International Journal of Psychophysiology 94 (2014) 407–419

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence

Cyma Van Petten

Department of Psychology, Binghamton University, 4400 Vestal Parkway East, Binghamton, NY 13902, United States

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form 23 October 2014
Accepted 25 October 2014
Available online 4 November 2014

Keywords:
Semantic similarity
Thematic relationship
N400
Concreteness
LSA
Mutual information

ABSTRACT

With increasing availability of digital text, there has been an explosion of computational methods designed to turn patterns of word co-occurrence in large text corpora into numerical scores expressing the "semantic distance" between any two words. The success of such methods is typically evaluated by how well they predict human judgments of similarity. Here, I examine how well corpus-based methods predict amplitude of the N400 component of the event-related potential (ERP), an online measure of lexical processing in brain electrical activity. ERPs elicited by the second words of 303 word pairs were analyzed at the level of individual items. Three corpus-based measures (mutual information, distributional similarity, and latent semantic analysis) were compared to a traditional measure of free association strength. In a regression analysis, corpus-based and free association measures each explained some of the variance in N400 amplitude, suggesting that these may tap distinct aspects of word relationships. Lexical factors of concreteness of word meaning, word frequency, number of semantic associates, and orthographic similarity also explained variance in N400 amplitude at the single-item level.

1. Introduction

Understanding the organization of semantic memory is one of the fundamental goals of cognitive science. Although some human knowledge may be poorly represented by language (e.g., procedural knowledge such as how one balances and steers a bicycle), many sources of data and many computational procedures have been brought to bear on quantifying the relationships among individual concepts expressed by words. One way of classifying data sources is by their generative versus receptive nature. Generative methods include free association to cue words (Jenkins and Russell, 1960), lists of the properties or features of words offered by standard language users (McRae et al., 1997), and word definitions offered by expert lexicographers (Miller and Fellbaum, 1991). Receptive methods include judgments of the degree of similarity or relatedness of a pair of words (Rubenstein and Goodenough, 1965), and semantic context effects in speeded tasks (i.e., the degree to which performance on a target word depends on the preceding cue, Meyer and Schvaneveldt, 1971). A third category of data is grounded in how often a pair of words co-occur in text; this source of data includes both a generative and receptive aspect, namely that writers have chosen to use two words within the same sentence or paragraph or document, and readers later encounter that correlated usage.

With ever-increasing computer power, the last decade has seen an explosion of computational methods designed to turn the raw data from these diverse sources into numerical scores expressing the "semantic distance" between any two words. The success of a new computational procedure is typically assessed by how well its distance metric can predict human performance, e.g., whether a procedure based on word co-occurrence in a text corpus can predict which words people will generate in a free association task (Griffiths et al., 2007; Ji et al., 2008), whether words with similar feature lists will also produce large RT benefits when paired in a lexical decision task (Vigliocco et al., 2004), whether overlap in expert word definitions can predict human ratings of word pair similarity (Finkelstein et al., 2002), etc. Notably lacking from these assessments are measures of brain activity, although the N400 component of the event-related potential has proven very sensitive to semantic relationships between words in pairs, and between larger contexts (sentences and discourse) and single words (see Kutas and Federmeier, 2011; Kutas et al., 2006; Van Petten and Luka, 2012 for reviews).

One reason for the lack of contact between the ERP literature and attempts to quantify semantic relations is fairly obvious, namely that standard psycholinguistic ERP methods depend on averaging brain responses to large sets of words, so that that experimental conditions are typically defined as categorical variables. In contrast, computational measures of semantic distance are continuous in nature, so that they are best compared to other continuous measures, i.e., measures that can be derived for individual word pairs. However, a handful of recent studies suggest that this methodological mismatch may be more apparent than

E-mail address: cvanpett@binghamton.edu.

http://dx.doi.org/10.1016/j.ijpsycho.2014.10.012
0167-8760/

real. For instance, Laszlo and Federmeier (2011) found that regressions on N400 amplitudes elicited by single words – averaged across 120 subjects – were successful in capturing variance associated with each word's orthographic similarity to other words in English. With smaller numbers of participants (16 to 24), Wlotko and Federmeier (2012, 2013) used similar methods to link the size of single-word N400s to the strength of semantic context imposed by a preceding sentence frame.

The measure of contextual-strength used by Wlotko and Federmeier was a generative one, namely the percentage of participants in a normative group who supply a particular word when asked to complete a sentence frame (*cloze probability*). The parallel generative measure for word pairs is *association strength*, the percentage of normative participants who supply Word B in response to Cue A. Here, I examine the correlation between N400 amplitude and association strength at the level of single words. Prior work using more conventional ERP methods of averaging across set of words leads to a strong prediction that N400 amplitude and association strength will be inversely correlated at the single-item level. However, below, I describe some limitations of free association norms which suggest that this correlation may be modest. The remainder of the Introduction provides brief descriptions of more recent methods proposed as metrics of semantic distance, in particular, methods based on the co-occurrence of words in text. The Methods section describes a sample of methods selected here in greater detail. The Results section examines the correlations between five different metrics of semantic distance and the amplitude of the N400 elicited by the second words of some 300 pairs, in addition to the influence of lexical variables like frequency of usage and concreteness of word meaning.

1.1. Free association

Since about 1910, free association norms have served as the gold standard for word-pair relationship strength, with strength expressed as the percentage of subjects who offer Word B in response to Cue word A, given instructions only to produce the first word that comes to mind (Jenkins and Russell, 1960; Kiss et al., 1973; Nelson et al., 2004; Postman and Keppel, 1970). Several ERP studies have compared sets of strongly-associated, weakly-associated, and unassociated cue-target pairs (i.e., categorical conditions that are explicitly ordered). All have shown that the target words of stronger pairs elicit smaller N400s than those of weaker pairs, which in turn elicit smaller N400s than words in unassociated pairs (Frishkoff, 2007; Kutas and Hillyard, 1989; Kandhadai and Federmeier, 2010; Luka and Van Petten, 2014a, 2014b; Ortu et al., 2013). Despite this promising result of graded N400 amplitudes in response to graded association strengths, my laboratory has been unable to demonstrate a three-way ERP amplitude difference when comparing three levels of non-zero association strength (Luka and Van Petten, 2014a, 2014b). This failure suggests that the relationship between N400 amplitude and association strength is somewhat noisy, so that fairly close levels of association strength (6% versus 12% in our experiments) are difficult to distinguish, although more distant levels can readily be discriminated (12% vs. 24%, 23% vs. 47%, and 9% vs. 40% in our experiments). The "noise" may arise from variability in semantic perceptions across individual participants, and from the standard source of noise in psycholinguistic ERP experiments, namely EEG activity that is unrelated to word comprehension. However, an additional possibility is that the free association task provides only a partial view of the conceptual links among words.

Some biases in free association norms have been well documented. One is a predominance of noun responses, with a consequent sparsity of verb and adjective responses (De Deyne and Storms, 2006). Superimposed on this noun bias is a tendency for responses to share part-of-speech with cues, at least for cues that are fairly common words (such that TALL elicits SHORT and SMOOTH elicits ROUGH as the most frequent responses, instead of typical objects that are tall or smooth; Deese, 1962). Both of these biases represent a departure from natural language use, which instead revolves around the actions, states or properties of entities (see Mollin, 2009 for extended discussion of qualitative differences between free association and natural discourse).

Other limitations on free association as a metric of semantic relationship arise from pragmatic factors that come with any laboratory procedure. In the most typical version of the task, each participant provides one response to a cue word; even "multiple response" versions are limited to two or three responses (De Deyne and Storms, 2008; Nelson et al., 2000). This procedure is likely to under-represent weak relationships; in theory, a word that occurred to every participant – but only as his second or fourth association to a given cue – would be completely absent in the final compilation of responses. The final limitation of free-association databases is also a purely practical one. The two databases for English word associations are very large for laboratory projects; the Edinburgh Associative Thesaurus contains responses to 8211 cue words, and the University of South Florida Word Association norms contains responses to 5019 cues (Kiss et al., 1973; Nelson et al., 2004). Nonetheless, these are far smaller than the average reader's vocabulary, and a researcher is not likely to find every word he or she might want in a free-association norm. Despite these limitations, association strength has been securely linked to brain electrical activity by multiple laboratories, so that measures from these large norms are included as predictors of target-word N400 amplitude in the current analyses.

1.2. Corpus measures: From co-occurrence to distributional similarity to latent semantic factors

1.2.1. Word co-occurrence

The increasing availability of digital text has spurred a different sort of non-laboratory data collection procedure that circumvents the size limitations of free-association norms. After assembling a large corpus of naturally-produced documents, a basic computation in corpus linguistics is to extract word *co-occurrence frequencies*, e.g., how often Word A and Word B are both present within some span of text — immediately adjacent, within a 5-word window, within the same document, etc. Aside from their possible value to linguists and psychologists, word co-occurrence frequencies are used to create search interfaces for information retrieval, methods for correcting errors in scanned text, and algorithms for machine translation between languages. For attempts to quantify conceptual links in human semantic memory, co-occurrence frequencies for pairs of words can only be meaningfully interpreted after some method of correcting for the base frequencies of the two words. Without such a correction, common words will necessarily have higher co-occurrence frequencies – with any other word – than less common words. Using a formula that divides co-occurrence frequency by the product of the two words' base frequencies, Recchia and Jones (2009) reported an average correlation of $r = .78$ between co-occurrence frequency (using a window of 10 sentences) and human judgments of word-pair similarity.

Separating base frequency-of-usage from co-occurrence frequency is particularly important when examining relationships between co-occurrence frequency and dependent measures that are known to be sensitive to base word frequency. For instance, Lapesa and Evert (2013) examined the correlation between co-occurrence frequencies for pairs of words and RTs to make a word/nonword judgment about the target word (lexical decision task). They reported an impressive correlation of $r = -.47$ between RT and log co-occurrence frequency, but it is likely that this result received a substantial contribution from the base frequency of the target word (see Balota et al., 2004 for the impact of word frequency on lexical decision time). For the current comparison to N400 amplitude, I collected co-occurrence frequencies in 5- and 9-word spans from the 450 million word *Corpus of Contemporary American English* (Davies, 2008-). Procedures for separating the influence of base word frequency from co-occurrence frequency are detailed in Methods.

1.2.2. Distributional similarity

Attempts to link statistical patterns in text to human knowledge rarely stop with co-occurrence frequencies, but instead implement at least one additional step to derive some measure of *distributional similarity* between words. The idea that words which occur in similar contexts must themselves be related in meaning is usually attributed to two linguists writing in the 1950s (Firth, 1957; Harris, 1954), and can be summarized by example: even if CLEAR and TRANSPARENT never occur in the same document, finding that both words tend to co-occur with SEE, COVER, MURKY, VAGUE, PROCEDURE, etc, would provide evidence that CLEAR and TRANSPARENT are themselves related. Rubenstein and Goodenough (1965) were perhaps the first psycholinguists to put this idea to an empirical test, and found that the degree to which words appeared in overlapping contexts predicted the likelihood that they would be judged as synonymous. For their computation of contextual overlap, Rubenstein and Goodenough were restricted to a very small corpus: one produced by laboratory subjects instructed to write a sentence using one member of a word pair (without seeing the other member of the pair).

Expanding resources for data storage and processing have produced an ever-expanding set of competing models for computing distributional similarity that are based on larger corpora of naturally-produced text not available to Rubenstein and Goodenough in 1965, with tens of millions of words. These models begin by creating a matrix with rows and columns representing each word in the text corpus, and then filling each cell of this matrix with numbers representing the times that a row-word and column-word co-occur within some moving window of N words. In the resulting matrix, each word (a row or column) is a vector representing the contexts in which that word was found, and distributionally similar words are those with similar vectors. Variable factors across competing models include the window size (typically 3–10 words), whether function words like prepositions and pronouns are excluded from the window, whether co-occurrence counts are weighted according to the number of intervening words, and perhaps more critically, on the exact formulas used to reduce the impact of overall word frequency and to assess vector similarity (see e.g., Burgess, 1998; Ji et al., 2008; Jones and Mewhort, 2007; Kolb, 2009; Lin, 1998; Rohde et al., unpublished; Shaoul and Westbury, 2010). Vector-similarity models have shown some success at mimicking human performance in forced-choice tasks such as choosing a word's closest synonym from a list of alternatives (Bullinaria and Levy, 2007), along with moderate correlations to association strength (Burgess, 1998; Spence and Owens, 1990), and higher correlations to ratings of the semantic similarity between paired words (Ji et al., 2008; Miller and Charles, 1991; Rohde et al., unpublished)[1].

There have been fewer attempts to predict online performance in word recognition tasks from vector similarities between the words of a pair. Lund and Burgess (1996) reported a significant correlation of $r = -.35$ between the vector similarity of paired words and lexical decision times to the second words of the pairs. However, this result may have received a substantial contribution from base word frequency (Shaoul and Westbury, 2006). With better control over the influence of word frequency, Shaoul and Westbury (2010) found no significant relationship between vector similarities and reaction time to decide if a word pair was related or unrelated.

Given the novelty of the current attempt to determine whether distributional similarity can predict N400 semantic context effects, no attempt to compare multiple models is made here. One point of general agreement is that the success of a model depends on the quality of the initial co-occurrence data, such that a larger text corpus is better than a smaller one, but also that higher-quality (i.e., edited) text is preferable to unedited text scraped from websites (Bullinaria and Levy, 2007; Recchia and Jones, 2009). Here, I examine distributional similarity as computed by one such model based on co-occurrence frequencies from 300,000 Wikipedia articles comprising 267 million words (Kolb, 2009).

1.2.3. Latent semantic factors

Other models of semantic distance include additional steps beyond computing co-occurrence matrices and comparing the resulting word vectors (Griffiths et al., 2007; Jones and Mewhort, 2007; Rohde et al., unpublished). Broadly, these include some compression of the “semantic space” into a smaller set of factors than the initially (very large) number of rows and columns in the initial co-occurrence matrix, by methods such as singular-value-decomposition (akin to a factor analysis). The earliest, and arguably the most popular model of this type is *Latent Semantic Analysis* (LSA; Landauer and Dumais, 1997; Landauer et al., 1998). In addition to compression of the semantic space to 300 factors, LSA differs from most other corpus-based methods in the use of a word-to-document co-occurrence matrix rather than word-to-word co-occurrence. The individual documents are chunks of text (250 to 325 words each) from a proprietary corpus of “textbooks, works of literature, and popular works of fiction and nonfiction used in schools and colleges throughout the United States” developed by an education corporation (Touchstone Applied Science Associates (TASA), now Questar Inc; Zeno et al., 1995). Like the simpler models described above, the LSA metric of semantic distance is reported to perform well in choosing the closest synonym in multiple-choice tasks (Landauer et al., 1998). Reported correlations between the LSA metric and human ratings of word-pair similarity range from null to as high as $r = .75$, depending on the word list used (Budiu et al., 2007; Recchia and Jones, 2009; Simmons and Estes, 2006). Partly because LSA distance measures are readily available on the web, this model has become a benchmark for the assessment of new computational models of semantic distance. However, much as for the simpler distributional similarity models, there have been only a few attempts to assess the relationship between semantic distance as computed by LSA and the immediate impact of semantic context on word processing. With a carefully selected set of 32 noun–noun pairs, Vigliocco et al. (2004) reported a correlation of $r = -.44$ between lexical decision RT and LSA distance, when controlling for word-specific factors via a partial correlation. In contrast, for a larger set of 600 pairs including adjectives and verbs as well as nouns, Hutchison and colleagues found a null relationship ($r = -.02$) between context effects in the lexical decision task and LSA distance (Hutchison et al., 2008). Jones and Golonka (2012) similarly found no significant ability to predict lexical decision RT from LSA measures.

Little effort has been devoted to examination of the relationship between ERP measures of semantic context and corpus-based models of semantic distance. Rhodes and Donaldson (2008) compared N400 amplitudes averaged across sets of target words in three types of word pairs: 1) closely related according to free association norms, 2) unassociated, but closely related according to a semantic distance model[2], and 3) unrelated according to either measure. As in other studies, the associated words elicited smaller N400s than the unrelated set, but the words deemed to be related according to the computational model elicited N400s as large as the unrelated set. Some published ERP studies have incorporated LSA measures of semantic distance, but only to support the claim that two sets of items were adequately matched for degree of semantic relatedness. These did not attempt to evaluate whether ERP measures actually show any sensitivity to the LSA metric of semantic distance (Chwilla et al., 2007; Davenport and Coulson, 2011; Ditman and Kuperberg, 2007; Kuperberg et al., 2011). One sentence-processing study compared critical words depending on LSA distance to a prior word in the sentence. In one of the two experiments,

[1] The Spence and Owens (1990) and Miller and Charles (1991) studies were based on a very small (one million word) corpus, leading to many zeroes in the word co-occurrence matrices, so that these early results should be regarded as preliminary.

[2] The specific semantic distance model is briefly described in one of Rhodes and Donaldson's (2008) citations (Huettig et al., 2006), but the website they cite for details of the model and its distance metric are no longer available.

two conditions with the same LSA distances elicited equivalent N400s, whereas two conditions with different LSA distances yielded N400s of different amplitudes. In the other experiment within this study, LSA distance was crossed with a manipulation that made some sentences pragmatically odd (e.g., "some people have lungs, …", not untrue, but unfelicitous given that all people have lungs). In this experiment, the results showed a complex interaction between LSA distance, pragmatic quality of the sentences, and pragmatic skill of the participants. Because the pragmatic manipulation was the focus of this study, the numerical LSA distances for the various conditions were not reported (Nieuwland et al., 2010). This last study provides a suggestion that LSA distance may influence N400 amplitude, but no comparison to any other measure of word relationship (see Coulson et al., 2005; Van Petten, 1993; Van Petten et al., 1997 for similar examinations of word-level context within sentences, using association strength as the metric for word-level relationships).

Most or all dependent measures used in psycholinguistic research (including N400 amplitude) are sensitive to characteristics of words aside from their semantic relationship to other words, such as frequency of usage, and orthographic similarity to other words. One experimental strategy for investigating semantic effects is thus to match sets of words as closely as possible on a variety of other lexical characteristics, or to present the same words in different contexts. A different strategy, primarily employed in behavioral research, is to tease apart the impact of various lexical and contextual variables via regression analyses (see Balota et al., 2004; Hutchison et al., 2008; Macizo and Van Petten, 2007 for behavioral examples and Laszlo and Federmeier, 2011, 2014 for rare ERP examples). The current goal of examining the relationship between N400 amplitude and continuous measures of "semantic distance" uses the latter strategy of comparing items that vary not only in semantic distance, but also in multiple lexical characteristics. The analyses here thus offer an opportunity to confirm prior results about the impact of word frequency, concreteness, and orthographic similarity on the N400 using continuous rather than categorical variables.

2. Methods

2.1. Participants

Thirty-two young adults were paid for their participation (17 men and 15 women). Their mean age was 23.8 years ($sd = 5.4$). All were native speakers of English with no history of neurological disorder, psychiatric disorder or learning disability by self report, nor any medications known to affect the central nervous system. All had some college education (mean years of formal education = 15.8 years, $sd = 1.8$, using a formula that assigns 12 years for a high school degree, 16 for a Bachelor's degree, and adds years up to a maximum of 5 for any post-graduate education). Data from three additional participants were not analyzed: one offered no behavioral response on roughly a third (32%) of the trials; reaction times for a second person were more than two standard deviations slower than the mean of the retained subjects; for a third person, more than 80% of the trials included non-EEG electrical artifacts.

2.2. Stimuli and procedure

Three hundred and twenty word pairs were initially constructed as control items for a sentence-processing study (Coulson et al., 2005). Half of the pairs were initially classified as semantically unassociated because they had forward association strengths of zero in the Edinburgh Associative Thesaurus (*EAT*; Kiss et al., 1973), meaning that none of the 100 subjects in that study offered the target word as a response to the cue word. The other half had positive association strengths. Associated and unassociated pairs shared context (or cue) words (e.g., ARMS-LEGS versus ARMS-TRUCK). The stimuli were split into two lists such that each subject viewed 80 cues followed by associated targets, 80 cues followed by unassociated targets, and 80 pairs comprised of words and pronounceable nonwords. Each item of a pair was presented for 200 ms in the center of a video monitor, with a 550 ms interstimulus interval, and a 4.7 s interval between trials. Subjects made speeded lexical decisions on the second item of each pair, signaled by button presses with the right and left thumbs. The mapping between right and left hands and word versus nonword decisions was counterbalanced across participants.

Results based on dividing the pairs into large sets of strongly related, weakly related, and unrelated conditions based on forward association strength (only) have been previously reported (Luka and Van Petten, 2014a, Exp. 1). Fig. 1 reprints those results obtained using conventional methods of averaging ERPs and RTs within subjects, but across sets of items.

From this initial set of items, relevant lexical and sublexical characteristics were available from published sources for 303 of the target words (161 associated words, 142 unassociated words); these comprise the stimuli analyzed here. The 303 pairs included some diversity in part-of-speech relations: 39 adjective-adjective (SMALL-LARGE, SQUARE-HAPPY), 31 adjective-noun (EMPTY-STOMACH, RIGHT-SPEECH), 5 adjective-verb (DIFFICULT-FOLLOW, SMALL-READ), 7 noun-adjective (THROAT-SORE, PEW-EMPTY), 159 noun–noun (HUSBAND-WIFE, TOAD-DESK), 18 noun-verb (GUN-SHOOT, DOOR-HURRY), 1 verb-adverb (CHEW-QUIETLY), 31 verb-noun (OBEY-VOWS, TAKE-SHOES), and 12 verb–verb (RISE-FALL, RIP-TEAR). The top half of Table 1 summarizes characteristics of the target words, which are described below.

2.2.1. Lexical and sublexical characteristics

The item-based data analyses include several aspects of the target words that are known or suspected to influence N400 amplitude and/or lexical decision time, in addition to the strength of their relationship to the cue words. One is frequency of usage, because multiple studies have shown that less common (low frequency) words elicit larger N400s and slower lexical decisions than higher frequency words (Balota et al., 2004; Smith and Halgren, 1987; Rugg, 1990; Van Petten and Kutas, 1990, 1991; Yap and Balota, 2009). Frequency of usage is expressed as the natural log of the target word's occurrence (per million) in the *Corpus of Contemporary American English* (COCA), which consists of roughly 450 million words evenly divided between spoken language (radio and TV shows), fiction, popular magazines, newspapers, and academic journals, all published or broadcast between 1990 and 2012 (Davies, 2008-).

A second factor is orthographic similarity to other English words, given that words which resemble many other words elicit larger N400s and faster lexical decision times than orthographically distinctive words (Balota et al., 2004; Holcomb et al., 2002; Laszlo and Federmeier, 2011; Molinaro et al., 2009; Vergara-Martinez and Swaab, 2012). Orthographic similarity was defined as the Levenshtein distance to the closest 20 neighbors, a better measure than the older metric provided by Coltheart's N when words exceed three or four letters in length (Yarkoni et al., 2008). Levenshtein distance is defined as the mean number of changes, additions or deletions of letters required to transform a word into other words, such that a low score indicates close similarity to other words and a high score indicates an unusual orthographic form[3]. Levenshtein distances for the current target words were retrieved from the English Lexicon Project (ELP, Balota et al., 2007). A single

[3] Although the Levenshtein measure is usually described as distance to "the closest 20 words", the actual number may be higher due to ties. For instance, if 10 words can be formed by changing one letter of the target word, and another 10 from changing two letters, the Levenshtein score will be 1.67 for this set of exactly 20 neighbors. However, if 10 words can be formed from changing one letter, and 50 from changing two letters, all 50 of the "two-change" words will be included, for a score of 1.83. Although Levenshtein distance is able to capture differences in orthographic similarity among words with zero neighbors according to the Coltheart-N measure, it remains strongly correlated with word length (Yarkoni et al., 2008), $r = .85$ for the current stimulus set. For this reason, word length was not considered as an independent variable in the current analyses.

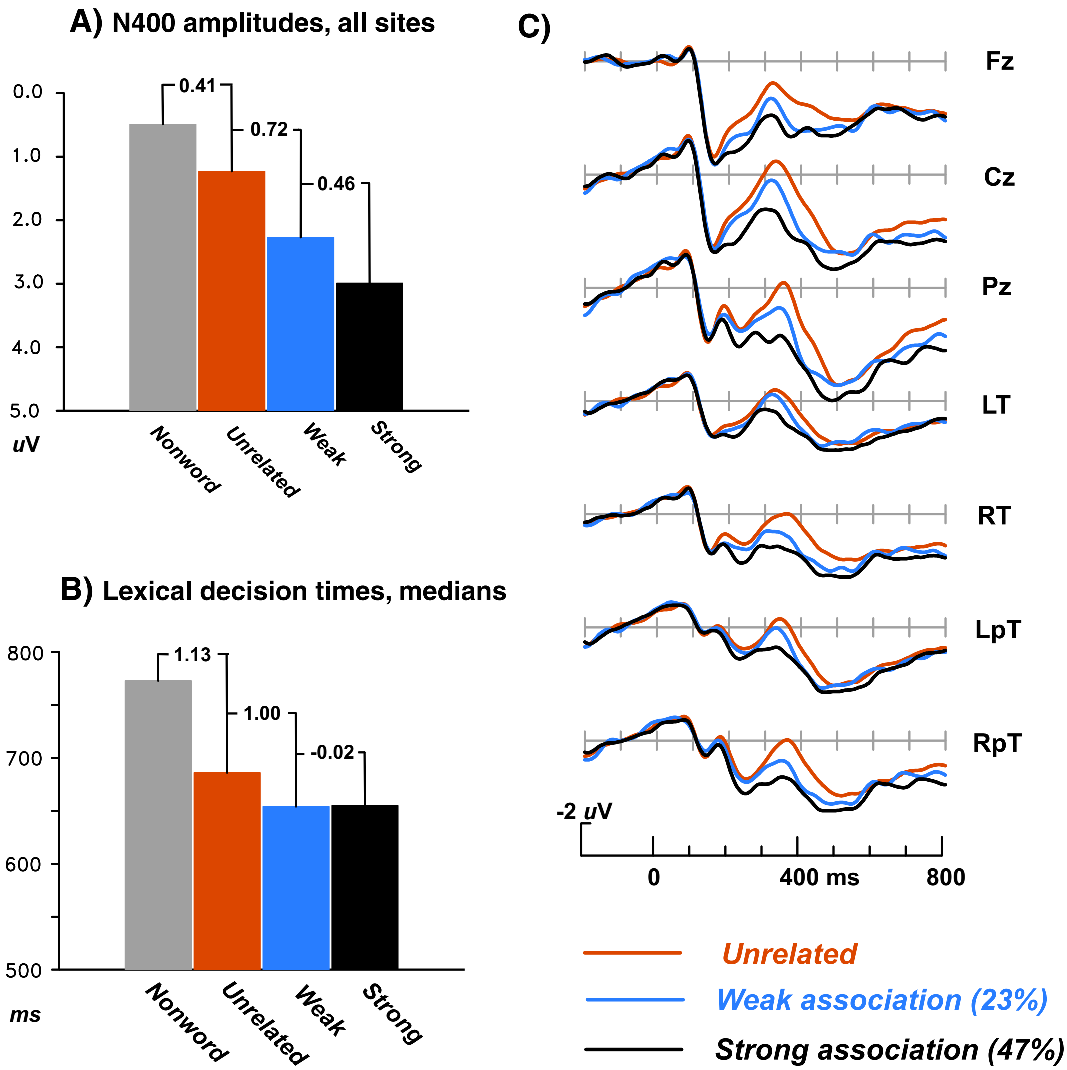

Fig. 1. Results of Experiment 1 of Luka and Van Petten (2014a), the dataset that is analyzed here. (A) The mean amplitudes of event-related potentials (ERPs) in the 250- to 450-ms latency range, averaged across sets of words with association strengths of zero (unrelated), a mean of 23% (weakly-associated) and a mean of 47% (strongly associated). Amplitudes were measured across all scalp sites and averaged in the conventional within-subject manner. Brackets between adjacent bars show the effect size of the difference between them, calculated as unbiased Cohen's d. (B) Median lexical decision times from the same 32 subjects. Brackets between adjacent bars show the effect size of the difference between them, as unbiased Cohen's d. (C) Grand average ERPs from midline frontal, central, and parietal scalp sites, along with left and right mid-temporal sites (LT, RT) and a pair of posterior-temporal sites (LpT and RpT). Copyright Psychonomic Society.

Table 1
Stimulus characteristics (N = 303).

	Mean (SD)	Range
Target words		
Word frequency	4.57 (1.26)	−0.21–7.20
Length	4.96 (1.47)	3–11
Orthographic similarity	1.72 (0.55)	1.00–4.87
Bigram frequency	8.05 (0.55)	6.00–9.00
Number morphemes	1.18 (0.35)	1–2
Number associates	40 (10)	12–64
Concreteness	4.07 (0.93)	1.61–5.00
Word pair relationship		
Forward association strength (FAS)	16.0 (19.9)	0–87
Backward association strength (BAS)	10.1 (18.4)	0–86
Pointwise mutual association (PMI)	2.38 (2.88)	−10.02–9.47
Distributional similarity (DISCO)	0.27 (0.21)	0.00–0.80
Latent semantic analysis (LSA)	0.33 (0.24)	−0.05–0.98

Note: See text for description of variables.

study (thus far) reports that bigram frequency also influences N400 amplitude (Laszlo and Federmeier, 2011), so that the natural log of mean bigram frequency was also collected from the ELP.

Many studies show that words with concrete meanings elicit larger N400s and faster lexical decision times than words with abstract meanings (Barber et al., 2013; Gulick et al., 2013; Kounios and Holcomb, 1994; Swaab et al., 2002). Concreteness ratings were drawn from a recent large-scale study in which subjects rated words on a scale from 1 (least concrete) to 5 (most concrete, Brysbaert et al., 2014).

For words presented in lists without an explicit manipulation of semantic context, two laboratories have found larger N400s for words which elicit a greater diversity of free-association responses as compared to those with fewer associates (Laszlo and Federmeier, 2011; Müller et al., 2010). Words with more associates also tend to elicit faster lexical decision times (Balota et al., 2004). Here, number of associates was tabulated as the number of distinct words offered when the item served as a cue in the Edinburgh Associative Thesaurus (Kiss et al., 1973).

Finally, number of morphemes was coded as 1 for root forms (HAMMER) and 2 for inflected words (STRIPES); the stimulus set did not include any compound words.

2.2.2. Strength of word-pair relationship

Five measures of the strength of relationship between the cue and target words of each pair were computed. These are summarized in the bottom half of Table 1.

2.2.2.1. Forward and backward association. Two measures of association strength were extracted from the Edinburgh Associative Thesaurus (EAT) and the University of South Florida Word Association norms (Kiss et al., 1973; Nelson et al., 2004): the percentage of subjects offering the target word in response to the cue (forward association strength, *FAS*) and the percentage of subjects offering the cue word in response to the target (backward association strength, *BAS*). Most pairs were represented in both norms so that association strength was averaged across them[4]; 11 pairs appeared only in the EAT. Many pairs with zero association strength were included in the stimuli (e.g., SPARE-PENCIL, CHIMNEY-PHONE), up to the highest FAS of 87 for HUSBAND-WIFE.

2.2.2.2. Co-occurrence frequency: pointwise mutual information. Word co-occurrence was examined in the COCA corpus (Davies, 2008-). Co-occurrence was summed across the base and inflected forms of the two words of each pair (such that HAND-FINGER, HANDS-FINGER, HAND-FINGERS and HANDS-FINGERS were summed) within a span of five words forward (HAND followed by FINGER) and backward (FINGER followed by HAND) and also within a span of nine words forward and backward. A window of five was selected because it is fairly typical of published co-occurrence measures, and a window of nine because it is the longest available for the COCA corpus. These four measures were very strongly intercorrelated, with r's ranging from .79 to .99. Based on a preliminary inspection of the relationship to N400 amplitude, the broadest window was adopted: the summed co-occurrence within 9 words forward and backward. For the 303 word pairs, only four had co-occurrence counts of 0 in this 18 word span (ZEBRA-GALLON, MUTTON-RACK, CLOUDY-FAINT, SCISSORS-BROKE); these were replaced by a value of .01. The highest count was some 54,000 for MAN-WOMAN. As noted in the Introduction, it is necessary to separate the co-occurrence frequency of a pair of words from their base frequencies. Most corpus linguists thus use some variant of a *pointwise mutual information* (PMI, Fano, 1961) formula to index the degree to which knowledge of one item reduces uncertainty about the occurrence of the other, beyond their baseline probabilities (Church and Mercer, 1993; Recchia and Jones, 2009). I used a typical version:

$$\log_2((ct * corpus\ size)/(c * t * span))$$

where *ct* is the co-occurrence count for cue and target words, *corpus size* is the total number of words (464,020,256 for the COCA corpus), *c* is the overall frequency of the first word of the pair in the same corpus, *t* is the overall frequency of the second word of the pair, and *span* is the window size for the co-occurrence count (18 words). A base of 2 for the logarithm is used so that the result indexes “bits of information” gained in predicting the second word from the occurrence of the first word. The lowest PMI score in the current stimulus set was −10.0 for SCISSORS-BROKE and the highest 9.5 for LOAF-BREAD; the large majority of the PMI scores were positive (232 of 303).

2.2.2.3. Distributional similarity. As outlined in the Introduction, word co-occurrence has been the raw input for a variety of computations aimed at providing measures of “semantic distance” between words. In these computational models, distance reflects not only the close proximity of two words in text, but also their likelihood of being found in similar contexts. Here, I examine a fairly simple metric that includes a small number of steps beyond gathering word co-occurrence frequencies. Kolb (2009) used a corpus consisting of 300,000 Wikipedia articles (267 million words) and then weighted co-occurrence frequencies in a three-word window by the number of intervening words, after excluding function words like THE, and IT. After using a PMI formula that included the base frequencies of both cue and target words, Kolb computed similarity between words by comparing their vectors in the co-occurrence matrix. Maximally similar words (1.00 on a scale from 0.00 to 1.00) are those that not only co-occur with each other, but are found in the company of the same other words. This model was dubbed *DISCO* for “*extracting DIStributionally similar words using COoccurrences*”. DISCO is one of many methods for computing word similarity based on shared contexts (Burgess, 1998; Ji et al., 2008; Jones and Mewhort, 2007; Lin, 1998; Rohde et al., unpublished; Shaoul and Westbury, 2010). It was selected here because it is based on a large corpus that is subject to some editorial supervision, and because an implemented calculator is freely and easily available. The lowest score in the current stimulus set was .002 for PAGE-PLUMBER and the highest .80 for LUNG-CANCER.

2.2.2.4. Latent semantic analysis. LSA measures of semantic distance were extracted from the LSA website (lsa.colorado.edu) under “pairwise comparison”, using 300 latent semantic factors derived from “general reading up to first year college” (the TASA corpus, Zeno et al., 1995). In theory, LSA distance scores range from −1.00 (distant) to 1.00 (close), but negative scores are rare. The lowest score in the current stimulus set was −.05 for PAGE-PLUMBER and the highest .98 for NEGATIVE–POSITIVE.

2.3. Electrophysiological methods

The electroencephalogram (EEG) was recorded with tin electrodes mounted in a commercially available elastic cap. Midline frontal (Fz), central (Cz) and parietal (Pz) recording sites were used, along with lateral pairs of electrodes over the posterior temporal (T5, T6) and occipital (O1, O2) scalp as defined by the 10–20 system (Jasper, 1958). Three additional lateral pairs were used: a fronto-temporal pair placed midway between F7–8 and T3–4, a mid-temporal pair placed 33% lateral to Cz (left and right *mid-temporal*), and a posterior temporal pair placed 30% of the interaural distance lateral and 12.5% of the inion–nasion distance posterior to Cz (left and right *posterior temporal*). Each scalp site was referred to the left mastoid during recording, and re-referenced to an average of the left and right mastoids prior to data analyses. Vertical eye movements and blinks were monitored via an electrode placed below the right eye referred to the left mastoid. Horizontal eye movements were monitored via a right to left bipolar montage at the external canthi. The EEG was amplified by a Grass Model 12 polygraph with half-amplitude cutoffs of 0.01 and 100 Hz, and digitized at a sampling rate of 250 Hz. Trials with eye movement, muscle, or amplifier blocking artifacts were rejected prior to averaging. After artifact rejection and exclusion of trials with incorrect lexical decisions, a mean of 14 trials for each target word were available for analysis (range 11 to 16).

2.4. Analytic methods

The mean amplitude of the EEG was measured in the N400 latency window of 250 to 450 ms after target word onset, relative to a 200 ms pre-stimulus baseline, from the three midline scalp sites (Fz, Cz, Pz) where conventional averaging and analyses showed the largest

[4] Four pairs had association strengths of zero according to the EAT, but low positive FAS in the USF norm; five showed the opposite pattern.

Table 2
Zero-order correlations, all word pairs (N = 303).

	A	B	C	D	E	F	G	H	I	J	K	L	M
A. Word frequency	–	−.13	−.03	−.16	.29	−.13	.03	−.05	−.35	−.06	−.01	.14	−.17
B. Orthographic similarity	–	–	−.02	.35	−.01	−.13	−.14	−.07	−.06	−.05	−.07	.11	.13
C. Bigram frequency	–	–	–	.09	.04	.04	.03	−.02	.01	.03	.02	−.10	.10
D. Number morphemes	–	–	–	–	.02	−.15	−.10	−.05	−.06	−.03	−.12	−.10	.19
E. Number associates	–	–	–	–	–	.02	−.16	−.42	.02	−.20	−.17	−.15	.10
F. Concreteness	–	–	–	–	–	–	.01	−.15	.19	.02	−.03	−.28	−.06
G. FAS	–	–	–	–	–	–	–	.67	.57	.63	.68	.36	−.25
H. BAS	–	–	–	–	–	–	–	–	.38	.51	.53	.32	−.22
I. PMI	–	–	–	–	–	–	–	–	–	.54	.62	.26	−.18
J. DISCO	–	–	–	–	–	–	–	–	–	–	.71	.26	−.23
K. LSA	–	–	–	–	–	–	–	–	–	–	–	.39	−.26
L. N400	–	–	–	–	–	–	–	–	–	–	–	–	−.27
M. LDT RT	–	–	–	–	–	–	–	–	–	–	–	–	–

Note: See text for descriptions of variables. Without correction for multiple tests, Pearson r's > .19 (positive or negative) are $p < .001$; $r > .23$ are $p < .0001$.

semantic context effects for these stimuli and subjects (Fig. 1)[5]. These measures were derived from single trials (words) from each participant.

To reduce variability associated with individual participants (large versus small EEG from different subjects), each single-trial amplitude was converted to a z-score based on all of the trials from that subject (artifact-free trials with correct lexical decisions). The resulting z-scores were then averaged across subjects for a given target word: the overall mean z for the entire set of words is 0.00, negative scores indicate larger (more negative) N400s and positive scores indicate smaller (more positive) N400s. This method of reducing the impact of inter-subject variability in amplitude is (to my knowledge) a novel one. The test of its viability will lie in the ability to replicate N400 effects that have proven robust under conventional averaging methods: smaller N400s for words with higher forward association strengths as compared to lower FAS, smaller N400s for words that are orthographically distinctive as compared to those with more orthographic neighbors, and smaller N400s for words with lower concreteness ratings as compared to higher ratings.

Reaction times in the lexical decision task that were more than two standard deviations away from the mean of a given subject were discarded; remaining RTs were converted to z-scores as above (following Hutchison et al., 2008).

Statistical procedures were correlations and regressions. Schematically, the analyses allowed the lexical and sublexical variables to account for as much variance in the dependent measures as possible, before examining the amount of additional variance associated with strength-of-relationship between the target word and the preceding cue. The linear relationship between N400 amplitude and the lexical/sublexical variables is of some interest given that only two previous studies (based on one dataset) have examined these for individual words (Laszlo and Federmeier, 2011, 2014). Of greater interest is whether the strength-of-relationship variables will prove to have a linear influence on N400 amplitude, and if so, which ones are most influential.

3. Results

3.1. N400 amplitude

An initial analysis confirmed that related words elicited smaller N400s than unrelated, when relationship was coded in a conventional binary fashion, as forward association strength of zero or greater than zero (mean amplitudes 1.8 versus 3.8 uV, $t(301) = 5.95$, $p < .001$). The remaining analyses are based on the z-score transform of N400 amplitudes, and use continuous rather than binary variables.

3.1.1. Zero-order correlations

Table 2 shows the correlations among the lexical/sublexical characteristics of the stimuli, the five measures of word-pair relationship strength, and the two dependent variables of N400 amplitude and lexical decision time. Notably, the five metrics of relationship strength were strongly intercorrelated, with r's ranging from .38 to .71. All five were significantly and positively correlated with N400 amplitude, such that closer "semantic distance" was associated with smaller N400s. The forward association strength, word co-occurrence (PMI), and latent semantic analysis (LSA) measures showed a somewhat stronger ability to predict N400 amplitude than backward association strength or distributional similarity from the DISCO model.

The scatterplot in Fig. 2 shows the relationship between N400 amplitude and forward association strength. A linear relationship between N400 amplitude is evident. However, the large spread of amplitudes

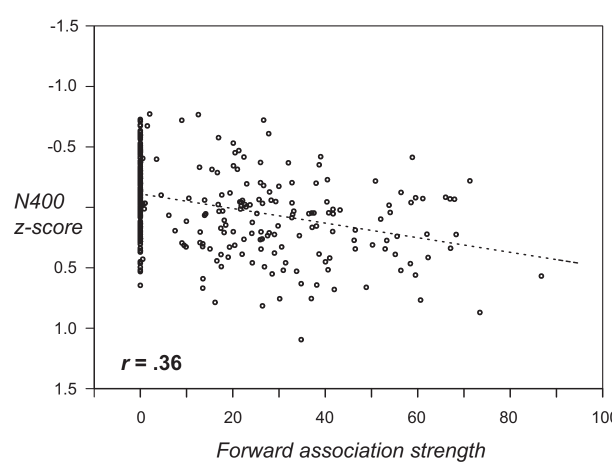

Fig. 2. Scatterplot of the relationship between N400 z-score amplitude for 303 target words and forward association strength (FAS) of cue-target relationship.

[5] Because each target item required a binary decision (word/nonword), the ERPs include a decision-related P300 component in addition to an N400. Studies with nonlinguistic stimuli show that P300 amplitude often varies with decision confidence (Hillyard et al., 1971; Paul and Sutton, 1972; Squires et al., 1975). Targets with higher relationship strength might elicit more confident "word" decisions and larger P300s. The 250–450 ms measurement interval used here was selected to capture primarily N400 amplitude and exclude much of the somewhat later but temporally overlapping P300. Prior results instill some confidence that this goal was achieved. When the current stimulus set was divided into sets of unassociated and weakly associated words (FAS of 0% versus 23%), Luka and Van Petten (2014a, Exp. 1) found an effect size of 0.72 (Cohen's d) for the 250–450 ms measurement interval. In a second experiment, we altered the assigned task so that no decisions were possible during the first 1500 ms of the ERP epoch, and thus no decision-related P300 in the timeframe of interest. In that second experiment, a comparison between words with forward association strengths of 0% versus 24% yielded an ERP effect size of 0.74 (Luka and Van Petten, 2014a, Exp. 2). The very close similarity of these effect sizes in experiments with and without the possibility of an overlapping P300 suggests that the 250–450 ms measurement interval is primarily a measure of N400 amplitude.

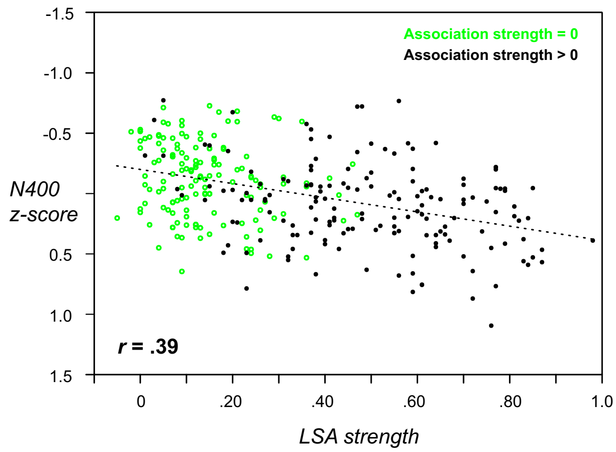

Fig. 3. Scatterplot of the relationship between N400 z-score amplitude for 303 target words and latent semantic analysis (LSA) strength of cue-target relationship.

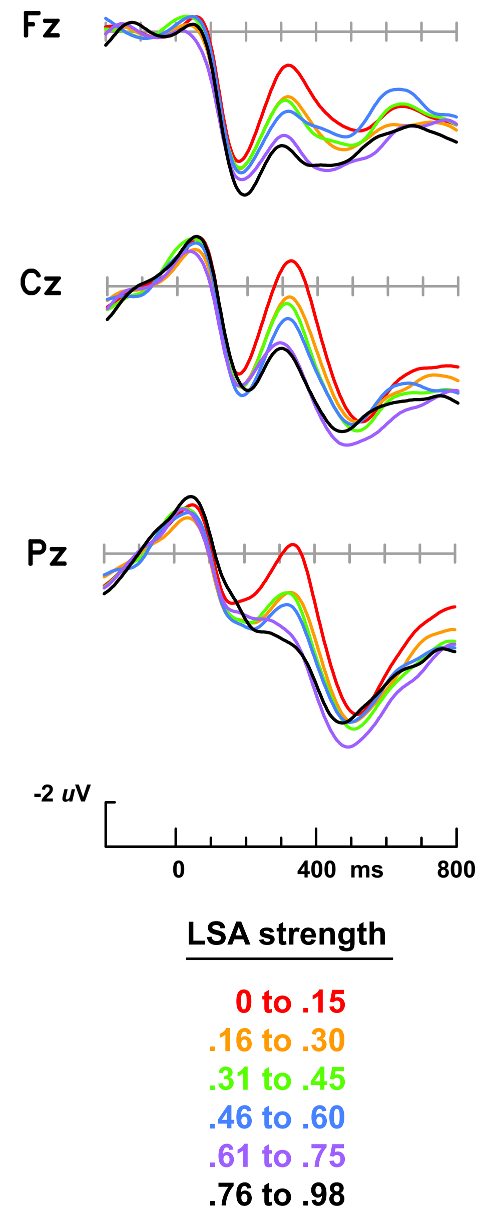

Fig. 4. Averaged ERPs for target words falling within six ranges of LSA strength-of-relationship to their preceding cue words, at frontal, central and parietal midline scalp sites. Note that the z-score transform to reduce inter-subject variability has not been applied here. A low-pass filter at 5 Hz has been applied for purposes of illustration only.

for pairs with zero association strength in Fig. 2 is also striking, which might suggest that a score of zero in association norms conceals some diversity in relationship strength.

In contrast, corpus-based measures (PMI, DISCO, LSA) are much less subject to the possibility of a floor effect given that co-occurrence counts of zero are rare in large text corpora. Fig. 3 shows scatterplots for the measure with the strongest correlation to N400 amplitude — the LSA measure of semantic distance. For visual confirmation of the LSA/N400 relationship, single EEG trials were averaged into six bins depending on LSA strength; Fig. 4 shows the resulting ERP waveforms.

3.1.2. Partial correlations

Table 2 shows that the strength-of-relationship variables had weak to moderate correlations with the lexical/sublexical variables of word frequency, orthographic similarity, bigram frequency, number of morphemes, number of associates, and concreteness. The next analytic step was thus to examine partial correlations between the various metrics of semantic distance and N400 amplitude after removing variance associated with these lexical/sublexical variables. The left side of Table 3 shows these partial correlations for N400 amplitude across all 303 word pairs. All five semantic measures remained significantly correlated with N400 amplitude; indeed, the partial correlations were not substantially different from the zero-order correlations. The partial *r's* for the three strongest measures of semantic distance – forward association strength, PMI word co-occurrence, and LSA – were compared to one another (Meng et al., 1992) – but no measure was significantly superior to the others in predicting N400 amplitude (largest *Steiger's* $z = 1.05$).

Because it is of some interest to know whether the corpus-based measures are able to reveal relationships that are subject to a floor effect in free-association norms, I also examined partial correlations separately for those words classified as "unrelated" or "related" by forward association strength. The middle column of Table 3 shows that the corpus-based measures had no significant ability to predict N400 amplitude for words that would be classified as unrelated by the traditional measure of association strength. However, the right column indicates that forward association strength, the PMI measure of word co-occurrence and LSA were correlated with N400 amplitude within the associated pairs.

3.1.3. Regressions

The correlational analyses above demonstrated that N400 amplitude was linearly related to all five of the semantic-distance measures examined here, and that the LSA measure was (nonsignificantly) the most powerful of the set. It remains possible, however, that the various metrics are capturing different aspects of the semantic relationships among the word pairs (Maki and Buchanan, 2008). Regressions were used to examine the possibility that a combination of measures would lead to greater predictive power than any single measure of strength of semantic relationship. The z-score measure of N400 amplitude was the dependent variable. Potential predictor variables were entered hierarchically. In Step 1, the six lexical and sublexical variables of word frequency, orthographic similarity to other words, bigram frequency, concreteness, number of semantic associates, and number of morphemes were entered as a group. In Step 2, the five metrics of relationship strength – forward and backward association strength, the PMI measure of word co-occurrence, distributional similarity as assessed by the DISCO metric, and LSA strength – were allowed to compete for entry in a stepwise fashion, with an alpha level of $p < .05$ for entry.

Table 4 shows that the full regression model accounted for roughly 31% of the variance in N400 amplitude, with a nearly equal division between lexical variables describing the target word per se, and semantic

Table 3
Partial correlations between N400 amplitude and word-pair relationship strength.

	All pairs		Unassociated		Associated	
	(N = 303)		(N = 142)		(N = 161)	
	r	*p*	*r*	*p*	*r*	*p*
FAS	.36	<.001	–	–	.20	.01
BAS	.23	<.001	−.05	ns	.07	ns
PMI	.34	<.001	.07	ns	.18	.02
DISCO	.27	<.001	−.05	ns	.11	ns
LSA	.38	<.001	.04	ns	.27	<.001

Note: Partial Pearson *r*'s after removing variance associated with the lexical and sublexical variables of word frequency, orthographic similarity to other words, bigram frequency, number of morphemes, number of associates and concreteness. FAS and BAS: forward and backward association strengths (Kiss et al., 1973; Nelson et al., 2004). PMI: pointwise mutual information from word co-occurrence. DISCO: Kolb's (2009) metric of distributional similarity. LSA: Latent Semantic Association word similarity score (Landauer and Dumais, 1997). Associated pairs are those with forward association strength (FAS) greater than zero; unassociated pairs are those with FAS of zero.

variables describing the strength of relationship between the target word and its preceding cue word.

For the lexical variables showing statistically significant contributions to the best-fit regression in Table 4, all were in the direction expected from prior work. Namely, words with lower frequency of occurrence were associated with larger N400s, as were words that are orthographically similar to a larger number of other English words, and words with a larger number of semantic associates. Concreteness of word meaning was the strongest predictor of N400 amplitude among the lexical variables, with larger N400s for more concrete words; Fig. 5 shows this relationship as a scatterplot.

As one might expect from the correlation analyses, LSA strength was the first semantic variable to enter the regression. Additional variance was explained by forward association strength and the PMI measure of word co-occurrence. Among the excluded variables, backward association strength and the DISCO metric of distributional similarity had very high alpha levels ($p = .88$ and .35, respectively), suggesting little likelihood of additional predictive power if they were allowed to enter the regression equation.

Given that association strength has, until now, been the only measure of semantic strength examined in ERP word-pair studies, it is worth examining the degree to which the model with a combination of relationship-strength measures – LSA, FAS and PMI – better predicts N400 amplitude than the traditional measure of FAS alone. A second

Table 4
Best-fit regression for N400 z-score.

	β	*t*	*p*
Step 1) Lexical/sublexical variables	$\Delta R^2 = .157$		
Word frequency	.188	3.24	.001
Orthographic similarity	.174	3.29	.001
Bigram frequency	−.080	1.63	.104
Number morphemes	−.108	2.09	.038
Number associates	−.112	2.11	.036
Concreteness	−.265	5.15	<.001
Step 2) LSA strength of relationship	$\Delta R^2 = .127$		
	.176	2.41	.016
Step 3) FAS association strength	$\Delta R^2 = .020$		
	.147	2.08	.038
Step 4) PMI word co-occurrence	$\Delta R^2 = .009$		
	.149	2.00	.046
Full model	$R^2 = .313$ (adjusted .292)		
	$F(9,293) = 14.8$, $p < .001$		

Note: ΔR^2 is the increase in R^2 when a variable (or set of variables) was entered into the regression. β is the standardized beta weight for each variable in the overall regression; *t* is the t-test for that variable's contribution, and *p* is the significance level for that variable's contribution.

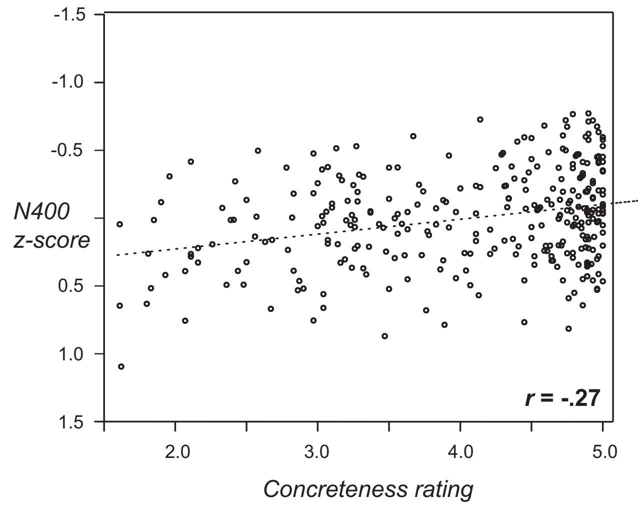

Fig. 5. Scatterplot of the relationship between concreteness of word meaning and N400 z-score amplitude for the 303 target words.

hierarchical regression was conducted with the same six lexical variables entered in Step 1, and forward association strength (only) in Step 2. This FAS model led to an R^2 of .27 as compared to .31 for the best-fit model with LSA, FAS and PMI. The predictive utility of the two regressions was compared; the combined model was significantly better (*Steiger's* $z = 2.07$, $p < .05$).

3.2. Lexical decision RT

Reaction times in the lexical decision task were analyzed in the same manner as N400 amplitude. An initial analysis confirmed that related words elicited faster lexical decisions than unrelated, when relationship was coded in a conventional binary fashion, as forward association strength of zero or greater than zero (mean RTs 660 versus 690 ms, $t(301) = 4.18$, $p < .001$).

Table 2 shows that all five of the measures of relationship strength were significantly correlated with RT z-scores, in the expected negative direction. The left side of Table 5 shows that all five measures remained significantly correlated with lexical decision RT after removing variance associated with the lexical and sublexical variables. The middle and right columns of Table 5 show weak correlations when the stimuli were divided into sets that would be classified as “unrelated” or “related” via a simple division based on forward association strength of zero versus greater than zero, with LSA performing the best for associated pairs, and PMI the best for unassociated pairs.

Table 5
Partial correlations between lexical decision RTs and word-pair relationship strength.

	All		Unassociated		Associated	
	(N = 303)		(N = 142)		(N = 161)	
	r	*p*	*r*	*p*	*r*	*p*
FAS	−.22	<.001	–	–	−.08	ns
BAS	−.18	<.001	−.10	ns	−.09	ns
PMI	−.22	<.001	−.16	.057	−.02	ns
DISCO	−.22	<.001	.03	ns	−.15	.054
LSA	−.24	<.001	.01	ns	−.16	.039

Note: Partial Pearson *r*'s after removing variance associated with the lexical and sublexical variables of word frequency, orthographic similarity to other words, frequency of orthographic neighbors, concreteness, bigram frequency, number of associates and number of morphemes. FAS and BAS: forward and backward association strengths (Kiss et al., 1973; Nelson et al., 2004). PMI: pointwise mutual information from word co-occurrence. DISCO: Kolb's (2009) metric of distributional similarity. LSA: Latent Semantic Association word similarity score (Landauer and Dumais, 1997). Associated pairs are those with forward association strength (FAS) greater than zero; unassociated pairs are those with FAS of zero.

Table 6
Best-fit regression for lexical decision RT z-score.

	β	t	p
Step 1) Lexical/sublexical variables	$\Delta R^2 = .095$		
Word frequency	−.198	3.40	.001
Orthographic similarity	.047	0.81	.416
Bigram frequency	.095	1.76	.080
Number morphemes	.103	1.75	.082
Number associates	.097	1.70	.090
Concreteness	−.081	1.45	.148
Step 2) LSA strength of relationship	$\Delta R^2 = .054$		
	−.237	4.32	<.001
Full model	$R^2 = .149$ (adjusted .129)		
	$F(7,295) = 7.39$, $p < .001$		

Note: β is standardized beta weight for each variable in the regression; *t* is the t-test for that variable's contribution, *p* is the significance level for that variable's contribution.

Table 6 shows the best-fit regression model for the z-score of lexical decision RT, which accounted for roughly 15% of the variance. Among the lexical variables, only word frequency was a significant predictor, with statistically marginal results in the expected directions for bigram frequency, number of morphemes, and number of associates. LSA strength-of-relationship accounted for an additional 5% of total variance. Among the excluded variables, the PMI measure of word co-occurrence had a *p* value to enter of .077; allowing this variable into the equation accounted for an additional 1% of variance.

4. Discussion

4.1. Strength of relationship and lexical decision RT

As in hundreds of other experiments using lexical decision or ERP methods, the second words of related pairs elicited faster lexical decisions and smaller N400s when the presence or absence of a relationship was coded in a binary fashion. However, as reviewed elsewhere (Luka and Van Petten, 2014a), graded influences of association strength are rarely observed in lexical decision RT. In three experiments, Luka and I observed graded N400s from unrelated to weakly-associated to strongly-associated words, but only a binary separation between related and unrelated pairs in lexical decision times (Fig. 1). We thus concluded that the semantic context effect on lexical decision times largely reflected a decision strategy, in which the detection of *any* relationship between the target and preceding cue — weak or strong — serves as an indication that the target letter string must be a word, allowing a fast "yes" response (an idea present in the behavioral literature for some decades). This conclusion is consistent with the current single-item analyses, in that RT was correlated with association strength only when pairs of zero strength were included, but not across the positive range of 0.5 to 87% FAS (Table 5).

In the current analyses, other metrics of relationship strength were compared to lexical decision times. As reviewed in the Introduction, previous attempts to map lexical decision times onto LSA strengths (or onto similarity between word-vectors as computed by other corpus-based models) have met with mixed success, with more null than positive results. The current results are, unfortunately, not entirely conclusive. Although LSA strength was a significant predictor of lexical decision time for the entire stimulus set, this result does not rule out the decision strategy that the association strength analyses suggest. A stronger test is whether LSA strength can predict lexical decision RT within pairs that would conventionally be classified as having *some* relationship, i.e., the associated pairs. Here, the correlation between LSA strength and reaction time (after removing variance accounted for by lexical variables) was significant at $p < .05$, but was not as persuasive as the stronger correlation observed in the parallel analyses for LSA strength and N400 amplitude (compare Tables 3 and 5). The relatively weak relationship between LSA strength and lexical decision RT present here will thus benefit from replication.

4.2. N400 amplitude: lexical variables

Although the first N400 reports focused on the influence of semantic context (Bentin et al., 1985; Kutas and Hillyard, 1980), subsequent research has documented the impact of single-word characteristics on N400 amplitude. The current results using continuous variables confirm prior reports that examined sets of words with extreme values. One confirmatory finding is larger N400s for words with more concrete meanings (Barber et al., 2013; Gulick et al., 2013; Kounios and Holcomb, 1994; Swaab et al, 2002). This "concreteness effect" is taken to reflect retrieval of more semantic information, specifically greater perceptual detail for concrete than abstract words. Although concrete/abstract is invariably coded as a binary variable in ERP studies, the current results suggest that it can be considered continuous, like other lexical characteristics.

Other confirmatory findings were larger N400s for words with a larger number of semantic associates (Müller et al., 2010), and larger N400s for words with greater orthographic similarity to other words (Holcomb et al., 2002; Molinaro et al., 2009; Vergara-Martinez and Swaab, 2012). These two effects have been attributed to the same underlying cause, namely partial activation of more words in the reader's vocabulary. Finally, the regression results also showed larger N400s for words with lower frequency of usage (Rugg, 1990; Smith and Halgren, 1987; Van Petten, 1993).

In two recent reports, Laszlo and Federmeier (2011, 2014) examined the impact of single-word characteristics on the N400 in a continuous fashion, like the current analyses, but with a much larger number of subjects (120 versus 32) and fewer legal words (75 versus 303). The two datasets show a great deal of convergence, but also two discrepancies. Convergent results include the impact of word frequency, orthographic similarity to other words, and number of semantic associates on N400 amplitude. Two discrepancies are that Laszlo and Federmeier found a significant effect of bigram frequency that was null in the current analyses, but were unable to isolate the effect of concreteness that made a substantial contribution to N400 amplitude here. Both of these discrepancies are likely to reflect the composition of the two stimulus sets. Laszlo and Federmeier were able to analyze a very wide range of bigram frequencies due to the addition of unpronounceable consonant strings and acronyms, which have much lower bigram frequencies than legal words. The null effect of bigram frequency in the current set of real words may thus be due to range restriction. Conversely, the large set of real words used here offered greater statistical power to detect a concreteness effect than Laszlo and Federmeier's smaller set of real words.

4.3. N400 amplitude: strength of relationship

The single-item analyses also confirmed the parallel between N400 amplitude and the strength of the preceding semantic context as assessed by the free association task, demonstrated in prior studies. In some respects, the general procedure in a word-pair study resembles that of a free-association task. The cue word is unpredictable and devoid of any specific context, so that a participant is free to interpret it as he/she likes. In the free-association task, the participant is asked to offer his/her first thought, overtly. In a word-pair paradigm, a participant may similarly generate a guess about the upcoming target (see Lau et al., 2013; Luka and Van Petten, 2014b for ERP evidence of such guessing), or may simply evaluate the relationship between the two words after both have been delivered. In either case, the exact interpretation of the cue word and the cue-target relationship will depend on the participant's prior experience with those words.

Average prior experience is exactly what corpus-based measures of semantic distance are designed to capture, so that a central goal of the current analyses was to evaluate the relationship between corpus-based measures of semantic distance and the immediate influence of semantic context seen in electrical brain activity. The correlation analyses showed that a simple measure of word co-occurrence (PMI) and a more complex measure based on co-occurrence (LSA) were roughly equivalent to free-association strength in their ability to predict N400 amplitude (their correlations with N400 amplitude were statistically indistinguishable). The observation that both co-occurrence counts from a large text corpus and LSA strength were as powerful as association strength in predicting N400 amplitude may be of some practical utility to researchers, given that these can be defined for any pair of words, whereas not all words have been subject to free association tests[6].

However, the regression analyses indicated that a combination of free-association, word co-occurrence, and LSA measures was the most powerful in predicting N400 amplitude. This outcome may suggest that the three measures capture somewhat different aspects of conceptual structure, all of which rapidly influence brain activity. Although the qualitative nature of the relationship between cues and responses in free association tasks shows a great deal of variety, the bias toward responses that share part-of-speech with the cue suggests that free association responses are somewhat biased toward relationships of similarity, such as categorically-related nouns (LION-TIGER, BROTHER-SISTER), near synonyms (RIVER-STREAM, STREET-ROAD), and antonyms (SHORT-TALL, EMPTY-FULL). [All examples from the current stimulus set.] Similarity relationships are likely to be less over-represented in pure word co-occurrence statistics like the PMI measure calculated here, where a writer may describe the habitat of the Siberian tiger, arguing with his brother, or meeting a tall woman, but not the alternatives involving lions, sisters, or short women. Instead, co-occurrence statistics are better suited to capture thematic relationships, such as those among actors, actions, and objects (e.g., CLOCK-TICK, PILOT-PLANE, GUN-SHOOT, DRIVE-CAR). The initial input for the LSA metric of relationship is also co-occurrence, but subsequent steps assign high scores to word pairs that share contexts, so that the resulting scores are likely to fall somewhere between the similarity relationships that dominate free-association and the thematic relationships that dominate pure word-occurrence measures. The slightly stronger link between LSA strength and N400 amplitude, as compared to the other measures, might reflect a more balanced sensitivity to both similarity and thematic relations[7].

The word pairs in the current stimulus set were not selected to exemplify particular varieties of relationship, and many are difficult to categorize qualitatively. One category that is easy to identify is the antonymic relationship between adjectives, instantiated in 27 of the current word pairs (e.g., EMPTY-FULL, OLD-YOUNG, SHORT-TALL, THICK-THIN, JUNIOR-SENIOR, HARD-SOFT). I compared forward association strength (FAS), the PMI measure of word co-occurrence, and LSA strength for this set of word pairs, after transforming the three measures into z-scores to normalize their different scales. The mean z-score was highest for FAS (1.13) and lowest for PMI (0.30), with LSA falling in between (0.99), $F(2,52) = 14.5$, $p < .001$, $\varepsilon = .82$, $\eta_p^2 = .36$. Paired t-tests showed that the PMI measure of relationship was significantly lower than both the FAS ($t(26) = 4.08$, $p < .001$) and LSA measures ($t(26) = 5.26$, $p < .001$) for this set of antonyms; the LSA and FAS measures did not differ from one another ($t(26) = 0.93$). This analysis considered only a small portion of the current stimulus set, but is consistent with the argument that free association and word co-occurrence are differentially sensitive to semantic similarity, with LSA occupying some middle ground. It would be useful, in future research, to more deliberately vary the type of semantic relation across word pairs in order to determine whether N400 amplitude is more closely correlated with one or another strength metric depending on the qualitative nature of word-pair relationship.

Overall, the observation of multiple lexical influences on N400 amplitude that were expected from prior research indicates that single-item analyses are feasible in data from a moderate number of participants, at least with large stimulus sets. In combination, the lexical and semantic-relationship variables examined here were able to explain about a third of the variance in N400 amplitude across single words. One can hope to obtain a higher R^2 with an increase in signal-to-noise ratio offered by a larger number of subjects, but can also expect that there will be some ceiling (now unknown) imposed by unobservable factors such as participants' idiosyncratic prior experiences with particular words and moment-to-moment fluctuations in arousal and attention to the experimental stimuli.

The success of both LSA and a less-processed measure of word co-occurrence (PMI) in predicting brain electrical activity suggests that it should be possible to evaluate and compare other corpus-based measures via ERP methods. Given this, it is worth re-visiting what one might hope to get from such measures. As noted in the Introduction, word co-occurrence statistics are currently applied to practical problems. For instance, autocompletions of Google search terms are based on both common completions by other users, and common phrases within web documents, i.e., two sources of co-occurrence data. This application of co-occurrence data is often successful in anticipating a user's current goal. But, like any attempt to predict a specific instance from an average, it also experiences failures: at this writing, the (proprietary) Google algorithm predicts that I seek information about latent tuberculosis, rather than latent semantic analysis.

Similarly, strength-of-relationship measures – whether derived from text corpora or free association tasks – are averages that can align well or poorly with specific instances of word use. Both behavioral and ERP results show that specific contexts can counter word-to-word relationships that are strong in the average case, such that PEPPER and SALT are perceived as strongly related in the default case, but as much less related if the specific context is about clearing an icy sidewalk (Coulson et al., 2005; Hess et al., 1995; see Barsalou, 1982; Federmeier and Kutas, 1999; Tabossi, 1988 for related work). These demonstrations of the potency of specific context are of interest because they support models of lexical semantics that view word meanings as sets of semantic features that are bound to orthographic/phonological word forms by links that vary in strength – tight links for typical features (salt as a flavoring agent) and weaker links for less typical features that require more contextual support to become active (salt as a chemical agent that raises melting point). In a featural framework, semantic context effects for pairs of words are best viewed as showing relationships among the underlying features that comprise word meaning. In the absence of a larger context, these will be dominated by the most typical features of the two words. Simple quantification of word-pair relationships thus provides an entrée to understanding the nature of these underlying features and their relationships, or the organization of semantic memory. Many methods are used to approach this core topic in cognitive science, including patterns of knowledge loss after brain damage (Saffran and Schwartz, 1994; Warrington and Shallice, 1984). The current results lend support to the idea that statistical patterns of word co-occurrence

[6] An important caveat to this point is that the current word pairs were constructed to be semantically related or unrelated according to the experimenter's judgment and forward association strength; their LSA (and PMI) scores were examined after the fact. It is unknown whether using high LSA scores *a priori* would always yield pairs that experimental participants perceived as highly related. Similarly, the PMI measure of word occurrence may become unstable for words with very low base frequencies.

[7] The DISCO model has some commonalities with LSA, in that both direct co-occurrence and shared contexts contribute to the DISCO metric of relationship strength. However, DISCO showed lower correlations with N400 amplitude than LSA, and did not contribute to the best-fit regression equation. Of the three corpus-based measures, DISCO uses the smallest window for co-occurrence counts (6 words, 3 forward and 3 backward), as compared to an 18-word window (9 forward and 9 backward) for the PMI measure computed here, and LSA's window of 250 to 325 words. A major difference between DISCO and LSA is that vectors based on co-occurrence counts are immediately compared to compute semantic distance in the DISCO model, but LSA employs an additional step of compressing the matrix to 300 "latent factors" before computing semantic distance. Either or both of these factors may have contributed to the superiority of LSA over DISCO in the current analyses.

in text are also useful for uncovering this organization, but that this source of data may be fruitfully complemented by free association norms.

References

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., Yap, M.S., 2004. Visual word recognition of single-syllable words. J. Exp. Psychol. Gen. 133, 283–316. http://dx.doi.org/10.1037/0096-3445.133.2.283.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., Treiman, R., 2007. The English Lexicon Project. Behav. Res. Methods 39, 445–459. http://dx.doi.org/10.3758/BF03193014 (website: http://elexicon.wust.edu, accessed 08/15/2014).

Barber, H.A., Otten, L.J., Kousta, S.-T., Vigliocco, G., 2013. Concreteness in word processing: ERP and behavioral effects in a lexical decision task. Brain Lang. 125, 47–53. http://dx.doi.org/10.1016/j.bandl.2013.01.005.

Barsalou, L.W., 1982. Context-independent and context-dependent information in concepts. Mem. Cogn. 1, 82–93. http://dx.doi.org/10.3758/bf03197629.

Bentin, S., McCarthy, G., Wood, C.C., 1985. Event-related potentials associated with semantic priming. Electroencephalogr. Clin. Neurophysiol. 60, 343–355. http://dx.doi.org/10.1016/0013-4694(85)90008-2.

Brysbaert, M., Warriner, A.B., Kuperman, V., 2014. Concreteness ratings for 40 thousand generally known English word lemmas. Behav. Res. Methods 46, 904–911. http://dx.doi.org/10.3758/s13428-013-0403-5.

Budiu, R., Royer, C., Pirolli, P.L., 2007. Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. Proceedings of the 8th Annual Conference of the Recherche d'Information Assistée par Ordinateur (RIAO). Centre des Hautes Études Internationales d'Informatique Documentaire, Pittsburgh, PA.

Bullinaria, J.A., Levy, J.P., 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. Behav. Res. Methods 39, 510–526. http://dx.doi.org/10.3758/bf03193020.

Burgess, C., 1998. From simple associations to the building blocks of language: modeling meaning in memory with the HAL model. Behav. Res. Methods Instrum. Comput. 30, 188–198. http://dx.doi.org/10.3758/bf03200643.

Church, K.W., Mercer, R.L., 1993. Introduction to the special issue on computational linguistics using large corpora. Comput. Linguist. 19, 1–24 (doi:).

Chwilla, D.J., Kolk, H.H.J., Vissers, C.T.W.M., 2007. Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. Brain Res. 1183, 109–123. http://dx.doi.org/10.1016/j.brainres.2007.09.014.

Coulson, S., Federmeier, K., Van Petten, C., Kutas, M., 2005. Right hemisphere sensitivity to word and sentence level context: evidence from event-related brain potentials. J. Exp. Psychol. Learn. Mem. Cogn. 31, 129–147. http://dx.doi.org/10.1037/0278-7393.31.1.129.

Davenport, T., Coulson, S., 2011. Predictability and novelty in literal language comprehension: an ERP study. Brain Res. 1418, 70–82. http://dx.doi.org/10.1016/j.brainres.2011.07.039.

Davies, M., 2008. The Corpus of Contemporary American English: 450 million words, 1990–present. http://corpus.byu.edu/coca/ (Accessed March–June, 2014).

De Deyne, S., Storms, G., 2006. Word associations: network and semantic properties. Behav. Res. Methods 40, 213–231. http://dx.doi.org/10.3758/BRM.40.1213.

De Deyne, S., Storms, 2008. Word associations: norms for 1,424 Dutch words in a continuous task. Behav. Res. Methods 40, 198–205. http://dx.doi.org/10.3758/brm.40.1.198.

Deese, J., 1962. Form class and the determinants of association. J. Verbal Learn. Verbal Behav. 1, 79–84. http://dx.doi.org/10.1016/s0022-5371(62)80001-2.

Ditman, T., Kuperberg, G.R., 2007. The time course of building discourse coherence in schizophrenia: an ERP investigation. Psychophysiology 44, 991–1001. http://dx.doi.org/10.1111/j.1469-8986.2007.00565.x.

Fano, R., 1961. The Transmission of Information. MIT Press, Cambridge, MA.

Federmeier, K.D., Kutas, M., 1999. A rose by any other name: long-term memory structure and sentence processing. J. Mem. Lang. 41, 469–495. http://dx.doi.org/10.1006/jmla.1999.2660.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E., 2002. Placing search in context: the concept revisited. ACM Trans. Inf. Syst. 20, 116–131. http://dx.doi.org/10.1145/503104.503110.

Firth, J.R., 1957. A synopsis of linguistic theory 1930–1955. Studies in Linguistic Analysis. Blackwell Publishers, Oxford, England, pp. 1–32.

Frishkoff, G.A., 2007. Hemispheric differences in strong versus weak semantic priming: evidence from event-related brain potentials. Brain Lang. 100, 23–43. http://dx.doi.org/10.1016/j.bandl.2006.06.117.

Griffiths, T., Steyvers, M., Tenenbaum, J., 2007. Topics in semantic representation. Psychol. Rev. 114, 211–244. http://dx.doi.org/10.1037/0033-295x.114.2.211.

Gulick, M.M., Mitra, P., Coch, D., 2013. Imagining the truth and the moon: an electrophysiological study of abstract and concrete word processing. Psychophysiology 50, 431–440. http://dx.doi.org/10.1111/psyp.12033.

Harris, Z., 1954. Distributional structure. Word 10, 146–162.

Hess, D.J., Foss, D.J., Carroll, P., 1995. Effects of global and local context on lexical processing during language comprehension. J. Exp. Psychol. Gen. 124, 62–82. http://dx.doi.org/10.1037/0096-3445.124.1.62.

Hillyard, S.A., Squires, K.C., Bauer, J.W., Lindsay, P.H., 1971. Evoked potential correlates of auditory signal detection. Science 172, 1357–1360. http://dx.doi.org/10.1126/science.172.3990.1357.

Holcomb, P.J., Grainger, J., O'Rourke, T., 2002. An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. J. Cogn. Neurosci. 14, 938–950. http://dx.doi.org/10.1162/089892902760191153.

Huettig, F., Quinlan, P.T., McDonald, S.A., Altmann, G.T.M., 2006. Models of high-dimensional semantic space predict language mediated eye movements in the visual world. Acta Psychol. 121, 65–80. http://dx.doi.org/10.1016/j.actpsy.2005.06.002.

Hutchison, K.A., Balota, D.A., Cortese, M., Watson, J.M., 2008. Predicting semantic priming at the item-level. Q. J. Exp. Psychol. 61, 1036–1066. http://dx.doi.org/10.1080/17470210701438111.

Jasper, H.H., 1958. Report of the committee on methods of clinical examination in electroencephalography. Electroenceph. Clin. Neurophys. 10, 370–375. http://dx.doi.org/10.1016/0013-4694(58)90053-1.

Jenkins, J.J., Russell, W.A., 1960. Systematic changes in word association norms: 1910–1952. J. Abnorm. Soc. Psychol. 60, 293–304. http://dx.doi.org/10.1037/h0042234.

Ji, H., Lemaire, B., Choo, H., Ploux, S., 2008. Testing the cognitive relevance of a geometric model on a word association task: a comparison of humans, ACOM, and LSA. Behav. Res. Methods 40, 926–934. http://dx.doi.org/10.3758/BRM.40.4.926.

Jones, L.L., Golonka, S., 2012. Different influences on lexical priming for integrative, thematic, and taxonomic relations. Front. Hum. Neurosci. 205. http://dx.doi.org/10.3389/fnhum.2012.00205.

Jones, M.N., Mewhort, D.J.K., 2007. Representing word meaning and order information in a composite holographic lexicon. Psychol. Rev. 114, 1–37. http://dx.doi.org/10.1037/0033-295x.114.1.1.

Kandhadai, P., Federmeier, K.D., 2010. Hemispheric differences in the recruitment of semantic processing mechanisms. Neuropsychologia 48, 3772–3781. http://dx.doi.org/10.1016/j.neuropsychologia.2010.07.018.

Kiss, G.R., Armstrong, C., Milroy, R., Piper, J., 1973. An associative thesaurus of English and its computer analysis. In: Aitken, A.J., Bailey, R.W., Hamilton-Smith, N. (Eds.), The Computer and Literary Studies. Edinburgh University Press, Edinburgh, Scotland, pp. 271–286.

Kolb, P., 2009. Experiments on the difference between semantic similarity and relatedness. In: Jokinen, K., Bick, E. (Eds.), Proceedings of the 17th Nordic Conference on Computational Linguistics, pp. 81–88 (Electronically published at Tartu University Library (Estonia). http://hdl.handle.net/10062/9206. website: http://www.linguatools.de/disco/disco_en.html, accessed 08/05/2014).

Kounios, J., Holcomb, P.J., 1994. Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. J. Exp. Psychol. Learn. Mem. Cogn. 20, 804–823. http://dx.doi.org/10.1037/0278-7393.20.4.804.

Kuperberg, G.R., Paczynski, M., Ditman, T., 2011. Establishing causal coherence across sentences: an ERP Study. J. Cogn. Neurosci. 23, 1230–1246. http://dx.doi.org/10.1162/jocn.2010.21452.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP), In: Fiske, S.T., Schacter, D.L., Taylor, S.E. (Eds.), Annual Review of Psychology, 10.1146/annurev.psych.093008.131123. vol. 62, pp. 621–647.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203–205. http://dx.doi.org/10.1126/science.7350657.

Kutas, M., Hillyard, S.A., 1989. An electrophysiological probe of incidental semantic association. J. Cogn. Neurosci. 1, 38–49. http://dx.doi.org/10.1162/jocn.1989.1.1.38.

Kutas, M., Van Petten, C., Kluender, R., 2006. Psycholinguistics electrified II: 1995–2005, In: Traxler, M., Gernsbacher, M.A. (Eds.), Handbook of Psycholinguistics, 2nd ed. Elsevier, New York, NY, pp. 659–724 http://dx.doi.org/10.1016/B978-012369374-7/50018-3.

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. Psychol. Rev. 104, 211–240. http://dx.doi.org/10.1037/0033-295x.104.2.211 (website: http://lsa.colorado.edu, accessed 08/10/2014).

Landauer, T.K., Foltz, P.W., Laham, D., 1998. An introduction to Latent Semantic Analysis. Discourse Process. 25, 259–284. http://dx.doi.org/10.1080/01638539809545028.

Lapesa, G., Evert, S., 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In: Dimberg, V., Levy, R. (Eds.), Proceedings of the Fourth Annual Association for Computational Linguistics Workshop on Cognitive Modelling and Cognitive Linguistics. Association for Computational Linguistics, Stroudsberg, PA, pp. 66–74.

Laszlo, S., Federmeier, K.D., 2011. The N400 as a snapshot of interactive processing: evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology 48, 176–186. http://dx.doi.org/10.1111/j.1469-8986.2010.01058.x.

Laszlo, S., Federmeier, K.D., 2014. Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. Lang. Cogn. Neurosci. 29, 642–661. http://dx.doi.org/10.1080/01690965.2013.866259.

Lau, E.F., Holcomb, P.J., Kuperberg, G.R., 2013. Dissociating N400 effects of prediction from association in single-word contexts. J. Cogn. Neurosci. 25, 484–502. http://dx.doi.org/10.1162/jocn_a_00328.

Lin, D., 1998. Automatic retrieval and clustering of similar words. ACL '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. volume 2, pp. 768–774. http://dx.doi.org/10.3115/980691.980696.

Luka, B.J., Van Petten, C., 2014a. Gradients versus dichotomies: how strength of semantic context influences event-related potentials and lexical decision times. Cogn. Affect. Behav. Neurosci. 14, 1086–1103. http://dx.doi.org/10.3758/s13415-013-0223-1.

Luka, B.J., Van Petten, C., 2014b. Prospective and retrospective semantic processing: prediction, time, and relationship strength in event-related potentials. Brain Lang. 135, 115–129. http://dx.doi.org/10.1016/j.bandl.2014.06.001.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28, 203–208. http://dx.doi.org/10.3758/bf03204766.

Macizo, P., Van Petten, C., 2007. Syllable frequency in lexical decision and naming of English words. Read. Writ. 20, 295–331. http://dx.doi.org/10.1007/s11145-006-9032-z.
Maki, W.S., Buchanan, E., 2008. Latent structure in measures of associative, semantic, and thematic knowledge. Psychon. Bull. Rev. 15, 598–603. http://dx.doi.org/10.3758/PBR.15.3.598.
McRae, K., de Sa, V.R., Seidenberg, M.C., 1997. On the nature and scope of featural representations of word meaning. J. Exp. Psychol. Gen. 126, 99–130. http://dx.doi.org/10.1037/0096-3445.126.2.99.
Meng, X., Rosenthal, R., Rubin, D., 1992. Comparing correlated correlation coefficients. Psychol. Bull. Rev. 111, 172–175. http://dx.doi.org/10.1037/0033-2909.111.1.172.
Meyer, D.E., Schvaneveldt, R.W., 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. J. Exp. Psychol. 90, 227–234. http://dx.doi.org/10.1037/h0031564.
Miller, G.A., Charles, W.G., 1991. Contextual correlates of semantic similarity. Lang. Cogn. Process. 6, 1–28. http://dx.doi.org/10.1080/01690969108406936.
Miller, G.A., Fellbaum, C., 1991. Semantic networks of English. Cognition 41, 197–229.
Molinaro, N., Conrad, M., Barber, H.A., Carreiras, M., 2009. On the functional nature of the N400: contrasting effects related to visual word recognition and contextual semantic integration. Cogn. Neurosci. 1, 1–7. http://dx.doi.org/10.1080/17588920903373952.
Mollin, S., 2009. Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. Corpus Linguist. Linguist. Theory 5, 175–200. http://dx.doi.org/10.1515/CLLT.2009.008.
Müller, O., Duñabeita, J.A., Carreiras, C., 2010. Orthographic and associative neighborhood density effects: what is shared, what is different? Psychophysiology 47, 455–466. http://dx.doi.org/10.1111/j.1469-8986.2009.00960.x.
Nelson, D.L., McEvoy, C.L., Dennis, S., 2000. What is free association and what does it measure? Mem. Cogn. 28, 887–899. http://dx.doi.org/10.3758/bf03209337.
Nelson, D.L., McEvoy, C.L., Schreiber, T.A., 2004. The University of South Florida free association, rhyme, and word fragment norms. Behav. Res. Methods Instrum. Comput. 36, 402–440. http://dx.doi.org/10.3758/bf03195588 (website: http://www.usf.edu/FreeAssociation/, accessed 07/15/2014).
Nieuwland, M.S., Ditman, T., Kuperberg, G.R., 2010. On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. J. Mem. Lang. 63, 324–346. http://dx.doi.org/10.1016/j.jml.2010.06.005.
Ortu, D., Allan, K., Donaldson, D., 2013. Is the N400 effect a neurophysiological index of associative relationships? Neuropsychologia 51, 1742–1748. http://dx.doi.org/10.1016/j.neuropsychologia.2013.05.003.
Paul, D.D., Sutton, S., 1972. Evoked potential correlates of response criterion in auditory signal detection. Science 177, 362–364. http://dx.doi.org/10.1126/science.177.4046.362.
Postman, L., Keppel, G., 1970. Norms of Word Association. Academic Press, New York.
Recchia, G., Jones, M.N., 2009. More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. Behav. Res. Methods 41, 647–656. http://dx.doi.org/10.3758/BRM.41.3.647.
Rhodes, S.M., Donaldson, D.I., 2008. Association and not semantic relationships elicit the N400 effect: electrophysiological evidence from an explicit language comprehension task. Psychophysiology 45, 50–59. http://dx.doi.org/10.1111/j.1469-8986.2007.00598.x.
Rohde, D. L. T., Gonnerman, L., & Plaut, D. C. (unpublished manuscript). An improved model of semantic similarity based on lexical co-occurrence. http://www.cnbc.cmu.edu/~plaut/papers/abstracts/RohdeGonnermanPlautSUB-CogSci.COALS.html, accessed June 2014.
Rubenstein, H., Goodenough, J.B., 1965. Contextual correlates of synonymy. Commun. ACM 8, 627–633. http://dx.doi.org/10.1145/365628.365657.
Rugg, M.D., 1990. Event-related brain potentials dissociate repetition effects of high- and low-frequency words. Mem. Cogn. 18, 367–379. http://dx.doi.org/10.3758/bf03197126.
Saffran, E.M., Schwartz, M.F., 1994. Of cabbages and things: semantic memory from a neuropsychological perspective — a tutorial review. In: Umilta, C., Moscovitch, M. (Eds.), Attention and Performance XV. MIT Press, Cambridge MA, pp. 507–536.
Shaoul, C., Westbury, C., 2006. Word frequency effects in high dimensional co-occurrence models: a new approach. Behav. Res. Methods 38, 190–195. http://dx.doi.org/10.3758/bf03192768.
Shaoul, C., Westbury, C., 2010. Exploring lexical co-occurrence space using HiDEx. Behav. Res. Methods 42, 393–413. http://dx.doi.org/10.3758/BRM.42.2.393.
Simmons, S., Estes, Z., 2006. Using latent semantic analysis to estimate similarity. Proceedings of the 28th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 2169–2173.
Smith, M.E., Halgren, E., 1987. Event-related potentials during lexical decision: effects of repetition, word frequency, pronounceability, and concreteness. Electroencephalogr. Clin. Neurophysiol. Suppl. 40, 417–421.
Spence, D.P., Owens, K.C., 1990. Lexical co-occurrence and association strength. J. Psycholinguist. Res. 19, 317–330. http://dx.doi.org/10.1007/bf01074363.
Squires, K.C., Squires, N.K., Hillyard, S.A., 1975. Decision-related cortical potentials during an auditory signal detection task with cued observation intervals. J. Exp. Psychol. Hum. Percept. Perform. 1, 268–279. http://dx.doi.org/10.1037/0096-1523.
Swaab, T.Y., Baynes, K., Knight, R.T., 2002. Separable effects of priming and imageability on word processing: An ERP study. Cogn. Brain Res. 15, 99–103. http://dx.doi.org/10.1016/S0926-6410(02)00219-7.
Tabossi, P., 1988. Effects of context on the immediate interpretation of unambiguous nouns. J. Exp. Psychol. Learn. Mem. Cogn. 14, 153–162. http://dx.doi.org/10.1037//0278-7393.14.1.153.
Van Petten, C., 1993. A comparison of lexical and sentence-level context effects and their temporal parameters. Lang. Cogn. Process. 8, 485–532. http://dx.doi.org/10.1080/01690969308407586.
Van Petten, C., Kutas, M., 1990. Interactions between sentence context and word frequency in event-related brain potentials. Mem. Cogn. 18, 380–393. http://dx.doi.org/10.3758/BF03197127.
Van Petten, C., Kutas, M., 1991. Electrophysiological evidence for the flexibility of lexical processing. In: Simpson, G. (Ed.), Word and Sentence. North Holland Press, Amsterdam, pp. 129–174.
Van Petten, C., Luka, B.J., 2012. Prediction during language comprehension: benefits, costs, and ERP components. Int. J. Psychophysiol. 83, 176–190. http://dx.doi.org/10.1016/j.ijpsycho.2011.09.015.
Van Petten, C., Weckerly, J., McIsaac, H.K., Kutas, M., 1997. Working memory capacity dissociates lexical and sentential context effects. Psychol. Sci. 8, 238–242. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00418.x.
Vergara-Martinez, M., Swaab, T.Y., 2012. Orthographic neighborhood effects as a function of word frequency: an event-related potential study. Psychophysiology 49, 1277–1289. http://dx.doi.org/10.1111/j.1469-8986.2012.01410.x.
Vigliocco, G., Vinson, D.P., Lewis, W., Garrett, M.F., 2004. Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. Cogn. Psychol. 48, 422–488. http://dx.doi.org/10.1016/j.cogpsych.2003.09.001.
Warrington, E.K., Shallice, T., 1984. Category specific semantic impairments. Brain 107, 829–854. http://dx.doi.org/10.1093/brain/107.3.829.
Wlotko, E.W., Federmeier, K.D., 2012. Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. Psychophysiology 49, 770–785. http://dx.doi.org/10.1111/j.1469-8986.2012.01366.x.
Wlotko, E.W., Federmeier, K.D., 2013. Two sides of meaning: the scalp-recorded N400 reflects distinct contributions from the cerebral hemispheres. Front. Psychol. 4, 181. http://dx.doi.org/10.3389/fpsyg.2013.00181.
Yap, M.J., Balota, D.A., 2009. Visual word recognition of multisyllabic words. J. Mem. Lang. 60, 502–529. http://dx.doi.org/10.1016/j.jml.2009.02.001.
Yarkoni, T., Balota, D., Yap, M., 2008. Moving beyond Coltheart's *N*: a new measure of orthographic similarity. Psychon. Bull. Rev. 15, 971–979. http://dx.doi.org/10.3758/PBR.15.5.971.
Zeno, S., Ivens, S., Millard, R., Duvvuri, R., 1995. The Educator's Word Frequency Guide. Touchstone, Brewster, NY.