9%, advanced = 1.8%) were excluded from the analyses.

Based on previous frequency intuition research, we hypothesized that both native and, to a lesser degree, nonnative speakers should exhibit more accurate (i.e., more corpora-like) judgments about collocation frequency in the extreme (high and very low) frequency bands than in the middle frequency

bands. Second, we expected native speakers to perform better than nonnative speakers, and advanced L2 learners to perform better than intermediate L2 learners.

Results

Mixed-Effects Modeling

In order to explore how L2 learners and native speakers compare against each other in their subjective judgments of collocation frequency, we used mixed-effects modeling (e.g., Baayen, Davidson, & Bates, 2008). The model was built using R version 3.0.2 (2013–09–25) and the R packages lme4 (version 1.0–6; Bates, Maechler, & Bolker, 2012), lmerTest (version 2.0–6) and languageR (version 1.4.1, Baayen 2008). The following predictors were included in the model: (a) collocation frequency (Perugia); (b) Word 1/Word 2 frequency (Perugia); (c) collocation frequency band (4 = *high*, 1 = *very low*); (d) the dispersion of the collocations, measured through the deviation of proportions (DP) value (e.g., Gries, 2008); (e) Word 1/Word 2 length (in characters); (f) proficiency (natives, intermediate nonnatives, advanced nonnatives); and (g) questionnaire version (1 or 2). In addition, we addressed the issue of collinearity between some of the predictors via residualization. Specifically, we identified highly correlated pairs of predictors (frequency and frequency band, $r = .6$) and moderately correlated pairs of predictors (Word 1 frequency and Word 1 length, $r = .3$; Word 2 frequency and Word 2 length, $r = .3$). We then residualized frequency band against frequency, Word 1 frequency against Word 1 length, and Word 2 frequency against Word 2 length. For example, by residualizing word frequency against word length, we obtained the effect of frequency that is not explained or predicted by word length, and removed collinearity from this relationship, such that there was no association between residualized Word 1 and Word 2 frequency ($r = –.08$) and Word 1 and Word 2 length ($r = .08$). Even without a significant correlation ($r = –.1$), we further residualized frequency against dispersion, in order to partial dispersion out of frequency and obtain the effect of frequency that is not already accounted for by dispersion. A summary of the variables used in the model can be found in Table 3.

Starting with a model that included the above predictors as independent variables, native and nonnative speaker judgments of collocation frequency as the dependent variable, and participants and items as random effects, we proceeded with a step-by-step backward model selection procedure (Manning, 2007), removing nonsignificant predictors and proceeding only if the likelihood ratio test was nonsignificant. Then, we added interactions between the

Table 3 Summary of variables used in mixed-effects modeling, with the adjusted range after residualization shown in parentheses

Variable	Range (adjusted)	*SD*	Median
Frequency	0–445 (–70.34–365.62)	84.67	–32.90
Frequency band	1–4 (–2.29–1.23)	0.83	–0.10
Word 1 frequency	16–21783 (–5795–15797)	4751.10	–1463.00
Word 2 frequency	19–11793 (–1704.2–10086.9)	1506.47	–315.40
Dispersion	0.23–0.75	0.14	0.43
Word 1 length	4–10	1.51	6
Word 2 length	4–12	1.99	7

Table 4 Summary of the model for native and (advanced and intermediate) nonnative speakers' judgments of collocation frequency

	Estimate	*SE*	*df*	*t*-value	*Pr*(>|*t*|)
(Intercept)	3.414e+00	2.108e−01	1.150e+02	16.200	<2e−16
Frequency band	2.535e−01	7.071e−02	5.600e+01	3.585	0.000709
Frequency	2.943e−03	5.418e−04	5.600e+01	5.432	1.26e-06
Word1 length	−7.191e−02	3.119e−02	8.000e+01	−2.306	0.023704
Proficiency:word 1 length	5.646e-02	1.756e-02	4.796e+03	3.215	0.001311

predictors. We looked, in particular, at the interactions of various predictors with proficiency (natives, intermediate nonnatives, advanced nonnatives). The interactions between the predictors were only included if the model fit was significantly better compared to the previous model, that is, if the likelihood ratio test was associated with a *p* value lower than .05 (e.g., Siyanova-Chanturia et al., 2011). Following this procedure, we obtained a final model with three significant predictors (frequency band, collocation frequency, Word 1 length), and a significant interaction between proficiency and Word 1 length. No significant interactions were found between proficiency and frequency band, or proficiency and collocation frequency. The coefficients of the fixed effects and their *p* values can be found in Table 4.

Thus, the analysis revealed that both natives' and nonnatives' judgments of collocation frequency were significantly affected by the frequency band of the collocations. Furthermore, the model showed that collocation frequency significantly influenced native and nonnative participants' judgments and that

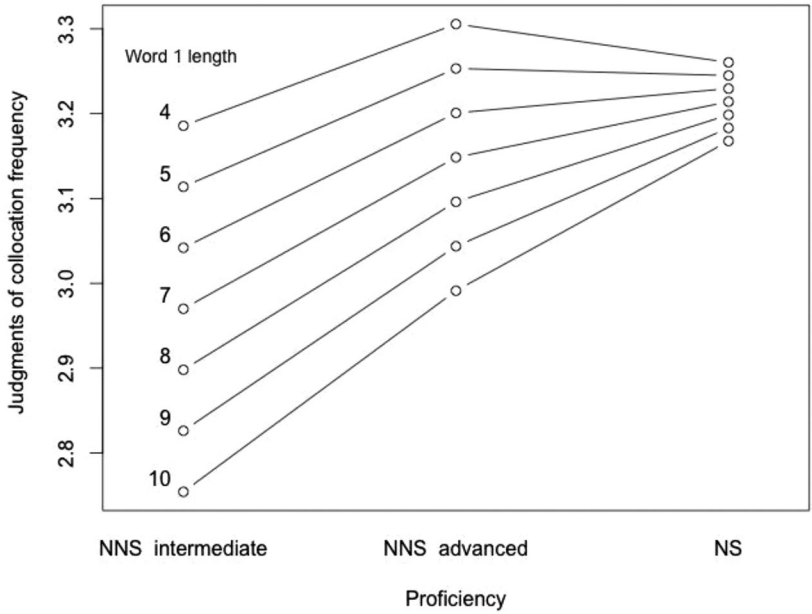

Figure 1 Participants' judgments of collocation frequency as a function Word 1 length (in characters).

this effect was not due to dispersion, because the effect of dispersion was partialed out of frequency. In addition, the significant interaction between proficiency and Word 1 length (plotted in Figure 1) suggested that natives' and (intermediate and advanced) nonnatives' judgments were affected differently by Word 1 length. As can be seen from Figure 1, native speaker judgments were clustered around 3.2, advanced learners' judgments had a wider range (from around 3 to around 3.3), while intermediate learners' judgments had the widest range, from around 2.8 to around 3.2. Interestingly, however, when the model was run only with natives, only with intermediate learners, and only with advanced learners (i.e., one proficiency group at a time), Word 1 length was not found to be a significant predictor in any of the three analyses (all ps $> .05$).

Finally, we wanted to know how strongly advanced and intermediate learner judgments correlated with the judgments by native speakers. Spearman correlation tests showed that advanced L2 judgments correlated more strongly with native speaker data ($r = .89$, $p < .001$) than did intermediate L2 judgments ($r = .72$, $p < .001$), although both correlations were highly significant.[4]

Table 5 Summary of Cohen's κ test statistics examining the agreement between native speakers' (NS) and nonnative speakers' (NNS) judgments and corpora frequency values for the four collocation frequency bands

Frequency band	NS ($n = 42$)	NNS ($n = 42$)	NNS intermediate ($n = 21$)	NNS advanced ($n = 21$)
High	20	20	18	20
Medium	1	1	3	1
Low	0	0	0	0
Very low	12	1	2	6

Note. The values indicate the total number of target items for which very strong ($\kappa = .80$–1.00), strong ($\kappa = .60$–$.79$), or moderate agreement ($\kappa = .40$–$.59$) was found between participants' judgments and corpora frequency bands. Full data are available in Appendix S1 in the Supporting Information online.

Agreement of L1 and L2 Judgments with Corpora

In order to measure the agreement between native and nonnative speaker judgments and the prior assumptions derived from the corpora (i.e., the adopted frequency bands, with 4 = high frequency, 3 = medium frequency, 2 = low frequency, and 1 = very low frequency), Cohen's kappa (κ) statistic was used. Cohen's κ coefficient is a statistical measure of inter-rater or inter-annotator agreement (Cohen, 1960).[5] In computational linguistics, Cohen's κ coefficient is often considered the standard test to measure inter-annotator agreement in corpora annotation tasks, where one needs to determine the consistency of different classifications of linguistic items, such as, for example, grammatical categories.

Cohen's κ coefficient was calculated individually for each of the 80 items (see Appendix S1 in the Supporting Information online), and a summary of this analysis is provided in Table 5. Native and nonnative speaker judgments agreed strongly with the corpora in the case of high frequency collocations. Table 5 shows that native speaker judgments exhibited almost perfect or very strong agreement for most of the 20 high frequency collocations. The same pattern of results was observed for advanced and intermediate learners (only 2 out of the 20 intermediate learner judgments revealed poor agreement with corpora frequency (i.e., for *posta elettronica* "e-mail" and *vita privata* "private life"). Conversely, no participant group showed accurate intuitions, relative to corpora frequency, for medium and low frequency collocations. For medium frequency items, moderate agreement was found only occasionally (native speakers = one item, advanced learners = one item, intermediate learners = three items), while

for the 20 collocations within the low frequency band, poor agreement with corpora was observed across all participant groups. As for very low frequency collocations, only native speakers showed moderate-to-strong agreement with corpora for the majority of the items (12 items). Advanced and intermediate L2 learners demonstrated moderate correlation with the corpora for six and two items, respectively (except for *finestra precoce* "early window," which showed strong agreement with corpora in advanced learner judgments). Agreement values for each of the 80 target items can be found in Appendix S1 in the Supporting Information online.

Discussion

By and large, researchers in the area of vocabulary acquisition agree that full mastery of a word necessitates a lot more than merely knowing a word's meaning. However, we still know relatively little about some aspects of the word knowledge, for example, word frequency intuitions in a L2 or intuitions about collocation frequency. In part, this is because it is extremely difficult to tap into intuitions of any kind (Schmitt & Dunham, 1999).

Thus far, research into frequency intuitions, albeit inconclusive, has focused primarily on the issue of native and, to a lesser degree, nonnative speaker intuitions about word frequency (e.g., Alderson, 2007; McCrostie, 2007; Ringeling, 1984; Schmitt & Dunham, 1999). This is unsurprising as a single word has traditionally been the major unit of vocabulary learning (e.g., Laufer, 1989, 1997a, 1997b; Nation, 1990, 2001; for an interesting discussion of the construct of word in applied linguistics, see Gardner, 2007). However, it has long been acknowledged that our mental lexicon is made not only of single words, but also of a large number of units above the word level (e.g., Jackendoff, 1995; Langacker, 1987; Tomasello, 2003). According to Jackendoff, the number of phrasal units in American English is comparable to the number of single words; thus, much of the language we encounter on a daily basis is formulaic.

Subjective Frequency Judgments

As was argued earlier in this article, an important aspect of Richards's (1976) assumption of word knowledge is that knowing a word also means knowing the degree of probability of encountering a word in speech or print (see also Nation, 1990; Schmitt & Meara, 1997; Schmitt, 1999). Richards further proposes that "for many words we also 'know' the sorts of words most likely to be found associated with the word" (p. 79). However, surprisingly little has been done

to empirically test subjective frequency estimates of units larger than a single word. What, to date, remains a largely unexplored area is native and nonnative speaker intuitions about collocation (i.e., phrasal) frequency. To bridge this gap, the present study investigated intuitions about collocation frequency in L1 and L2 Italian. We asked a group of native speakers and (advanced and intermediate) L2 learners to judge the frequency of 80 Italian noun-adjective collocations as high, medium, low, or very low frequency. Data analyses were conducted such that native speaker judgments were compared against nonnative speaker ones, using mixed-effects modeling; and native and nonnative speaker judgments were correlated with corpora-derived frequency information.

Both native speakers and (advanced and intermediate) nonnatives were sensitive to the frequency of collocations; their judgments were found to be affected by corpus frequency as well as the frequency bands. Interestingly, our analysis suggested comparable native and nonnative speaker intuitions about collocation frequencies. On the contrary, the individual word frequencies were not found to be significant predictors. Further, a significant interaction between proficiency and Word 1 (noun) length suggested that the three proficiency groups were differently affected by the variation in Word 1 length. Native speaker judgments did not appear to be influenced by Word 1 length, such that their scores were similar for longer and shorter words. In contrast, advanced and, especially, intermediate learner judgments seemed to have been differently affected by Word 1 length, with shorter nouns receiving higher scores than longer ones. We observed no significant interactions between proficiency and Word 2 (adjective) length.

Although mixed-effects modeling revealed comparable native and nonnative intuitions about collocation frequencies, correlation analyses showed some interesting differences. We found very strong correlations between both native and nonnative speaker judgments of high frequency collocations and the prior assumptions of collocation frequency derived from the corpora; correlations between intermediate learners and corpora were slightly lower, but still very strong. Further, we found poor agreement between corpora values and both native and nonnative speaker intuitions of medium and low frequency collocations. With respect to the very low frequency band, only for native speakers did we observe moderate-to-strong agreement between frequency judgments and the prior assumptions of collocation frequency for the majority of the items judged. As such, these findings provide support for our hypothesis that the more experienced language users should perform better (relative to a reference corpus) than the less experienced ones.

The finding of different degrees of agreement with the corpora for the various frequency bands provides evidence in support of our hypothesis that native speakers should exhibit more accurate judgments about collocation frequency in the extreme (high and very low) frequency bands than in the two middle frequency bands. Our results suggest that native speakers exhibited good intuitions only in the case of the two extreme frequency bands (although the intuitions were better for the high frequency items). L2 learners were also most accurate in the high frequency band; in fact, their intuitions were as good as those of native speakers. When the L2 group was split into advanced and intermediate, it was found that in the very low frequency band, moderate agreement was observed for six out of 20 items in advanced learners, and for two out of 20 items in intermediate learners. Thus, advanced learners also appeared (relatively) most accurate in the two extreme frequency bands (high and very low frequency) just as their native speaker counterparts. These findings are in line with McCrostie (2007), whose participants were found to have difficulties judging the frequency of words in the middle frequency range, but were able to accurately judge word frequency in the very high and very low frequency bands. Although McCrostie investigated word frequency intuitions, while we looked at intuitions about collocation frequency, the common pattern of results observed in the two studies is noteworthy and allows us to draw parallels between the mechanisms involved in subjective word and collocation frequency intuitions.

A few more words need to be said about the accuracy of native speaker intuitions—a topic that has stirred some controversy among researchers (e.g., McGee, 2008, 2009). By and large, corpus and applied linguists have questioned the ability of native speakers to accurately judge language frequencies (e.g., Biber et al., 1996; Hoey, 2000; Hunston, 2002; Sinclair, 1991; Stubbs, 1995, 1996; Wray, 2002). As Hunston argues, it is impossible to be consciously aware of the relative frequency of words, phrases, and other structures; and as Hoey further elaborates, intuitions even of trained linguists are likely to be flawed. And so, it is often assumed that language corpora are the only reliable source of frequency information. However, some researchers (e.g., McGee, 2008, 2009; Siyanova & Schmitt, 2008) have challenged the idea of inadequacy of native speaker frequency intuitions. Siyanova and Schmitt argued that, although corpora are indeed very useful in identifying the most frequent and representative collocations to be incorporated in teaching materials, language teachers nevertheless should be able to trust their intuitions about collocation frequency.

The results of the present investigation are not straightforward, as native speakers exhibited good intuitions in the case of the extreme, but not middle, frequency bands. As pointed out by Alderson (2007), the level of frequency to be judged is indeed an important variable. This finding suggests that a language user (a native speaker or a L2 learner) may find it easier to accurately judge the frequency of something very frequent or very infrequent, because such items are salient, in that they appear at the far ends of the frequency continuum. Highly frequent or infrequent items may strike a language user as something they have heard many times or, perhaps, never. Items that are in the middle of a frequency continuum might be more difficult to judge accurately precisely because they are less salient and less striking than highly frequent or infrequent items. Although tentative, this proposition appears to be in line with usage-based and exemplar-based theories (discussed below in greater detail), according to which speakers' mental representations are determined by language use. The more frequent an item, the stronger its mental representation; conversely, the less frequent an item, the weaker its mental representation. Items that are highly infrequent (e.g., zero frequency in a large corpus) may have a very weak mental representation and may thus stand out.

However, the observed selectively good intuitions may imply that, perhaps, quantitative analyses alone, such as those conducted in the present study, cannot satisfactorily answer the question of whether or not native speakers (and proficient nonnatives) are successful in judging subjective frequencies (and why this might be the case), as the accuracy of their judgments seems to depend on what is being judged—high, low, or medium frequency items. It may be that a combination of quantitative analyses together with qualitative techniques (e.g., retrospective interviews) can shed more light on the nature of these selectively accurate subjective frequency judgments and on the possible strategies employed by participants during the task (for a similar proposition, see Alderson, 2007).[6] This, however, will remain to be addressed in future research.

Theoretical Implications

As has been argued throughout the paper, the mental lexicon of native and proficient nonnative speakers consists not only of single words and highly idiosyncratic phrases (such as idioms), but also of thousands of phrasal elements varying in length, frequency, internal structure, degree of fixedness, abstractness, and other factors (Langacker, 1987). That is, humans have the ability to store large numbers of frequent phrases alongside single words (but

see Siyanova-Chanturia, 2015, for a discussion of "holistic" storage of frequent phrases). It has been proposed that such units form chunks in long-term memory (Ellis, 2001)and that it is easier and more economic to learn and use language in chunks rather than as a combination of single words (Langacker, 1987; Wray, 2002, 2008). Indeed, recent psycholinguistic research (e.g., involving behavioural and eye-tracking data and event-related brain potentials [ERPs]) has demonstrated that frequent phrases are processed differently from novel phrases. Processing studies with collocations, binomials, lexical bundles, complex prepositions, and idioms have shown that frequent phrases enjoy quantitatively faster processing times compared to novel strings (Siyanova, 2010; Siyanova-Chanturia, 2013; Siyanova-Chanturia & Martinez, 2014). This finding, emphasizing the role of phrasal frequency, has important implications for the nature of the mental lexicon and theories of language acquisition, processing, and use. Specifically, the processing advantage for frequent phrases over control phrases provides empirical evidence that argues against the traditional distinction between the lexicon, a collection of memorized forms, and grammar, a collection of rules (Arnon & Snider, 2010; Siyanova-Chanturia et al., 2011). Importantly, it has been taken to support a number of usage-based and exemplar-based models according to which the basic unit of language acquisition is a construction (Abbot-Smith & Tomasello, 2006; Bod, 2006; Borensztajn, Zuidema, & Bod, 2009; Bybee, 1998; Goldberg, 2006; Pierrehumbert, 2001; Tomasello, 2003).[7] In line with usage-based and exemplar-based theories, our mental representations are determined solely by language use, in other words, by frequency of occurrence (Abbot-Smith & Tomasello, 2006; Bod, 1998, 2006; Bybee, 1985, 1995, 1998, 2006; Croft, 2001; Goldberg, 1995, 2006; Langacker, 1987; Pierrehumbert, 2001; Tomasello, 2003, 2006; also see Ellis, 2011, 2012). As Bod (2006) noted, language should be viewed not as a set of grammar rules, but as a statistical accumulation of experiences that changes every time a particular utterance is encountered.

In line with Bod (2006), language users continuously "tag" each and every occurrence of a form, and should thus be sensitive to the frequency of occurrence of various linguistic events (at the word or phrase level). This sensitivity can manifest itself in a number of ways, for example, in online experiments.[8] Reaction times studies generally show faster reading/reaction times for higher frequency items compared to controls (e.g., Arnon & Snider, 2010). Similarly, eye-tracking studies show fewer and shorter fixations on more frequent chunks than control phrases (e.g., Siyanova-Chanturia, Conklin, & Schmitt, 2011; Siyanova-Chanturia et al., 2011). ERP studies show that higher frequency items (e.g., idioms and collocations) enjoy not only a

quantitative advantage compared to novel phrases, but also a qualitative one; that is, they exhibit easier semantic integration than matched novel phrases (e.g., Laurent, Denhières, Passerieux, Iakimovac, & Hardy-Baylé, 2006; Siyanova, 2010; Strandburg et al., 1993; Vespignani, Canal, Molinaro, Fonda, & Cacciari, 2010; for an overview, see Siyanova-Chanturia, 2013).

Clearly, the sensitivity to frequency distributions in language can also manifest itself offline, when language users are explicitly asked to judge (or rank) the relative frequency of a linguistic event (e.g., a word or a phrase). In one early study on subjective frequency estimates, Tryk (1968) proposed that "people carry with them a kind of subjective 'yardstick' of word frequency enabling them to measure the 'magnitude' of words in a dimension of word frequency" (p. 170). Although Tryk's study focused on word frequency, his proposition seems to also be true of units above the word level. In line with Bod (2006) and Bybee (1998), this "subjective yardstick" changes every time a particular word or phrase is encountered. If this is so, then, an important role in developing frequency intuitions clearly belongs to frequency of exposure. More exposure will lead to more accurate (more representative) intuitions, less exposure will result in poorer intuitions. The use of advanced and intermediate L2 learners, alongside native speakers, allowed us to investigate the role of frequency of exposure. We found that adult native speakers, who have accumulated a sufficient amount of experience with Italian noun-adjective pairings of various frequencies, exhibited better intuitions about the very low frequency items than did adult nonnative speakers, whose exposure to Italian has not been as rich. Indeed, 12 native speaker judgments correlated with the corpus data, while only one nonnative speaker judgment correlated with the corpus data, when nonnatives were considered as one group. It also appears that more experienced (advanced) learners had an advantage compared to less experienced (intermediate) learners in their judgments of the very low frequency items. Six advanced learner judgments correlated with the corpus data, while only two intermediate learner judgments correlated with the corpus data, when nonnatives were considered as two proficiency groups. In addition, our finding of a significant interaction between proficiency and Word 1 length further illustrates clear differences between natives and nonnatives of different proficiencies, with advanced learners being in the middle of an intuition–proficiency continuum and intermediate learners and native speakers being at the far ends.

Conclusion

In conclusion, our data showed that both native and nonnative Italian speakers had difficulties judging the frequency of collocations in the two middle ranges. It seems that it is almost impossible to answer the question of whether or not language users have accurate intuitions about collocation frequency; it all depends on the frequency range in question, with intuitions about high and very low (but not medium and low) frequency items correlating more strongly with the corpus data. As such, the present study has reaffirmed the need for a more nuanced approach to the investigation of collocation frequency and the nature of frequency intuitions. Our findings, albeit not straightforward, should be viewed as a step closer to making more sense of the complexities involved in subjective frequency estimates. It is hoped that this study has presented the case for why more research is needed not only into the framework of word knowledge (Schmitt, 1999; Schmitt & Meara, 1997), but also into the framework of collocation knowledge. After all, if units above the word level are an integral part of the lexicon, then intuitions about phrasal frequency should be just as important and relevant to the study of the mental lexicon as intuitions about single word frequency.

Final revised version accepted 26 July 2014

Notes

1 The Perugia corpus (Spina, 2014; http://perugiacorpus.unistrapg.it) is a collection of about 26 million words of written and spoken Italian (written: 22 million words, spoken: 4 million words), divided into 10 textual genres: academic prose (1,113,590 words), administrative texts (1,160,334 words), school essays (1,257,842 words), literary fiction (3,619,472 words), nonfiction texts (2,384,059 words), spoken texts (2,158,522 words), television transcriptions (1,147,151 words), web texts (7,359,419 words), press (5,772,170 words), and film dialogues (626,487 words). At the time the study was conducted, the Perugia corpus sections "web texts," "press," and "literary fiction" were not completed, and the corpus measured about 14 million words. The corpus covers the years between 1995 and 2012.

2 La Repubblica corpus (http://dev.sslmit.unibo.it/corpora/corpus.php?path = &name = Repubblica) is a written collection of about 380 million words that contains all the articles published by the Italian national daily newspaper *La Repubblica* between 1985 and 2000 (Baroni, Bernardini, Comastri, Piccioni, & Volpi, 2004).

3 Although the Web is not a perfect corpus, many researchers have argued that the large amount of data available online outweighs potential problems (e.g., Keller & Lapata, 2003; see also the 2003 special issue of *Computational Linguistics* dedicated to using the Web as a corpus).

4 Correlation analyses were performed on all 80 items (rather than separately for each frequency band), in order to obtain a bigger picture.
5 Studies that measure the agreement between two or more observers generally include a statistic that takes into account the fact that observers may agree or disagree simply by chance. The κ coefficient is a commonly used statistic for this purpose. A κ of 1 indicates perfect agreement, a κ of 0 indicates no agreement (Viera & Garrett, 2005).
6 One of the reviewers noted that the participants, although having been asked to judge the frequency of cooccurrence, might have been judging the collocation strength between the two words. Future qualitative analyses may also want to address this concern.
7 Usage-based and exemplar-based accounts, also known as "empiricist" theories, define a construction as "associations between a semantic frame and a syntactic pattern, for which the meaning or form is not strictly predictable from its component parts" (Borensztajn et al., 2009, p. 175).
8 We consider online processing as one happening in real time, under significant time pressure. In online studies, reaction times, eye movements, and/or brain activity are recorded while participants perform a task in a laboratory setting. Online studies contrast with offline ones (e.g., the present investigation) where a given task is performed under no immediate time pressure (studies employing written corpora, questionnaires, and surveys).

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*, 383–409.

Alexander, R. (1982). What in a four-letter word? Word meaning in English and second language meaning. *Die Neueren Sprachen*, *81*, 219–224.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82. doi:10.1016/j.jml.2009.09.005

Baayen, H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69–83). New York: Routledge.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Backman, J. (1976). Some common word attributes and their relations to objective frequency counts. *Scandinavian Journal of Educational Research*, *20*, 175–186.

Backman, J. (1978). Subjective structures in linguistic recurrence. ERIC Document ED180195.

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 & Performance*, *10*, 340–357.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. et al. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*, 1771–1774.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375–35. Retreived from http://CRAN.R-project.org/package!lme4

Bates, D. M., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0–6.

Biber, D., Conrad, S., & Reppen, R. (1996). Corpus based investigations of language use. *Annual Review of Applied Linguistics*, *16*, 115–135.

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

Bod, R. (1998). *Beyond grammar: An experience-based theory of language*. Stanford, CA: Center for the Study of Language and Information.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from exemplars. *The Linguistic Review*, *23*, 291–320.

Bod, R., J. Hay, & S. Jannedy. (2003). 'Introduction' in R. Bod, J. Hay, & S. Jannedy (eds): *Probabilistic Linguistics* (pp. 1–10). The MIT Press.

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age-evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, *1*, 175–188.

British National Corpus. (2000). *World edition [CD-ROM]*. Oxford, UK: Humanities Computing Unit of Oxford University.

Bybee, J. (1985). *Morphology: Study of relation between meaning and form*. Amsterdam: Benjamins.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

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

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

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior*, *10*, 722–729.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford, UK: Oxford University Press.

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

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). Cambridge: Cambridge University Press.

Ellis, N. C. (2011). Frequency-based accounts of SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 193–210). London: Routledge.

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. Th. Gries & D. S. Divjak (Eds.), *Frequency effects in language learning and processing* (Vol. *1*, pp. 7–34). Berlin, Germany: Mouton de Gruyter.

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, *20*, 29–62.

Forster, K. (1976). 'Accessing the mental lexicon' in R. Wales & E. Walter (eds): *New Approaches to Language Mechanisms* (pp. 257–284). North Holland Publishing.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75–93). Harlow, UK: Longman.

Frey, E. (1981). Subjective word frequency estimates and their stylistic relevance in literature. *Poetics*, *10*, 395–407.

Gardner, D. (2007). Validating the construct of *word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*, 241–265.

Goldberg, A. (1995). *Constructions*. Chicago: University of Chicago Press.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403–437.

Hoey, M. (2000). A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocation* (pp. 224–243). Hove, UK: Language Teaching Publications.

Hoffman, S., & Lehmann, H.-M. (2000). Collocational evidence from the British National Corpus. In J. M. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora* (pp. 17–32). Amsterdam: Rodopi.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology* (pp. 161–186). Oxford, UK: Clarendon Press.

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

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linded, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Hillside, NJ: Erlbaum.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*,29, 459–484.

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

Langacker, R. (1987). *Foundations of cognitive grammar* (Vol. *1*). Stanford, CA: Stanford University Press.

Laufer, B. (1989). A factor of difficulty in vocabulary learning: Deceptive transparency. *AILA Review*, *6*, 10–20.

Laufer, B. (1990). Why are some words more difficult than others? Some intralexical factors that affect the learning of words. *International Review of Applied Linguistics*, *28*, 293–307.

Laufer, B. (1997a). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In M. McCarthy & N. Schmitt (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 140–155). Cambridge, UK: Cambridge University Press.

Laufer, B. (1997b). The lexical plight in second language reading: Words you don't know, words you think you know and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge, UK: Cambridge University Press.

Laurent, J., Denhières, G., Passerieux, C., Iakimovac, G., & Hardy-Baylé, M. (2006). On understanding idiomatic language. *Brain Research*, *1068*, 151–160.

Manning, C. (2007). *Generalized linear mixed models*. Course handout from Stanford University. Retrieved from http://nlp.stanford.edu/˜manning/courses/ling289/GLMM.pdf

McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *RELC Journal*, *38*, 53–66.

McGee, I. (2006). *Lexical intuitions and collocation patterns in corpora*. Unpublished Ph.D. dissertation, Cardiff University, UK.

McGee, I. (2008). Word frequency estimates revisited—A response to Alderson (2007). *Applied Linguistics*, *29*, 509–514.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences and the whys and wherefores. *Corpus Linguistics and Linguistic Theory*, *5*, 79–103.

Monsell, S., Doyle, M., & Haggard, P. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*, 43–71.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder.

Perugia Corpus. 2013. Università per Stranieri di Perugia. Retrieved from http://perugiacorpus.unistrapg.it

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Pollio, H., Barlow, J., Fine, H., & Pollio, M. (1977). *Psychology and the poetics of growth: Figurative language in psychology, psychotherapy, and education*. Hillsdale, NJ: Erlbaum.

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

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC Journal*, *5*, 69–84.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*, 77–89.

Ringeling, T. (1984). Subjective estimates as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin*, *8*, 59–69.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing*, *16*, 189–216.

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*, 389–411.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.

Shapiro, B. J. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior*, *8*, 248–251.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Siyanova, A. (2010). *On-line processing of multi-word sequences in a first and second language: Evidence from eye-tracking and ERP*. Unpublished Ph.D. thesis, the University of Nottingham.

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *Mental Lexicon*, *8*, 245–268.

Siyanova-Chanturia, A. (2015). On the "holistic" nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, doi: 10.1515/cllt-2014-0016.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). 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. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multi-word sequences. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *37*, 776–784.

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*, doi: 10.1093/applin/amt054.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*, 429–458.

Sorhus, H. (1977). To hear ourselves—Implications for teaching English as a second language. *ELT Journal*, *31*, 211–221.

Spina, S. (2010). The dictionary of italian collocations: Design and integration in an online learning environment. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis et al. (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 3202–3208). Paris: European Language Resources Association.

Spina, S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In R. Basili, A. Lenci, & B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014* (pp. 354–359). Pisa: Pisa University Press.

Strandburg, R., Marsh, J., Brown, W., Asarnow, R., Guthrie, D., & Higa, J. (1993). Event-related potentials in high-functioning adult autistics. *Neuropsychologia*, *31*, 413–434.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*, 23–55.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford, UK: Blackwell.

Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, *22*, 149–172.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*,*7*, 215–244.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teacher's College, Columbia University.

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

Tomasello, M. (2006). Acquiring linguistic constructions. In W. Damon, R. Lerner, D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology* (6th ed., Vol. 2, pp. 255–298). New York: Wiley.

Tryk, H. E. (1968). Subjective scaling of word frequency. *American Journal of Psychology*, *81*, 170–177.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C . (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, *22*, 1682–1700.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*, 360–363.

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford, UK: Oxford University Press.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Cohen's *K* Test Statistics Examining the Agreement Between Native Speakers' (NS) and Non-Native Speakers' (NNS) Judgments and Corpus Frequency Values for the Four Collocation Frequency Bands.
Appendix S2: Verbatim English Translation of the Instructions Accompanying the Questionnaire.