Is it you you're looking for?

Personal relevance as a principal component of semantics

Chris Westbury and Lee H. Wurm
University of Alberta | Gonzaga University

Previous evidence has implicated personal relevance as a predictive factor in lexical access. Westbury (2014) showed that personally relevant words were rated as having a higher subjective familiarity than words that were not personally relevant, suggesting that personally relevant words are processed more fluently than less personally relevant words. Here we extend this work by defining a measure of personal relevance that does not rely on human judgments but is rather derived from first-order co-occurrence of words with the first-person singular personal pronoun, *I*. We show that words estimated as most personally relevant are recognized more quickly, named faster, judged as more familiar, and used by infants earlier than words that are less personally relevant. Self-relevance is also a strong predictor of several measures that are usually measured only by human judgments or their computational estimates, such as subjective familiarity, age of acquisition, imageability, concreteness, and body-object interaction. We have made all self-relevance estimates (as well as the raw data and code from our experiments) available at https://osf.io/gdb6h/.

Keywords: self-relevance, semantics, word2vec, co-occurrence, word processing, age of acquisition, subjective familiarity

The issue of what factors influence how people access individual words has commanded an enormous amount of research attention within psycholinguistics. One predictor that has received much attention is subjective familiarity (e.g. Gernsbacher, 1984; Balota et al. 2001; Bird et al. 2001; Bonin et al. 2003; Cortese and Khanna 2008; Ferrand et al. 2008, 2003; Flieller and Tournois 1994; Marques et al. 2007; Ghyselinck et al. 2000; Stadthagen-Gonzalez and Davis 2006; Westbury, 2014). In her ground-breaking work, Gernsbacher (1984) showed that subjective familiarity could differ from objective frequency for low frequency

https://doi.org/10.1075/ml.20031.wes | Published online: 18 March 2022
The Mental Lexicon 17:1 (2022), pp. 1–33. ISSN 1871-1340 | E-ISSN 1871-1375

words and argued that accounting for subjective familiarity could resolve a number of inconsistent findings related to lexical access. A major problem with using subjective frequency as a predictor of lexical access behaviour is precisely that it is *subjective*, which limits its value as an explanatory principle. If we do not know how judges are making their subjective judgments, then using those judgments as explanatory constructs simply correlates one mystery (the behavior to be explained) with another (the subjective judgments; see discussion in Westbury, 2016). Brown and Watson (1987) showed that judgments of age of acquisition were the major predictor of subjective frequency, followed by objective frequency measures and word length. All the substantive collections of age of acquisition judgments (most notably, the largest collection of Kuperman, Stadthagen-Gonzalez, and Brysbaert, 2012) are themselves subjective judgments, as were the judgments used by Brown and Watson. Explaining subjective familiarity judgments with subjective age of acquisition judgments piles one mystery on top of another.

After noting that many of the words rated as highly familiar (e.g., *hungry, beer, cash, kiss, lecture*) had referents that seemed particularly relevant to the students who had rated the words, Westbury (2014) presented evidence to support the claim that subjective familiarity judgments don't reflect lexical properties per se (other than objective frequency) but can rather be explained in terms of the subjective relevance of a word's referent. Estimates derived from a computational model of affective force based on a co-occurrence model of semantics were able to account for as much variance in subjective familiarity judgments as independent subjective familiarity judgments: that is, they correlated as strongly with familiarity judgments as familiarity judgments correlated with themselves. The computed estimates of affective force could also predict lexical access as well as those familiarity judgments.

Previous work has implicated subjective relevance in many domains of psychology. For example, Sui, He, and Humphreys (2012) showed that attaching self-relevant labels such as 'you' or 'mother' to arbitrary stimuli facilitated performance in a perceptual matching trial. This is one of a number of studies showing that judgments are quicker for self-relevant stimuli than for other stimuli (e.g. Frings & Wentura, 2014; Schäfer, Frings, and Wentura, 2016; Schäfer, Wenturam & Frings, 2015; Golubickis, Falben, Cunningham, & MacCrae, 2018). Alexopoulos, Muller, Ric, and Marendaz (2012) showed that self-relevant stimuli (a person's own name) automatically captured attention. Many experiments have shown that memory performance is enhanced for self-relevant stimuli (for a review, see Symons & Johnson, 1997). The neural underpinnings of self-relevance judgments have been identified (Northoff, Heinzel, De Greck, Bermpohl, Dobrowolny, & Panksepp, 2006; Schmitz & Johnson, 2007). As the examples

listed suggest, most of the cognitive studies on self-relevance have relied on small set of intuitively-derived stimuli. In this paper we re-formulate and extend this work by presenting evidence for the relevance to lexical access of a continuous measure of personal relevance that is directly computable for all words from patterns of word co-occurrence.

Study 1. Psychometric properties of a measure of personal relevance

We used the word2vec continuous skip-gram model of semantics (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This model uses a neural network with one hidden layer to predict the neighbours of every word in several billions of words of text in the Google news corpus (available from https://github.com/mmihaltz/word2vec-GoogleNews-vectors). The weights on 300 hidden units are adjusted to minimize prediction error. After training, each word is therefore represented by a vector of length 300 that encodes that word's context. Similarity between contexts is represented by similarity in vectors. This measure is called *second-order co-occurrence* because it is not a measure of how often words occur close together (*first-order co-occurrence*) but rather a measure of how often words occurred in similar contexts. Since words that have highly similar contexts tend to have highly similar meanings, the vectors can stand in, to some extent, for a word's meaning.

We take advantage of the fact that a vector composed of the averaged skip-gram vectors of many semantically related words (which we call *a category-defining vector* [CDV]) can serve as a measure of relatedness of the common semantic category of those words. For example, if we average the vectors for the words *jacket, pants, suit*, and *shirt*, the closest neighbors to that CDV that are not in its component list are *sweatshirt, polo shirt, blazer, trousers, shirts, hoodie, sweater, jeans, windbreaker, jackets*, and *tracksuit*. As a rule of thumb (based on results in Westbury and Hollis, 2018), CDVs composed of 100 words tend to define very stable categories.

We defined a CDV for words that are most likely to be personally relevant based on their first-order co-occurrence with the first-person singular personal pronoun, *I*. We extracted every word in the 560 million word *Corpus of Contemporary American English* (COCA; Davies, 2010) that either directly followed the word 'I' or followed 'I' with one word separating it. There were 70,048 such word tokens, comprised of 10,556 word types. We eliminated word types that had occurred less than 20 times in the COCA to focus on the most common personally-relevant words, reducing the word set to 465 types. From these, we removed all words that had word frequencies (from Shaoul and Westbury, 2006)

above 500 occurrences per million, as very common words do not differentiate well between contexts since they occur in so many contexts. This left 338 words, whose vectors we averaged to define a personal relevance CDV. We will refer to cosine distance from this CDV as PersonalRelevance.

The relationships between the vectors of the 200 words closest to this CDV are shown in Figure 1. The closest words to the personal relevance CDV are strongly dominated by verbs (e.g. *wants, thinks, figured, looked*), along with several adjectives (i.e., *confident, thankful, fortunate, afraid*). As the examples above and in Figure 1 suggest, the CDV has good face validity, since the verbs and adjectives are almost entirely limited to personal actions and qualities.

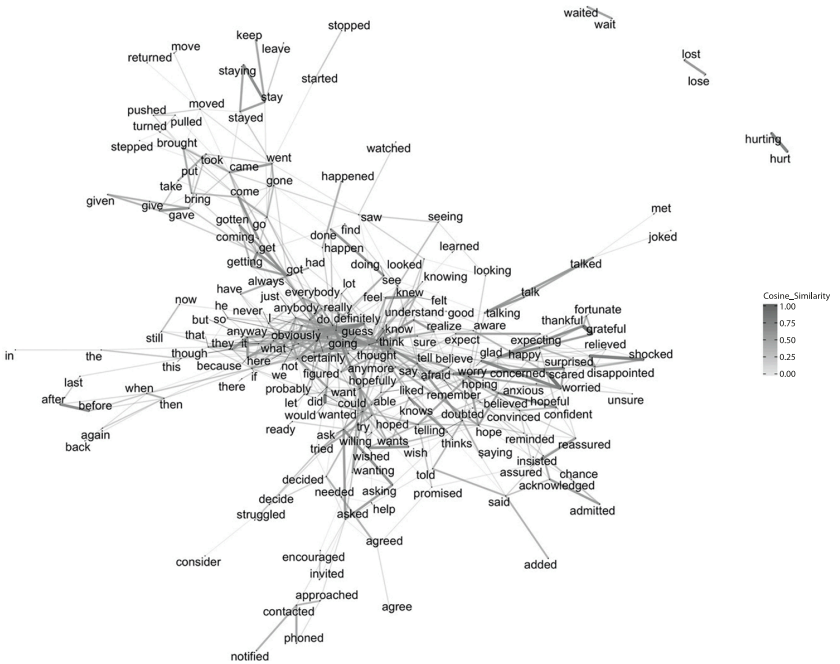

Figure 1. The 192 of 200 words closest to the personal relevance CDV that have vectors with cosine similarities > 0.5 with at least one other word's vector. Any words with vectors that did not have a cosine similarity with any other word above that threshold are not shown. Distances between unconnected clusters are arbitrary and have no interpretation.

There are almost no nouns in the words most strongly associated with personal relevance. We addressed the lack of nouns by defining three measures closely related to PersonalRelevance. To extend the corpus-based measure of personal relevance to nouns, we selected the 500 verbs (of any form) closest to the PersonalRelevance CDV. We then found all of the nouns in the COCA

that followed any of those verbs after a determiner (article, quantifier, possessive, demonstrative), as in the phrases *accepting a **bribe**, receive two **doses**, find his **niche***, and *recalled that **moment***. There were 178,613 such noun tokens, which collapsed to 10,746 types. Of those, 1783 occurred at least 20 times in the COCA. We averaged the vectors of these these 1783 nouns to define a CDV for a measure we call RelevantNoun. To make this more specific, we also defined a concrete and abstract version (RelevantConcrete and RelevantAbstract) from the most and least concrete 500 words in that set, as estimated using the extrapolated concreteness norms from Hollis, Westbury, and Lefsrud (2017). Over the entire dictionary, the ten words closest to the RelevantConcrete CDV (i.e., the concrete words that are suggested by the model to be most personally relevant) are *car, backpack, wallet, plastic_bag, billfold, truck, house, vehicle, suitcase*, and *stuffed_animal*. The ten words closest to the RelevantAbstract CDV (i.e., the abstract words that are suggested by the model to be most personally relevant) are *commitment, opportunity, motivation, decision, concerns, desire, concern, involvement, importance*, and *responsibility*. We judge these examples to have high face validity.

We measured the convergent validity of these noun-based measures by examining their ability to predict human ratings of body-object interaction. Tillotson, Siakuluk, and Pexman (2008) provided human ratings of this measure, the extent to which an object afforded direct interaction for 1618 nouns, of which two (the presumably mis-spelled *guaze* and *har*) did not appear in our dictionary and were ignored. If BOI and personal relevance were related, we would expect to see a negative correlation between PersonalRelevance and the BOI ratings because more personally relevant words (closer to the PersonalRelevance CDV) should have higher body-object interaction ratings. However, the correlation is not different than chance ($r = 0.038$, $p > 0.05$), suggesting that the PersonalRelevance measure does not capture this aspect of nouns. The RelevantNoun CDV distances correlate with 1616 BOI judgments at $r = -0.097$ ($p < 0.0001$), a small correlation reflecting that this CDV includes both irrelevant abstract and relevant concrete nouns. Distances from the RelevantAbstract CDV correlate with the BOI judgments at $r = 0.167$ ($p < 0.0001$), a direction (more abstract = lower BOI) that would be expected considering the inherently concrete focus of the BOI measure. Most importantly for the present purposes, distances from the RelevantConcrete CDV correlate with the 1616 BOI judgments at $r = -0.574$ ($p < 0.0001$), thereby accounting by themselves for about 32% of the variance in those judgments.

The cosine relationships between the vectors of the 200 words closest to the RelevantConcrete CDV are shown in Figure 2. Along with several smaller clusters, there are many clear large clusters of concrete objects: one of personal relationships (e.g., *mother, boyfriend, neighbor, aunt*), one of clothing (e.g. *bandanna, sweatshirt, jacket, shirt*), one of tools (e.g., *knife, screwdriver, gun, crowbar*), one for means of transport (e.g., *truck, motor_scooter, golf_cart, bus*), one of everyday

businesses (e.g., *convenience_store, restaurant, shop, hotel*), one for parts of a house (e.g., *driveway, sidewalk, front_porch, mailbox*), one for rooms (e.g. *bathroom, bedroom, restroom, hallway*), and one for personal containers (e.g., *baggie, backpack, wallet, purse*). The categories defined by these clusters also seem to have high face validity in terms of being both concrete and personally relevant.

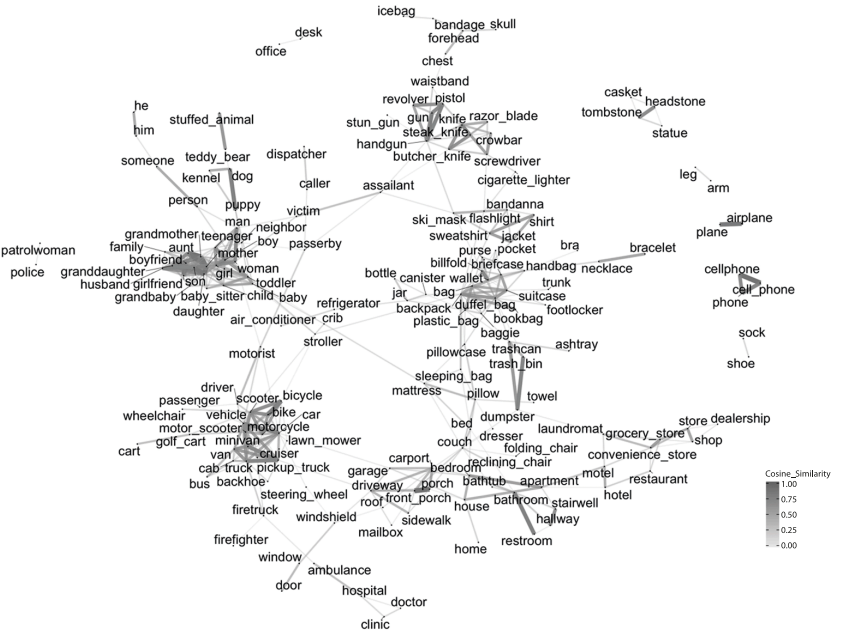

Figure 2. The 173 of 200 words closest to RelevantConcreteCDV with vectors with cosine similarities > 0.5 with at least one other vector. Any words with vectors that did not have a cosine similarity with any other word's vector above that threshold are not shown. Distances between unconnected clusters are arbitrary and have no interpretation.

Table 1 shows the linear correlations between the four measures of personal relevance and human judgments of age of acquisition (Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012; $N = 29900$), subjective familiarity (Stadthagen-Gonzalez & Davis, 2006), $N = 1508$), imageability (also from Stadthagen-Gonzalez and Davis, 2006; $N = 1508$), as well as computational estimates of valence, arousal, dominance, and concreteness (all from Hollis, Westbury, and Lefsrud, 2017; $N = 78278$), logged written word frequency (Shaoul & Westbury, 2006, $N = 35282$), logged spoken word frequency from the British National Corpus [BNC] (Leech, Rayson, & Wilson, 2001); the logged ratio of spoken BNC frequency to written BNC frequency, word length, and the body-object interaction judgments discussed above. These correlations provide both convergent and divergent

validity for the measures. RelevantConcrete is strongly negatively correlated with imageability judgments ($r = -0.534$, $p < 0.0001$) and concreteness estimates ($r = -0.654$, $p < 0.0001$), whereas RelevantAbstract is strongly positively correlated with those measures (Imagebility: $r = 0.406$, $p < 0.0001$; Concreteness: $r = 0.374$, $p < 0.0001$).

Table 1. Linear correlations with distance from the four CDVs associated with personal relevance. All $p < 0.0001$ except where otherwise stated

Measure	Type	N	PersonalRelevance	RelevantNoun	RelevantAbstract	RelevantConcrete
LogWordFreq	Empirical	67426	−0.458	−0.504	−0.430	−0.314
Log(SpokenFreq/ WrittenFreq)	Empirical	4068	−0.428	−0.372	−0.230	−0.192
Age of Acquisition	Judgment	29900	0.441	0.402	0.184	0.492
Familiarity	Judgment	1508	0.431	−0.342	−0.262	−0.180
Imageability	Judgment	1508	0.301	0.077 [$p < 0.05$]	0.406	−0.534
Arousal	Estimate	78278	−0.198	−0.132	−0.170	0.03
Dominance	Estimate	78278	−0.168	−0.191	−0.144	−0.183
LogSpokenFreq	Empirical	4068	−0.184	−0.126	−0.087	−0.049 [$p < 0.005$]
Concreteness	Estimate	78278	0.128	−0.027	0.374	−0.654
Length	Empirical	78278	0.053	−0.036	−0.202	0.165
Valence	Estimate	78278	−0.045	−0.119	0.034	−0.194
Body-Object Interaction	Judgment	1616	0.038 [$p > 0.05$]	−0.097	0.167	−0.574

Given the strong zero-order correlations of RelevantConcrete and RelevantAbstract with the extrapolated concreteness ratings, we constructed a generalized additive model (GAM) to predict those concreteness ratings from those two self relevance measures, alone and in interaction ($N = 38521$). The model produced estimates that correlated with those extrapolated ratings at $r = 0.89$ ($p < 0.0001$). This is significantly better than the correlation reported in Hollis, Westbury, and Lefsrud (2017) between two independently-gathered sets of human concreteness ratings ($N = 3937$; $r = 0.835$, $p < 0.0001$; Fisher's r-to-$z = 12.99$, $p < 0.0001$). As discussed further below, concreteness ratings are well-known to correlate with lexical decision (LD) reactions times (RTs). The raw correlation between those ratings and 38521 LD RTs from the English Lexicon Project (ELP; Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007) was −0.21, i.e. more concrete words are generally recognized

faster. When we removed the variance attributable to the GAM estimates from the concreteness judgments, the correlation between those residuals and LDRT dropped to zero ($r = 0.006$, $p = 0.24$), further supporting the claim that the variance attributable to human concreteness judgments can be entirely predicted from our algorithmically-computed self-relevance measures.

RelevantConcrete is significantly more strongly correlated with age of acquisition judgments ($r = 0.492$, $p < 0.0001$) than RelevantAbstract ($r = 0.184$, $p < 0.0001$; Fisher's r-to-$z = 43.1$, $p < 0.0001$), as we might expect since many words learned early are concrete (though see Westbury and Nicoladis, 1998, for evidence that children's first 10 words are often abstract).

To follow up on this correlation with age of acquisition, we examined PersonalRelevance for the 30 words uttered in all three of the first three months of language production among 45 infants, as reported in Hart (1991). Two exclamations in that list (*uhoh, uhuh*) did not occur in our dictionary. The average normalized PersonalRelevance among the remaining 28 words was −2.16z. Fourteen [50%] of the 28 words had self-reference values below −2z, and all 28 had self-reference values below the mean of 0z. This further supports the conclusion suggested by the correlation with the age of acquisition results: that children's early words are strongly self-relevant.

The three largest magnitude correlations with PersonalRelevance are those with written word frequency, logged ratio of spoken to written word frequency, and age of acquisition. As shown in Figure 3, there is a negative correlation between logged written BNC frequency and distance from the PersonalRelevance CDV, suggesting that words that are more personally relevant (closer to the PersonalRelevance CDV) are more frequent. The same figure shows a similar relation between distance from the PersonalRelevance CDV and the logged ratio of spoken BNC frequency to written BNC frequency: words closer to the PersonalRelevance CDV are more likely to have higher ratios of spoken to written frequency, suggesting that spoken words are more likely to be personally relevant than written words.

There is also a near-linear negative correlation between PersonalRelevance and judgments of subjective familiarity (Figure 4). Words that are more personally relevant (closer to the PersonalRelevance CDV) are judged to be more familiar than words that are less personally relevant.

Since the computational estimates and written word frequency are well-defined in virtue of being grounded in word use, it is of interest to see how well they (+ word length) can predict the human judgments whose grounding is opaque, as we did for concreteness judgments above. Table 2 shows the GAM for predicting age of acquisition judgments. To develop the model, we first split the data randomly into two halves, to have a development and validation data

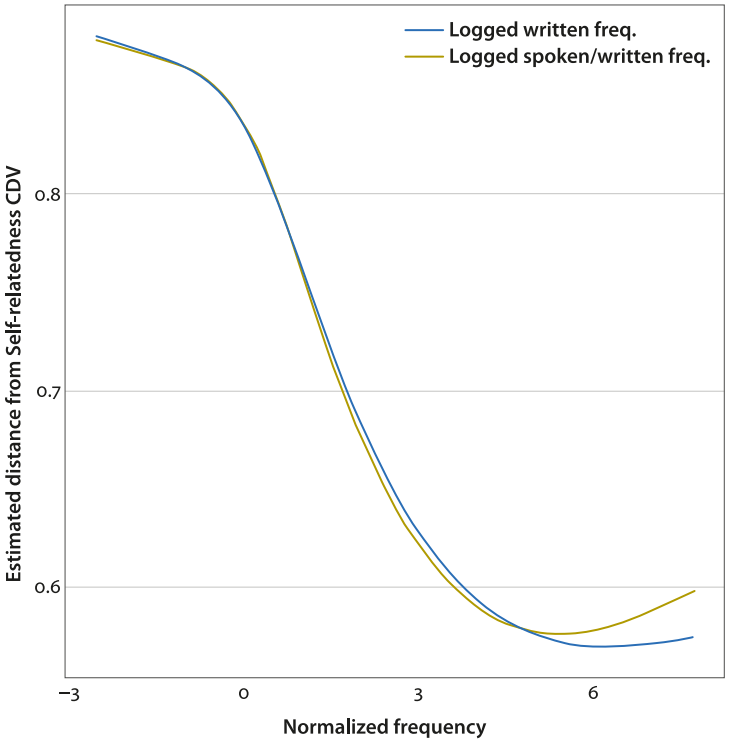

Figure 3. GAM-estimated distance to PersonalRelevance CDV x Logged written word frequency and Logged(spoken/written frequency), using BNC frequencies. 95% confidence intervals are graphed but not visible

set. We developed the model on the development set ($N = 10526$) by entering all predictors in a GAM, both singly and in interaction with normalized logged word frequency, and removing individual predictors or interactions by decreasing t-value until all remaining interactions or single predictors contributed with $p <= 0.05$. We entered all other predictors before entering the four personal relevance measures, both by themselves and interaction with LogFrequency. All four measures of personal relevance entered the model, reducing the AIC from 460413 to 44226, a very large reduction of 1815, suggesting that the model including self-relevance was much more likely to minimize information loss.

The model produced estimates that account for 53.1% of the variance in the age of acquisition judgments in the development set and 52.5% of the variance in the validation set ($N = 10227$). By comparison, the correlation between 1464 age of acquisition measures from Kuperman, Stadthagen-Gonzalez and Brysbaert (2012) and Stadthagen-Gonzalez and Davis (2006) is $r = 0.830$ ($r^2 = 0.69$), which is much better (Fisher's r-to-z test: $z = 9.44$, $p < 0.0001$). Nevertheless, the model

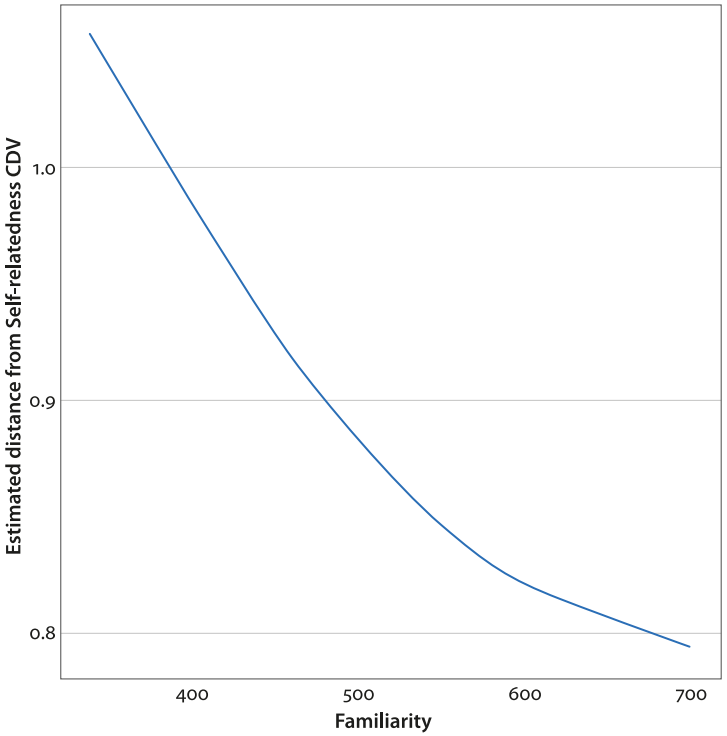

Figure 4. GAM-estimated distance to PersonalRelevance CDV x Familiarity. 95% confidence intervals are graphed but not visible

in Table 2 is able to account for most (52.5/69 = 76.1%) of the variance seen in different human age of acquisition judgments of the same words.

The interaction between LogFrequency and PersonalRelevance in predicting AofA is shown in Figure 5. The words lowest on AoA are words that are close to the PersonalRelevance CDV, with little effect of frequency. The words that are highest on AoA are words that are far from the PersonalRelevance CDV, again with little effect of frequency.

We have many fewer data points for predicting familiarity judgments. The best validation-set model ($N = 767$) before adding any measures of personal relevance included Dominance, Valence, LogFrequency, and phonological neighbourhood (PN). That model accounted for 32.7% of the variance in those estimates (AIC = 7744) and 23.7% of the variance in the validation set ($N = 745$). After adding the personal relevance measures by themselves and in interaction with LogFrequency, PN dropped out, and PersonalRelevance, NounRelevance, and AbstractNounRelevance entered in. That model accounted for 41.5% of the variance (AIC = 7652, a decrease of 92, suggesting the the second model is better

Table 2. GAM for predicting age of acquisition judgments in 10527 words. Predictors are ordered in terms of decreasing effective df. $R^2 = 0.53$. Terms involving personal relevance are bolded

Predictor	Effectivedf	F	P-value
Length	7.66	15.49	< 0.0001
RelevantNoun	**7.23**	**15.49**	**< 0.0001**
LogFreq* RelevantNoun	**7.2**	**5.95**	**< 0.0001**
LogFreq*PersonalRelevance	**7.07**	**6.21**	**< 0.0001**
Arousal	6.31	2.01	0.046
LogFreq*Length	5.9	8.84	< 0.0001
Valence	5.29	58.43	< 0.0001
LogFreq*Syllables	5.14	6.77	< 0.0001
LogFreq*Valence	4.97	2.72	0.008
LogFreq	4.5	129.5	< 0.0001
LogFreq*Concreteness	4	22.77	< 0.0001
PN	3.94	6.02	< 0.0001
PersonalRelevance	**3.93**	**128.4**	**< 0.0001**
Concreteness	3.05	13.91	< 0.0001
RelevantAbstract	**2.24**	**5.63**	**0.00074**
ON	2.21	7.07	0.00018
Syllables	1.73	84.91	< 0.0001
RelevantConcrete	**1**	**220.82**	**< 0.0001**
Dominance	1	80.63	< 0.0001

at minimizing information loss). It accounted for 32.5% of the variance in the validation set. As reported in Westbury (2014), independent human judgments of subjective familiarity for 699 words (from Stadthagen-Gonzalez and Davis, 2006 and Gilhooly and Logie, 1980) correlated at $r = 0.70$ ($r^2 = 0.49$). By Fisher's r-to-z test, the correlations between the validated model estimates and the familiarity judgments are worse than the correlations between the two sets of human judgments ($z = 4.16$, $p < 0.0001$), i.e., the model does not predict subjective familiarity as well as independent human judgments of subjective familiarity do. However, it again accounted for the majority ($32.5/49 = 66.4\%$) of the variance seen in different human familiarity judgments of the same words.

As another check of the self-relevance measures, we computed the correlation between PersonalRelevance and the early principal components (PCs) in the

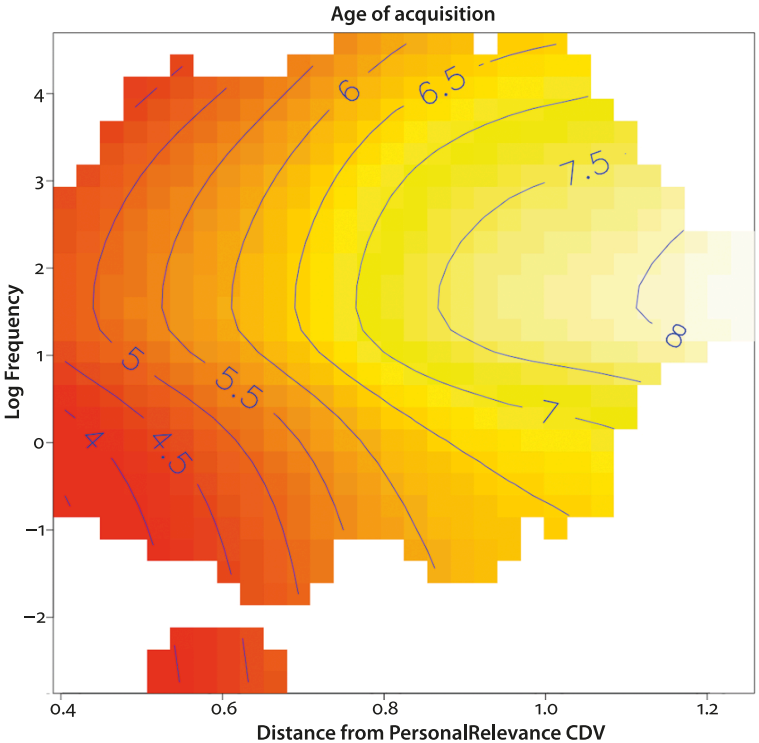

Figure 5. Interaction between LoggedFrequency and distance from PersonalRelevance CDV in predicting AofA

Google news corpus matrix. We extracted the vectors from that matrix for the 78,278 words for which we had estimates of personal relevance. We then performed principal components analysis on this subsetted matrix. The correlations between the predictors we consider in this paper and the first five PCs are shown in Table 3. Among the measures considered in that table, PersonalRelevance has the highest magnitude correlations with the first three PCs ($r = 0.47$, $r = 0.46$, and $r = -0.50$ respectively; $p < 0.0001$), suggesting that it plays a major role in the structure of semantic space. In their analysis of the principal components of a (smaller) word2vec matrix, Hollis and Westbury (2016) had suggested that PC1 was mainly associated with word frequency, reporting a non-cross-validated correlation of 0.42 (much higher than the $r = -0.12$ here) between the PC1 values and logged frequency. A GAM model constructed only with logged frequency accounted for just 1.5% of the variance in PC1 values in the development set ($N = 18087$) and 1.3% in the validation set ($N = 17924$). Adding all four personal relevance measures alone and in interaction with frequency produced a GAM model that accounted for 51.6% of the variance of in PC1 values in the develop-

ment set and the same amount in the validation set. In contrast, a GAM model built with all predictors in Table 3 except the self-relevance measures accounted for 42.4% of the variance in PC1 values in the development set and 41.8% in the validation set. The cross-validated performances of the two models suggest that the model built using only self-relevance measures correlates significantly better with the PC1 values than the model built using every predictor except the self-relevance measures (Fisher's r-to-z: $z = 12.9$, $p < 0.0001$). The first principal component of the word2vec matrix largely reflects personal relevance.

Table 3. Linear correlations between the first five principal components (PC) of the Google news matrix, and the predictors considered in this paper. All $p < 0.0001$. Terms involving personal relevance are bolded

Predictor	PC1	PC2	PC3	PC4	PC5
PersonalRelevance	**0.467**	**0.463**	**−0.495**	**−0.137**	**−0.057**
RelevantNoun	**0.193**	**0.552**	**−0.471**	**0.051**	**−0.133**
RelevantConcrete	**−0.128**	**0.055**	**−0.761**	**−0.006**	**−0.075**
RelevantAbstract	**0.268**	**0.715**	**−0.119**	**0.054**	**−0.179**
Valence	0.167	0.160	0.217	0.051	0.224
Arousal	−0.385	−0.022	−0.068	−0.041	0.056
Dominance	0.053	0.032	0.215	0.162	0.118
Concreteness	0.443	0.385	0.586	−0.045	−0.053
AoA	0.121	−0.087	−0.438	−0.157	0.044
Length	0.016	−0.263	−0.252	−0.084	0.259
WestburyLogFreq	−0.122	−0.269	0.175	−0.019	−0.128

RelevantAbstract is correlated very strongly with PC2 ($r = 0.715$, $p < 0.0001$) and RelevantConcrete is correlated very strongly with PC3 ($r = -0.761$, $p < 0.0001$). The latter correlation is a significantly higher magnitude correlation than the correlation of *concreteness* with PC3, $r = 0.57$ (Fisher's r-to-z: $z = 33.2$, $p < 0.0001$). GAM models using only the single predictors accounted for 61.0% of the variance in PC2 and 62.6% of the variance in PC3 in the development set ($p < 0.0001$). The models cross-validated well, accounting for 60.4% of the variance in PC2 and 62.3% of the variance PC3 in the validation set. PC2 appears to load largely on personally-relevant abstract concepts, especially verbs. The top ten words are *notify, reevaluated, notified, expedited, evaluating, disclose, evaluated, refile*, and *recused*. PC3 appears to largely reflect not concreteness itself, but more specifically the personal relevance of concrete objects. Although not all of the first 50 words

are nouns, those words include the nouns *sandbags, wire_cutters, backhoes, sleds, squeegees, hydrants, tarp, duffel_bag, billfold*, and *electrician*,

We have so far presented evidence showing that personal relevance measures extracted directly from patterns of word use are significant predictors of age of acquisition, subjective familiarity judgments, children's first words, body-object interaction judgments, and the early principal components in the word2vec matrix. We now turn to considering how well these measures can predict behavioral measures of lexical access.

Study 2. Lexical Decision and Word Naming

Background

Table 4 shows GAM-estimated correlations between PersonalRelevance and several database measures of lexical access: lexical decision (LD) reaction times (RTs) and accuracy from both the English Lexicon Project (ELP; Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007) and the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012), as well as word naming accuracy and RTs from the English Lexicon Project only. All reaction times here and in the analyses below have been transformed as −1000/RT, which is directly proportional to response rate (how often per time unit a person could respond to that word with that RT), with the sign flipped to keep the correlation in the same direction as untransformed RTs for the sake of ease of interpretation. By itself, PersonalRelevance is a significant predictor of all of these measures ($p < 0.0001$ in all cases), accounting for between 12.1% of the variance (ELP Naming accuracy) and 23.0% of the variance (BLP LD accuracy).

Table 4. GAM-estimated correlations between distance from PersonalRelevance CDV and measures of lexical access from the English Lexicon Project (ELP) and the British Lexicon Project (BLP). Reaction times have been transformed as −1000/RT

	LD	LD accuracy	Naming	Naming accuracy
ELP r	0.386	0.391	0.378	0.348
ELP r^2	0.149	0.153	0.143	0.121
ELP n	38251	38521	38529	38528
BLP r	0.392	0.480	N/A	N/A
BLP r^2	0.54	0.230	N/A	N/A
BLP n	27312	27312	N/A	N/A

In order to assess whether personal relevance more broadly construed (using all four measures) accounts for any unique variance over and above that explained by widely recognized predictors of lexical access, we used GAMs. We constructed tentative base models that entered word length, valence, arousal, dominance, concreteness, AoA, phonological neighbourhood size, number of syllables, and orthographic neighbourhood size [ON], as well as each in interaction with LogFreq. We removed elements (interactions or single predictors) that contributed with $p > 0.05$.

The final model for predicting ELP LD RT is shown in Table 5. Before adding in the measures of personal relevance, the model accounted for 58.1% of the variance variance in ELP LDRTs in the development dataset ($N = 18076$; AIC = −21428). After adding the four personal relevance measures, alone and in interaction with logged frequency, the model accounted for 59.6% of the variance (AIC = −22053, a large decrease of 625, suggesting that adding PersonalRelevance decreased the probability of information loss). On the cross-validation set ($N = 17916$), the model accounted for 58.9% of the variance.

Table 5. Summary of GAM model for predicting ELP lexical decision RTs. Predictors are ordered in terms of decreasing effective df. $R^2 = 0.589$. Terms involving personal relevance are bolded

Predictor	Effective df	F	P-value
LogFreq*Valence	11.12	4.33	< 0.0001
LogFreq*Length	10.65	12.04	< 0.0001
LogFreq*RelevantNoun	**10.52**	**6.07**	**< 0.0001**
RelevantNoun	**8.08**	**13.33**	**< 0.0001**
LogFreq*PersonalRelevance	**6.98**	**4.34**	**< 0.0001**
LogFreq	6.82	305.39	< 0.0001
LogFreq*Syllables	6.4	4.03	0.00012
Length	5.52	97.55	< 0.0001
PN	5.18	9.14	< 0.0001
LogFreq*RelevantConcrete	**4.93**	**2.91**	**0.0076**
LogFreq*Concreteness	4.69	2.70	0.01
Arousal	4.28	9.18	< 0.0001
ON	3.86	22.34	< 0.0001
RelevantAbstract	**3.67**	**3.26**	**0.0084**
LogFreq*RelevantAbstract	**3.54**	**4.8**	**0.00086**
Valence	3.32	67.72	< 0.0001
PersonalRelevance	**2.15**	**15.57**	**< 0.0001**
Syllables	1.91	212.93	< 0.0001
RelevantConcrete	**1**	**121.37**	**< 0.0001**

We constructed an analogous model that also included the logged ratio of BNC spoken to written word frequency, although we have many fewer data points for this (development set $N = 1919$) since the number of words in the BNC spoken frequency database is small. Both the spoken word ratio and all four self-relevance measures entered the model with $p < 0.05$. However, the model on this small subset of data was much worse than the full model above, accounting for just 43% of the variance on the development set and 39.4% in the validation set ($N = 1869$).

To predict ELP LD accuracy, we converted the accuracy rates to Ns, using the average number of participants who saw each word ($N = 34$, as reported in Balota, Yap, et al., 2007) and analyzed those averages using a binomial family GAM. The model for predicting ELP LD accuracy, including only written word frequencies, is shown in Table 6. Before adding the personal relevance measures, the model accounted for 39.4% of the variance across the 18076 words in the development set (AIC = 140516). Adding the four personal relevance measures, alone and in interaction with LogFreq, substantially improved the model ($r^2 = 0.431$; AIC = 134584), with the very large AIC reduction of 5931 again suggesting that the model was much more likely to minimize information loss than the model without it. The model accounted for 41.9% of the variance in the cross-validation set ($N = 17916$).

Table 6. Summary of GAM for predicting ELP lexical decision accuracy. Predictors are ordered in terms of decreasing effective df. $R^2 = 0.432$. Terms involving personal relevance are bolded

Predictor	Effectivedf	Chi.sq	P-value
LogFreq*Valence	15.89	261.49	< 0.0001
LogFreq*Dominance	15.88	166.98	< 0.0001
LogFreq*PersonalRelevance	**15.33**	**311.83**	**< 0.0001**
LogFreq*RelevantNoun	**15.15**	**149.01**	**< 0.0001**
LogFreq*RelevantAbstrac	**14.6**	**268.52**	**< 0.0001**
LogFreq*Length	12.96	245.33	< 0.0001
LogFreq*Arousal	12.88	139.88	< 0.0001
LogFreq*Concreteness	12.15	260.2	< 0.0001
LogFreq*RelevantConcrete	**11.32**	**172.7**	**< 0.0001**
LogFreq*Syllables	10.57	191.92	< 0.0001
PN	8.94	493.52	< 0.0001
RelevantNoun	**8.86**	**245.17**	**< 0.0001**
Concreteness	8.84	161.35	< 0.0001
Dominance	8.61	712.28	< 0.0001
PersonalRelevance	**8.18**	**405.19**	**< 0.0001**

Table 6. *(continued)*

Predictor	Effectivedf	Chi.sq	P-value
Length	7.99	4943.17	< 0.0001
Arousal	7.46	244.04	< 0.0001
LogFreq	7.16	9072.16	< 0.0001
RelevantAbstract	**6.82**	**59.81**	**< 0.0001**
ON	6.7	955.35	< 0.0001
RelevantConcrete	**5.94**	**487.02**	**< 0.0001**
Syllables	2	558.08	< 0.0001

We undertook the same analyses for ELP naming RT and naming accuracy.

The best model for predicting naming RTs ($N = 18076$) while including the personal relevance measures (Table 7; AIC = −22460) accounted for 53% of the variance, as compared to 51.1% without the personal relevance measures (AIC: −21789, a difference of 671). Including the personal relevance measures again resulted in a very significant improvement in the model. The model accounted for 52.3% of the variance in the validation set ($N = 17916$).

Table 7. Summary of GAM for predicting ELP naming RTs ($N = 18{,}076$). Predictors are ordered in terms of decreasing effective df. $R^2 = 0.527$. Terms involving personal relevance are bolded

Predictor	Effective df	F	P-value
LogFreq*RelevantNoun	**8.95**	**3.52**	**< 0.0001**
RelevantNoun	**8.35**	**10.82**	**< 0.0001**
PN	7.68	7.48	< 0.0001
PersonalRelevance	**6.88**	**15.00**	**< 0.0001**
LogFreq*RelevantConcrete	**6.58**	**3.86**	**0.00014**
LogFreq*PersonalRelevance	**6.55**	**2.79**	**0.0046**
LogFreq	5.75	189.69	< 0.0001
LogFreq*Syllables	5.24	21.85	< 0.0001
Valence	5.1	18.98	< 0.0001
ON	4.55	24.78	< 0.0001
LogFreq*Valence	4.45	3.87	0.00071
LogFreq*Length	4.16	9.60	< 0.0001
Length	3.22	72.73	< 0.0001
Syllables	1.59	259.75	< 0.0001
Arousal	1.52	19.48	< 0.0001
RelevantConcrete	**1.19**	**90.41**	**< 0.0001**
Concreteness	1	7.25	0.0071

There was also an effect of personal relevance on ELP naming accuracy (Table 8). The best model ($N = 18076$) that did not include any of the personal relevance measures accounted for 30.1% of the variance (AIC = 89556). Adding the personal relevance measures (by themselves and in interaction with LogFreq) increased the amount of variance accounted for to 33.9% (AIC = 86219, a large decrease of 3337). The same model accounted for 34.3% of the variance in the validation set ($N = 17918$).

Table 8. Summary of GAM for final model for predicting ELP naming accuracy ($n = 18{,}076$). Predictors are ordered in terms of decreasing effective df. $R^2 = 0.339$. Terms involving personal relevance are bolded

Predictor	Effectivedf	Chi.sq	P-value
LogFreq*Arousal	15.19	111.79	< 0.0001
LogFreq*RelevantAbstract	**14.96**	**92.69**	**1.07e-12**
LogFreq*Syllables	14.64	135.47	< 0.0001
LogFreq*PersonalRelevance	**14.52**	**106.78**	**1.05e-15**
LogFreq*Dominance	13.42	110.54	< 0.0001
LogFreq*RelevantNoun	**12.41**	**101.4**	**5.83e-15**
LogFreq*Valence	11.66	98.3	3.63e-15
LogFreq*Length	9.79	102.74	< 0.0001
LogFreq*Concreteness	9.73	56.12	4.77e-08
LogFreq*RelevantConcrete	**9.37**	**101.5**	**< 0.0001**
PN	8.95	257.43	< 0.0001
Length	8.9	1241.71	< 0.0001
LogFreq	8.81	3054.77	< 0.0001
Concreteness	8.71	76.89	4.35e-13
Dominance	8.45	242.16	< 0.0001
RelevantNoun	**8.39**	**85.81**	**7.11e-15**
PersonalRelevance	**8.35**	**415.48**	**< 0.0001**
RelevantConcrete	**8.32**	**121.38**	**< 0.0001**
ON	8.01	386.61	< 0.0001
RelevantAbstract	**7.13**	**34.24**	**6.42e-05**
Arousal	6.92	203.04	< 0.0001
Syllables	1.78	1086.59	< 0.0001

Analogous modeling results for the BLP LD RT and accuracy are shown in Tables 9 and 10, respectively.

The best development set model of BLP LD RTs ($N = 9224$; AIC = −11511; Table 9) that included the measures of personal relevance accounted for 51.1% of the variance, as compared to 47.4% without those measures (AIC: −10974, a large reduction of 537). The model accounted for 49.2% of the variance in the validation set ($N = 9001$).

Table 9. Summary of GAM for predicting BLP LD RT ($n = 9224$). Predictors are ordered in terms of decreasing effective df. $R^2 = 0.511$. Terms involving personal relevance are bolded

Predictor	Effective df	F	P-value
LogFreq*PersonalRelevance	**12.84**	**6.53**	**<0.0001**
LogFreq*Length	8.62	6.43	<0.0001
PersonalRelevance	**7.98**	**15.73**	**<0.0001**
RelevantNoun	**7.94**	**9.59**	**<0.0001**
LogFreq	6.7	231.83	<0.0001
LogFreq*RelevantNoun	**5.69**	**4.23**	**<0.0001**
PN	4.63	3.65	0.0017
ON	4.57	6.37	<0.0001
RelevantAbstract	**4.14**	**3.82**	**0.0015**
LogFreq*RelevantConcrete	**4**	**13.29**	**<0.0001**
Length	3.21	3.11	0.014
RelevantConcrete	3.08	34.83	<0.0001
Dominance	3.02	7.14	<0.0001
LogFreq*Valence	2.45	5.5	0.001
Arousal	2.33	15.76	<0.0001
LogFreq*Dominance	2.22	5.21	0.0032
LogFreq*RelevantAbstract	**1**	**9.46**	**0.0021**
Valence	1	59.62	<0.0001

The best development set model of BLP LD accuracy that included the measures of personal relevance (Table 10) accounted for 51.2% of the variance ($N = 9226$; AIC = 131129), as compared to 45.3% without the measures of personal relevance (AIC: 143658 for a large reduction in AIC of 12529 from entering the

personal relevance measures). In the validation dataset ($N = 9005$), the same model accounted for 49.5% of the variance.

Table 10. Summary of GAM for predicting BLP LD accuracy ($N = 9226$). Predictors are ordered in terms of decreasing effective df. $R^2 = 0.512$. Terms involving personal relevance are bolded

Predictor	Effective df	Chi.sq	P-value
LogFreq*Dominance	16	479.69	< 0.0001
LogFreq*RelevantNoun	**15.99**	**270.28**	**< 0.0001**
LogFreq*RelevantAbstract	**15.98**	**313.54**	**< 0.0001**
LogFreq*Valence	15.94	614.68	< 0.0001
LogFreq*Concreteness	15.73	300.29	< 0.0001
LogFreq*PersonalRelevance	**14.89**	**577.58**	**< 0.0001**
LogFreq*Length	14.32	329.22	< 0.0001
LogFreq*RelevantConcrete	**14.15**	**280.17**	**< 0.0001**
LogFreq*Arousal	14.02	289.89	< 0.0001
ON	8.96	1142.54	< 0.0001
PN	8.94	782.76	< 0.0001
RelevantConcrete	**8.91**	**1040.03**	**< 0.0001**
PersonalRelevance	**8.89**	**1219.2**	**< 0.0001**
Arousal	8.81	767.63	< 0.0001
RelevantNoun	**8.8**	**699.28**	**< 0.0001**
Concreteness	8.75	109.67	< 0.0001
Dominance	8.55	969.33	< 0.0001
RelevantAbstract	**8.51**	**97.55**	**< 0.0001**
Length	8.21	5925.97	< 0.0001
LogFreq	7.94	14279.08	< 0.0001

Discussion

These results provide evidence of a clear role for measures of personal relevance in accuracy and RT for both the lexical decision and word naming task. Adding measures of personal relevance to generalized additive models predicting RT and accuracy in two different datasets and across two tasks substantially improved the models' prediction in every case.

To follow up this data mining analysis, we conducted our own experiments, asking participants to make both lexical and word familiarity decisions for words closely matched on nine predictors associated with variance in RT, in three blocks estimated to be high, medium, and low on PersonalRelevance. By tightly matching these three sets of words to control on many extraneous sources of variance in LDRT, we hoped to to focus on the effects of personal relevance alone.

Method

Stimuli

We began with 22,944 words for which we had all the measures discussed below. We normalized the personal relevance estimates and divided the words into three groups: high personal relevance (zPersonalRelevance < −2z [the sign reflects that lower values are closer to the CDV, i.e. more personally relevant]; 301 words), mid personal relevance (−0.5z < zPersonalRelevance < 0.5z; 7296 words) and low personal relevance (zPersonalRelevance > 2z; 2042 words).

We matched a subset of the selected words on nine measures: valence, arousal, dominance, concreteness, logged frequency, ON, length, number of syllables, and number of morphemes. To make the match, we normalized all nine measures and found the closest match between the high and low words by exhaustive search of all possible pairs. The search first found the two words from each of the personal relevance categories that were most closely matched (i.e. had the smallest average distance between each of the nine normalized measures). We then removed these two words and repeated the procedure to find the next most closely matched pair, continuing until the average difference between the two words >= 0.4z. Under this constraint, we ended up with 139 words pairs. We repeated the matching to match the 139 high-self-relevance words to mid-self-relevance words. Since there were so many of these, we were able to match them within the difference cut-off, with a maximal average difference of 0.28z.

We thus ended up with 139 matched triplets. As shown in Table 11, there was just one significant difference on any of the nine lexical characteristics between any pairs of the three categories: high and medium personally-relevant words were higher than low personally-relevant words on estimated arousal ($p < 0.05$). This reflects the tightness of the matching more than a difference that might be of any practical concern for our purposes. The maximal difference of 0.016 between the average estimated arousal of high self-relevance and low self-relevance words is a small difference of just 0.23 SDs, covering about 2.8% of the entire range of the measure.

Table 11. Values of nine predictors in words high, medium, or low on self-relevance (139 words per category). Significant pairwise t-test differences between self-relevance categories are shown in bold

Measure	Self-relevance category	Self-relevance	Valence	Arousal	Dominance	Concreteness	Frequency	ON	Length	Syllables	Morphemes
Average	High	0.67	0.56	0.470	0.59	0.55	2.86	2.91	7.73	2.21	2.07
	Mid	0.93	0.55	0.472	0.58	0.56	2.68	2.63	7.68	2.24	2.06
	Low	1.09	0.56	0.457	0.57	0.55	2.57	2.22	7.46	2.37	2.04
SD	High	0.03	0.09	0.05	0.06	0.10	2.62	3.29	1.84	0.93	0.61
	Mid	0.03	0.09	0.05	0.06	0.11	2.37	3.53	1.87	0.91	0.58
	Low	0.03	0.10	0.05	0.06	0.13	2.49	3.10	2.03	0.92	0.67
T-test p	H-M	**0.00**	0.78	0.69	0.41	0.34	0.54	0.50	0.80	0.79	0.92
	H-L	**0.00**	0.73	**0.033**	0.057	0.67	0.33	0.076	0.24	0.15	0.64
	M-L	**0.00**	0.53	**0.011**	0.25	0.22	0.70	0.30	0.36	0.24	0.70

The nonwords were constructed beginning with the base set of 8000 NWs from Westbury, Hollis, Sidhu, and Pexman (2018). The strings are available from http://www.psych.ualberta.ca/~westburylab/. Because many of the words in the matched self-relevance triplets were affixed and some were long, we affixed these 8000 NWs in eight ways:

i. By suffixing with *ing* (*fluce* --> *flucing*)
ii. By suffixing with *s* (*blut* --> *bluts*)
iii. By suffixing with *ed* (*jurv* --> *jurved*)
iv. By prefixing with *un* (*pleck* --> *unpleck*)
v. By suffixing with *ies* (*nurth* --> *nurthies*)
vi. By both prefixing with *un* and suffixing with *ed* (*lirph* --> *unlirphed*)
vii. By compounding two random NWs (*stulk* + *tanch* --> *stulktanch*)
viii. By prefixing a compound NW with *dis (karkborch* --> *diskarkborch)*

We regularized some resultant problematic strings (*rette* --> *retteing* --> *retting*) and removed others that seemed problematic for any reason, to end up with the 58,455 NWs from which we randomly drew for each participant's LD stimuli.

We used these three lists of 139 words and the list of 58,455 NWs as a stimulus pool in two experiments, a lexical decision experiment and familiarity judgment experiment. In both experiments, stimuli were drawn randomly for each participant. All participants participated in both experiments, with the lexical decision task always preceding the familiarity judgment task. The experiments were carried out using three Apple G4 Macintosh Minis with 17.1-in. monitors running

Apple OS 10.15 using custom-written software. The screens' resolutions were set to 1,280 × 1,024 pixels.

In the lexical decision experiment, each participant saw 198 stimuli, comprised of 99 NWs and 99 words selected (though not presented) as 33 matched word triplets. The words and NWs that each participant saw were matched exactly on length only. Subjects were told they would see strings one at a time on the screen and asked to hit one button ('c' for correct) if it was a word and another button ('x' for incorrect) if it was not a word, using the first and second fingers of their dominant hand. The strings were presented in 75-point Times font in with an ISI of 1000 ms. Each stimulus was preceded by a '+' shown for a random time drawn uniformly from a distribution between 250 and 500 ms, which participants were informed was there to prepare them to decide about the following string.

During the familiarity judgment task each participant saw 33 random matched triplets that had not been included their lexical decision task stimuli. Participants saw each triplet in three rows with two columns of radio buttons, one labeled 'Most familiar' and the other labeled 'Least familiar'. They were asked to click on one radio button from each column, with selections forced to be mutually exclusive (i.e. selections in one column would turn off any selection for that word in the other column). When two words had been selected, participants were instructed to click on a button marked 'Next' to move to the next trial.

All elements of this study were evaluated and accepted by the University of Alberta research ethics board.

Participants

Participants were 32 (23 female, 9 male) self-reported native English speakers who participated in the experiments for partial course credit. They have average [SD] age of 18.3 [0.9] years and an average [SD] of 13.5 [0.8] years of education. All but one reported themselves to be right-handed. All participated after giving written informed consent.

Results: Lexical decision RT

One male participant's data were removed before analysis of the lexical decision results because he made only 67.2% correct decisions, more than two standard deviations below the average correct response rate. After removing his data, the average [SD] correct decision rate was 92.4% [4.8%] (Words: 95.0% [4.0%]; NWs: 89.8% [9.0%]).

As with the published RTs analyzed above, we transformed the RTs as −1000/RT. We analyzed the correct RTs from our experiment with GAM mixed-effect models. We defined an initial base model by entering only random effects of participant and stimulus order. Both effects entered with $p < 0.001$, accounting for 1.52% of the variance. With the random effects fixed, we entered effects of word length, valence, arousal, dominance, concreteness, AoA, phonological neighbourhood size, number of syllables, and orthographic neighbourhood size [ON], and each of these in interaction with LogFreq. We then removed fixed effects (interactions or single predictors) that contributed with $p > 0.05$. We compared models using the Aikake Information Criterion (Aikake, 1973, 1974).

The stimulus-matching had the effect of narrowing the range of many of these predictors, presumably because the closest matches are possible for the most common values. As a result, the predictors had little predictive power. For example, the correlation of LogFreq with 36,011 transformed ELP LD RTs is −0.61, accounting for 37.6% of the variance, while the correlation of LogFreq with transformed RT over the 416 words in our experiment was only −0.30, accounting for just 9.0% of the variance after collapsing by word, less than a quarter as much as in the ELP. Similarly, the correlation of word length with the 519 ELP LD RTs is −0.55, accounting for 31% of the variance, while the correlation of LogFreq with the RT over the words in our experiment was −0.26, accounting for just 6.9% of the variance after collapsing by word, only about 22% as much as in the ELP. While this has the salutary effect of almost completely controlling for the effects of the nine variables on which we matched our stimuli across the three levels of PersonalRelevance, it also suggests (as a reviewer pointed out) that we have restricted "the dataset to a potentially very atypical and weird sector of lexical space". Just as we have attenuated the variables we controlled for in order to isolate the effects of a predictor of interest, we may have attenuated the effects of that predictor of interest.

The base model included random effects of stimulus order and participants, as well as fixed effects of LogFreq, ON, Length, Arousal, Dominance, Valence, and Concreteness. It accounted for only 8.33% of the variance in the transformed RTs (AIC = 2769). Adding the personal relevance measures by themselves and in interaction with LogFreq knocked many of the predictors out of the model. The short final model, shown in Table 12, included fixed effects of PersonalRelevance, RelevantConcrete, RelevantAbstract, LogFreq, Length, and Concreteness. The variance explained increased to 8.62%, reducing the AIC by 124 to 2644, suggesting a robust effect of the three measures of personal relevance.

Table 12. Summary of GAM fixed effects for predicting experimental LDRTs. Predictors are ordered in terms of decreasing effective df. $R^2 = 0.086$. Terms involving personal relevance are bolded

Predictor	Effectivedf	Ref.df	F	P-value
RelevantAbstract	**15**	**1**	**21.72**	**< 0.0001**
Concreteness	15	1	4.61	0.032
Length	3.68	4.6	8.11	< 0.0001
RelevantConcrete	**2.92**	**3.71**	**6.97**	**< 0.0001**
LogFreq	2.83	3.54	12.49	< 0.0001
PersonalRelevance	**2.12**	**2.62**	**3.32**	**0.034**

Results: Lexical decision error rates

We analyzed the errors in our dataset using the same procedure as we used with the RTs, using a binomial family mixed-effects GAM. After adding the personal relevance measures, the final model (AIC = 911, accounting for 18.8% of the variance) was a slight improvement over the base model (AIC = 921, accounting for 18.6% of the variance). The final model is shown in Table 13. Figure 6 shows the estimated RTs across the three categories of PersonalRelevance.

Table 13. Summary of GAM fixed effects for predicting experimental LD accuracy. Predictors are ordered in terms of decreasing effective df. $R^2 = 0.19$. Terms involving personal relevance are bolded

Predictor	Effective df	Chi.sq	P-value
Length	13	6.97	0.0083
RelevantNoun	**11**	**6.45**	**0.011**
LogFreq	5.71	51.51	< 0.0001
Valence	4.05	12.37	0.031
LogFreq*RelevantConcrete	**3.08**	**12.07**	**0.016**
LogFreq*Syllables	2.89	15.71	0.0029
RelevantConcrete	**2.48**	**36.92**	**< 0.0001**

Discussion: LD

The results from our lexical decision data implicate personal relevance in correct LD RT and accuracy, replicating the analyses above, albeit with the caveat that the

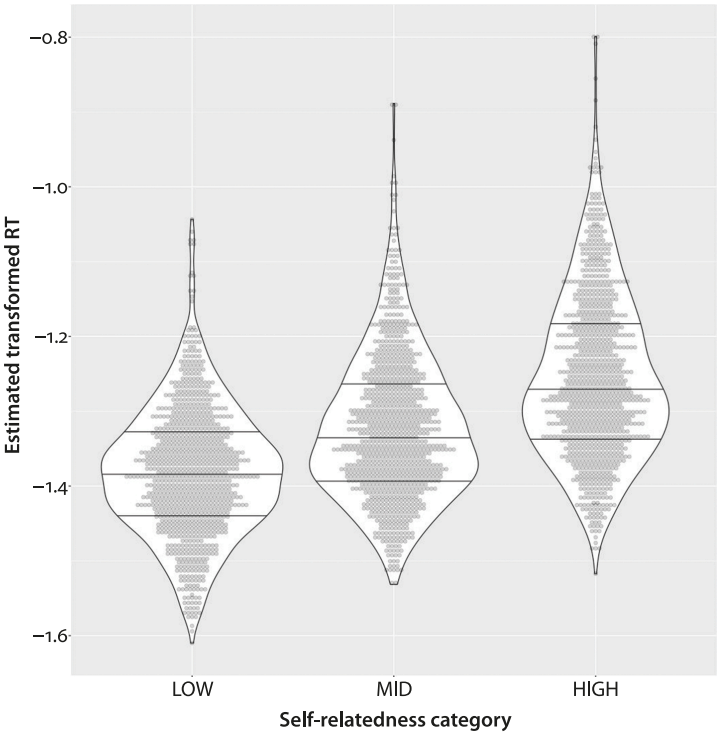

Figure 6. GAM-fitted transformed RTs for lexical decision from our experiment, by distance from the PersonalRelevance CDV. Individual points are observations. Black lines indicate quartiles

way we matched our words may have resulted in a non-representative sample of words for our experiment.

Results: Familiarity decision

To analyze the familiarity decision data, we coded words chosen most and least familiar as 1 and 0, respectively, and analyzed the data with a binomial family GAM, using the same predictor set as above in the base model. We also included random effects of participant and stimulus order, but neither reached significance. The final model is shown in Table 14. The base model without any of the measures of personal relevance had an AIC of 2555 ($r^2 = 0.13$). Adding the personal relevance measures decreased the AIC by 142, to 2413, again suggesting a strong effect of those measures on reducing information loss.

The fitted values are shown by PersonalRelevance category in Figure 7. The pattern of results is as predicted: participants judge words closer to the Personal-

Table 14. Summary of GAM for predicting familiarity judgment Predictors are ordered in terms of decreasing effective df. $R^2 = 0.18$. Terms involving personal relevance are bolded

Predictor	Effective df	Chi.sq	P-value
LogFreq*RelevantAbstract	**10.77**	**39.59**	**0.00011**
LogFreq*RelevantConcrete	**8.96**	**25.58**	**0.0079**
RelevantConcrete	**6.94**	**36.46**	**< 0.0001**
RelevantNoun	**5.77**	**49.92**	**< 0.0001**
Valence	5.52	19.17	0.006
PersonalRelevance	**4.84**	**69.93**	**< 0.0001**
LogFreq*Arousal	3.54	18.66	0.00053
Concreteness	3.12	13.54	0.0087
LogFreq	2.02	17.98	0.00033
Length	1	8.51	0.0035

Relevance CDV as most familiar and words further from the PersonalRelevance CDV as least familiar.

Discussion: Familiarity

The results of this experiment replicate the main finding from Westbury (2016): that words that are more personally self-relevant are recognized as more familiar than closely-matched words that are less self-relevant.

Conclusion

The results we have reported highlight the well-known difficulty of comparing LDRTs obtained with different nonword backgrounds and different populations. After collapsing by word, the SD of the RTs we collected ($SD = 92$ ms) and the RTs from the ELP ($SD = 91$ ms) were much larger than the SDs from the BLP ($SD = 55$ ms), which has much shorter RTs than either of the other two datasets.

Westbury (2014) showed that words estimated to be high on affective force were rated more familiar than matched words with low affective force. In the present study, we controlled for affective force by matching words on estimated valence, arousal, and dominance. Personal relevance is not simply synonymous with affective force, but rather captures an aspect related to affective force that is not captured by the affect estimates themselves.

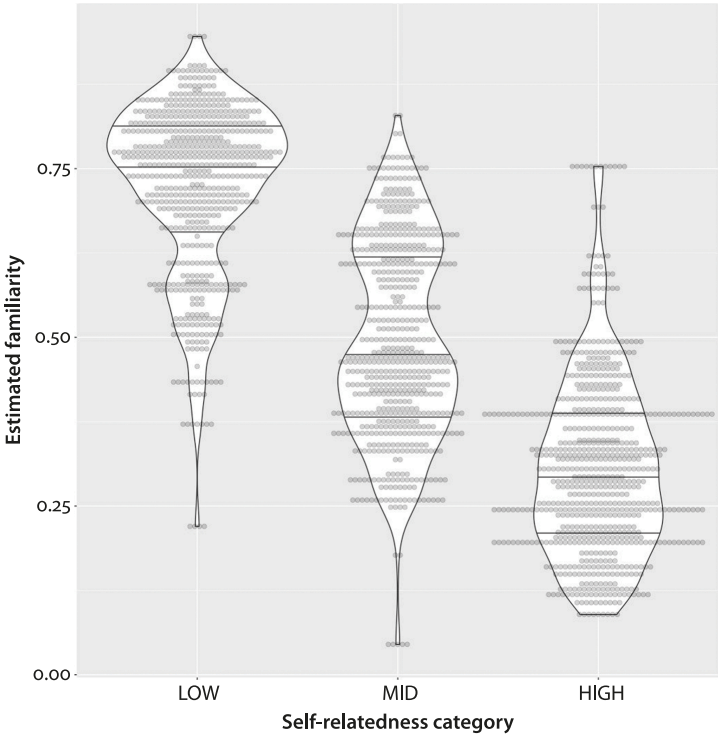

Figure 7. GAM-fitted familiarity estimates, by distance from the PersonalRelevance CDV. Individual points are observations. Black lines indicate quartiles

We have made all self-relevance estimates (as well as raw data and R code from the experiments described above, and the pool of 58,455 nonwords) available at the Open Science Foundation, https://osf.io/gdb6h/.

We have demonstrated across multiple datasets that an algorithmically-defined measure of self-relevance is implicated in speed and accuracy of lexical access. Words deduced to be more self-relevant based on their patterns of use are recognized as words more quickly and named faster. Self-relevance is also a strong predictor of several measures that are usually measured only by human judgments or their computational extrapolations, such as subjective familiarity, concreteness, age of acquisition, imageability, and body-object interaction.

There is a tendency in psychology to adjudicate the often-implicit issues of construct validity using just a single weak criterion: temporal precedence. Although we have never seen it stated bluntly, we have often encountered the suggestion that a construct cannot be scientifically admissible if it is correlated 'enough' (whatever that might mean) with pre-defined constructs. Of course, this is a not a useful criterion for adjudicating between constructs. There are many

cases of accepted constructs being shown to be correlates of psychometrically stronger constructs. For example, Baayen (2010) argued that word frequency disappeared as a predictor of LDRT if one computed local syntactic and morphological co-occurrence probabilities, suggesting that word frequency is a proxy for learned relationships between linguistic elements. Hollis (2020) recently presented evidence suggesting that the much-studied construct of context diversity was confounded with a nonlinear transformation of word frequency and had little or no effect as usually defined. In these and other cases, the admissibility of a construct has depended not on which construct happened to be named first, but rather on which construct was most convincingly able to tie a set of observations (the *explanandum*) into a separate domain (the *explanans*) that allows us to predict those observations (Hempel & Oppenheim, 1948; see discussion in Westbury, 2016). If we accept the idea that our constructs should be grounded in a domain *other than the one in which they are used as explanatory elements* (i.e., frequency effects explained using learning theory), then we have a better tool for adjudicating between constructs than historical happenstance. Judgments made by humans of measures such as subjective familiarity, age of acquisition, and body-object interaction have often been used (at least, implicitly) as *causal* elements in explaining lexical access. However, these constructs are just names given to observed correlations between measures we understand equally little: human judgments, on the one hand, and measures of lexical access, on the other hand (see Baayen, Feldman & Schreuder, 2006, p. 303–304).

Psychometricians have to learn to let go of their explanations, because to fall in love with a construct is (to borrow a phrase from the 13th century Zen Buddhist Wu-men Hui-hai) to tie one's self with false rope. We have offered multiple arguments in this paper for why we believe that personal relevance is a psychometrically more convincing construct than age of acquisition, subjective familiarity, or body-object interaction: because *it is grounded in something other than its own definition*. Although we do not wish to argue that this means it is 'true', we do argue that our analysis suggests that the role for age of acquisition or subjective familiarity or body-object interaction is smaller than previously thought. The observed correlations between human judgments and measures of lexical access cannot trump an explanation that is grounded in an independent explanans. They must yield to it. To be precise: to us it makes more sense to say that (say) age of acquisition is less important than some thought it was because we can explain about 20% of its variance using empirically-grounded measures than it does to say that personal relevance is unimportant because it happens to be correlated with peoples' opinions about when they first learned a word.

This is not intended to imply that personal relevance, as we have defined it here, is 'correct'. There are free parameters in our definitions of the personal

elevance measures that make it easy to question whether it is optimally defined. It is certainly possible that there is a 'better' (more general; better motivated; simpler; less parameterized; more epistemologically integrative) explanation than personal relevance for the partial variance in early language learning, human judgments, reaction times, and decision accuracy that we have tried to explain here. However, we offer the construct here because we believe that constructs that are not merely average measures of personal opinion are more useful than constructs that *are* merely average measures of personal opinion. The construct of personal relevance is then, perhaps, a small step in the right direction.

Funding

This work was supported by a grant to the first author from the Natural Sciences and Engineering Research Council (NSERC) of Canada

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski. (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alexopoulos, T., Muller, D., Ric, F., & Marendaz, C. (2012). I, me, mine: Automatic attentional capture by self-related stimuli. *European Journal of Social Psychology*, *42*, 770–779. https://doi.org/10.1002/ejsp.1882

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461. https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313. https://doi.org/10.1016/j.jml.2006.03.008

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639–647. https://doi.org/10.3758/BF03200465

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B. & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bird, H., Franklin, S., and Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, *33*, 73–79. https://doi.org/10.3758/BF03195349

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 158–167. https://doi.org/10.3758/BF03195507

Brown, G. D., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & cognition*, *15*(3), 208–216. https://doi.org/10.3758/BF03197718

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40*(3), 791–794. https://doi.org/10.3758/BRM.40.3.791

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods*, *40*(4), 1049–1054. https://doi.org/10.3758/BRM.40.4.1049

Ferrand, L., Grainger, J., & New, B. (2003). Normes d'âge d'acquisition pour 400 mots monosyllabiques [Age-of-acquisition norms for 400 monosyllabic French words]. *L'Année Psychologique*, *103*, 445–467. https://doi.org/10.3406/psy.2003.29645

Flieller, A., & Tournois, J. (1994). Imagery value, subjective and objective frequency, date of entry into the language, and degree of polysemy in a sample of 998 French words. *International Journal of Psychology*, *29*(4), 471–509. https://doi.org/10.1080/00207599408246553

Frings, C., & Wentura, D. (2014). Self-priorization processes in action and perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1737.

Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256. https://doi.org/10.1037/0096-3445.113.2.256

Ghyselinck, M., De Moor, W., & Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four-and five-letter nouns. *Psychologica Belgica*, *40*(2), 77–98. https://doi.org/10.5334/pb.958

Gilhooly, K.J., and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research, Methods Instruments, and Computers*, *12*, 395–427. https://doi.org/10.3758/BF03201693

Golubickis, M., Falben, J. K., Cunningham, W.A., & Macrae, C. N. (2018). Exploring the self-ownership effect: Separating stimulus and response biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(2), 295.

Hart, B. (1991). Input frequency and children's first words. *First Language*, *11*(32), 289–300. https://doi.org/10.1177/014272379101103205

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104–146. https://doi.org/10.1016/j.jml.2020.104146

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619. https://doi.org/10.1080/17470218.2016.1195417

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(*1*), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Leech, G., Rayson, P. & Wilson, A. (2001). Companion website for: Word Frequencies in Written and Spoken English: based on the British National Corpus. http://ucrel.lancs.ac.uk/bncfreq/

Marques, J.F., Fonseca, F.L., Morais, S., & Pinto, I.A. (2007). Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables. *Behavior Research Methods*, *39*(3), 439–444. https://doi.org/10.3758/BF03193013

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

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 (Neural Information Processing Systems Conference, 2013)*; pp. 3111–3119).

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain – a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440–457. https://doi.org/10.1016/j.neuroimage.2005.12.002

Schäfer, S., Wentura, D., & Frings, C. (2015). Self-prioritization beyond perception. *Experimental Psychology*. https://doi.org/10.1027/1618-3169/a000307

Schäfer, S., Frings, C., & Wentura, D. (2016). About the composition of self-relevance: Conjunctions not features are bound to the self. *Psychonomic Bulletin & Review*, *23*(3), 887–892. https://doi.org/10.3758/s13423-015-0953-x

Shaoul, C. & Westbury, C. (2006). USENET Orthographic Frequencies for 111,627 English Words. (2005–2006) Edmonton, AB: University of Alberta (downloaded from http://www.psych.ualberta.ca/~westburylab/downloads/wlfreq.download.html

Schmitz, T.W., & Johnson, S.C. (2007). Relevance to self: A brief review and framework of neural systems underlying appraisal. *Neuroscience & Biobehavioral Reviews*, *31*(4), 585–596. https://doi.org/10.1016/j.neubiorev.2006.12.003

Stadthagen-Gonzalez, H., & Davis, C.J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*, 598–605. https://doi.org/10.3758/BF03193891

Sui, J., He, X., & Humphreys, G.W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1105–1117. https://doi.org/10.1037/a0029792

Symons, C.S., & Johnson, B.T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin*, *121*(3), 371. https://doi.org/10.1037/0033-2909.121.3.371

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body-object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075–1078. https://doi.org/10.3758/BRM.40.4.1075

Westbury, C. (2014). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, *43*(5), 631–649. https://doi.org/10.1007/s10936-013-9266-2

Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, *11*.3, 350–374. https://doi.org/10.1075/ml.11.3.02wes

Westbury, C., & Hollis, G. (2018). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, 1–28.

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language*, *99*, 122–150. https://doi.org/10.1016/j.jml.2017.09.006

Westbury, C., & Nicoladis, E. (1998). Meaning in children's first words: Implications for a theory of lexical ontology. In *Proceedings of the 22nd Annual Boston University Conference on Language Development* (pp. 768–778). Cascadilla Press: Somerville, MA.

Address for correspondence

Chris Westbury
Department of Psychology
University of Alberta
P217 Biological Sciences Building
Edmonton, AB, T6G 2E9
Canada

chrisw@ualberta.ca

Co-author information

Lee H. Wurm
Department of Psychology
Gonzaga University

wurm@gonzaga.edu

Publication history

Date received: 26 August 2020
Date accepted: 7 February 2022
Published online: 18 March 2022