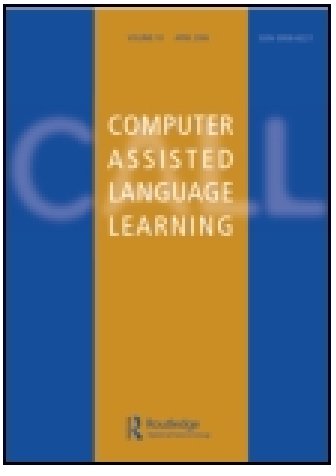

Computer Assisted Language Learning
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ncal20

Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing

Joe Geluso [a]
[a] Department of International Communication , Kanda University of International Studies , Chiba , Japan
Published online: 14 Dec 2011.

To cite this article: Joe Geluso (2013) Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing, Computer Assisted Language Learning, 26:2, 144-157, DOI: 10.1080/09588221.2011.639786

To link to this article: http://dx.doi.org/10.1080/09588221.2011.639786

Computer Assisted Language Learning, 2013
Vol. 26, No. 2, 144–157, http://dx.doi.org/10.1080/09588221.2011.639786

Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing

Joe Geluso*

Department of International Communication, Kanda University of International Studies, Chiba, Japan

Usage-based theories of language learning suggest that native speakers of a language are acutely aware of formulaic language due in large part to frequency effects. Corpora and data-driven learning can offer useful insights into frequent patterns of naturally occurring language to second/foreign language learners who, unlike native speakers, are not privy to a lifetime of input and fine-tuning. Recently, the use of the web in combination with the Google search engine as an accessible corpus and concordancer has received much attention. This article describes an experiment which tests the hypothesis that native speakers of English perceive learner-generated phrases to be more natural after learners have searched the phrases on Google and modified them in light of the frequency of search results. The findings indicate that native speakers perceive phrases that generated more results in Google searches to be more natural.

Keywords: phraseology; formulaic sequences; data-driven learning; Google; frequency; naturalness

Introduction

Patterns can be seen everywhere in our world. Look up and one can see patterns within a flock of birds moving seemingly as one through the air as if their migratory flight was a well-rehearsed performance; from a bird's-eye view above a city, patterns in the flow of traffic as people commute can be seen; and if one were to view our world from even higher up, patterns of weather systems circulating through the atmosphere could be seen. These patterns are emergent and dynamic; they cannot be predicted but can be subsequently explained by chaos/complexity theory (Larsen-Freeman, 1997, 2002; Larsen-Freeman & Cameron, 2008).

Language is no different. There are synchronic and diachronic patterns at all levels (e.g. phonetic, morphological, lexical, syntactic, semantic, pragmatic, and discourse) (Ellis, 2008a). As Ellis observes, these patterns are not preordained by a higher power, such as a deity, human policy, or genes. Rather, what brings about the emergence of linguistic patterns in language learners, first or second language (L1/L2), can be explained by the various usage-based theories of language learning within the realm of cognitive linguistics that take frequency effects into account.

*Email: geluso.joe@gmail.com

These include, but are not limited to, constructionist theories, connectionism, emergentism, and applied linguistics influenced by chaos/complexity theory (see Broeder & Plinkett, 1994; Bybee, 2008; Ellis, 2003; Goldberg, 2006; Larsen-Freeman & Cameron, 2008; Robinson & Ellis, 2008). In essence, language learners repeat and manipulate the words and patterns they hear, and the more a learner hears and reproduces a word or pattern of words, the more likely it is the word or pattern will be remembered (Ebbinghaus, 1885).

Linguistic patterns are abundant in both written and spoken language, and for the language learner pursuing insights into the frequency of linguistic patterns in writing, a corpus can be a useful tool. As will be discussed below, few corpora can match the size and accessibility of the web empowered by a search engine such as Google. A language learner with access to the Internet can quickly check the frequency of occurrence of any given phrase on the web by performing a simple search of the phrase in double quotation marks. To illustrate with an example, consider the following two phrases: (1) *the hamburger has a good taste* and (2) *the hamburger tastes good*. In September 2011, a search of both phrases enclosed in double quotation marks on Google generated four results for the first phrase and 231 for the second. Just as an individual learning English as an L1 is likely to have experienced the second phrase at a much higher frequency, the Google search reveals that the second phrase occurs at a higher frequency on the web as well. The experiment described in this article was designed to test the hypothesis that native speakers of English perceive learner-generated phrases to be more ''natural'' after the phrases have been searched on Google and modified in light of the search results.

This article begins by establishing a theoretical framework to help define the notion of ''natural language''. From there, major concepts and previous work in corpus-based learning and the web as corpus and language-learning tool are reviewed. The focus then shifts to the methodology and analysis used in addressing the hypothesis. Finally, the article concludes with a discussion of the findings, limitations of the study, and possible paths for future research.

Phraseology

Lexis and syntax, or vocabulary and grammar, have traditionally been viewed as discrete aspects of language in teaching (Hoey, 2005; Romer, 2009), but a growing number of scholars from a variety of theoretical camps within applied linguistics and second language acquisition (SLA) argue that the two are in fact inseparable (e.g. cognitive linguists, constructionists, and corpus linguists). Ellis (2008b), for example, maintains that children learning their L1 ''learn words from phrases as much as phrases from words'' (p. 5). Hoey's (2005) theory of *Lexical priming* offers a compelling argument to support the notion that ''lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure'' (p. 1). To put it another way, grammar emerges from the collocational fields of discrete vocabulary, or the patterns around individual words. This view of vocabulary and grammar as inseparable is reflected in the relatively recent sub-field of applied linguistics known as *phraseology* (Granger & Meunier, 2008). Granger and Paqout (2008) define phraseology as ''the study of word combinations'' and define a phraseological unit as being made up of ''at least two words'' (p. 32).

One area that is a source of some confusion in the field of phraseology, and should thus be noted here, is the abundant terminology for multi-word units and

their corresponding definitions. *Formulaic sequences (FSs)/language* (Schmitt, 2004; Wray, 2002) seem to have emerged as cover terms encompassing various patterns of words, including collocations (Liu, 2010), lexical bundles (Biber, Conrad, & Cortes, 2004), and idioms. (See Wray (2002) for a more exhaustive list of these terms.) Wray (2002) defines a *FS* as:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (p. 9)

Meanwhile, Sinclair (2008), a pioneer in the field of Corpus Linguistics and phraseology, offers perhaps the safest suggestion in terms of terminology, arguing that the simple pre-theoretical term "phrase" is best for describing "a string of words whose status is not determined" (p. 407). For the purposes of this article, multi-word units will be referred to as *FSs/language* and *phrases* interchangeably.

Regardless of how phrases are labeled, it is now widely recognized that natural language makes considerable use of them (Ellis, 2008b). Sinclair (1991) notes the conspicuous role of formulaic language in his principle of idiom which states that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110). Estimates about the percentage of formulaic language that English speakers typically access and use range as high as 80% (Altenberg, 1998). Erman and Warren (2000) offer a slightly more moderate estimate, calculating that FSs constitute 52.3% and 58.6% of written and spoken discourse, respectively. Needless to say, the employment of formulaic language in written and spoken discourse constitutes a substantial portion of natural discourse in the English language.

Grammaticalness, naturalness, and frequency

Given the fact that speakers of a language make such frequent use of prepackaged formulaic language, Pawley and Syder's (1983, p. 193) famous observation that "native speakers do not exercise the creative potential of syntactic rules to anything like their full extent" should not come as much of a surprise. Without question, there is an innumerable amount of grammatically correct phrases one could utter that would sound completely unnatural. Instead, as Pawley and Syder (1983) so aptly point out, language users tend to use utterances they have heard and used themselves before. To clarify this idea, let us return to our previous example phrases: (1) *the hamburger has a good taste* and (2) *the hamburger tastes good*. Both are *grammatical*, but the latter is more *natural* than the former. A learner of English as a second/foreign language (ESL/EFL), even one who is familiar with English grammar and has a large vocabulary, is likely to produce either utterance. However, a learner of English as an L1 in the United States (or another country where English is commonly spoken) is more likely to produce the more natural sounding second option, as that is the construction he or she will hear most frequently.

Ellis (2002) asserts that, "'rules' of language ... are structural regularities that emerge from learners' lifetime analysis of the distributional characteristics of the language input" (p. 144). This being the case, it can sometimes be difficult for

language teachers to explain why a phrase appears as it does to the inquisitive learner. As Wray (2000) maintains, there are times when there is no apparent explanation other than "we just say it that way" (p. 463). ESL/EFL learners simply lack the necessary input to induce the "rules" that a native speaker of a language learns implicitly. Often it is formulaic language that eludes even the advanced ESL/EFL learner in the quest to dominate the language. This is where corpora can be of value. Gries (2008) asserts:

> This fact alone – that Corpus Linguistics is basically all about frequency – would already provide for a strong affinity of Cognitive Linguistics and Corpus Linguistics: Corpus Linguistics provides exactly the kind of data that are at the heart of Cognitive Linguistics. (p. 412)

Data-driven learning

Data-driven learning (DDL), or corpus-based learning, is a natural pedagogical extension of usage-based theories of language learning which maintain that the majority of L1 acquisition happens implicitly by inductive means over the course of a learner's lifetime. Because learners of a second or foreign language usually do not have similar contact time and experience with the target language to fall back on, DDL can be a useful means of study. Hunston (2002) explains that the theory behind DDL is that by studying concordance lines, or snippets of authentic language "students act as 'language detectives' discovering facts about the language they are learning for themselves" (p. 170). DDL facilitates the inductive discovery of patterns in the target language, an approach that more accurately mimics naturalistic L1 acquisition. Hunston goes on to contend that DDL supports learning because "students are motivated to remember what they have worked to find out" (p. 170).

Recently, there have been numerous studies that have examined the incorporation of corpora and DDL in foreign language curricula (see Chambers, 2005; Kennedy & Miceli, 2010; Sun & Wang, 2003; Yoon, 2008; Yoon & Hirvela, 2004). These studies have reported largely positive findings in terms of learners' perceptions of the utility of DDL. Participants in Chambers' (2005) study, for example, believe that hunting for grammatical patterns in concordance lines better facilitates memorization of problematic aspects of the target language than being "spoonfed" the rules (p. 120). Participants in Yoon's (2008) study reported becoming more cognizant of the interface between lexis and syntax as a result of increased corpus use. As one participant related, "Basically, what we learned as grammar [in traditional EFL/ESL classes] is all related to collocation" (p. 41).

Sun and Wang (2003) provide statistical evidence that supports the pattern-hunting approach of DDL. Their study compared how inductive and deductive teaching approaches affected student uptake of collocations. One group representing each approach took a pre- and posttest in which they had to correct sentences for grammar errors that featured keywords. The treatment differed in that the inductive group was charged with searching for instances of the keywords on a web-based concordancer and noting the patterns around the word, while the deductive group was simply taught the grammar rules necessary to correct the sentences. The inductive group performed significantly better on the posttest, lending credence to the effectiveness of pattern hunting through DDL.

While there are numerous advantages to DDL and corpus-based learning, there are a number of drawbacks that also need to be noted. Participants in Chambers' (2005) study, for example, expressed concern about the validity of their findings in the corpus "and remained faithful to the grammar book as the ultimate authority" (p. 120). Chambers and O'Sullivan (2004) and Yoon and Hirvela (2004) also noted that participants in their respective studies felt that DDL was too time consuming and that corpora were not always readily available. In addition, there is a definite learning curve to using corpora, which is why scholars in the field urge sufficient training for students before consulting a corpus (Chambers, 2005; Chambers & O'Sullivan, 2004; Yoon & Hirvela, 2004).

Google as corpus and concordancer

Interest in the web as a corpus and concordancer (i.e. the web accessed by a search engine such as Google) has reached new heights recently. This is evidenced in the surge of publications devoted to the examination, innovation, and evaluation of Google as a corpus and concordancer (Chinnery, 2008; Killgrarrif & Grefenstette, 2003; Robb, 2003; Sha, 2010; Shei, 2008; Wu, Franken, & Witten, 2009). Before continuing, though, it would be wise to first address the question as to whether the web can be considered a "corpus" at all. Hunston (2002) defines a corpus at its most basic level as "a collection of naturally occurring examples of a language" (p. 2). Sha (2010) points out that while corpus linguists may not agree unanimously on the definition of a corpus, they do "agree that the dispensable component of a corpus is a collection of machine-readable texts" (p. 377). The web satisfies these two basic requirements, and if further convincing is needed, Kilgrarrif and Grefenstette (2003) provide a cogent argument on the matter, ultimately concluding that the web can indeed be classified as a corpus.

As discussed above, from the perspective of cognitive linguistics, much of what is perceived as *natural* language has to do with phraseology as informed by frequency effects. The web, simply due to its sheer size, is well prepared to offer insights on frequency of occurrence of FSs. As Shei (2008) expounds, "Google can offer solutions to many of the research questions in phraseology which even a billion-word corpus can hardly handle" (p. 70). Indeed, the Google search engine is used with such prolificacy to search for information that the word *google* now appears in the dictionary as a verb (Jewell & Abate, 2010). Variant spellings of words can be "googled" to see which version generates the most results, which can then be assumed to be the "correct" option. Competing phrases enclosed within double quotation marks can also be googled for insights into grammaticality and naturalness.

The use of Google as corpus and concordancer is not without its naysayers, however. Wu et al. (2009) note that because the web is not vetted, search results have the potential of being inaccurate or "dirty". The publicly available *Corpus of Contemporary American English* (COCA) (Davies, 2008–) observes that while Google is a great search engine, there are many things that a corpus specifically designed by corpus linguists can do which Google cannot. For example, COCA can measure changes in phrases over time, provide lists of frequent collocates of keywords, and perform other criteria-based searches, while Google can only search for specific words and, when enclosed in double quotation marks, phrases.[1]

Nevertheless, there are numerous advantages in using Google as a corpus. In addition to its size, the dynamic nature of the web is an often-noted advantage (Shei, 2008; Wu et al., 2009). Like a language, the web is "alive" so to speak and constantly changing. Sha (2010) notes that dynamic corpora, as opposed to static corpora like the British National Corpus (BNC), are constantly growing in size.[2] The number of results, or "hits", Google retrieves from a search of the web is growing at an incredible rate. This growth is depicted in Figure 1, which builds on search results from the summers of 2007 and 2009 that were reported by Sha (2010). We can see that the occurrences of the phrase *is kind of* retrieved by Google grew from 3.31 million to 123 million between the summers of 2007 and 2011, a factor greater than 37. Meanwhile, the phrases *place to go* and *place to go to* grew by factors of 29 and 4, respectively, over the same approximately four-year span. Another advantage of Google cited by Sha (2010) was the speed in which results are generated. Forty separate searches of distinct words and phrases in Google and the BNC revealed that Google took an average of 0.3 seconds to complete a search while the BNC took "several" seconds (Sha, 2010, p. 385).

The idea of using Google as corpus and concordancer is far from novel. Shei (2008), for example, offers an intriguing method of investigating the formulaicity of FSs through Google. Shei (2008) suggests using the base-two logarithm ($\log_2$) of the raw frequency Google searches generate to guide one's choice, as it produces more comprehensible figures that still accurately reflect frequency of occurrence on the web. An example from Shei's (2008) article illuminates the process. The phrases *have been found to be contaminated with* and *have been found to be polluted with* are contrasted by comparing the $\log_2$ of Google search results of the first word of the phrase followed by each consecutively lengthened sequence. Obviously, the longer the search query the fewer results that will be generated. So, searching *have* yields more results than *have been, have been* yields more results than the subsequent *have been found*, and so on. When one reaches the fork in the phrase, so to speak, at *polluted* or *contaminated*, the option that generates the number of results closer to the previous search can be said to be the more formulaic option. The phrase *have been*

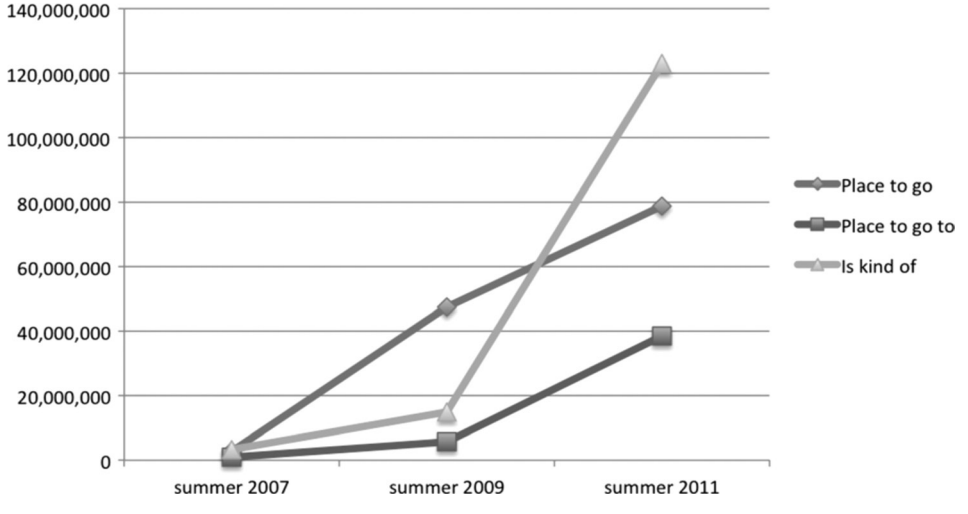

Figure 1. Google growth between summer 2007 and summer 2011.

found to be generated a base-two logarithm of 21, adding *contaminated* to the sequence drops it to 14, while *polluted* lowers it to 10. Therefore, one can assume that *have been found to be contaminated with* is more formulaic than *have been found to be polluted with*. This method is innovative and indubitably more informative than simply performing a double quotation mark search of competing phrases, but it escapes the accessibility and simplicity of the latter. And, this accessibility is precisely what make double quotation mark searches so attractive to ESL/EFL learners looking for guidance in their selection of FSs.

While there are studies that investigate the use of the web and Google as corpus and concordancer (Robb, 2003; Shei, 2008), Google as a ''pedagogical tool'' (Chinnery, 2008), static corpora derived from Google searches of the web (Guo & Zhang, 2007; Wu et al., 2009), and comparisons of Google to more traditional corpora (Sha, 2010), there are very few, if any, studies that provide a statistical analysis of teachers' perceptions of learner-generated phrases that have been informed by the process of conducting simple double quotation mark Google searches, or ''Google-informed'' phrases. The aim of this study is to fill that gap.

Method

To answer the research question of whether or not native speakers of English perceive learners' Google-informed phrases as more natural than non-Google-informed phrases (i.e. the learner's initial version of the phrase, before consulting Google), a mixed sample of 334 phrases was collected and given to native speakers of English to rate for naturalness. To look at a likely example of how a Google-informed phrase was arrived at in this study, suppose a learner is vacillating between the phrases *he was riding on the car* and *he was riding in the car*, unsure about which preposition to use. Obviously, the context makes a difference in terms of which phrase the learner actually wants to use, but a quick Google search can provide some general insights. The learner searches *he was riding on the car* in Google, which generated about six hits in September 2011. The learner then makes a slight modification to the phrase and searches *he was riding in the car*, which generated approximately 291,000 hits. Running on the assumption that frequency corresponds with naturalness, the results suggest that the latter phrase, or the Google-informed phrase, is the more natural phrase, and the learner thus includes it in his or her writing.

Data

Data for this study were collected from 25 students enrolled in a freshman-writing course in the Department of International Communication over one semester (April–July 2010) at a foreign language university in Japan. The average age of the students at the time was 18, and the class consisted of 5 males and 20 females. The students' Test of English for International Communication (TOEIC) scores ranged from 295 to 805, with an average score of 496. The course met twice a week for 90-minute sessions for 14 weeks. Students used a writing textbook from a major publishing company that focused on the development of paragraph writing in the first semester. Starting in the third week, students were required to submit nine written paragraphs over the remainder of the semester. It is from these paragraphs that the data were collated.

Approximately 40 minutes of class time spread over three separate class meetings in the first few weeks was spent training students to conduct Google searches of FSs

in double quotation marks in order to investigate frequency of occurrence on the web. Students were instructed to select ''questionable phrases'', or phrases they suspected might be ungrammatical or unnatural, from their writing to search in Google. In order to illustrate the process, which was dubbed ''Google-drafting'', for the students, the instructor handed out a worksheet outlining instructions and an example. The first part of the worksheet explained each step of the Google-drafting process in a bullet point list:

- Go to www.google.com (make sure you are on the English site)
- Enter the phrase in question with ''quotation marks'' around it
- Record the number of results
- Enter a slightly different version of the phrase that you think may be better (don't forget the quotation marks)
- Record the number of results
- Repeat the process 3–5 times, or until you feel you have found the best phrase
- Use the phrase that generated the most results

The second half of the worksheet featured an example of how one might improve a phrase or sentence by Google-drafting. The example featured a sentence which contained a ''questionable'' phrase in bold. Beneath the sentence was a table with different variations of the questionable phrase on the left-hand side, and the number of results each variation generated in Google searches on the right-hand side. Following the Google-drafting process, the instructor advised that the variation of the phrase that generated the most hits be selected and included in the writing assignment. Google-drafting was subsequently made a mandatory component of the students' writing process for the remainder of the course.

After receiving the first set of paragraphs with Google-informed phrases, it was noticed that some students were reporting astronomical numbers for rather long phrases. For example, one student reported 19.3 million results for the phrase *how long have you been learning English?* The researcher searched the same phrase and received 220,000 results. The conclusion was drawn that students needed to be consistently reminded to use double quotation marks around their search queries. From that point on, instances of phrases in which students reported suspiciously high results were searched again to verify whether or not students were using double quotation marks. In instances where search results differed by millions, the phrases were disregarded from the study.

Additionally, phrases that students searched in Google but did not change (perhaps because search results suggested that the student's initial phrase was correct) were also disregarded. The last issue was length of the FSs. Chen and Baker (2010) note that four-word sequences are the most researched length of FSs in writing studies as they are within a manageable size. For this study, the researcher decided to disregard any FS longer than seven units. Ultimately, pre- and post-versions of 167 phrases were collected from the paragraphs, for a total of 334 phrases.

Raters

Four English language lecturers were recruited from the university where the study took place – all were native English speakers from the United States. Raters were

compensated with a small stipend for time spent in a norming session and rating the 167 items for naturalness.

Procedure

The four raters participated in a norming session in which the interface between naturalness and grammaticality was discussed and defined. As mentioned above, ''rules'' of a language are actually structural regularities that emerge from learners' lifetime analysis of the language input. Therefore, ''the knowledge of a speaker/hearer cannot be understood as a grammar, but rather as a statistical ensemble of language experiences that changes slightly every time a new utterance is processed'' (Ellis, 2003, pp. 63–64). It can thus be concluded that phrases perceived as *natural* by a native speaker may not necessarily be *grammatical*. Ellis (2002) exemplifies this idea with the following three sentences:

(a) Tom is one of those clumsy people who cuts himself shaving.
(b) Tom is one of those clumsy people who cuts themselves shaving.
(c) Tom is one of those clumsy people who cut themselves shaving. (p. 161)

While only one of these examples is technically grammatical, Ellis (2002) notes that, ''the combined cue strengths do not fall clearly one way or another'' (p. 161). The fact that none of these phrases is apt to be noticed and labeled ''unnatural'' by a native speaker of English can be explained by frequency effects, which in turn can be accounted for by the various camps in the field of SLA that subscribe to usage-based theories of language learning. The scale that raters were given to guide them when rating the phrases is shown in Table 1. A number of sample sentences were rated (phrases that had been disregarded from the study) by the raters and researcher in the norming session and subsequently discussed.

Google-informed and non-Google-informed versions of the 167 learner-generated FSs (total of 334) were randomly divided and entered into two separate Excel files to avoid raters from being influenced by similar versions of a phrase. This way, the raters were rating 167 FSs based solely on their own merits, not in comparison with another version of the phrase. As can be seen in Figure 2, phrases to be judged were underlined and embedded in context deemed to be necessary to understand the intended meaning. The adjacent column featured a numerical rating scale in a drop-down menu. The four participants emailed the completed Excel sheets to the researcher.

Table 1. Scale for rating phrases.

Rating	Description
5	Perfectly natural
4	Somewhat unnatural but still perfectly intelligible. A native speaker might even produce something similar
3	Unnatural, but intelligible
2	Unnatural, but you are fairly certain that you can guess what the speaker is trying to say
1	So unnatural that it is unintelligible

Results and discussion

To answer the research question, whether or not native speakers perceive learners' phrases to be more natural after searching them on Google, a one-way analysis of variance (ANOVA) was performed on each participant's ratings. As can be seen in Tables 2 and 3, the hypothesis was confirmed: mean scores from each rater were higher for Google-informed phrases than for non-Google-informed phrases. Scores were also found to be statistically significant for each rater at $p < .01$.

As mentioned above and outlined in Table 2, a total of 334 phrases were rated: participants A and B rated 167 randomly selected pre- or post-versions of the phrases (85 non-Google-informed phrases and 82 Google-informed phrases); participants C and D then rated the 167 counterparts, 82 non-Google-informed phrases and 85 Google-informed phrases. The mean scores for Google-informed phrases assigned by raters A, B, C, and D were 1.22, 0.91, 0.57, and 0.63, higher than the mean score awarded to non-Google-informed phrases, respectively. On average,

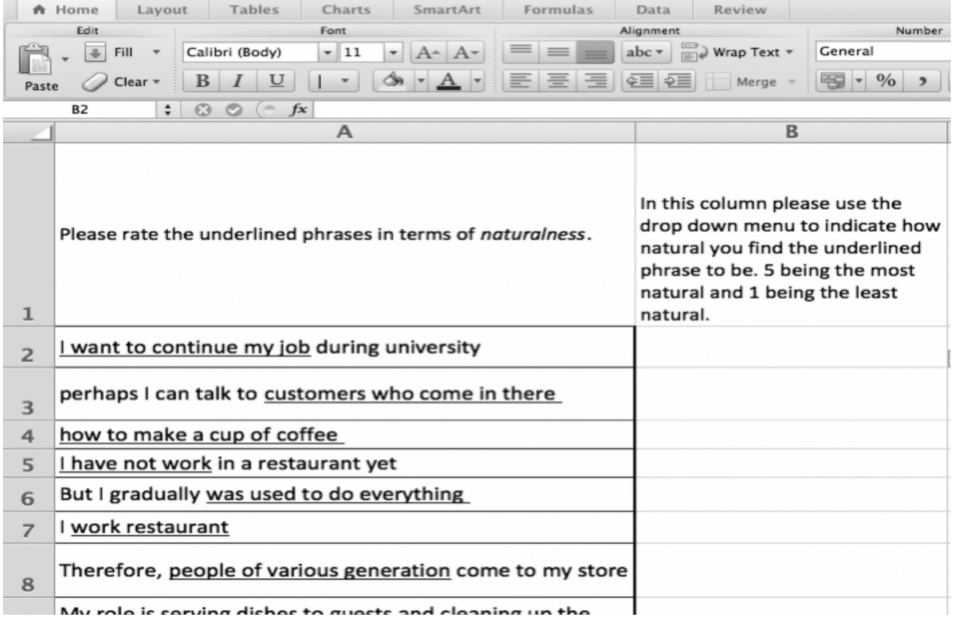

Figure 2. Excel file with phrases to be rated for naturalness.

Table 2. Descriptive statistics.

	Google-informed phrases						
	Pre			Post			
Raters	*k*	*M*	SD	*k*	*M*	SD	Post *M* − pre *M*
A	85	3.05	0.99	82	4.27	0.94	1.22
B	85	3.59	0.93	82	4.50	0.81	0.91
C	82	3.24	0.98	85	3.81	1.17	0.57
D	82	3.39	1.18	85	4.02	1.09	0.63

Table 3. One-way ANOVA.

Raters	*k*	df	*F*	*p*	Effect size
A	167	1	65.88	.000*	.285
B	167	1	45.77	.000*	.217
C	167	1	11.56	.001*	.065
D	167	1	12.94	.000*	.073

Notes: *$p < .01$; effect size = Eta squared.

Google-informed phrases scored 0.83 points higher than the non-Google-informed phrases.

An ANOVA assumes equal variance, but upon closer inspection of Table 2, it might be noticed that the difference in standard deviation of the pre- and post-ratings of raters B and C is greater than 0.10, at 0.12 and 0.20, respectively. In fact, rater B just passed Levene's Test of Equality of Error Variance with a p value of exactly .050, and rater C failed it entirely at .000. However, the assumption of equal or unequal variance makes no meaningful difference in this study. For the sake of precaution, though, all raters' F values were recalculated *without* the assumption of equal variance and all still proved significant at $p < .01$.

The results indicate that there is a relationship between frequency of occurrence on the web, as revealed by Google searches, and perceived naturalness among native speakers of English. This "perceived naturalness" is likely related to processing advantages that have been associated with using FSs in speech and writing (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Conklin & Schmitt, 2008; Jiang & Nekrsova, 2007; Millar, 2011). Millar's (2011) study, in particular, illustrates that "learner deviation from target language formulaicity places an increased processing burden on native speaker addressees" (p. 142). It is not unreasonable to assume that phrases with a higher frequency of occurrence on the web are more formulaic and in turn easier to process by native speakers, leading them to be perceived as more natural.

On a pedagogical level, the results suggest that training learners to use double quotation mark searches on Google is a worthwhile endeavor. The training process is relatively easy and time efficient when compared to training learners to use the more intricate features of traditional corpora, no matter how "user-friendly" they may be. As was mentioned in the "Method" section, training a group of 25 students in Google-drafting, including practice time, was accomplished in about 40 minutes. The process results in raising learner awareness of potentially problematic and unnatural language and gives them a means to further interact with and hopefully improve their writing.

Conclusion

A native speaker of a language accesses an implicitly constructed "grammar" based on the lifetime distributional frequency of input when making online judgments at the lexical and syntactic level, and as such is acutely aware of formulaic language. While ESL/EFL learners do not have similar linguistic experience with the target language to fall back on, they can mimic the process to an extent by using corpora to test hypotheses about the target language. One of the main obstacles to DDL and

corpus use, though, is the training and specialized knowledge needed to use most corpora. This is exactly where the use of Google and the web is advantageous: accessibility and ease of use. Through simple searches, learners are able to quickly and accurately investigate the frequency of occurrence of FSs on the web, which can provide insights into the formulaicity and naturalness of phrases.

However, there are some limitations with this study that should be addressed. First, while there were 167 items judged for their naturalness, there were only four raters. Therefore, this study can be seen as a pilot study and should be replicated at a larger scale, i.e. with more raters. Second, perceptions of what constitutes "natural English" will vary from individual to individual, and most certainly from country to country. While all raters were native speakers of English, all were from the United States. Future studies may aim to have a more diverse selection of English speakers share their intuitions as to what typifies "natural English". Along these lines, it may be interesting to investigate whether or not using Google to corroborate one's suspicions of formulaicity leads to more "American English" as opposed to British, Australian, or another variety of English.

Despite the limitations noted above, the results reported here are powerful and suggest that by using the web as a corpus and Google as a concordancer, students can improve the naturalness of their writing. While the web and Google are not designed to be corpus and concordancer, respectively, they can be defined as such given their characteristics and functionality. Training students to perform double quotation mark searches on Google is a relatively simple matter, and considering the ubiquitous nature of the search engine, many students may already be engaging in such behavior. This study lends evidence to the utility of such searches and showed that native speakers did perceive Google-informed phrases to be more natural than non-Google-informed phrases at a level of statistical significance.

Acknowledgements

I would like to thank my many colleagues for their constant support and insightful feedback, and the anonymous reviewers for their helpful comments on earlier versions of this article.

Notes

1. It is the case now that one can track the occurrence of words and phrases through time on Google books' "Ngram Viewer" (http://books.google.com/ngrams).
2. Sha (2010) provides an informative deconstruction of how Google works as a search engine and an interesting comparison between Google as a corpus and the BNC.

Notes on contributor

Joe Geluso is currently a lecturer of English in the Department of International Communication at Kanda University of International Studies in Chiba, Japan. His research interests include corpus studies, phraseology and usage-based theories of language learning.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101–122). Oxford: Oxford University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at … Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*, 245–261.

Broeder, P., & Plinkett, K. (1994). Connectionism and second language acquisition. In N.C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 421–453). London: Academic Press.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 216–236). New York and London: Routledge.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning and Technology*, *9*(2), 111–125.

Chambers, A., & O'Sullivan, I. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, *16*, 158–172.

Chen, Y.H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*, 30–49.

Chinnery, G.M. (2008). You've got some GALL: Google-assisted language learning. *Language Learning and Technology*, *12*(1), 3–11.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*, 72–89.

Davies, M. (2008–). *The Corpus of Contemporary American English: 425 million words, 1990-present*. Retrieved September 25, 2011, from http://corpus.byu.edu/coca/

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.). New York, NY: Teachers College, Columbia.

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

Ellis, N.C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M.H. Long (Eds.), *Handbook of second language acquisition* (pp. 63–103). Oxford: Blackwell.

Ellis, N.C. (2008a). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal*, *92*, 232–249.

Ellis, N.C. (2008b). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 1–13). Amsterdam: John Benjamins.

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

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

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdiscplinary perspective*. Amsterdam: John Benjamins.

Granger, S., & Paqout, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Gries, S.T. (2008). Corpus-based methods in analyses of second language acquisition data. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 406–431). New York and London: Routledge.

Guo, S., & Zhang, G. (2007). Building a customised Google-based collocation collection to enhance language learning. *British Journal of Educational Technology*, *38*, 747–750.

Hoey, M. (2005). *Lexical priming*. New York, NY: Routledge.

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

Jewell, E.J., & Abate, F.R. (Eds.). (2010). *New Oxford American Dictionary* (3rd ed.). Oxford: Oxford University Press.

Jiang, N., & Nekrsova, T.M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*, 433–445.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28–44.

Killgrarrif, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Association for Computer Linguistics*, *29*, 333–347.
Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics*, *18*, 141–165.
Larsen-Freeman, D. (2002). Language acquisition and language use from a chaos/complexity theory perspective. In C. Kramsch (Ed.), *Language acquisition and language socialization* (pp. 33–46). London: Continuum.
Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.
Liu, D. (2010). Going beyond patterns: Involving cognitive analysis in the learning of collocations. *TESOL Quarterly*, *44*, 4–30.
Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*, 129–148.
Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.
Robb, T. (2003). Google as a quick 'n dirty corpus tool. *TESL-EJ, Teaching English as a Second or Foreign Language*, *7*. Retrieved from http://tesl-ej.org/ej26/int.html
Robinson, P., & Ellis, N.C. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.
Romer, U. (2009). The inseparability of lexis and grammar. *Annual Review of Cognitive Linguistics*, *7*, 141–163.
Schmitt, N. (Ed.). (2004). *Formulaic sequences*. Amsterdam: John Benjamins.
Sha, G. (2010). Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. *Computer Assisted Language Learning*, *23*, 377–393.
Shei, C.C. (2008). Discovering the hidden treasure on the internet: Using Google to uncover the veil of phraseology. *Computer Assisted Language Learning*, *21*, 67–85.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (2008). Envoi. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdiscplinary perspective* (pp. 407–410). Amsterdam: John Benjamins.
Sun, Y., & Wang, L. (2003). Concordances in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16*, 83–94.
Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463–489.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wu, S., Franken, M., & Witten, I.H. (2009). Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning*, *22*, 249–268.
Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology*, *12*(2), 31–48.
Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, *13*, 257–283.