Simple co-occurrence statistics reproducibly predict association ratings

Markus J. Hofmann*

Department of Psychology, University of Wuppertal

Chris Biemann

Department of Computer Science, University of Hamburg

Chris Westbury

Department of Psychology, University of Alberta

Mariam Murusidze, Markus Conrad, & Arthur M. Jacobs

Department of Psychology, Free University Berlin

*Corresponding author: Dr. Markus J. Hofmann, Room Z.01.06, Max-Horkheimer Str. 20, 42119 Wuppertal, Germany. Phone: +49 202 4392340; E-mail: mhofmann@uni-wuppertal.de

Manuscript accepted for publication in Cognitive Science

Keywords: association strength, co-occurrence statistics, semantic long-term memory, associative read-out model, interactive activation model.

Abstract

What determines human ratings of association? We planned this paper as a test for association strength (AS) that is derived from the log likelihood that two words co-occur significantly more often together in sentences than is expected from their single word frequencies. We also investigated the moderately correlated interactions of word frequency, emotional valence, arousal, and imageability of both words (r's =< 0.3). In three studies, linear mixed effects models revealed that AS and valence reproducibly account for variance in the human ratings. To understand further correlated predictors, we conducted a hierarchical cluster analysis and examined the predictors of four clusters in competitive analyses: Only AS and word2vec skip-gram cosine distances reproducibly accounted for variance in all three studies. The other predictors of the first cluster (number of common associates, (positive) point-wise mutual information, and word2vec CBOW cosine) did not reproducibly explain further variance. The same was true for the second cluster (word frequency and arousal); the third cluster (emotional valence and imageability); and the fourth cluster (consisting of joint frequency only). Finally, we discuss emotional valence as an important dimension of semantic space. Our results suggest that a simple definition of syntagmatic word contiguity (AS) and a paradigmatic measure of semantic similarity (skip-gram cosine) provide the most general performance-independent explanation of association ratings.

Though it was originally conceptualized as a variable dependent on individual human behavior (Jung, 1905), psychologists have often used free association performance to account for other human performance (e.g., Hutchison, Balota, Cortese, & Watson, 2008; Roediger, Watson, McDermott, & Gallo, 2001). Similarly, human definitions of semantic systems are commonly used in computational linguistics. In WordNet (Miller, 1990), for example, the creators define "addition" as "math operation" or "building extension". These senses are grouped into synsets with the same meaning, such as {purse, wallet}, connected mainly by human-selected taxonomic (e.g. addition IS-A math operation) and semantic relations (e.g. PART-OF). Though Jiang and Conrath (1997) used WordNet to account for variance in the rated association of 30 word pairs (Miller, 1990; Miller & Charles, 1991), it is questionable whether human performance on one task is a very useful explanation for human performance on another task. To proceed from a mere description towards an explanation, psychological models must be able to generate non-circular and generalizable predictions of human performance (e.g. Hofmann & Jacobs, 2014; Perry, Ziegler, & Zorzi, 2007, 2010; Pitt, Myung, & Zhang, 2002).

This same circularity problem arises with association ratings. While such ratings are useful for constraining the functional locus of priming (e.g., Dimigen, Kliegl, & Sommer, 2012; Lucas, 2000), it is questionable how useful they can be in predicting other human performance measures (Jacobs & Grainger, 1994; Pitt, Myung, & Zhang, 2002; Westbury, 2016). Along with circularity inherent in predicting priming effects by association ratings, such a prediction introduces the practical problem of placing strong constraints on the number of associations that can be defined. For instance, when aiming to provide a general definition for the semantic associations of the more than 300,000 word forms of the CELEX lexical database, over 90,000,000,000 ratings would be required from several participants (Baayen, Piepenbrock, & Gulikers, 1995).

In this paper, we compare several association measures computed from large text corpora with respect to their ability to predict subjective association ratings on a 7-point rating scale. These

co-occurrence measures will be introduced in the next section. In the second section, we outline why the experiential variables of emotional valence, arousal and imageability should account for additional variance. In the third section, we sketch the analytical strategies for the three studies we conducted.

Measures derived from word co-occurrence in sentences

Association Strength. To offer a general definition of associations in semantic long-term memory, a very simple predictor of performance on tasks such as association judgment, associative priming, and false and veridical recognition memory is *association strength (AS)*. We can define two words as associated when they occur significantly more often together than is predictable from their single occurrence frequency in a large sentence corpus (Franke, Roelke, Radach & Hofmann, 2017; Hofmann & Jacobs, 2014; Hofmann, Kuchinke, Biemann, Tamm, & Jacobs, 2011; Kuchinke et al., 2013; Quasthoff et al., 2006; Rapp & Wettler, 1991, Roelke et al., 2018; Stuellein, Radach, Jacobs, & Hofmann, 2016). To quantify the strength of this relationship, the log-likelihood can be calculated from the ratio of the observed co-occurrence of two words divided by the co-occurrence predicted by chance (Dunning, 1993; Evert, 2005). If they show a statistically reliable association ($p \leq 0.01, \chi^2 \geq 6.63$), the resulting χ^2 value is log10-transformed to generate AS; otherwise, AS is defined as zero (Hofmann et al., 2011).

Number of common associates. Because AS is computed over co-occurrence probabilities, it can be characterized by de Saussure's (1959) term "syntagmatic" (cf. Hofmann & Jacobs, 2014), which refers to the association between words due to their direct co-occurrence. De Saussure (1959) also proposed a second type of relation, the paradigmatic relation. Two words have a paradigmatic relationship when they "have something in common [and so] are associated in the memory" (p. 123). A standard approach of computational linguists quantifies paradigmatic relationship strength by the number of common associates: "For example, the semantic similarity of the words *red* and

blue can be derived from the fact that they both frequently co-occur with words like *color, flower, dress, car, dark, bright, beautiful*, and so forth" (Rapp, 2002, p. 1). The number of common associates successfully predicts semantic priming effects (Franke et al., 2017; Roelke et al., 2018). To define semantic feature overlap, we here used the 1000 words with the largest χ^2 values as the semantic features of these words (Stuellein et al., 2016), and simply count the number of common associates. To constrain the number of common associates to words relatively diagnostic for a particular meaning, we excluded the 100 most frequent words (cf. e.g. Griffiths, Steyvers & Tennenbaum, 2007).

Point-wise mutual information. Another co-occurrence-based measure defined by the relation of joint frequency to the single-word frequencies is log-normalized pointwise mutual information (PMI; e.g. Bouma, 2009; cf. Lund & Burgess, 1996; McKoon & Ratcliff, 1992; Westbury et al., 2013; Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015). PMI is calculated by the probability that the words co-occur in sentences, divided by the product of the single occurrence probabilities of the words. PMI is based on the first word co-occurrence measure that accounted for semantic priming in absence of a performance-based definition of semantic association (McKoon & Ratcliff, 1992), and it is still used in co-occurrence-based representational approaches of semantics (e.g. Bullinaria & Levy, 2007; Westbury et al., 2013; Westbury et al., 2015).

From a statistical point of view, AS can be considered an interaction term. On the one hand, either there is a significant log likelihood that two words occur more often together than predictable by word frequency, or not. This can be defined as a dummy-coded predictor variable (0/1). On the other hand, if there is a significant association, AS is defined by the log-transformed χ^2 value. This is formally equivalent to an interaction (0/1 * log10(χ^2); cf. Hofmann et al., 2011). Consequently, AS has a very different distribution than the continuous variable PMI, which renders their direct comparison problematic. Therefore, we also took a thresholded PMI measure into account. Many

researchers use only positive scores of PMI (e.g. Bullinaria & Levy, 2007; Shaoul & Westbury, 2006). We here also explore *positive pointwise mutual information* (PPMI).

Joint Frequency. The log-likelihood-based approach has become standard in computational linguistics (Evert, 2005, p. 137). However, there is a somewhat simpler alternative when aiming to define two of the associative laws of the Aristotelian tradition (McKeon, 1941), *the law contiguity* and *the law of frequency*. The law of contiguity suggests that "the experience or recall of one object will elicit the recall of things that were originally experienced along with that object". The law of frequency suggests that "the more frequently two things are experienced together, the more likely it will be that the experience or recall of one will stimulate the recall of the second" (cf. Olson & Hergenhahn, 1982, p. 35)[1]. Rather than defining an association over and above word frequency as for AS, joint frequency simply gives the number of sentences in which the words co-occur (cf. e.g., Evert, 2005). To obtain a predictor variable comparable to the previous predictors, we log-transformed joint frequency.

Word2Vec cosine. In contrast to these measures based simply on joint and single-word frequencies, we also used a recent approach based on a connectionist model. In his seminal work, Elman (1990) proposed a network consisting of symbolic input and output units for each stimulus that are connected by a layer of hidden units. To obtain a representation of the language context, the state of this hidden layer is copied to a contextual layer when the next stimulus is presented. This network is trained to predict the respectively next word. Mikolov, Chen, Corrado, and Dean (2013; see also Mikolov, Yih, & Zweig, 2013) used this basic model architecture to define two language models, *the continuous bag of words (CBOW)* model and *the skip-gram* model. In CBOW, a word is predicted from the context of the two preceding and two following words. The skip-gram model is the inverse, in which a word is used to predict its context, two words in either direction. This

[1] Retrieved from https://principlesoflearning.wordpress.com/dissertation/chapter-3-literature-review-2/the-behavioral-perspective/associationism-aristotle-–-350-b-c-e/#_ftn1 on March 1, 2018.

model has been successfully used by Bhatia (2017) to account for association-based judgments in different tasks. In Tversky and Kahneman's (1983) 'Linda problem', for instance, participants are given the background information that "Linda is 31 years old, (…) majored in philosophy (…) was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations" (Bhatia, 2017, p. 2). Tversky and Kahneman (1983) showed that participants were more likely to judge Linda to be a feminist bank teller rather than a bank teller, even though the conjunction of two probabilities must be less probable than either individual probability (the conjunction fallacy). When the semantic distances between questions and answers are used to account for such associative judgments, word2vec can nicely account for this conjunction fallacy. Hofmann, Biemann and Remus (2017) also relied on Mikolov, Chen et al.'s (2013) model when accounting for human cloze completion probabilities, as well as event-related potentials and eye movement parameters previously accounted for by cloze completions. The CBOW model not only accounted for N400 effects, but also for single-fixation durations, i.e. for the duration of a single fixation that was sufficient for successfully recognizing a word. Finally, Mandera, Keulers and Brysbaert (2017) used word2vec to successfully predict primed lexical decision and naming times, as well as semantic similarity and relatedness ratings.

Experiential measures

In addition to the word frequencies of both words, we also addressed the interaction of other single-word features of the two stimulus words. We used the three rating variables of the BAWL-R (Võ et al., 2009): emotional valence, emotional arousal, and imageability. By measuring the extent to which the interaction of these three features could account for variance in association ratings, we were able to test Bower's (1981) proposal that positive or negative emotion is itself a memory unit that is associated with a word. This associative spreading-activation approach predicts that two words will be judged to have a greater association when they are of a consistent (positive or

negative) valence (Jacobs et al., 2015; Kuhlmann, Hofmann, Briesemeister, & Jacobs, 2016). Because the product of emotional valence is largest when both words have a high valence with the same sign, this interaction approach tests for the hypothesis that the consistent valences of both words can account for variance in the association rating task. This prediction can also be derived from Schröder and Thagard's (2013) proposal that two concepts are associated when they are evaluated with the same positive or negative valence. Because their activity dimension reflects arousal ratings, they would also predict that two words are judged to be associated when they both have a high level of arousal. Finally, Paivio's (1969) dual coding theory would suggest that in addition to the association between the verbal representations, the association between highly imageable words can rely on mental imagery representations. Therefore, two highly imageable words may elicit greater association ratings (see discussion in Johns & Jones, 2015).

The present studies

Because averaging across items has become standard for the evaluation of computational models (Spieler & Balota, 1997), we evaluated the predictor variables based on the amount of explained item-level variance in mean association ratings. As the amount of reproducible variance should be comparable across studies, and because this kind of evaluation depends on the number of subjects (Rey, Courrieu, Schmidt-Weigand, & Jacobs, 2009), we decided to use approximately the same number of participants as in all previous AS-based behavioral studies (N ~ 30; e.g., Hofmann et al., 2011; Hofmann & Jacobs, 2014; Stuellein et al., 2016).

While the item level is well-suited for calculating regression coefficients (Lorch & Myers, 1990), it discards any between-subject variability. To avoid false positive statistical results, particularly due to effects close to the significance threshold (Quene & Van den Bergh, 2008), we additionally used linear mixed effect models (LME) on the single-trial data (e.g., Baayen, Davidson, & Bates, 2007; Kliegl, Risse & Laubrock, 2007).

We planned three studies to test whether AS and the interactions of the single-word features (i.e. word frequency, emotional valence, arousal and imageability) could reproducibly account for unique variance in the association rating task. Study 1 tests this directly. In study 2, we additionally tested whether PMI could also account for unique variance. Therefore, we selected word pairs in which these variables provided a low to moderate correlation, i.e. r's ~ 0.3 (Cohen, 1988). Because many studies obtain the best results using PPMI (e.g. Bullinaria & Levy, 2007), in study 2 we further selected stimuli with a *positive* PMI score only. As studies 1 and 2 contained only words with a length of up to nine letters, study 3 was conducted to generalize to words of a length of up to 18 letters (Hofmann & Jacobs, 2014; Jacobs & Grainger, 1994; Pitt et al., 2002).

In more exploratory analyses, we considered other computational measures of association. As these predictors often provided a high correlation with the other predictors (see Table 1 below for the intercorrelations between the variables), we conducted a hierarchical cluster analysis, and then examine the LME model predictor variables in a competitive cluster-wise fashion, as we describe in more detail below.

General method

Procedure

All three studies were implemented online as HTML and PHP script. The experimenter started it either in a Safari browser (Mac computers) or in a Firefox browser (Windows computers). On the introductory screen, participants entered an identification number, as well as their sex and age. Participants were instructed verbally and in written form to rate the intensity of association of word pairs on a 7-point scale. To reduce inter-subject variability with respect to the definition of the term "Assoziationsstärke" [association strength], we constrained the meaning of this term, by using the quasi-synonymous definitions of "wie stark ist der Bedeutungszusammenhang" [how strong is the meaning relatedness], and "gedanklich verknüpft" [mentally linked] in the instructions.

Furthermore, we gave standardized examples of high association pairs (“Butter” [butter] and “Brot” [bread], and “Pudel” [poodle] and “Hund” [dog]); and low association pairs (“Pyramide” [pyramid] and “Schnee” [snow], “Yoga” [yoga] und "Imprägnierspray" [waterproofing spray]) to minimize decision criteria variability. Two lists presenting the two words in inverted order were generated. Participants were randomly assigned to one of the lists. The sequence of the word pairs was randomized for each participant separately. Each trial consisted of a stimulus and a rating box. Each stimulus consisted of a 30-character string with one word at the beginning and the other word at the end. The middle positions were filled with dashes (‘-’). Below each stimulus, we presented a box containing the rating scale. The box contained the question “Wie stark sind diese beiden Wörter assoziiert? (1 – nicht assoziiert… 7 – stark assoziiert): ” [How strong is the association between both words? (1- not associated… 7- strongly associated):], as well as seven fields with the numbers 1-7 given below. Participants were instructed to tick one of these fields, and then press the “OK” button in the lower right corner of the box. Though the total time available for the ratings was not constrained, participants were instructed that there are no wrong answers and that they should respond as spontaneously as possible. In all, the studies took 30 minutes or less.

Materials

All single-word and sentence co-occurrence data were taken from the German corpus of the Leipzig Wortschatz project (status: December 2006; Quasthoff et al., 2006). The corpus is largely composed of online newspapers (1992–2006), and consists of 800 million tokens in 43 million sentences.

AS was based on whether two words occur more often together in the sentences than is predictable from the single occurrence frequency of the words, using the method described above. For the **number of common associates**, we used the 1000 words with the most significant log likelihoods (Dunning, 1993), while excluding the 100 most frequent words. **PMI** was calculated

from the probability that the words co-occur in sentences, divided by the product of the single occurrence probabilities of the words. The resulting term was log-normalized (for details see Bouma, 2009). For **PPMI**, values lower than zero were set to zero. **Joint frequency** is a count of the number of sentences in which both words co-occur. In addition to the log-transformation, we set the value of the predictor to zero when the words did not co-occur at all.

To train **CBOW** and **skip-gram** models, we used gensim 3.0.0 (Rehurek & Sojka, 2010; cf. Mikolov, Chen et al., 2013). We used 1000 hidden units, excluded words with a frequency of 10 or lower, and trained the model in 10 iterations to predict the words of the 43 million sentences, respectively. To adjust for the influence of the most frequent words, we relied on negative sampling (k = 10; Mikolov, Sutskever Chen, Corrado, & Dean, et al., 2013). As all other co-occurrence measures were based on a log-transformed scale, we explored this possibility for the word2vec predictors as well. Because the cosine distances can become negative, we added their minima to make all predictors positive. We log10-transformed the resulting values. When we examined the item-level variance based on all 900 stimuli, log-transformed CBOW distances accounted for 25% of the item-level variance, which is substantially lower than the 35% of the raw CBOW measure. Log-transformed skip-gram data accounted for 29%, in contrast to 43% item-level variance explained by the raw skip-gram model. Since the nonlinearity provided by the word2vec model seems to provide interval scaling properties that make an additional log transform unnecessary, we did not log-transform these predictors.

Word frequency measures were taken from the same corpus. To facilitate the comparison of corpora of different sizes, Leipzig word frequency classes relate the frequency of each word to the frequency of the most frequent word using the definition that the most common German word "der" is 2^{class} more frequent than the word of which the frequency is given (Quasthoff et al., 2006). Using this measure, more frequent words are assigned to lower word frequency classes. To make

this measure more intuitive (i.e. so that more frequent words are given higher frequency values), we multiplied this class by -1.

The experiential measures of study 1 were taken from the revised Berlin Affective Word List (BAWL-R; see Võ et al., 2009, for details). Stimulus selection for studies 2 and 3 was based on an extended version of the BAWL including about 9,000 words (Conrad et al., in prep.; cf. Võ et al., 2009). The emotional **valence** ratings are based on a 7-point scale ranging from -3 (very negative) through 0 (neutral) to 3 (very positive). The 7-point **imageability** scale ranges from 1 (low imageability) to 7 (high imageability). **Arousal** ratings are based on a 5-point scale ranging from 1 (low arousal) to 5 (high arousal).

We selected 300 noun pairs as stimuli for each study. For the predictor variables derived from the single-word features of frequency, valence, arousal, and imageability, we computed the interaction of the respective feature values of both words as the product of those values. As can be seen in rows and columns 1-5 of Table 1, AS, word frequency, emotional valence, arousal, and imageability provided a maximum correlation of 0.3 with each other predictor variable. For the exploratory predictor variables in rows and columns 6-11, there are larger correlations. The distributional features of the predictor variables can be examined in Table 2.

--- insert Tables 1 and 2 about here ---

Data analysis

In each study, after testing how much variance in participant responses could be explained by each of the single predictors alone, we used LME modeling on the single-trial data as implemented by the lme4 r-package (version 1.1-12). In addition to including random intercepts for subjects and items, we tested whether the trial number should be included in the baseline model to

account for fatigue or familiarization with the task. All independent variables were centered before the analyses, except for those providing an interpretable 0 score (i.e. AS, PMI, and PPMI).

In the planned analyses for the three studies, we tested whether AS, and the interaction of the single-word features (frequency, emotional valence, arousal and imageability) could account for unique variance in the association rating task. We selected the stimuli of the three studies so that these predictor variables provided a low to moderate correlation, i.e. r values ~ 0.3 (Cohen, 1988). Fulfilling this criterion assures that the tolerance is 0.91 or larger, and the variance inflation is 1.09 or lower, ensuring that multicollinearity was not critical (O'Brian, 2007). We started by entering all the predictors, and then removed those that did not enter into the model significantly, using two standard errors as significance criterion (i.e. $t \geq 2$; cf. Baayen et al., 2008, footnote 1; Masson & Kliegl, 2013). If a predictor did not surpass that threshold, we excluded it from the subsequent model. After each predictor was entered or removed, we tested whether there was a significant difference in the log likelihood of the resultant model compared to the previous model, keeping more complex models if they show a significant difference and keeping less complex models if they did not.

In addition to examining qq-plots and the residual distribution, we used Kolmogorov-Smirnov (KS) tests to decide whether the residuals significantly deviate from normality. If they did, we trimmed the models by excluding trials in which the residuals deviated more than two standardized residuals from the mean.

The amount of explained variance for the fixed effects and the total variance explained were taken from the marginal and conditional r^2 values, as computed by the *r.squaredGMM* function from the *MuMin* R package (version 1.40.0).

In exploratory analyses, we additionally tested the predictive power of the number of common associates, (P)PMI, joint frequency, and CBOW and skip-gram cosine distances. The inclusion of these exploratory predictors led to some large intercorrelations between these variables

(see Table 1). To interpret the intercorrelational patterns of the predictors, we conducted a cluster analysis (Baayen, 2007; Montefinese, Ambrosini, Fairfield, & Mammarella, 2014), by pooling the 900 word pairs of the three studies and computing the Spearman correlations between the predictor variables (Baayen, 2008). Then we performed an agglomerative hierarchical cluster analysis using R package *hclust* (version 3.2.2). The dendrogram can be examined in Figure 1.

--- insert Figure 1 about here ---

The largest of the four clusters consisted of the co-occurrence-based measures of AS, common associates, (P)PMI, and CBOW and skip-gram cosine distances. The second cluster consisted of word frequency and arousal. The third cluster contained the experiential predictors of emotional valence and imageability. The fourth cluster contained joint frequency only. In performing the exploratory analyses we proceeded by entering variables from each of the clusters, from the largest to the smallest. The model selection procedure was identical to that described above for the planned analyses.

Study 1

Participants

Thirty-four native German subjects participated in study 1 (24 female; age: $M = 27.65$ years, range 19 to 57, SE = 1.67). The participants reported no language or speech impairment, and either participated voluntarily or received course credits.

Results and Discussion

Item-level analyses. The top predictors of the first cluster were skip-gram cosine distances, accounting for 55% of the item-level variance, followed by CBOW cosine accounting for 48%.

Correlation coefficients of the predictors with the mean association ratings can be examined in the last column of Table 1. The two word2vec predictors were strongly correlated ($r = 0.93$). AS accounted for 46% of the variance in the mean association ratings. PPMI scored slightly better than PMI with r-squared values of 41% vs. 40%, followed by the number of common associates with 28%.

While the predictor variables of the second cluster, i.e. word frequency and arousal, showed no significant correlation, the predictors of the third cluster showed significant correlations. Emotional valence and imageability accounted for 12% and 5% of the item-level variance, respectively. Finally, joint frequency accounted for 38% of the item-level variance.

Planned LME analyses. With 300 word pairs and 34 participants, the analyses were based on 10,200 rows of data. The modeling is summarized in Table 3. The best model (Model M3 in Table 3) included significant effects of all fixed factors, i.e. AS, word frequency, emotional valence, arousal, and imageability. When testing for the normality of the residuals using a KS-test, we found a significant deviation from normality ($D = 0.025$, $p < 0.001$). Therefore, we removed all 548 outlier trials deviating more than two standardized residuals from the mean (e.g. Montefinese, Zanino, & Ambrosini, 2015). When we re-fit the best model based on the remaining 9,652 trials, the KS-Test revealed no significant deviation from normality ($D = 0.010$, $p = 0.24$). AS, word frequency, emotional valence, arousal and imageability remained significant predictors (see Table 4). The fixed effects predictors accounted for 43% of the variance. Together with the random terms, the full model accounted for 80% of variance.

--- insert Table 3 and 4 about here ---

Exploratory LME analyses. The model selection process revealed that model 8 was the best model (model M8 in Table 3), which included AS, number of common associates, skip-gram cosine

distances, emotional valence, and imageability. The KS-Test revealed a significant deviance from normality ($D = 0.025, p < 0.001$). We therefore removed trials deviating more than two standardized residuals from the mean, leaving us with 9,566 trials. The KS-Test revealed no significant deviation from normality in the final model ($D = 0.011, p = 0.23$; see Table 5). The fixed effects of this final model accounted for 50% of the variance, while the full model including random intercepts for subject and items accounted for 80%.

--- insert Table 5 about here ---

Study 2

Participants

In study 2, we recorded data from 31 German native speakers (17 female; age: $M = 27.65$; range 20 to 61; $SE = 1.84$). reporting no language or speech impairment. They either participated voluntarily or received course credits.

Results and Discussion

Item-level regressions. Word2vec skip-gram cosine distances accounted for the greatest portion (48%) of item-level variance in the first cluster, followed by AS with 37%. The CBOW distances accounted for 35%, while the number of common associates accounted for 28%. Across the words pairs of study 2 with a positive PMI score, PPMI accounted for 6% of the item-level variance. In the second cluster, frequency accounted for 6% as well, and arousal accounted for a significant portion of 1%. Emotional valence in the third cluster did not account for a significant portion of variance, while imageability accounted for 6%. Joint frequency, the only predictor of the last cluster, accounted for 44% of the item-level variance.

Planned LME analyses. The analyses, based on 9,300 rows of data, are summarized in Table 6. The best model included effects of frequency, valence, and imageability, but no effect of arousal (Model M4 in Table 6). The KS-Test revealed that the residuals were not normally distributed (D = 0.029, $p < 0.001$). Thus, we excluded all trials which deviated more than two standardized residuals from the mean. After the exclusion of these 518 trials, the KS-Test still indicated a non-normal distribution of the residuals ($D = 0.018$, $p = 0.0052$). When we examined the data more closely, we found that there was an unusually high number of '1' (not associated) responses that resulted in skewed response and residual distributions. We believe that this is because the stimulus selection constraint of a low correlation of AS and PPMI resulted in the selection of many words that are not associated. To test whether this constrained the interpretability of our LME, we excluded 3,005 trials in which the response '1' (not associated) was given, leaving 5,996 trials for analyses. The KS-Test revealed no significant deviance from normality of the residuals in these trials ($D = 0.013$, $p = 0.27$). The model did not change (see Table 4). The fixed effects in the final model accounted for 25% of the variance. The total amount of variance explained was 70%.

--- insert Table 6 about here ---

Exploratory LME analyses. For this analysis, we used only PPMI as predictor, because the stimuli were selected such that all PMI values were positive. The model selection procedure is shown in Table 6. The best model included AS, skip-gram cosine, and joint frequency (Model M15 in Table 6). The KS-test revealed a significant deviance from normality (D = 0.030, $p < 0.001$), which even remained when removing trials exceeding the criterion of more than two standardized residuals from the mean ($D = 0.019$, $p = 0.0038$). To correct for the unusually strong left-skewed response distribution, we then removed all trials in which the participants responded '1', which left us with 5,930 trials for the final analyses. The same predictors remained significant, and the KS-test

revealed no significant deviance from normality (D = 0.016, p = 0.08). The fixed effects accounted for 26% of the variance, while the full model accounted for 59% of the variance (see Table 5).

Study 3

Participants

17 of the 31 participants were female (age: M = 29.00; range: 19 to 57; *SE* = 1.72). All participants were native German speakers, reported no language or speech impairment, and either participated voluntarily or received course credits.

Results and Discussion

Item-level regressions. The item level benchmarks of the third study revealed that skip-gram cosine is the strongest predictor of association ratings, explaining 50% of the variance. Among the other predictors of the first cluster, number of common associates accounted for 41% of the variance, directly followed by AS and CBOW distances, both accounting for 40%. PPMI could account for 30% of the variance, while PMI accounted for 10% only. Within the second cluster, word frequency did not account for variance, but arousal accounted for 2%. Of the other experiential predictors in the third cluster, emotional valence accounted for 7%, but imageability failed to reach significance. Finally, joint frequency accounted for 27% of the item-level variance.

Planned LME analyses. The planned analyses, summarized in Table 7, were conducted on 9,300 rows of data. The best model (Model M4 in Table 7) included AS, emotional valence and imageability. Since the KS-Test indicated non-normality of the residuals ($D = 0.021, p < 0.001$), we excluded trials in which the residual deviated more than two standardized values from the mean, leaving 8,834 trials for analysis. The KS-test revealed no significant deviation from normality (D =

0.013, $p = 0.089$). The fixed effects accounted for 26% of the variance, while the full model accounted for 70% of the variance (see Table 4).

--- insert Table 7 about here ---

Exploratory LME analyses. The exploratory analyses of study 3 are summarized in Table 7. The best model (Model M9 in Table 7) included AS, skip-gram cosine, and arousal. A KS-Test indicated non-normality ($D = 0.021$, $p < 0.001$), thus we removed trials deviating more than two standardized residuals from the mean. After that the KS-Test revealed no significant deviance from normality ($D = 0.014$, $p = 0.07$). In all, the model accounted for 70% of the variance, while the fixed effects accounted for 33% of the variance (see Table 5).

General Discussion

In the item-level benchmarks, skip-gram cosine accounted for approximately 50% of the variance in all three studies. AS performed as well as CBOW cosine, reproducibly accounting for approximately 40% of the item-level variance. In study 3, the number of common associates performed at a similar level, while its performance dropped to around 30% in the others. PMI and PPMI showed the largest variability in item-level performance, ranging from around 40% for both predictors in study 1 down to 6% for PMI in study 2. Joint frequency reproducibly accounted for around 30% of the variance. The only experiential predictor that was significant in the item-level analyses of all studies was arousal accounting for at most 2% of the variance. Moreover, the correlation was positive in study 1 and study 3, and negative in study 2, thus supporting the notion that it is fitting error variance.

When selecting the stimuli with a low to moderate correlation between AS and the single-word features for the planned multiple-predictor analyses, our LME analysis revealed that not only AS, but also emotional valence reproducibly accounted for variance. The occurrence of arousal and imageability effects depended on the specific selection of stimuli. Emotional valence, in contrast, reproducibly accounted for additional association rating performance, which makes it unlikely that the result was merely obtained by (over-)fitting the models to error variance (e.g., Pitt et al., 2002). The inclusion of this predictor suggests that two words are more likely to elicit high association ratings when both words are rated to be highly emotional with a consistent positive or negative valence. However, when included in a competitive cluster-wise analysis, emotional valence failed to reach significance in two of three studies. Rather, we found that only AS and skip-gram cosine reproducibly account for variance in the association rating task.

Experiential-affective word features and semantic cohesion

The planned LME analyses consistently revealed emotional valence effects in all three studies (e.g. Andrews, Vigliocco, & Vinson, 2009; Jacobs et al., 2015; Montefinese et al., 2015). The added value of emotional valence confirms Bower's (1981) proposal that positive or negative emotional valence can be theoretically captured by a memory unit associated with a word. When both words have been consistently rated as positive or negative, they also provide greater association ratings. These affective consistency effects are consistent with a recent theory of affective evaluation (Schröder & Thagard, 2013).

We further tested whether high imageability words relied not only on associations between verbal representations, but also on imageability (Paivio, 1969). This prediction was confirmed in the planned analyses of study 1 and study 2. We speculate that the missing imageability effect in study 3 may result from its relation to emotional valence, which was our most robust rating-based predictor of association rating performance. Though there may also be concreteness attributable to

other sensory modalities (e.g., Barsalou, 1999), imageability ratings quantify the concreteness only of the visual modality. However, because it correlates strongly with concreteness ratings (e.g., r = .85 in Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011), imageability can be treated as tentative empirical proxy to concreteness. Vigliocco and colleagues (2014) found that emotionally valenced words provide lower concreteness ratings, i.e. tend to be more abstract. Further, Kousta and colleagues (2011) showed that word recognition effects of concreteness can be absorbed by emotional valence. Our reproducible effects of emotional valence in absence of imageability effects supported this conclusion. Kousta et al. (2011) suggest that abstract mental representations develop from words referring to the internal world. As this internal world becomes very concrete in affective bodily expressions (Barrett, 2006), the semantic distances to basic emotion terms referring to facial expressions can nicely account for emotional valence, imageability and word recognition variance (e.g., Ekman, Sorenson, & Friesen, 1969; Westbury et al., 2013, 2014). Therefore, emotional valence can be considered a semantic superfeature that is abstracted by generalizing over specific emotions (cf. Jacobs et al., 2015).

When skip-gram cosine was included in the exploratory analyses, valence effects disappeared in two out of three studies. There are three potential reasons for this finding. First, emotional valence effects are relatively weak. During word recognition, for instance, Kuperman and colleagues (2014) found that it accounts for about 2% of the variance (cf. e.g. Hofmann et al., 2009; Larsen, Mercer, Balota & Strube, 2008). Second, when trying to account for other human performance, predictor variables derived from other human performance, such as ratings, provide the disadvantage that they contain error variance (e.g., Westbury et al., 2013). Though valence and arousal ratings provide good reliability scores ($r \sim .9$; see Kanske & Kotz, 2010), the error variance obtained during a first rating study adds to the error variance obtained during the subsequent study. Thus, when considering whether experiential and/or corpus-based approaches are a better explanation (Andrews et al., 2009), we suggest that researchers should try to get as far as possible to

explain the subjective (e.g. ratings) by the non-subjective, because we can understand better what an algorithm does than what a human rater does. The subjective is the explanandum of psychology, i.e. the thing to be explained. Using subjective judgments as explanans (i.e. as an independent variable) causes circularity (cf. Hempel & Oppenheim, 1948; Westbury, 2016). Co-occurrence-based approaches represent a deeper level of explanation because they provide the key components of memory, i.e. a learning history, and a formal account how this "experience" translates to microstructure, as well as retrieval of the "remembered" long-term association. The third reason why valence effects may be absorbed by co-occurrence-based approaches is the semantic cohesiveness hypothesis of affective word processing (Phelps et al., 1998; Maratos, Allen & Rugg, 2000; cf. Hofmann & Jacobs, 2014, for an overview). This hypothesis proposes that at least some word recognition variance previously ascribed to affective word features can be more parsimoniously explained by the fact that strongly valenced words have a larger semantic cohesiveness. Thus, if a model contains semantic relations, there may be less need for an additional affective evaluation mechanism (Koch, Alves, Krüger & Unkelbach, 2016; Hofmann & Jacobs, 2014). The present study supported a strong version of the semantic cohesion hypothesis of affective word processing. Rather than seeing a confound between emotion and semantics, we favor the theoretical explanation that emotional valence is one of the cardinal dimensions constituting semantic space (e.g., Osgood, Suci, & Tannenbaum, 1957).

From contiguity to semantic structure

A very simple approach to word co-occurrence is provided by AS, which is used within a localist connectionist model of semantic processes (Hofmann et al., 2011; Hofmann & Jacobs, 2014). While all dual-route models propose a semantic layer at a pre-quantitative level of theorizing (Coltheart et al., 2001; Perry et al., 2007), the Associative Read-Out Model (AROM) is the first interactive activation model that provides a fully implemented semantic layer and thus a complete

whole-word route (Coltheart et al., 2001; Perry et al., 2007). It was designed as a general approach to all tasks activating semantic long-term representations (Collins & Loftus, 1975; Hofmann et al., 2011; McNamara, 2005). Initially, the AROM was introduced to account for episodic memory: When a word has many associated items in a recognition memory task, participants are more likely to respond that the word was an old word learned in the study phase. Associative connections not only account for a semantically induced false memory effect (Roediger & McDermott, 1995), but they also successfully predict a boost of recognition memory performance in studied words (Hofmann et al., 2011). Recently, we showed that associative connections within the stimulus set predict P200 effects of lexical access, as well as semantic integration effects at the level of the N400 (Stuellein et al., 2016).

The first word-pair-based test of AS was a re-analysis of Forgász et al.'s (2012) data (Hofmann & Jacobs, 2014). It revealed that the AS between the nouns of a compound can make parametric predictions of the hemodynamic response in the left inferior frontal gyrus (Hofmann & Jacobs, 2014). The most significant voxels crossed a statistical threshold of $p < 0.005$ after full Bonferroni correction for more than 90,000 statistically independent comparisons. The present results complement this account by providing reproducible predictions of a behavioral performance measure, suggesting that the predictive power of AS can not only be generalized vertically across behavioral and neurocognitive levels of analysis, but also horizontally across different stimulus sets (Jacobs & Grainger, 1994; Jacobs & Hofmann, 2013; Hofmann & Jacobs, 2014).

To bridge the gap between models addressing eye movement control during natural reading and models visual word recognition (Engbert et al., 2005; Reichle, Rayner, & Pollatsek, 2003; Reilly & Radach, 2006), Radach and Hofmann (2016, Fig. 2) recently demonstrated how the aggregation of associative energy of multiple prime words may account for the cloze-completion-based predictability of a target word from sentence context. There is also empirical evidence that

high AS to preceding words shortens first fixation and first-pass gaze durations during sentence reading (Hölscher, 2017).

These results suggest that rather than tediously collecting different types of association performance pre-experimentally, future studies can simply use AS to quantify associations between words. It is well known that such a simple statistical approach to within-sentence co-occurrence can account for free association performance in adults (Rapp & Wettler, 1991). Moreover, recent research revealed that AS predicts more reliable priming effects at a 1000 ms lag between prime and target, thus mirroring the classic finding of greater, free association-based priming at such a long stimulus onset asynchrony (SOA; Lucas, 2000; Roelke et al., 2018). At a short, 200 ms SOA, in contrast, more reliable effects of the number of common associates are obtained in adults. In all, this response pattern suggests that associative and semantic priming during lexical decision can be predicted by AS and the number of common associates, respectively (Roelke et al., 2018). A free association-based definition of associative priming, however, is particularly problematic, because there is a developmental shift from syntagmatic to paradigmatic free association responses from the first to the fifth grade in school (Nelson, 1977). Therefore, primed lexical decisions of fourth graders profit from the number of common associates, while the number of common associates is associated with increased error rates in the second grade (Franke et al., 2017). A direct association between prime and target, in contrast, decreases error rates for both samples, though at the expense of response speed in second graders. Franke and colleagues (2017) propose that semantic structure develops first from simple, syntagmatic contiguity learning as captured by AS, while higher-order structure is later developed by generalizing across many common associates. When comparing the number of common associates to a classic definition of semantic priming (Lucas, 2000), synonyms can simply be found, for instance, by searching for two words providing many common associates, and excluding word pairs that directly co-occur (Rapp, 2000), because synonyms are typically embedded in similar sentence contexts (Rubenstein & Goodenough, 1965).

While the number of common associates did not reproducibly add explained association rating variance in the present studies, the skip-gram model as an alternative paradigmatic measure of semantic similarity did (e.g. Frank & Willems, 2017). A potential reason for this may be sought in the reduction of the dimensionality (e.g. Griffiths et al., 2007; Landauer & Dumais, 1997). Rather than using only those contextual feature words that are apparent, the skip-gram model searches for a set of latent representations that predict the contextual words by a target. Therefore, the skip-gram model generates "latent semantic dimensions" that rely not only on the words that are apparent, but also on similar words.

One type of criticism of co-occurrence based accounts is that the computational linguist selects several parameters, e.g. the number of latent semantic dimensions or connections from and to hidden units, or other 'free parameters' such as the size of the contextual window, which hampers the generality of the results and explains a part of the success of word2vec models (Levy, Goldberg, & Dagan 2015; Mandera, Keulers, & Brysbaert, 2015). The selection of an appropriate number of dimensions is an optimization problem that must be addressed by dimension-reducing approaches (e.g., Griffiths et al., 2007; Landauer & Dumais, 1997; Mikolov, Chen, et al., 2013; Mandera et al., 2015). AS, however, has no latent semantic dimensions or hidden units to be optimized, thus reducing the number of free parameters. For AS, the decision to use the sentence level as the next larger contextual window was theoretically motivated, because we see it as the logically next higher contextual increment on top of the orthographic word layer of interactive activation models (Jacobs & Grainger, 1996; McClelland & Rumelhart, 1981). After visual features, letters and orthographic word forms, the sentence level logically seems to be the next larger grain size (cf. Ziegler & Goswami, 2005). A paragraph and discourse level for including long-range semantics might be the next increments (Biemann et al., 2015).

When thinking about how semantic information is represented in the cognitive system, the symbolic approach of the AROM successfully complements the subsymbolic approach of the skip-

gram model (e.g., Griffith et al., 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996), at least when accounting for association rating variance. While the latter may be used to conceptualize semantic memory in theories in which the memory representation is distributed across subsymbolic units (e.g., Harm & Seidenberg, 2004; McClelland & Chappel, 1998; Steyvers et al., 2006), direct links between word units can be used to define semantic long-term associations in symbolic models (Anderson, 1983; Anderson et al., 2004; Hummel & Holyoak, 2004; McClelland & Rumelhart, 1981).

Conclusions

We conclude that the association rating task is a useful task to benchmark differential computational approaches to the association between two words. Our results show that a syntagmatic approach such as AS, together with a paradigmatic approach such as the skip-gram model reproducibly accounts for a great deal of association rating variance. While we think that emotional valence is still a useful theoretical construct, its association rating variance can be often explained by co-occurrence-based approaches, probably because sharing the internal world with others is one of the reasons why language evolved.

Acknowledgements

We like to thank the students that helped with data collection, as well as Andre Roelke, Chris Vorstius and Lars Kuchinke for useful comments. This paper was part of a grant of the Deutsche Forschungsgemeinschaft to Markus Hofmann (HO 5139/2-1 and 2-2).

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036-1060.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–98.

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

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Barrett, L.F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28-58.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(4), 637-660.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Biemann, C., Remus, S., & Hofmann, M. J. (2015). Predicting word 'predictability' in cloze completion, electroencephalographic and eye movement data. *Proceedings of Natural Language Processing and Cognitive Science* (pp.1-10), Krakow, Poland.

Bullinaria, J., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, *39*(3), 510–26.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of GSCL* (pp. 31–40).

Bower, G. H. (1981). Mood and memory. *The American psychologist*, *36*(2), 129–48.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–56.

De Saussure, F. (1959). *Course in General Linguistics*. New York: Philosophical Library.

Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading : A study with fixation-related brain potentials. *Neuroimage*, *62*, 381–393.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*, 61–74.

Ekman, P. Sorenson, E.R. & Friesen, W.V. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science*, *164*(3875), 86-88.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *211*, 1–28.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777-813.

Evert, S. (2005). *The Statistics of Word Cooccurrences Word Pairs and Collocations*. Universität Stuttgart.

Forgács, B., Bohrn, I., Baudewig, J., Hofmann, M. J., Pléh, C., & Jacobs, A. M. (2012). Neural correlates of combinatorial semantic processing of literal and figurative noun noun compound words. *NeuroImage*, *63*(3), 1432–42.

Franke, N., Roelke, A., Radach, R., & Hofmann, M. J. (2017). After braking comes hasting: reversed effects of indirect associations in 2nd and 4th graders. In *Proceedings of the Cognitive Science Society* (pp. 2025-2030), London, UK.

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192-1203.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review*, *103*(3), 518–65.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation, *Psychological Review, 114*(2), 211–244.

Harm, M., & Seidenberg, M. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135-175.

Hofmann, M. J., Biemann, C., & Remus, S. (2017). Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements. In: Sharp, B., Sedes, F., Lubaszewski, W. (Eds.). Cognitive Approach to Natural Language Processing, ISTE press, Elsevier.

Hofmann, M. J., & Jacobs, A. M. (2014). Interactive Activation and Competition Models and Semantic Context: From Behavioral to Brain Data. *Neuroscience and Biobehavioral Reviews, 46*, 85-104.

Hofmann, M. J, Kuchinke, L., Biemann, C., Tamm, S., & Jacobs, A. M. (2011). Remembering words in context as predicted by an associative read-out model. *Frontiers in Psychology*, *2*, 252.

Hofmann, M. J., Kuchinke, L., Tamm, S., Võ, M. L., & Jacobs, A. M. (2009). Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(4), 389-397.

Hölscher, U. (2017). Assoziative und semantische Effekte beim natürlichen Lesen: Eine Blickbewegungsstudie. Master thesis, University of Wuppertal.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220.

Hutchison, K. A., Balota, D. A., Cortese, M. J. & Watson, J. M. (2008). Predicting semantic priming at the item-level, *Quarterly Journal of Experimental Psychology, 61,* 1036-1066.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1311–1334.

Jacobs, A. M., & Hofmann, M. (2013). Neurokognitive modellierung. *Enzyklopädie der Psychologie. Affektive und Kognitive Neurowissenschaft*. Hogrefe, Göttingen, 431-447.

Jacobs A. M., Võ, M. L.-H., Briesemeister, B. B., Conrad, M., Hofmann M. J., Kuchinke L., Lüdtke, J, & Braun, M. (2015). 10 years of BAWLing into affective and aesthetic processes in reading: what are the echoes? *Frontiers in Psychology*, *6*, 1-15.

Johns, B. T., & Jones, M. N. (2015). Generating Structure From Experience: A Retrieval-Based Model of Language Processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*. *69*(3), 233–251

Jiang, J. J. & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference Research on Computational Linguistics* (pp. 1–15). Taiwan.

Jung, C.G. (1905). *Ueber das Verhalten der Reaktionszeit beim Assoziationsexperimente*. Leipzig: Ambrosius Barth.

Kanske, P., & Kotz, S. A. (2010). Leipzig affective norms for German: A reliability study. *Behavior Research Methods*, *42*(4), 987-991.

Kliegl, R., Risse, S. & Laubrock, J. (2007). Preview Benefit and Parafoveal-on-Foveal Effects from Word N+2. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1250–1255.

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A General Valence Asymmetry in Similarity: Good Is More Alike Than Bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1171-1192.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The Representation of Abstract Words: Why Emotion Matters. *Journal of Experimental Psychology: General*, *140*, 14–34.

Kuchinke, L., Fritzemeier, S., Hofmann, M. J., & Jacobs, A. M. (2013). Neural Correlates of Episodic Memory: Associative Memory and Confidence Drive Hippocampus Activations. *Behavioural Brain Research, 254*, 92-101.

Kuhlmann, M., Hofmann, M. J., Briesemeister, B. B., & Jacobs, A. M. (2016). Mixing positive and negative valence: Affective-semantic integration of bivalent words. *Scientific Reports*, *6*, 1-7.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. Journal of Experimental Psychology: General, *143*(3), 1065-1081.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not All Negative Words Slow Down Lexical Decision and Naming Speed: Importance of Word Arousal. *Emotion*, *8*(4), 445-452.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225.

Lucas, M. (2000). Semantic priming without association: a metaanalytic review. *Psychonomic Bulletin & Review*, *7*, 618–630.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203-208.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables?. *Quarterly Journal of Experimental Psychology*, *68*(8), 1623-1642.

Maratos, E.J., Allan, K., & Rugg, M.D. (2000). Recognition memory for emotionally negative and neutral words: an ERP study. *Neuropsychologia, 38*, 1452–1465.

Masson, M. E., & Kliegl, R. (2013). Modulation of Additive and Interactive Effects in Lexical Decision by Trial History. *Learning, Memory*, *39*(3), 898-914.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–60.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings. *Psychological Review*, *5*, 375–407.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: mediated priming revisited. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *18*, 1155–72.

McNamara, T. P. (2005). Semantic priming: Perspectives from memory and word recognition. *Psychology Press*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 1-12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, G. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography, 3,* 245-264.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751).

Miller, W., & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, *6*, 1–28.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). Semantic significance: a new measure of feature salience. *Memory & Cognition, 42*, 355-369.

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research, 79*(5), 785-794.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: a review of research and theory. *Psychological Bulletin*, *84*(1), 93.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5), 673–690.

Olson, M. H., & Hergenhahn, B. R. (1982). *An introduction to theories of learning*. Prentice Hall.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76*(3), 241-263.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review*, *114*, 273–315.

Phelphs, E.A., LaBar, K.S., Anderson, A.K., O'Connor, K.J., Fullbright, R.K., & Spencer, D.D. (1998). Specifying the contributions of the human emotional memory: a case study amygdala to. *Neurocase 4*, 527–540.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of LREC-06* (pp. *1799-1802*). Genova, Italy.

Quené, H. & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413-425.

Radach, R. & Hofmann, M. (2016). Graphematische Verarbeitung beim Lesen von Wörtern. In U. Domahs & B. Primus, Laut, Gebärde, Buchstabe (Handbuch Sprachwissen, Band 2), pp. 455-473, De Gruyter.

Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Association for Computational Linguistics, Proceedings of the 19th International Conference on Computational Linguistics* – Volume 1, pp. 1–7.

Rapp, R., & Wettler, M. (1991). Prediction of Free Word Associations Based on Hebbian Learning. In *Proceedings of the Joint Conference on Neural Networks* (pp. 25–29).

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4), 445-476.

Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, 7(1), 34-55.

Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A. M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin & Review*, *16*, 600-608.

Roediger, H. L. I., & McDermott, K. B. (1995). Creating False Memories : Remembering Words Not Presented in Lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *21*, 803–814.

Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8,* 385-407.

Roelke, A., Franke, N., Biemann, C., Radach, R., Jacobs, A. M., & Hofmann, M. J. (in press). A novel co-occurrence based approach to predict pure associative and semantic priming. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-018-1453-6

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627-633.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review, 120*, 255-280.

Shaoul, C., & Westbury, C. F. (2006). Word frequency effects in high-dimensional co-occurrence models : a new approach. *Behavior Research Methods*, *38*, 190–195.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411-416.

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*(7), 327–34.

Stuellein, N., Radach, R., Jacobs, A. M., & Hofmann, M. J. (2016). No one way ticket from orthography to semantics in recognition memory: N400 AND P200 effects of associations. *Brain Research, 1639,* 88-98.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reason- ing: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: the role of emotion. *Cerebral Cortex*, *24*(7), 1767-1777.

Võ, M. L.-H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*, 534–8.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M J. & Jacobs, A. M. (2015). Avoid violence, rioting and outrage; Approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology*. *68*, 1599-1622.

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it , now you don ' t : on emotion , context , and the algorithmic prediction of human imageability judgments. Frontiers in Psychology, *4*(991), 1–13.

Westbury, C. (2016). Pay no attention to that man behind the curtain. *The Mental Lexicon*, *11*(3), 350-374.

Westbury, C. , Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*(991), 1–13.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*, 3-29.

Figure Caption

Figure 1 shows the result of the hierarchical cluster analysis of the 11 predictors of the present study.

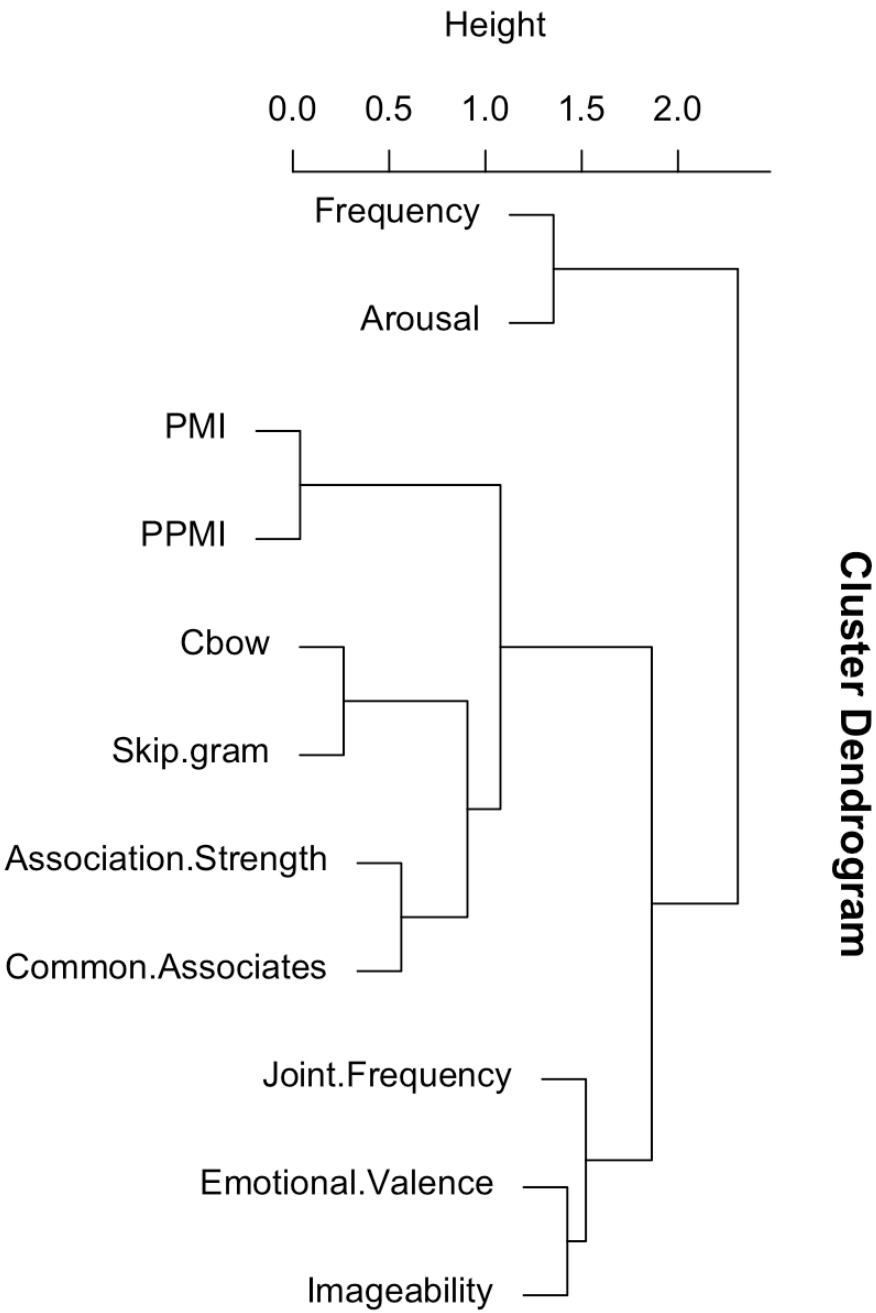
Height
0.0
0.5
1.0
1.5
2.0
Frequency
Arousal
PMI
PPMI
Cbow
Skip.gram
Association.Strength
Common.Associates
Joint.Frequency
Emotional.Valence
Imageability
Cluster Dendrogram

Table 1. Correlation coefficients (top triangle) and two-tailed significance (bottom triangle) of the moderately correlated planned (columns 1-5), and the exploratory predictors (columns 6-11) and the mean association ratings (column 12) of study 1/ 2/ 3.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Association Strength		-0.27/ -0.21/ -0.20	0.18/ -0.06/ 0.18	0.02/ -0.26/ -0.01	0.16/ 0.20/ 0.08	0.72/ 0.52/ 0.77	0.76/ 0.31/ 0.49	0.79/ 0.31/ 0.80	-0.95/ 0.80/ 0.84	0.61/ 0.46/ 0.49	0.75/ 0.59/ 0.67	0.68/ 0.61/ 0.63
2. Frequency	***/ ***/ ***		-0.07/ -0.06/ -0.09	-0.01/ -0.25/ -0.06	-0.07/ 0.07/ -0.15	0.28/ 0.59/ 0.31	-0.31/ -0.32/ 0.48	-0.25/ -0.32/ -0.16	0.41/ 0.57/ 0.57	-0.14/ 0.06/ -0.08	-0.01/ 0.13/ -0.06	-0.03/ 0.25/ 0.02
3. Emotional Valence	**/ 0.32/ **	0.21/ 0.32/ 0.14		0.15/ 0.32/ 0.18	0.07/ -0.19/ 0.07	0.18/ 0.00/ 0.18	0.19/ 0.05/ 0.06	0.19/ 0.05/ 0.22	-0.15/ -0.07/ 0.11	0.24/ 0.05/ 0.19	0.25/ 0.05/ 0.28	0.35/ 0.06/ 0.26
4. Arousal	0.71/ ***/ 0.83	0.86/ ***/ 0.33	**/ ***/ **		-0.12/ -0.23/ -0.22	0.10/ -0.24/ 0.09	0.03/ 0.01/ -0.08	0.03/ 0.01/ 0.00	-0.01/ -0.30/ -0.05	0.10/ -0.14/ 0.14	0.10/ -0.13/ 0.13	0.12/ -0.12/ 0.15
5. Imageability	**/ ***/ 0.17	0.2/ 0.24/ **	0.24/ ***/ 0.22	*/ ***/ ***		0.06/ 0.16/ 0.02	0.21/ 0.06/ -0.02	0.20/ 0.06/ 0.12	-0.16/ 0.25/ 0.02	0.16/ 0.34/ 0.02	0.15/ 0.31/ 0.05	0.23/ 0.24/ 0.02
6. Common Associates	***/ ***/ ***	***/ ***/ ***	**/ 0.96/ **	0.1/ ***/ 0.12	0.31/ **/ 0.73		0.54/ 0.05/ 0.39	0.55/ 0.05/ 0.52	-0.75/ 0.73/ 0.70	0.70/ 0.50/ 0.68	0.81/ 0.61/ 0.82	0.53/ 0.53/ 0.64
7. PMI	***/ ***/ ***	***/ ***/ ***	**/ 0.39/ 0.28	0.57/ 0.9/ 0.16	***/ 0.31/ 0.75	***/ 0.37/ ***		0.98/ 1.00/ 0.53	-0.72/ 0.24/ 0.58	0.66/ 0.31/ 0.16	0.73/ 0.38/ 0.21	0.63/ 0.25/ 0.32
8. PPMI	***/ ***/ ***	***/ ***/ **	***/ 0.39/ ***	0.63/ 0.9/ 0.96	***/ 0.31/ *	***/ 0.37/ ***	***/ ***/ ***		-0.74/ 0.24/ 0.52	0.66/ 0.31/ 0.42	0.74/ 0.38/ 0.58	0.64/ 0.25/ 0.55
9. Joint Frequency	***/ ***/ ***	***/ ***/ ***	**/ 0.26/ 0.06	0.83/ ***/ 0.37	**/ ***/ 0.77	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***		-0.57/ 0.50/ 0.37	-0.72/ 0.66/ 0.51	-0.62/ 0.66/ 0.52
10. CBOW Cosine	***/ ***/ ***	*/ 0.31/ 0.16	***/ 0.42/ ***	0.09/ */ *	**/ ***/ 0.75	***/ ***/ ***	***/ ***/ **	***/ ***/ ***	***/ ***/ ***		0.93/ 0.89/ 0.91	0.69/ 0.59/ 0.63
11. Skip-gram Cosine	***/ ***/ ***	0.9/ */ 0.31	***/ 0.41/ ***	0.07/ */ *	**/ ***/ 0.39	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***		0.74/ 0.69/ 0.71
12. Association Rating	***/ ***/ ***	0.65/ ***/ 0.75	***/ 0.29/ ***	*/ */ *	***/ ***/ 0.78	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	***/ ***/ ***	

* P < 0.05; ** P < 0.01; *** P < 0.001

Table 2. Distributional features of study 1/ 2/ 3.

	M	SD	Min	Max
Association Strength	1.09/ 1.18/ 1.12	1.17/ 0.82/ 1.36	0/ 0/ 0	4.01/ 3.24/ 5.00
Frequency	12.12/ 14.01/ 11.47	2.10/ 2.02/ 2.76	18/ 20/ 20	7/ 8/ 4
Emotional Valence	0.00/ -0.18/ 0.32	1.36/ 1.13/ 1.14	-2.90/ -3.00/ -2.8	2.90/ 2.56/ 2.9
Arousal	2.96/ 2.93/ 2.75	0.73/ 0.67/ 0.66	1.28/ 1.33/ 1.29	4.71/ 4.55/ 5.61
Imageability	4.19/ 4.81/ 4.53	1.38/ 1.23/ 1.45	1.44/ 1.78/ 1.29	6.88/ 6.89/ 6.88
Common Associates	86.50/ 50.22/ 141.81	82.93/ 61.49/ 138.53	2/ 0/ 0	431/ 391/ 683
PMI	0.19/ 0.27/ -10.20	0.15/ 0.13/ 24.10	-0.17/ 0/ -60.64	0.58/ 0.50/ 11.26
PPMI	0.20/ 0.27/ 1.93	0.13/ 0.13/ 2.28	0/ 0/ 0	0.58/ 0.50/ 11.26
Joint Frequency	51.29/ 10.85/ 223.61	179.04/ 30.88/ 1668.25	1/ 1/ 0	1641/ 405/ 28228
CBOW Cosine	0.25/ 0.26/ 0.39	0.17/ 0.15/ 0.19	-0.04/ -0.04/ -0.13	0.84/ 0.76/ 0.94
Skip-gram Cosine	0.21/ 0.21/ 0.30	0.14/ 0.10/ 0.18	-0.01/ 0.01/ 0.03	0.78/ 0.58/ 0.90
Word Length	6.11/ 6.11/ 6.69	1.23/ 1.34/ 2.28	4/ 4/ 3	8/ 8/ 18

Table 3: LME model summary for study 1. The final accepted models for planned and exploratory analyses are shown in bold.

Model Name	Type	Model Description	df	χ^2	p	ACCEPTED
M1	Planned	(1 | Subjects) + (1 | Items)				Baseline
M2	Planned	M1 + Trial Number	1	3.31	0.07	No
M3	**Planned**	**M1 + AS + Frequency + Valence + Arousal + Imageability**	**5**	**253.77**	**< 0.001**	**Yes**
M4	Exploratory	M1 + AS + Common Associates + (P)PMI + CBOW + Skip-gram	6	398.57	< 0.001	Yes
M5	Exploratory	M1 + AS + Common Associates + Skip-gram	3	5.43	0.14	Yes [Simpler]
M6	Exploratory	M1 + Frequency + Arousal	2	4.31	0.12	No
M7	Exploratory	M1 + Valence + Imageability	2	53.92	< 0.001	Yes
M8	**Exploratory**	**M5 + Valence + Imageability**	**2**	**23.82**	**< 0.001**	**Yes**
M9	Exploratory	M1 + Joint Frequency	1	145.17	< 0.001	Yes
M10	Exploratory	M8 + Joint Frequency	1	2.51	0.11	No

Table 4. Linear mixed effects model results for significant planned predictor variables of the three studies.

	Study 1			**Study 2**			**Study 3**		
Random Effects	**σ^2**	**SD**		**σ^2**	**SD**		**σ^2**	**SD**	
Item Intercept	1.576	1.255		1.199	1.095		1.644	1.282	
Subject Intercept	0.247	0.497		0.353	0.594		0.534	0.731	
Residual	1.002	1.001		1.046	1.023		1.507	1.225	
Fixed effects	**b**	**SE**	**t-value**	**b**	**SE**	**t-value**	**b**	**SE**	**t-value**
(Intercept)	2.823	0.134	21.033	2.894	0.159	18.157	2.847	0.164	17.404
Association Strength	1.128	0.067	16.623	0.974	0.082	11.818	0.767	0.056	13.653
Frequency	-0.011	0.002	-4.757	0.007	0.002	3.233			
Emotional Valence	0.173	0.037	4.673	0.169	0.064	2.659	0.139	0.049	2.866
Arousal	0.0522	0.022	2.362				0.067	0.025	2.700
Imageability	0.021	0.008	2.726	0.017	0.007	2.349			

Table 5. Linear mixed effects model results for the exploratory analyses.

	Study 1			**Study 2**			**Study 3**		
Random Effects	**σ^2**	**SD**		**σ^2**	**SD**		**σ^2**	**SD**	
Item Intercept	1.246	1.116		0.779	0.883		1.287	1.135	
Subject Intercept	0.250	0.500		0.339	0.582		0.533	0.730	
Residual	1.003	1.001		1.448	1.203		1.502	1.225	
Fixed effects	**b**	**SE**	**t-value**	**b**	**SE**	**t-value**	**b**	**SE**	**t-value**
(Intercept)	3.723	0.155	24.042	3.946	0.179	21.994	3.294	0.165	19.995
Association Strength	0.567	0.090	6.313	0.240	0.110	2.175	0.363	0.067	5.416
N Common Associates	-0.007	0.001	-4.748						
Skip-gram Cosine	9.360	0.899	10.414	4.859	0.744	6.534	4.893	0.514	9.525
Emotional Valence	0.143	0.033	4.342						
Arousal							0.044	0.022	2.002
Imageability	0.015	0.007	2.098						
Joint Frequency				0.286	0.070	4.076			

Table 6: LME model summary for study 2. The final accepted models for planned and exploratory analyses are shown in bold.

Model Name	**Type**	**Model Description**	**df**	χ^2	**p**	**ACCEPTED**
M1	Planned	(1 | Subjects) + (1 | Items)				Baseline
M2	Planned	M1 + Trial Number	1	2.73	0.098	No
M3	Planned	M1 + AS + Frequency + Valence + Arousal + Imageability	5	162.09	< 0.001	Yes
M4	**Planned**	**M1 + AS + Frequency + Valence + Imageability**	**1**	**1.31**	**0.25**	**Yes [Simpler]**
M5	Exploratory	M1 + AS + Common Associates + PPMI + CBOW + Skip-gram	5	234.1	< 0.001	Yes
M6	Exploratory	M1 + AS + Skip-gram	3	4.54	0.29	Yes [Simpler]
M7	Exploratory	M1 + Frequency + Arousal	2	20.88	< 0.001	Yes
M8	Exploratory	M1 + Frequency	1	1.43	0.23	Yes [Simpler]
M9	Exploratory	M6 + Frequency	1	9.12	0.0025	Yes
M10	Exploratory	M1 + Valence + Imageability	2	21.44	< 0.001	Yes
M11	Exploratory	M1 + Imageability	1	3.27	0.07	Yes [Simpler]
M12	Exploratory	M9 + Imageability	1	0.17	0.68	No
M13	Exploratory	M1 + Joint Frequency	1	173.34	< 0.001	Yes
M14	Exploratory	M9 + Joint frequency	1	5.20	0.023	Yes
M15	**Exploratory**	**M1 + AS + Skip-gram + Joint frequency**	**1**	**0.18**	**0.68**	**Yes [Simpler]**

Table 7: LME model summary for study 3. The final accepted models for planned and exploratory analyses are shown in bold.

Model Name	**Type**	**Model Description**	**df**	χ^2	**p**	**Accepted**
M1	Planned	(1 | Subjects) + (1 | Items)				Baseline
M2	Planned	M1 + Trial number	1	0.02	0.96	No
M3	Planned	M1 + AS + Frequency + Valence + Arousal + Imageability	5	175.55	< 0.001	Yes
M4	**Planned**	**M1 + AS + Valence + Imageability**	**2**	**3.99**	**0.14**	**Yes [Simpler]**
M5	Exploratory	M1 + AS + Common Associates + (P)PMI + CBOW + Skip-gram	6	248.52	< 0.001	Yes
M6	Exploratory	M1 + AS + Skip-gram	4	8.77	0.07	Yes [Simpler]
M7	Exploratory	M1 + Frequency + Arousal	2	5.40	0.07	No
M8	Exploratory	M1 + Arousal	1	4.93	0.036	Yes
M9	**Exploratory**	**M6 + Arousal**	**1**	**3.9**	**0.048**	**Yes**
M10	Exploratory	M1 + Valence + Imageability	2	22.6	< 0.001	Yes
M11	Exploratory	M1 + Valence	1	0.034	0.85	Yes [Simpler]
M12	Exploratory	M9 + Valence	1	2.68	0.1	No