Taming big data: Applying the experimental method to naturalistic data sets

Eyal Sagi¹

Published online: 8 January 2019
© The Psychonomic Society, Inc. 2019

Abstract

Psychological researchers have traditionally focused on lab-based experiments to test their theories and hypotheses. Although the lab provides excellent facilities for controlled testing, some questions are best explored by collecting information that is difficult to obtain in the lab. The vast amounts of data now available to researchers can be a valuable resource in this respect. By incorporating this new realm of data and translating it into traditional laboratory methods, we can expand the reach of the lab into the wilderness of human society. This study demonstrates how the troves of linguistic data generated by humans can be used to test theories about cognition and representation. It also suggests how similar interpretations can be made of other research in cognition. The first case tests a long-standing prediction of Gentner's natural partition hypothesis: that verb meaning is more subject to change due to the textual context in which it appears than is the meaning of nouns. Within a diachronic corpus, verbs and other relational words indeed showed more evidence of semantic change than did concrete nouns. In the second case, corpus statistics were employed to empirically support the existence of phonesthemes—nonmorphemic units of sound that are associated with aspects of meaning. A third study also supported this measure, by demonstrating that it corresponds with performance in a lab experiment. Neither of these questions can be adequately explored without the use of big data in the form of linguistic corpora.

Keywords Corpus statistics · Big data · Semantic change · Representation · Phonesthemes

Traditional lab-based studies provide a great degree of control. This control enables experimental designs that can be used to explore subtle effects. However, that level of control also means that some lab results do not readily replicate in other, less controlled circumstances, most notably in real-world situations. This article supports the proposal that some of the approaches and methods that have proven so useful in the lab can be applied to more naturalistic data sets gathered from external sources, primarily the internet and other collections of big data. By applying such methods to more naturalistic data, researchers can strike a new balance between internal and external validity in their pursuit of furthering our understanding of cognition and behavior.

This study was designed to establish the efficacy of these methods by applying them to investigate two related questions regarding the representation of word meaning—whether verb representations are more relational than those of nouns, and

the relationship between word form and its meaning. Both of these cases involve hypotheses regarding the variability of meaning across words and uses, whether grouped by grammatical category or phonetic similarity. Moreover, the dependent measures in both studies involve the textual context in which the words appear and its variability. As a result, the same overall methodological approach can be applied in both cases, with some modifications.

The first study demonstrates how big data can allow researchers to develop new approaches for testing existing question by examining patterns that unfold over long periods of time. In particular, this study tests the hypothesis that the meaning of verbs changes more quickly than the meaning of nouns. The second study shows how, by replacing participants with text, researchers can test hypotheses that are larger in scope and replace some of the reliance on participants' intuitions and judgment with objective statistical measures. Specifically, the second study explores proposed relationships between phonetic clusters and the meaning of words incorporating them. Both of these studies illustrate how the combination of large corpora and traditional hypothesis testing designs enables researchers to conduct naturalistic studies with external validity and high statistical power. In particular, using

✉ Eyal Sagi
esagi@stfrancis.edu

¹ University of St. Francis, 500 Wilcox St, Joliet, IL 60435, USA

large datasets enables research to approach problems from a different perspective, allowing questions that are difficult, or perhaps even impossible, to explore in the lab to be answered. These difficulties can arise out of the limitations of the lab (Study 1), or because collecting a similar quantity and quality of data from participants is difficult and expensive (Study 2).

The experimental method and the study of cognition

Psychological researchers have customarily focused on lab-based experiments to test their theories and hypotheses. The lab provides many advantages for research in psychology, and especially for investigations of cognition. Primary among these is the important role control plays in experiments. By controlling the environment, researchers can eliminate many possible confounds and other threats to the validity of their conclusions. This results in studies with a high degree of internal validity and provides a dramatic increase in the statistical power available for testing hypotheses at the cost of reduced external validity. Additionally, the degree of control available at the lab means studies are also easier to replicate, although the success of such efforts at replication has recently come under scrutiny (Aarts et al., 2015).

Nevertheless, conducting research in the lab has its disadvantages. In particular, questions often arise regarding the *external* validity of lab-based results. That same level of control and care that researchers exercise in the lab can result in studies whose results *depend* on the particular conditions of the study. Small variations in those conditions, such as the addition of noise or ambiguity in language, might cause the effects observed in the lab to be greatly reduced, or even disappear.

To assuage these concerns, researchers also conduct studies in more natural settings. This can be achieved either by endeavoring to recreate such settings within the lab or by venturing outside of the lab to conduct studies in less controlled environments. The advantage of the former is that it allows the researcher to maintain a high degree of control over the study. Its disadvantage is that it is nearly impossible to faithfully recreate a natural setting within a controlled environment and such settings tend to present a compromise between a fully controlled lab study and a study conducted in a natural setting.

In contrast, although studying behavior in a natural setting seems like an ideal avenue for conducting studies in psychology, it complicates the study designs and limits the possible manipulations an experimenter can employ, as well as the types and precision of the quantitative measurements that can be collected. Therefore, although the inspiration of theories and hypotheses is often found outside of the lab, researchers frequently start their scientific investigation by

conducting rigorous, precise, and controlled lab studies. Once the phenomenon is understood better in such controlled settings, researchers will turn to support the results by providing convincing, if less conclusive, evidence that their theories also predict behavior outside the lab.

Using big data to conduct studies

The vast amounts of data now available to researchers can be a valuable resource. By incorporating this new realm of data and translating it into traditional laboratory methods, we can expand the reach of the lab into the wilderness of human society. This can allow researchers to conduct research that has more external validity than traditional lab studies, while maintaining, or even improving, the available statistical power. The first step toward such translations is the realization that data from outside the lab, although less controlled, can also be analyzed using the same methods employed for analyzing data gathered in the lab.

Studies conducted in the lab are often concerned with the effect of one or more independent variables (IVs) on the outcome as measured by a dependent variable (DV). Commonly, such studies are designed as experiments, in which the IVs are intentionally manipulated by the researcher. The effect of such manipulations on the measured DV are then explored using inferential statistics such as *t* tests and ANOVAs. In most studies, several different manipulations are used, each giving rise to a different experimental condition and differences in the DV due to the condition in which they are measured provide evidence of the effect of the manipulation (hence, the IVs) on the DV. The extensive control that researchers have in a lab setting manifest themselves as a reduction in variability caused by extraneous factors and therefore increases statistical power.

In contrast with lab studies, studies using *big data*, large amounts of data obtained from outside the lab in forms that defy traditional methods of analysis, do not include a direct manipulation of the IVs because no new data is being collected. Nevertheless, in practice, the quantity of data helps offset issues of control and manipulation by providing alternative means of increasing statistical power. Instead of minimizing variance due to random error to increase the likelihood that trends and regularities can be identified, a larger number of samples helps separate them from randomness without sacrificing external validity.

Structured versus unstructured sources of data

When using big data for research purposes, it is important to note that there are two large classes of data—structured and

unstructured. Most lab studies carefully collect structured data, in which each measurement is classified and categorized according to the conditions under which it is collected. The data are therefore annotated and structured on the basis of relevant variables and conditions. Likewise, many existing datasets, for example those that are often used for marketing and business purposes, are structured. Each datum is provided with contextual information that relates it to the dataset in relevant, and often important, ways.

However, plenty of data are also available that are unstructured—that is, data that are provided on their own, with very little relevant contextual information. This lack of relevant contextual information means that the researcher needs to supply a structure in which to contextualize the provided data and facilitate analysis.

Text is perhaps the best known of these unstructured data sets. The context provided by text is often included in the text itself. Nevertheless, even texts are often accompanied by some structural information. For instance, the date the text was written, as well as the identity of its authors, is often available. However, for most uses text is largely unstructured, because it is difficult to convert the provided textual information into quantifiable measurements. This makes textual data difficult to analyze using standard statistical methods.

Quantifying language

Multidimensional spaces and vector arithmetic

One common approach to producing quantified information out of text focuses on the analysis of the contexts in which words appear. These approaches ignore the structures of language and follows the premise that the distribution of words in a text is primarily governed by its content. This premise, succinctly identified by Firth (1957) when he postulated “You shall know a word by the company it keeps,” has proven resilient and useful in many studies. It forms the basis for some of the most frequently employed methods used to quantify textual data, such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), topic models (Griffiths, Steyvers, & Tenenbaum, 2007; Steyvers & Griffiths, 2007), and machine-learning-based approaches such as Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013a), skip-gram (Mikolov, Chen, Corrado, & Dean, 2013b), and GloVe (Pennington, Socher, & Manning, 2014). These approaches extract patterns of word co-occurrence as a proxy to their semantic content.

In all cases, the methods attempt to estimate the similarity of word meanings based on their proximity of appearance within a text. Figure 1 shows a two-dimensional depiction of the space for the words representing some mammals (“dog”

and “cat”) and birds (“dove” and “eagle”), as well as two associated motion verbs (“walking” and “flying”). As the figure illustrates, related terms (e.g., mammals) appear in relative proximity, and distinct terms are separated by space. It is also relatively straightforward to represent phrases and sentences by combining the representations of the words of which they are composed, via methods such as vector addition.

Researchers have combined these techniques with other methods from natural language processing to explore a variety of applications, including answering questions (Mohler & Mihalcea, 2009), summarizing texts (Yeh, Ke, Yang, & Meng, 2005), automatic grading (Foltz, Laham, & Landauer, 1999; Graesser et al., 2000), and translating between languages (Tam, Lane, & Schultz, 2007). More importantly, in the context of psychological research, measurements of word similarity based on these methods also correlate with human performance in related tasks, such as judgments of similarity and semantic priming (Günther, Dudschig, & Kaup, 2016; Landauer & Dumais, 1997). Iliev, Dehghani, and Sagi (2015) have reviewed some of the methods of textual analysis used in psychology and related disciplines.

Conducting studies on quantified language data

A measure of textual similarity is surprisingly useful when it comes to testing psychological theories using texts. It provides a basic quantified measurement of difference that is amenable to statistical analyses and designs that are common in psychology. Even more importantly, the underlying representations used to generate this measure are already quantities, although they involve vector representations rather than scalars. Specifically, we can calculate the central tendency and variability of the vectors representing a group of related texts. The distances between pairs of vectors, whether they be representations of individual texts or central tendencies, are scalars (i.e., single numbers). The similarity measure mentioned above is an example of such a measure of distance. Consequently, we can use these vector representations as a basis for conducting a variety of statistical tests, such as *t* tests, analyses of variance (ANOVAs), and regression models.

This becomes of particular interest for psychological research when we consider that texts are produced by people. As such, texts can be considered as representing the individuals who created them. When comparing texts created by individuals that differ on specific attributes, such as gender, culture, or moral values, we are essentially comparing how these individuals use language. If a theory predicts some differences between individuals based on those attributes, we might be able to predict how the texts would differ, and consequently look for such differences as a test of the theory. Essentially, these attributes could play the role of the IVs in our studies, while textual similarity would be the DV, and the

Fig. 1 Sample two-dimensional representation of the relative positions of the words *eagle*, *dove*, *cat*, *dog*, *walking*, and *flying*. The positions of the words were generated using multidimensional scaling, to reduce a 100-dimensional space based on co-occurrence patterns in the British National

Corpus (BNC Consortium, 2007). The terms cluster by their category (bird, mammal, verb) and are also related by semantic properties (e.g., *flying* is closer to the birds, while *walking* is closer to the mammals)

method itself would follow the same pattern as more common approaches to hypothesis testing in psychology (see also Sagi, 2018).

It is relatively straightforward to consider how such studies can be used in conjunction with lab-based studies. For instance, after observing an effect in the lab, we might be able to test for a similar effect on texts collected from online sources. This can provide researchers with an accessible approach for examining the external validity of their results.

However, there are also cases in which it is better to test a hypothesis using collected texts first. This often occurs when bringing the participants of interest to the lab is particularly difficult. For instance, we might predict that the outlawing of slavery following the Civil War in the United States changed the reasoning that individuals apply toward issues such as freedom and racial differences. However, it is difficult to test this theory, because all the individuals currently alive were born after those changes took place. Nevertheless, various textual artifacts have been left over from the Civil War period, such as books, letters, and journals. If we have theory-based predictions regarding how such individuals would consider slavery, we can collect textual evidence generated by individuals prior to the Civil War as well as similar evidence generated after the Civil War, and compare how individuals in these different periods treated slavery. One theory that can provide us with such predictions is Haidt and Joseph's (2004) moral foundations theory, and Sagi and Dehghani (2014) have described how it can easily be applied to comparing the style of moral reasoning about particular concepts in texts.

Moreover, in some cases psychological theories make predictions that are difficult to test because they require the analysis of trends that are difficult to generate in the lab. Below, a number of texts are used to analyze two such cases related to the representation of meaning and its link to the variability in how words are used: The first case tests a long-standing prediction of Gentner's (1982) natural partition hypothesis: that verb meaning is more subject to change due to the textual context in which it appears than is the meaning of nouns (e.g., Gentner & France, 1988; Gillette, Gleitman, Gleitman, & Lederer, 1999). Drawing on theories of semantic change, such variability should lead to a higher rate of semantic change for verbs than for concrete nouns. Although testing the context specificity of verbs and nouns can be reasonably achieved via lab experiments, semantic change takes shape over decades, and consequently is difficult to re-create in the lab. Using a diachronic corpus, we can demonstrate that relational words, such as verbs, show more evidence of semantic change than do concrete nouns.

Two additional studies were conducted, demonstrating that similar language-based analyses can be used to empirically support phonesthemes—nonmorphemic units of sound that are associated with aspects of meaning (e.g., the English prefix *gl-* is associated with the visual modality, as in *glimpse*, *glow*). A large number of phonesthemes have been proposed (see Hutchins, 1998) but are difficult to support empirically. Corpus statistics were employed to gauge the likelihood that each proposed phonestheme is indeed associated with meaning. This concept was also supported by demonstrating that it corresponds with participants' performance in a lab experiment.

Study 1: Semantic change and relational representations

Relations in the representation of word meaning

Dedre Gentner and her colleagues (Asmuth & Gentner, 2005, 2017; Gentner, 1982; Gentner & France, 1988) proposed that, whereas many nouns denote specific entities (e.g., *dog*, *lion*, *man*), the meaning of verbs is inherently relational. For example, the notion of *buying* can only be exemplified using an entity such as a woman, who is performing the action on a different entity, such as a computer. That is, the verb *buy* can only be used in reference to other entities, frequently denoted using nouns. The denoted action can therefore be thought of as identifying a relationship between the entities involved. More generally, verbs denote relations between entities. For instance, Gentner (2006) argues that concrete nouns are easier to learn because they are inherently individuated and more easily separable from the environment. In contrast, the relational nature of verbs makes their meaning more dependent on the context in which they appear. Likewise, Gentner and France proposed that this contextual sensitivity explains why participants in their studies preferred to adjust the meaning of the verbs than those of concrete nouns when paraphrasing sentences such as “The lizard worshipped.” Asmuth and Gentner (2005, 2017) demonstrated that such results can be seen not only when contrasting nouns and verbs, but also when contrasting concrete nouns (such as *lion*) with nouns that denote relational meanings (such as *threat*).

Interestingly, this hypothesis has implications to the structure of language and to our expectations regarding the uses of nouns and verbs more generally. In particular, such adjustments are an essential aspect of metaphors. The hypothesis that verbs are more relational than nouns can therefore be used to predict that it is easier to use verb metaphorically than it is to use concrete nouns. Moreover, linguistic theories on semantic change have long argued that metaphorical uses are one of the primary avenues through which the meaning of words is changed and extended (Traugott & Dasher, 2001).

Consequently, we can hypothesize that a word that is more contextually sensitive should appear in a greater variety of contexts, and, more importantly, change its meaning more over time. However, such changes take place over long periods of time, and are likely to be infrequent. It is therefore difficult to observe and measure such changes in the context of lab studies. In contrast, we have access to a variety of textual sources that were created over long periods of time. Statistical methods can thus be applied in order to test the hypotheses that particular classes of words, such as verbs, vary more across these texts than do words that arguably are less relational, such as concrete nouns. We could then examine the variability of context within a period, as well as how it varies across periods.

Measuring semantic change in a diachronic corpus

Semantic change has traditionally been measured on a word-by-word basis. Researchers identify a word whose meaning they are interested in tracing. They collect the contexts in which it appears over a period of time (often centuries) and record its use in each case. The hypothesis of semantic change can then be tested by examining trends in its uses over time. One such famous example was the rise of periphrastic *do*, which was traced by (Ellegård, 1953). In this case, the word *do* used to have a specific verb meaning in Old and Middle English—it denoted a causative relation (e.g., “did him gyuen up,” the Peterborough Chronicle, ca. 1154). In modern English, *do* is more frequently used as a grammatical function word (e.g., “Do you like it?”).

Although *do* started out as a verb with a meaning that was quite relational, its meaning was still less context-sensitive in Middle English than it is in Modern English. By measuring the variety of contexts in which *do* appears, Sagi, Kaufmann, and Clark (2012) demonstrate this shift using corpus statistics, with results that correspond to Ellegård’s hand-coded measures. The measure they use is essentially a measure of the variability of contexts in which the word appears within each period. The contextual variability in the uses of *do* exhibits a marked incline between the 15th and 16th centuries.

However, not all changes in meaning necessarily result in its *broadening* as was the case for periphrastic *do*. In many cases, such broadening is limited to the addition of a handful of new uses, or the depreciation of an old use. It might therefore be useful to also examine the change in use as a shift in the contexts rather than simply an increase in their variability. One possible source of such shifts, through metaphoric extension, might arise out of the effects of conceptual framing, such as the framing of terror as an act of war instead of as a crime following the events of September 11, 2001 (see Lakoff, 2009; for a related computational method see Sagi, Diermeier, & Kaufmann, 2013).

The similarity measure used in LSA and other methods of corpus statistics provide one possible method for tracing these changes. In particular, the more similar the uses of a word in one period are to its uses in another, the less likely it is that semantic change has occurred. Conversely, if a word has undergone a shift in its meaning or how it is used, we might expect it to appear in a different set of contexts in the new period than it did in the old. For example, the word *computer* used to mean a person who computes. This meaning was largely replaced by its current use, referring to a class of machines. Therefore the vectors representing the new contexts should be farther away from the vectors representing the old contexts than for a word whose meaning did not change (or changed to a lesser degree). By examining these measures across a large number of words, we can use statistics to identify trends in semantic change.

It is important to distinguish between two distinct sources of variability in contexts of use over time. In the first case, words with broader meanings (periphrastic *do* presents an extreme case of such words), are likely to be used in a wide variety of ways. Consequently, their uses will vary greatly within a time period, as well as across time periods. Semantic change presents a second source of variability of contexts over time. In this case, a drift, or change, in the meaning of a word, such as *computer*, or the addition of new meanings, result in a change in the contexts that a word is used in over time. Importantly, without semantic change, the broadness of application of a word remains constant over time, and therefore can be expected to be largely constant regardless of the time span involved. In contrast, drifts in meaning can be expected to accumulate over time and therefore show an increase in contextual variability as the time span examined increases. That is, whereas the effect of broadness of meaning on contextual variability does not depend on the length of time between uses, semantic change should show an increase in variability for longer time periods. For example, variability in contexts due to the broadness of applicability of a word should be the same whether measured over 25 years or 50 years, whereas semantic change would be expected to result in higher variability when measured over 50 years than when measured over 25 years.

Finally, an additional source of variability in the context of use of words over time comes from changes in the use of other words. That is, a shift in which the word *man* appears frequently with the word *silly* at one time period, but more frequently with *blessed* in another, might occur not because the meaning of *man* has changed, but because of pejoration in the meaning of the word *silly*, whose uses are then replaced by *blessed*. For the purposes of the present study, since each word was analyzed in isolation by comparing its uses in one period to its later uses, these changes will be treated as statistical noise, under the assumption that they are uniformly distributed across the corpus and do not vary by grammatical category.

Method

Materials

Corpus To identify changes in word meaning in modern English, a corpus of 19th-century texts were collected from Project Gutenberg (www.gutenberg.org; Lebert, 2011), using the bulk of the English-language literary works available through the project's website. This resulted in a corpus of 4,034 separate documents, consisting of over 240 million words. The Gutenberg Project preamble was removed from the books prior to analysis. Infomap (2007; Takayama, Floumoy, Kaufmann, & Peters, 1998) was used to generate a semantic space based on this corpus, using default settings

(the 20,000 most frequent content words for the analysis with a co-occurrence window of ± 15 words and generating a 100-dimension space) and its default stop list.

Dating texts from the 19th century is difficult, because publication dates are often not readily available. Moreover, when considering language change, the publication date might not be the relevant date to use because the manuscript might have been written years earlier. The analysis was based on the birth dates of the authors instead, because they were easily obtained and are relevant from a linguistic perspective—since much of language learning occurs within the first few years of life. For the purposes of the analysis below, texts were also aggregated into 25-year periods. Consequently, the text from 3,490 books were written by authors born in the 19th century were used in the present analysis (1800–1824, 887 books; 1825–1849, 1,020 books; 1850–1874, 1,243 books; 1875–1899, 340 books).

Nouns and verbs The nouns and verbs used in the study came from two sources: First, high-frequency nouns and verbs were collected from the 500 most frequent words in the corpus. This procedure contrasts words that are in frequent use and have a relatively stable meaning. The grammatical categories of the high-frequency words were determined on the basis of the MRC2 database (Wilson, 1988).¹ Nouns that were only rarely used as verbs were counted as nouns, and vice versa. Out of the 500 words, 168 nouns and 95 verbs followed this selection criterion (roughly 52.6% of the high-frequency words examined). Although this list includes more nouns than verbs, this is to be expected when examining high-frequency words. Nevertheless, these nouns and verbs are relatively equally interspersed among the list of the 500 most frequent words in the corpus. Specifically, the mean frequency of the nouns was 46,486 ($SD = 2,884.78$), and the mean frequency of the verbs was 43,077 ($SD = 3,289.40$). The two conditions did not significantly differ in frequency, $t(258) = 0.744$, $p = .46$.

Second, the study employed a list of frequency-matched relational nouns, entity nouns, and verbs obtained from Dedre Gentner, which was based on the lists used in previous studies (primarily from Asmuth & Gentner, 2005). This list comprised 70 entity nouns (e.g., *emotion*, *fruit*), 81 relational nouns (e.g., *game*, *marriage*), and 76 frequency-matched verbs (e.g., *buy*, *explain*).

Procedure

Calculating context vectors This study was designed to compare the variability of different word classes across uses and

¹ Although it is possible to categorize specific uses of each word as a noun or a verb, in this study the interest was in examining how a word's common grammatical class might affect its contextual variability and its rate of contextual change. As such, focus was placed on words that were consistently associated with a specific grammatical role.

time. This analysis was based on the precomputed semantic space generated from a corpus of texts from the 19th century, as described above. This space provides vector representation for 20,000 words. However, these vectors are computed as aggregates over the entire corpus.

We can employ vector arithmetic to compute the vectors representing the use of a word, such as *man*, in a particular subset of a corpus (such as a particular book, an author, or a time period).² This is done by aggregating the contextual representations of the word and essentially averaging them together. Specifically, we can calculate the *context vector* of each occurrence of the target word by summing up the vectors of the words that appear in its context and normalizing this resulting vector to a unit length. Following the convention used in Infomap, the contexts used here comprised the 15 words that preceded the target word and the 15 words that followed it, for a total of 30 words. After computing a context vector for each appearance of the target word (e.g., *man*) in the selected subset, we can average the resulting context vectors together using the same vector addition and normalization process. The resulting vector represents the centroid of the vectors it aggregates and is functionally equivalent to the mean in a scalar context.

Measuring vector similarity We can gauge the similarity of two vector representations by examining the angle between them—similar vectors will point in similar directions and will have a small angle whereas differing vector will point in different directions and therefore exhibit a larger angle. In vectors of unit length, the cosine of the angle is equivalent to the Pearson correlation between the components of each vector, which is the basic measure of similarity used in this article.

Computing contextual variability The variability associated with an aggregate vector (such as a vector that represents a word in a subset of the corpus as described above) can be conceptualized as the variability of its vectors constituting it. We can therefore measure such variability by examining the similarity of each constituent vector to the centroid, in a similar fashion to how variance of individual data points is measured with relation to the mean. Importantly, since the correlation measure is higher for context vectors that are closer to the centroid (with a maximal value of 1 for vectors that are identical to the centroid), higher numbers indicate *less* variability. To aid with interpretation, below this correlation will

be referred to as a measure of *uniformity*. Finally, its distribution is asymmetric, with maximal uniformity at one end and maximum variability at the other. As a result, there is no need to use the absolute value of the differences or to square them in order to get an accurate estimate of uniformity for comparisons.

Results

Nouns and verbs

The first analysis was based on the variability in context and change in word use over time between high-frequency nouns and verbs. For each word, the vectors representing the contexts in which the word appeared were examined in groups of texts spanning 25-year-long periods, based on the birth date of the authors. Two hypotheses were tested: first that verbs are used in more varied textual contexts than nouns, and second that the contexts in which verbs appear change more rapidly over time than do the contexts in which nouns appear. To demonstrate this, changes in the mean textual contexts were compared over two time scales—25 and 50 years. Importantly, if we were to observe higher variability over periods of 50 years than over periods of 25 years, we could demonstrate that change in use accumulates over time. Since the number of authors whose works are out of copyright (and therefore can be provided by Project Gutenberg) drops sharply at the beginning of the 20th century, the starting periods in this analysis were limited to authors born from 1800 to 1825 and from 1825 to 1850. The means and standard deviations of the correlations between context vectors across time can be found in Fig. 2.

Variability within a time period was measured by averaging the correlation of the contexts of each term to the centroid representing it for the particular time period. That is, first the average vector of all of the contexts was calculated for a particular word (e.g., *man*), and then the correlation of each context to this centroid. The resulting measure informed as to the uniformity of contexts—if all contexts were identical, the average of these correlations would be 1. The more variability there was in the contexts, the lower the average correlation of the individual vectors to the centroid would be. This measure of uniformity was then averaged across all the 25-year time periods, to calculate the overall uniformity of the context for each word. The overall uniformity of use of nouns and verbs could then be compared using a one-way ANOVA. As predicted, nouns ($M = .433$, $SD = .037$) were more uniform than verbs ($M = .399$, $SD = .033$), $F(1, 258) = 55.32$, $MSE = 0.0013$, $p < .001$, $\eta_p^2 = .18$.

A two-way ANOVA was used to analyze change over time. In this analysis, the basic dependent measure was the correlation of the centroids between periods. That is, the centroid of each word from one time period (e.g., 1800–1825) was

² Although it would be possible to apply this method with similar results using a variety of semantic spaces (such as those generated from the BNC), using an Infomap space that was generated from the same corpus being analyzed provides an intuitive interpretation of the process. It essentially recreates a representation of a subset of the corpus within the semantic space of the corpus as a whole. This process is not inherently circular, because the semantic space used is static and does not vary; it simply provides the backdrop for the comparisons between particular terms and their contexts. The only variables are the contexts that are computed and compared.

Fig. 2 Measures of semantic change in nouns and verbs over time. Higher values indicate more similarity in meaning over time. Error bars represent standard errors of the means

correlated to its centroid in a second time period (e.g., 1825–1850 for a 25-year span, or 1850–1875 for a 50-year span). Grammatical category (noun vs. verb) and length of time elapsed (25 vs. 50 years) were the IVs, and the similarity of meaning was the DV (measured as the correlation between the centroids of a word for two time periods that began either 25 or 50 years apart). Where grammatical category was a between-subjects variable (since different words counted as subjects in this study), the elapsed length of time was a within-subjects variable.

As predicted, a significant main effect of grammatical category emerged, in which the meaning of nouns was more similar over time than was the meaning of verbs, $F(1, 258) = 50.59$, $MSE = 0.00032$, $p < .001$, $\eta_p^2 = .16$. Unsurprisingly, there was also a main effect of time, in which the correlation between the centroids was lower for the 50-year than for the 25-year periods, $F(1, 258) = 2,523.75$, $MSE = 0.000037$, $p < .001$, $\eta_p^2 = .91$. More importantly, the predicted interaction was also observed—verbs showed more change in their centroids over time than did nouns, $F(1, 258) = 37.28$, $MSE = 0.000037$, $p < .001$, $\eta_p^2 = .13$.

It is important to consider that grammatical classes differ not only in relationality, but also in qualities such as concreteness and familiarity. In particular, nouns tend to denote more concrete entities than verbs do. To test whether concreteness and familiarity accounted for the differences observed, all the words in the high-frequency study that had MRC2 concreteness and familiarity ratings were collected, and a median split was used to identify low- and high-rating words. Concreteness significantly correlated with context similarity for both the 25-year span ($r = .165$, $p < .05$) and the 50-year span ($r = .266$, $p < .01$). Similarly, familiarity also correlated with context similarity for the 25-year span ($r = .195$, $p < .01$) and the 50-year span ($r = .211$, $p < .01$). The analysis above was repeated on this reduced set of words (135 nouns and 60 verbs), with concreteness and familiarity as covariants, and replicated the

above effect. Importantly, the interaction observed earlier was still significant, even after controlling for the effects of concreteness and familiarity, $F(1, 191) = 5.29$, $MSE = 0.000033$, $p < .05$, $\eta_p^2 = .03$. Nevertheless, the effect size was reduced, suggesting that the effects of grammatical class might be partially, but not completely, explained as differences in concreteness and familiarity between the two classes.

Entity nouns and relational nouns

Next, we turn to a comparison of entity nouns and relational nouns. As was mentioned earlier, if the likelihood of semantic change is higher for relational words, we should expect the higher rate of change to be evident not only for verbs, but for other relational words, such as relational nouns. The means and standard deviations of the correlations between context vectors across time for the entity nouns, relational nouns, and frequency-matched verbs used in the analysis can be found in Fig. 3.

As before, first I computed the average uniformity of use for each word. A one-way ANOVA was used to test whether relational nouns and verbs showed more variability in use than entity nouns. As predicted, entity nouns ($M = .36$, $SD = .056$) exhibited more contextual uniformity than either relational nouns ($M = .33$, $SD = .043$) or verbs ($M = .29$, $SD = .048$), $F(2, 224) = 45.10$, $MSE = .002$, $p < .001$, $\eta_p^2 = .29$. Tukey's HSD test showed that all three classes of words were different from each other in their uniformity. That is, relational nouns were more variable than entity nouns, and verbs exhibited less uniformity than either class of nouns.

For analyzing change in context over time, the same overall procedure was followed that had been used previously. As before, a small but significant main effect of grammatical category emerged, $F(2, 224) = 6.33$, $MSE = .001$, $p < .01$, $\eta_p^2 = .053$. The difference between the centroids also increased over time, $F(2, 224) = 1,004.21$, $MSE = .0001$, $p < .001$, $\eta_p^2 = .818$.

Fig. 3 Measures of semantic change in entity nouns, relational nouns, and frequency-matched verbs over time. Higher values indicate more similarity in meaning over time. Error bars represent standard errors of the means

More importantly, the expected interaction was observed, in which this increase over time was greater for relational nouns and for verbs than for entity nouns, $F(2, 224) = 11.08$, $MSE = .0001$, $p < .01$, $\eta_p^2 = .09$.

Because this effect might have been driven primarily by the change in verbs, a planned analysis was also conducted that did not include the verbs. This analysis resulted in a similar pattern, with entity nouns showing less overall evidence of change in their centroid than relational nouns, $F(2, 149) = 4.80$, $MSE = .0008$, $p < .05$, $\eta_p^2 = .031$. The rate of change also increased over time, $F(2, 149) = 760.66$, $MSE = .00001$, $p < .001$, $\eta_p^2 = .836$. Most importantly, the observed interaction, in which relational nouns showed an increased rate of change over time as compared to entity nouns, was also preserved, $F(2, 149) = 6.98$, $MSE = .00001$, $p < .01$, $\eta_p^2 = .045$.

Discussion

In this study, patterns of language change were compared for English nouns and verbs. This analysis revealed that nouns showed less contextual variability within each time period than verbs. Likewise, the centroids representing nouns changed more slowly over time than verbs, and entity nouns change more slowly than relational nouns. These results are in line with theories that argue that verbs, and relational nouns, are represented using relations whereas entity nouns are represented as direct denotations.

These results also demonstrate the utility and efficacy of corpus statistics as a tool for observing large scale trends in language use. Whereas in the lab we observe and record the behavior of an individual or a small number of individuals at a time, focusing on the details of their behavior, corpora provide us with an overview of the behavior of large groups of humans. Converging evidence from both methodologies is likely to provide researchers with more confidence in the validity and reliability of their results.

Study 2: Phonesthemes in text

The case for phonological correlates of meaning

It is a popular intuition that words with similar sounds also mean similar things. There is a long tradition of belief in the association between phonetic clusters and semantic clusters going back at least as far as Wallis's grammar of English (Wallis, 1699). Morphemes form one such well-known cluster, but other submorphemic phonetic clusters that contribute to the meaning of the word as a whole have also been hypothesized (Firth, 1957; Jakobson & Waugh, 1979). Anthropologists have documented sound symbolism in many languages (Blust, 2003; Nuckolls, 1999; Ramachandran & Hubbard, 2001), but its role as a purely linguistic phenomenon is still unclear. Moreover, the Saussurean notion of the arbitrary relationship between the sign's form and its referent is a matter of dogma for most linguists (de Saussure, 1916/2011; Hockett & Hockett, 1960). This makes the study of words that *do* participate in predictable sound–meaning mappings all the more important, since, under the framework of contemporary linguistics it is difficult to explain how these patterns come to be, or why they might survive despite the obvious benefits of arbitrary sound–meaning mappings. What is meant by “sound–meaning mapping” here is not purely sound symbolism, however, nor is it morphology. The present study offers a statistical, corpus-based approach to *phonesthemes*, or submorphemic units that have a predictable effect on the meaning of a word as a whole. These nonmorphological relationships between sound and meaning have not been explored thoroughly in behavioral or computational research, with some notable exceptions (e.g., Bergen, 2004; Hutchins, 1998).

Monaghan, Chater, and Christiansen (2005) and Farmer, Christiansen, and Monaghan (2006) studied the diagnosticity of phonological cues for lexical category membership. They performed a regression analysis on over 3,000 monosyllabic

English words and demonstrated that certain phonological features are associated with an unambiguous interpretation as either a noun or a verb. An associated series of experiments demonstrated reaction time, reading time, and sentence comprehension advantages for phonologically “noun-like nouns” and “verb-like verbs.”

Bergen (2004) used a morphological-priming paradigm to test whether there was a processing advantage for words containing phonesthemes over words that shared only semantic or only formal features, or that contained “pseudo-phonesthemes.” He found a difference in reaction times between the phonestheme condition and the other three conditions by comparing primed reaction times to RTs to the same words in isolation, drawn from Washington University’s English Lexicon Project. He demonstrated both a facilitation effect for word pairs containing a phonestheme and an inhibitory effect for word pairs in which the prime contained a pseudo-phonestheme. His use of corpus-based methods (in this case, latent semantic analysis; Landauer et al., 1998) was limited to ensuring that the list of words used in meaning-only priming pairs did not have any higher semantic coherence than the list of words used in phonestheme priming pairs.

Finally, Hutchins (1998, Studies 1 and 2) examined participants’ intuitions about 46 phonesthemes, drawn from nearly 70 years of speculation about sound–meaning links in the literature. In her studies, participants ranked phonestheme-bearing words’ perceived coherence with a proposed gloss or definition meant to represent the meaning uniquely contributed by the phonestheme. Participants also assigned candidate definitions to nonsense words containing phonesthemes at rates significantly above chance, whereas words without phonesthemes were assigned particular definitions at rates not significantly different from chance. She also examined patterns internal to phonesthemes: strength of sound–meaning association, regularity of this association, and “productivity,” defined as likelihood that a nonword containing that phonestheme will be associated with the definition of a real word containing that phonestheme.

A big-data approach to studying phonesthemes

Previous studies of phonesthemes relied on the intuitions of participants to verify the sound–meaning relationships of interest (e.g., Bergen, 2004; Hutchins, 1998). These methods are at their best when testing only a limited number of phonesthemes. As a result, such studies have often constrained their examination to only a handful of phonesthemes. Even in the most extensive of these works, Hutchins, who identified over 100 phonesthemes previously indicated in the literature, used only 46 of them in her experiments. Big data, and in particular textual data in the form of corpora, provides an alternative source of information on the

meanings of words. As was described earlier, statistical approaches such as LSA, topic models, and Word2Vec can be used to extract measures that correlate with participants’ performance on a variety of semantic similarity measures. Consequently, we can use a corpus to examine the hypothesis that words sharing a particular phonestheme also share a similarity in meaning.

Because the phonestheme as a construct necessarily involves a partial overlap in meaning beyond that generally found in language, the hypothesis here was that words sharing a phonestheme would exhibit greater semantic relatedness than words chosen at random from the entire corpus. This computational approach to the problem has two distinct advantages over the experimental methods commonly found in the literature. First, this method is objective and does not rely on intuition on either the part of the experimenter (e.g., in choosing particular examples and glosses for a phonestheme) or the participants (e.g., in Study 1, Hutchins, 1998, asked participants to rate the fit between glosses and words).³ Second, it is possible to use the method to test a large number of candidate phonesthemes without requiring us to probe each participant for hundreds of linguistic intuitions at a time.

Method

Materials

Proposed phonesthemes in English The bulk of the candidate phonesthemes used were taken from the list used by Hutchins (1998), with the addition of two possible orthography-based clusters that seemed interesting. The materials also included a couple of letter combinations that seemed unlikely to be phonesthemes, in order to test the method’s capacity to discriminate between phonesthemes and nonphonesthemes. The study was based on 149 possible phonesthemes collected by Hutchins. Of these, 46 were taken from the list Hutchins used in her first study, two were candidates that were considered plausible orthographic clusters (*kn-* and *-ign*), and two were chosen as phonemic sequences that seemed unlikely to be phonesthemes (*br-* and *z-*). After examining this corpus, I decided to drop 43 of the 149 possible phonesthemes because each of those candidates had six or fewer types in the corpus, and therefore were not suitable for statistical analysis (e.g., the prefixes *str_p-*, *sp_t-*, *spl-*, and the suffixes *-asp*, *-awl*, and *-inge*). Therefore, a final list of 106 candidate phonesthemes were tested (33 of which had also been used in Hutchins’s first study).

³ At present, the experimenters choose which phonetic clusters to test, meaning that intuition is still part of the process. However, whether or not any phonetic cluster qualifies as a valid phonestheme is entirely statistically determined.

Cooper	Poop	Swoop
Droop	Sloop	Troop
Hoop	Stoop	Whoop
Loop		

Fig. 4 List of words ending with the phonestheme *-oop*

For each phonestheme, all instances of that phonestheme were collected from the 20,000 most frequent content words, based on orthographic match. For each individual word stem, all but one occurrence of the stem were removed from the list (e.g., in the list for the phonestheme *-ash*, the words *dashed* and *dashes* were removed, while the word *dash* itself was retained). Likewise, morphemic uses of particular phonesthemes, such as *-er*, were also eliminated (e.g., *bigger*, *thinner*). Preference was given to retaining the stem itself whenever it was available in the list. Finally, I verified that all the words within a particular phonesthemic cluster shared the same phonetic pronunciation of the phonestheme. A sample list of words is given in Fig. 4.

Corpus This analysis was based on the same set of texts, publicly available from Project Gutenberg, used in the previous analysis.

Procedure

One of the primary results from studies correlating semantic vector space representations is that the distance between words in such spaces correlates well with the performance of participants in semantic similarity tasks. This property of semantic spaces was used to test the hypothesis that pairs of words sharing a phonestheme are more likely to share some aspect of their meaning than pairs of words chosen at random.

Measures of the semantic relatedness of each cluster were taken by randomly sampling 1,000 pairs of words from the cluster and averaging the cosine similarity of these pairs.⁴ A one-way, single-sample *t* test, based on the above average and its variability, tested whether the word cluster representing each candidate phonestheme exhibited a level of semantic relationship that was significantly higher than that demonstrated among pairs of words selected at random from the entire corpus. Because this analysis involved 106 comparisons, a

⁴ It should be noted that this method oversamples the smaller clusters, and in those cases it is virtually identical to an exhaustive calculation of the similarity of all possible word pairs. However, a limited number of random samples were chosen, to provide an upper bound on the computation and a consistent approach for all materials, so as to make the results more comparable and easier to follow.

Bonferroni correction was used, and alpha was adjusted to .000485. As an estimate of the degrees of freedom, the number of types identified was adopted as the effective sample size of each phonestheme. The relevant critical *t* scores ranged from 3.34 (*-er* with 229 types) to 5.99 (e.g., *-oom* with seven types).

Results

First, the baseline of the semantic relationship between randomly selected words in the corpus was calculated, using 1,000 randomly chosen word pairs. This provided a baseline estimate of the expected similarity distribution for unrelated terms ($M = .021$, $SD = .11$). As was described above, the *strength* of each phonestheme was calculated as the average of the pair-wise correlation of 1,000 randomly selected pairs of words that share the phonestheme. It is possible to interpret this strength measure as an effect size measure. In particular, using Cohen's *d*, any phonestheme exhibiting a strength measurement greater than 0.043 could be argued to have a small relationship to the meaning of the words including it ($d' > 0.2$), whereas a measured strength greater than 0.076 ($d' > 0.5$) would indicate a phonestheme with a medium-strength relationship to the meaning of the words including it. A list of the results for each of the tested phonesthemes can be found in the [Appendix](#).

Next, single-sample *t* tests, with a population mean of .021 as measured above, were used to test whether each candidate phonestheme exhibited more semantic cohesiveness than pairs of words chosen at random from the corpus (the *t* scores are also provided in the [Appendix](#)). Among the 106 potential phonesthemes tested, the evidence provided statistical support for 61 (57%). Among Hutchins's original list of 33 possible phonesthemes, 24 proved to be statistically reliable candidates (73%). Overall, the results were in line with the empirical data collected by Hutchins (1998). By way of comparing the two data sets, the present measure of phonestheme strength correlated well with Hutchins's average rating measure ($r = .51$, $p < .01$). Neither of the unlikely phonestheme candidates was statistically supported by the test ($t_{br} = 2.03$, $t_{z} = -2.47$), whereas both of the newly hypothesized orthographic clusters were statistically supported ($t_{kn} = 9.22$, $t_{gn} = 9.54$).

Interestingly, a negative correlation ($r = -.32$, $p < .001$) was apparent between the number of tokens for a given phonestheme and its significance frequency. However, it is important to note that this correlation is not unique to the present method, since it is also evident in the results reported by Hutchins (1998; e.g., $r = -.44$, $p < .05$, between the number of types in the present study and the average rating in Hutchins's Study 1).

Discussion

This study provided statistical evidence for over 50% of the proposed phonesthemes. Given the wide range of phones proposed and their overall relatively high level of support, it seems likely that some aspects of meaning might be related to sound after all. Although this might appear at first to be a significant blow to the hypothesis that the assignment of meaning to words is arbitrary, it is important to remember that much of this relationship might have historical roots and be a result of the nonarbitrariness of semantic change, like the change discussed in the previous study. In particular, as Boussidan, Sagi, and Ploux (2009) have demonstrated, it is possible to connect phonesthemes such as *gl-* to specific phonetic clusters that are hypothesized to be part of the reconstructed Proto-Indo-European language.

Given such extensive historical roots for at least some phonesthemes, it is possible that there are some perceptual links between specific phonesthemes and their meanings. This possibility is akin to suggesting that phonesthemes might originate from onomatopoeias. On the other hand, it is possible that, over time, semantic change might result in clusters of words that share both phonetic and semantic aspects (e.g., through borrowing a set of words from a different language). Importantly, these two hypotheses are not contradictory, and it is likely that even phonesthemes whose origin is onomatopoeic would exhibit some change and drift over time.

One possibility that is particularly intriguing in this regard is that phonesthemes might provide individuals with clues to the meanings of unfamiliar words. For instance, when a child encounters a word such as *glamorous* for the first time, the child might try to understand the word's meaning from the context in which it occurs (e.g., “the actress was glamorous”). However, it is possible that in such cases the child would not limit him- or herself to the immediate context, but would also consider other possible sources of information. Phonesthemes might provide such a source.⁵ In particular, if that child already knew the words *glisten*, *gleam*, and *glow*, this regularity in sound might influence the child toward interpretations of *glamorous* that involve visual aspects of the actress rather than her behavior. The following study was designed to further explore this hypothesis, by examining whether phonesthemes affect participants' interpretations of nonce words in context.

The next study examined one such process that might give rise to phonesthemes. The hypothesis was that phonesthemes would influence participants' guesses as to the meaning of unknown words. This hypothesis was tested by presenting participants with a fill-in-the-blank task and asking them to

choose the best-fitting word among three nonce words. In line with the previous results, the prediction was that participants would prefer an option with a phonestheme that fit the context over ones that did not. For example, when asked to complete the sentence “The stone's _____ flashed from under the leaves,” which provides a context that is largely visual, participants should choose completions that involved the vision-related *gl-* phonestheme more frequently than words that included other phonesthemes, such as *-oop*.

Study 3: Phonesthemes in the lab

Method

Participants

Nineteen native English-speaking participants from a major Midwestern university participated in the study in exchange for course credit.

Materials

Six phonesthemes were selected that had been well supported in Study 2: *gl-*, *sn-*, *kn-*, *-ign*, *-oop*, and *-ump*. Six nonce words were created using each phonestheme (e.g., for *gl-*, the stimuli were *glaim*, *glandor*, *glatt*, *glay*, *glunst*, and *glybe*), as well as 18 additional nonce words that did not involve any phonestheme (e.g., *coffle*, *fane*, and *argol*). Importantly, the nonce words representing a phonestheme did not exhibit any other phonestheme.

Also, 36 sentences were generated. Each phonestheme was congruent with the blank in six sentences, based on its associated meaning (e.g., the blanks were best fit with words with visual meaning for *gl-*). In addition, a different phonestheme was identified that was not congruent with the blank in each sentence. These matches were further verified by comparing the word vector representing the sentence to the aggregate vector for the phonestheme in the same corpus that had been used in Study 2. The overall correlation between the congruent phonesthemes and their relevant context sentences was $r = .32$. The overall correlation between the incongruent phonesthemes and their matched context sentences was $r = .004$. These correlations were significantly different from each other, according to a paired-samples t test, $t(35) = 9.3$, $p < .0001$. Each phonestheme was matched with three other phonesthemes, resulting in 18 phonestheme pair matches. Within each pairing, each phonestheme was congruent in two sentences and incongruent in another two sentences. The nonce words were randomly assigned to each of these 18 pairs.

⁵ This is not unlike using morphology to identify the stem of a word—to relate *encyclopedic* to *encyclopedia*, for example.

Procedure

As described above, each sentence was associated with a congruent and incongruent phonestheme, as well as a nonphonesthemic nonce word. The order in which the words were presented for each particular sentence was randomized, so that four randomly determined orders were used (two random orders and their inverses). The study was presented in a pen-and-paper format. A sample sentence is provided in Fig. 5. Participants were asked to circle the word that made the most sense to them as the missing word in each sentence.

Results

The mean rate of each choice is presented in Fig. 6. Participants chose congruent phonesthemes 43% of the time ($M = 15.4$, $SD = 2.69$) and the incongruent phonestheme 23% of the time ($M = 8.26$, $SD = 1.76$). Single-sample t tests compared these rates with the expected base rate of random choice (33%, or 12 out of 36). As predicted, participants chose words incorporating the congruent phonestheme more frequently than would be expected by chance, $t(18) = 3.42$, $p < .001$, $d' = 1.27$. Likewise, participants chose words incorporating the incongruent phonestheme less frequently than would be expected by chance, $t(18) = -3.74$, $p < .001$, $d' = 2.12$.

Discussion

What is the role of phonesthemes in language? At first blush, phonesthemes are at odds with the widely held Saussurian argument that the relationship between words and their meaning is arbitrary. However, as these results have demonstrated, there are some limits to this arbitrariness. In Study 2, 61 phonesthemes were identified as predicting some aspect of the meaning of the words that incorporate them. Moreover, Study 3 demonstrated that phonesthemes affect our predictions as to the meaning of unknown words. Taken together, these results suggest that although phonesthemes do not encapsulate meaning in the same manner that words and morphemes do, they do affect some cognitive processes related to associating words with meaning.

Nevertheless, while individuals are frequently explicitly aware of the meaning of words and the function of

morphemes such as *un-* and *-ing*, this does not seem to be the case with phonesthemes. It is therefore more reasonable to hypothesize that phonesthemes are only implicitly associated with meaning, possibly through our inherent sensitivity to the statistical cues inherent in language (e.g., Hutchinson & Louwerse, 2014; Saffran, 2003; Saffran, Aslin, & Newport, 1996). This hypothesis is further strengthened by the fact that the findings here provided support for phonesthemes by examining one source of such cues; the statistical method employed was based on exploiting the nonrandom nature of the distribution of words and their patterns of co-occurrence.

Interestingly, this suggests that linguistic processing might also be influenced by nonphonesthemic parts of words. For instance, it is possible, and perhaps likely, that in the processing of unknown words, we attempt to draw on other words that sound similar. This suggests that the role that phonesthemes played in Study 3 was not qualitatively different from that other similar-sounding words might play. However, phonesthemes are quantitatively more likely to be associated with meaning than are arbitrarily chosen phonetic clusters that are not morphemic or phonesthemic.

The historical roots of phonesthemes It is also useful to consider that these distributional cues might, at least in part, be due to gradual shifts in the meaning of words over years and generations. Such shifts can result in phonesthemes in two ways: First, one historical root can be responsible for multiple related, but distinct, words. That is the case for many words that can be traced back to Proto-Indo-European (e.g., Boussidan et al., 2009; Watkins, 2000). For example, as Boussidan et al. noted, the phonestheme *gl-* is related to the Proto-Indo-European root **ghel* (“to shine”). Many words in English that begin with *gl-* directly relate to the visual modality, and some others can be demonstrated to have historically been derived from terms associated with vision (e.g., *global*, which is derived from *globe*). It is important to note that this explanation presupposes the existence of particular roots—it is possible that these roots might not be arbitrarily associated with their respective meaning. For example, it is possible that the sound *gl* is cognitively associated with particular experiences, in the same vein investigated by Ramachandran and Hubbard (2001), who demonstrated that participants have particular expectations for the meanings of the nonce words *Bouba* and *Kiki*. In such cases, from a historical perspective, the meaning associated with a particular phonestheme might also not be determined arbitrarily.

Second, it is possible that an existing phonestheme can influence the interpretation of words that are unfamiliar, either because of their low frequency or because they are recently borrowed from a different language. A possible example of this is the word *glance*. The modern definition of the word generally involves some visual aspect (e.g., “a brief or hurried look”; retrieved April 4, 2018, from <http://en.>

That old actress's _____ is fading.

1. drell
2. noop
3. glybe

Fig. 5 A sample item from Study 3. The congruent word is *glybe* (*gl-*), the incongruent word is *noop* (*-oop*), and the nonce word is *drell*

Fig. 6 Percentages of congruent, incongruent, and neutral responses in Study 3. The dashed line represents chance responding. Error bars represent standard errors of the means

oxforddictionaries.com/definition/glance). However, in Middle English, *glacen* means “to graze,” a meaning that is still maintained in the English uses such as *a glancing blow*. Etymologically, it is likely that the word was borrowed from the Old French word *glacier* (“to slip”). Since the phonestheme *gl-* traces to Old English and earlier (from Proto-Indo-European), it seems reasonable to hypothesize that English speakers, upon first hearing the word *glacen* when it was newly borrowed, understood its intended meaning of *slip* from context, but also connected it to the existing cluster of words starting with *gl-*. As a result, its meaning quickly shifted to its modern equivalent, which is essentially “a look that slips.”

General discussion

This article has demonstrated how we can employ the methodology of hypothesis testing to measurements acquire from corpora in a similar fashion to how laboratory studies apply the method to data that is collected in lab experiments. For testing psychological hypotheses, this new application essentially uses texts as a proxy to the individuals that produced them. By coding and quantifying these texts, we can therefore analyze such texts similarly to how we would analyze lab-generated data.

However, quantifying texts is not a trivial endeavor. When there are large bodies of such texts, as is frequently the case when texts are collected from the internet or other sources of big data, it is feasible to explore patterns of co-occurrence within these texts as a mean of quantitatively measuring the overall similarity of words and phrases within them. These measurements then form the backbone of analyses such as those that were carried out in this article.

When conducting such studies, it is important to keep in mind that the data, although produced by individuals, do not comprise a direct measurement. In particular, all data collected

and analyzed in this fashion were mediated through linguistic expressions. This type of mediation might affect the data collected and the possible effect of linguistic processing on the content needs to be considered as part of the design. However, such considerations are also important in lab studies in which the participants produce written or spoken responses. More generally, it is difficult to conduct studies of higher-level cognition without some language-based interaction with the participants during the collection of data. Although controlling for such influences in existing corpora is more difficult, the greater quantity of available data often allows for greater statistical power that can be employed to overcome some of these issues.

Experiments inside and outside the lab

Lab-based studies have always struck a balance between the need for maximizing the internal validity of the study, and the need to produce results that have external validity and apply in a wide range of circumstances. In the lab, researchers can exercise a great degree of control and achieve high levels of internal validity. However, this level of control can lead to results that do not replicate well outside of the lab.

In contrast, in studying data collected outside of the lab, as is frequently the case with big-data and corpora studies, researchers are electing to greatly limit the degree of control they have over the data and its collection. This results in more variability in the data, which makes statistical analysis more difficult. At the same time, the larger quantity of data often compensates for this and can provide greater statistical power than would have been possible in lab-based studies. Nevertheless, the threats to internal validity cannot all be mitigated by mere quantity of data. In particular, it is rare that existing data includes a manipulation that is relevant to the researcher’s hypothesis. Consequently, the results from such studies are quasi-experimental at best and care needs to be taken in their interpretation.

The value of big data in supporting lab research

It is probably best to consider studies based on big data and corpora as complementary to lab studies, and a complete research program can often benefit from including both components. In many cases, researchers might want to begin by exploring their hypothesis in a tightly controlled lab study, and then extend this result by examining its manifestation in larger datasets collected outside of the lab. However, as demonstrated in Study 3, it is not uncommon for a study of corpora to provide insights and predictions that can be further refined in a lab experiment (e.g., Dehghani et al., 2016). More generally, big data can be an important addition to the arsenal of psychological research. Such data can provide an avenue for studies that will synergize with lab research, leading to better theories and, ultimately, to a deeper understanding of cognition and human behavior.

Appendix—Detailed results from Study 2

Statistically supported phonesthemes and orthographic clusters are in bold.

Table 1 Prefix phonesthemes from Hutchins (1998)

Cluster	Strength	<i>d'</i>	<i>t</i> Score	Types
bl-	0.047	0.24	5.89	42
cl-	0.033	0.11	2.97	62
cr-	0.023	0.02	0.41	64
dr-	0.046	0.23	8.67	41
fl-	0.052	0.28	7.12	53
fr-	0.024	0.03	0.81	51
gl-	0.120	0.90	27.52	22
gr-	0.028	0.06	3.40	66
qu-	0.027	0.05	1.89	48
sc-/sk-	0.038	0.15	3.94	72
scr-	0.050	0.26	6.72	16
sl-	0.044	0.21	3.78	40
sm-	0.048	0.25	8.39	17
sn-	0.080	0.54	18.69	16
sp-	0.023	0.02	0.53	69
spr-	0.121	0.91	24.57	8
squ-	0.038	0.15	7.44	11
st-	0.028	0.06	1.89	139
str-	0.051	0.27	7.72	38
sw-	0.045	0.22	4.07	28
th-	0.025	0.04	2.17	52
thr-	0.045	0.22	6.37	17
tr-	0.033	0.11	4.00	84
tw-	0.058	0.34	9.44	23
wh-	0.045	0.22	5.86	25
wr-	0.067	0.42	12.49	22

Table 2 Suffix phonesthemes from Hutchins (1998)

Cluster	Strength	<i>d'</i>	<i>t</i> Score	Types
-ab	0.037	0.15	3.76	8
-ack	0.056	0.32	8.23	23
-ag	0.072	0.46	12.48	11
-ail	0.043	0.20	5.05	17
-ain/-ein	0.040	0.17	4.36	48
-ake	0.033	0.10	2.38	20
-ale	0.046	0.23	6.68	15
-am	0.064	0.39	7.37	17
-amp	0.011	0.09	− 2.28	9
-an	0.032	0.10	2.97	33
-and	0.042	0.19	5.30	20
-ane	0.014	− 0.06	− 1.31	16
-ang	0.080	0.54	12.87	12
-ank	0.035	0.13	3.15	14
-ap	0.060	0.35	8.15	18
-ar	0.028	0.06	1.41	45
-are	0.034	0.12	3.16	26
-art	0.014	− 0.06	− 1.35	15
-ash	0.052	0.28	8.76	14
-at	0.067	0.42	10.46	19
-ay	0.024	0.02	0.74	33
-eat/-et	0.028	0.06	2.43	89
-eck/-ek	0.035	0.12	3.22	7
-eek/-eak	0.034	0.12	3.11	18
-eel	0.064	0.39	11.30	10
-eep	0.117	0.87	25.10	8
-eer	0.026	0.04	1.11	17
-eet/-eat	0.029	0.08	1.82	26
-ell	0.073	0.47	9.11	11
-er	0.019	− 0.02	− 0.55	229
-ere	0.074	0.48	14.66	10
-est/-east	0.028	0.06	1.94	26
-ew	0.031	0.09	2.01	21
-ick	0.067	0.42	15.01	18
-ide	0.047	0.24	5.34	21
-iff	0.062	0.37	5.30	9
-ig	0.093	0.65	11.63	10
-ile/-uile	0.026	0.05	0.81	31
-ill	0.022	0.01	0.29	17
-im	0.045	0.22	5.81	13
-ime	0.047	0.24	8.30	10
-ine	0.027	0.06	1.47	31
-ing	0.140	1.08	30.08	11
-ink	0.060	0.35	10.92	12
-ip	0.064	0.39	11.10	20
-ir/-ur	0.013	− 0.08	− 1.82	15
-it	0.031	0.09	2.68	50
-le	0.028	0.06	1.69	158
-Vng	0.035	0.13	2.67	36

Table 2 (continued)

Cluster	Strength	d'	t Score	Types
-nk	0.037	0.15	3.63	33
-oast/-ost	0.017	− 0.04	− 0.83	12
-ob	0.047	0.23	7.75	8
-ock	0.029	0.07	2.02	19
-od	0.123	0.93	24.61	11
-oil	0.048	0.25	7.51	8
-ol	0.011	− 0.09	− 2.43	8
-one	0.037	0.15	3.82	11
-ook	0.030	0.08	2.23	7
-oom	0.030	0.08	5.65	8
-oon	0.060	0.35	5.45	12
-oop	0.055	0.31	5.91	10
-oot	0.036	0.13	4.14	7
-op	0.095	0.67	8.10	14
-ope	0.038	0.15	4.88	8
-ore	0.038	0.15	5.18	18
-os	0.036	0.13	2.85	7
-ough	0.081	0.55	13.65	8
-ow	0.044	0.21	4.14	52
-sk	0.029	0.07	2.71	15
-ub	0.068	0.43	8.14	9
-uck	0.063	0.38	9.06	13
-ude	0.051	0.27	7.14	9
-uff	0.142	1.10	36.77	9
-ug	0.077	0.51	25.86	11
-ump	0.095	0.67	19.41	11
-um/-umb	0.066	0.41	20.78	14
-unk	0.089	0.62	18.83	9
-ush	0.063	0.39	12.54	14
-ust	0.028	0.06	1.86	17
-ute	0.032	0.10	2.10	27

Table 3 Additional orthographic clusters tested

Cluster	Strength	d'	t Score	Types
kn-	0.060	0.35	9.22	15
-ign	0.059	0.35	9.54	14
br-	0.029	0.07	2.03	68
z-	0.011	0.09	− 2.47	8

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Barnett-Cowan, M. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <https://doi.org/10.1126/science.aac4716>
- Asmuth, J., & Gentner, D. (2005). Context sensitivity of relational nouns. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 163–168). Mahwah, NJ: Erlbaum.
- Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology*, 70, 2007–2025.
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80, 290–311.
- Blust, R. A. (2003). The phonestheme η in Austronesian languages. *Oceanic Linguistics*, 42, 187–212.
- BNC Consortium. (2007). British National Corpus, version 3 (BNC XML ed.). Oxford, UK: Oxford University Computing Services. Retrieved from www.natcorp.ox.ac.uk
- Boussidan, A., Sagi, E., & Ploux, S. (2009). Phonaesthetic and etymological effects on the distribution of senses in statistical models of semantics. In *Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts (DiSCo 2009)* (pp. 35–40). Austin, TX: Cognitive Science Society.
- de Saussure, F. (2011). Nature of the linguistic sign. In *Course in general linguistics* (pp. 65–70). New York, NY: Columbia University Press. (Original French work published in 1916)
- Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., . . . Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145, 366–375. <https://doi.org/10.1037/xge0000139>
- Ellegård, A. (1953). The auxiliary do: The establishment and regulation of its use in English (Vol. 2). Stockholm, Sweden: Almqvist & Wiksell.
- Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103, 12203–12208.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis* (pp. 1–31). Oxford, UK: Blackwell.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1, 939–944.
- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning (Technical Report No. 257). Urbana, IL: Center for the Study of Reading.
- Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Michnick Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). Oxford, UK: Oxford University Press.
- Gentner, D., & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence* (pp. 343–382). San Mateo, CA: Kaufmann.
- Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73, 135–176.
- Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 129–147.

- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211–244. <https://doi.org/10.1037/0033-295X.114.2.211>
- Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, 69, 626–653.
- Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133, 55–66.
- Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203, 88–97.
- Hutchins, S. S. (1998). *The psychological reality, variability, and compositionality of English phonesthemes* (Ph.D. dissertation). Emory University, Atlanta, GA.
- Hutchinson, S., & Louwerse, M. M. (2014). Language statistics explain the spatial–numerical association of response codes. *Psychonomic Bulletin & Review*, 21, 470–478. <https://doi.org/10.3758/s13423-013-0492-2>
- Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 7, 265–290.
- Infomap [Computer Software]. (2007). Stanford, CA. Retrieved from <http://infomap-nlp.sourceforge.net/>
- Jakobson, R., & Waugh, L. R. (1979). *The sound shape of language*. Bloomington, IN: Indiana University Press.
- Lakoff, G. (2009). *The political mind: A cognitive scientist's guide to your brain and its politics*. New York, NY: Penguin.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240. <https://doi.org/10.1037/0033-295X.104.2.211>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284. <https://doi.org/10.1080/01638539809545028>
- Lebert, M. (2011). *The EBook is 40 (1971–2011)*. Project Gutenberg.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. ArXiv preprint. ArXiv: 1301.3781
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). *Efficient estimation of word representations in vector space*. Paper presented at the International Conference on Learning Representations Workshop, Scottsdale, AZ.
- Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567–575). Stroudsburg, PA: Association for Computational Linguistics.
- Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96, 143–182. <https://doi.org/10.1016/j.cognition.2004.09.001>
- Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28, 225–252.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies*, 8, 3–34.
- Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12, 110–114.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Sagi, E. (2018). Developing a new method for psychological investigation using text as data. *Sage Research Methods Cases*. <https://doi.org/10.4135/9781526442604>
- Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, 32, 132–144. <https://doi.org/10.1177/0894439313506837>
- Sagi, E., Diermeier, D., & Kaufmann, S. (2013). Identifying issue frames in text. *PLoS ONE*, 8, e69185. <https://doi.org/10.1371/journal.pone.0069185>
- Sagi, E., Kaufmann, S., & Clark, B. (2012). Tracing semantic change with latent semantic analysis. In K. Allan & J. A. Robinson (Eds.), *Current methods in historical semantics* (pp. 161–183). Berlin, Germany: Walter de Gruyter.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Mahwah, NJ: Erlbaum.
- Takayama, Y., Flounoy, R., Kaufmann, S., & Peters, S. (1998). *Information mapping: Concept-based information retrieval based on word associations*. Stanford, CA: CSLI Publications.
- Tam, Y.-C., Lane, I., & Schultz, T. (2007). Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21, 187–207.
- Traugott, E. C., & Dasher, R. B. (2001). *Regularity in semantic change*. Cambridge, UK: Cambridge University Press.
- Wallis, J. (1999). *Grammar of the English language*. Oxford, UK: Lichfield.
- Watkins, C. (2000). *The American heritage dictionary of Indo-European roots* (2nd ed.). Boston, MA: Houghton Mifflin Harcourt.
- Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20, 6–10. <https://doi.org/10.3758/BF03202594>
- Yeh, J.-Y., Ke, H.-R., Yang, W.-P., & Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management*, 41, 75–95.