THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2016
http://dx.doi.org/10.1080/17470218.2016.1195417

Extrapolating human judgments from skip-gram vector representations of word meaning

Geoff Hollis[a], Chris Westbury[a] and Lianne Lefsrud[b]

[a]Department of Psychology, University of Alberta, Edmonton, AB, Canada; [b]Department of Material & Chemicals Engineering, University of Alberta, Edmonton, AB, Canada

ABSTRACT

There is a growing body of research in psychology that attempts to extrapolate human lexical judgments from computational models of semantics. This research can be used to help develop comprehensive norm sets for experimental research, it has applications to large-scale statistical modelling of lexical access and has broad value within natural language processing and sentiment analysis. However, the value of extrapolated human judgments has recently been questioned within psychological research. Of primary concern is the fact that extrapolated judgments may not share the same pattern of statistical relationship with lexical and semantic variables as do actual human judgments; often the error component in extrapolated judgments is not psychologically inert, making such judgments problematic to use for psychological research. We present a new methodology for extrapolating human judgments that partially addresses prior concerns of validity. We use this methodology to extrapolate human judgments of valence, arousal, dominance, and concreteness for 78,286 words. We also provide resources for users to extrapolate these human judgments for three million English words and short phrases. Applications for large sets of extrapolated human judgments are demonstrated and discussed.

ARTICLE HISTORY
Received 22 February 2016
Accepted 17 May 2016

KEYWORDS
Affect; Co-occurrence models; Human judgment; Semantics; Skip-gram; Word2vec

High-quality word norm sets are of broad value to research within psychology. For example, norm sets for affective measures of valence, arousal, and dominance have been used in research aimed at understanding lexical access (Larsen, Mercer, & Balota, 2006; Larsen, Mercer, Balota, & Strube, 2008; Võ et al., 2009; Westbury et al., 2015), processing of political concepts (Lodge & Taber, 2005), the function of the amygdala (Hamann & Mao, 2002) and the role of cortisol levels in recall (Abercrombie, Kalin, Thurow, Rosenkranz, & Davidson, 2003). Developing large norm sets is an essential part of supporting quantitative and experimental approaches within psychology. However, running large numbers of participants and validating norms is an involved process. Recent crowdsourcing efforts through services like Mechanical Turk are an important methodological advancement for reducing many of the burdens that come with large-scale data collection (e.g., Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Warriner, Kuperman, & Brysbaert, 2013) but, even so, financial considerations in combination with the large variety seen both between and within languages conspire to ensure that purely brute force approaches have practical limitations.

One solution to the tractability problem is to computationally extrapolate human judgments from norms that already exist. Numerous researchers have been recently tackling this problem, particularly within the context of affective measures of valence, arousal, and dominance (e.g., Bestgen & Vincze, 2012; Mandera et al., 2015; Recchia & Louwerse, 2015; Westbury et al., 2015). One benefit of such

CONTACT Geoff Hollis hollis@ualberta.ca
Supplemental material is available via the "Supplemental" tab on the article's online page (http://dx.doi.org/10.1080/17470218.2016.1195417)

work is clear when numbers of rated items are compared. The often-used ANEW norm set (Bradley & Lang, 1999) for valence, arousal, and dominance contains 1034 items. In comparison, Westbury et al. (2015) released a dataset of extrapolated valence and arousal values for over 70,000 words. This order of magnitude larger scope affords researchers unprecedented control over selecting for specific stimulus properties when designing experiments.

The utility of extrapolated human judgments hinges on the validity of the extrapolation procedure. Mandera et al. (2015) suggest two main reasons to be pessimistic about current practices. First, the methodologies being used for extrapolation are not well suited for estimations of the extremes of human judgments. Second, some important validation criteria are not applied in most of the previously reported research.

Much of the effort to extrapolate human semantic judgments has relied on co-occurrence models of semantics and has been focused primarily on extrapolating the affective properties of valence, arousal, and dominance (e.g., Bestgen & Vincze, 2012; Mandera et al., 2015; Recchia & Louwerse, 2015; Westbury et al., 2015). One way to approach the problem is with the *k-nearest-neighbours* estimation technique: The semantic properties of a word are estimated by averaging over relevant values for a set of words similar to it in meaning (as measured by a co-occurrence model) that human estimates are available for. Suppose the valence of *exuberant* is unknown, but it is similar in meaning to *joyous* and *party*. Furthermore, human judgments of valence are available for both *joyous* and *party*. The valence of *exuberant* can be estimated by averaging the human judgments available for *joyous* and *party*.

The *k-nearest-neighbours* approach allows small norm sets of human judgments to be used to extrapolate values for larger sets of unjudged words. It is the most common and successful method for extrapolating human judgments from co-occurrence models of semantics (e.g., Bestgen & Vincze, 2012; Mandera et al., 2015; Recchia & Louwerse, 2015; but see Westbury et al., 2015, for an alternate approach). However, since the k-nearest-neighbours approach is relying on averaging across similar words, it is inherently limited in its ability to extrapolate to extreme values. The highest and lowest value words will always, by necessity, be mis-estimated due to the use of averaging. The practical limitations of k-nearest-neighbours approaches still need to be rigorously empirically tested. For the moment, it remains largely a theoretical concern. Regardless, this concern motivates the search for new methodologies that allow for extrapolation without the averaging of human judgments.

The second problem with extrapolating human judgments lies in the error component of such extrapolations. If extrapolated judgments are to be of any use for psychological research, they need to carry information about the semantic judgments they are modelling. This is the portion of the problem that most researchers have previously been focused on. However, it is also important that the error component does not contain systematicities. For example, if a human judgment is estimated by using word frequency measures because word frequency happens to correlate with that human judgment, it becomes a point of concern whether or not systematic variation having to do with word frequency now contaminates estimates. If the answer is "yes", then such estimates lose much of their utility for experimental research and computational modelling.

The issue is well illustrated by Mandera et al. (2015), who built predictive models of valence, arousal, and dominance that correlate with actual human judgments at $r = .69$, $r = .49$, and $r = .59$, respectively. Their models have high predictive validity of actual human judgments. However, Mandera et al. showed that actual human judgments and predictions of those human judgments differ in critical ways. For example, there are strong relationships between estimates of dominance and word frequency ($r = .78$), between estimates of valence and word frequency ($r = .97$), and between estimates of arousal and word length ($r = .51$). These values are all well above what is observed between actual human judgments and lexical variables: $r = .16$, $r = .17$, and $r = .10$, respectively. These disparities in correlation strength indicate that the error component of human judgment estimates reported by Mandera et al. systematically introduces variation having to do with word frequency and word length. This reduces the validity of these estimates for application within psychological research where word frequency or word length might also be relevant variables. Current work on extrapolating human judgments needs validation criteria that go beyond maximizing variance accounted for in the prediction target.

One way validation criteria can be improved is by ensuring that predictions of human judgments share the same correlational structure with other

lexical and semantic properties of words as do the human judgments themselves. This helps rule out the possibility that estimates contain systematic error. Researchers should consider validating extrapolations against lexical properties that are known to influence lexical access: as a minimum set, word length, log word frequency, and orthographic neighbourhood size, all of which play a central role in benchmarking models of lexical access (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013). Researchers should also consider validating against behavioural measures of lexical access itself: lexical decision times and word naming times. Estimates for each of these variables for over 40,000 words can be downloaded via the English Lexicon Project (ELP) (Balota et al., 2007), and estimates for nearly 30,000 words can be downloaded from the British Lexicon Project (BLP) (Keuleers, Lacey, Rastle, & Brysbaert, 2012). Effective validation techniques allow researchers to extrapolate human judgments in a way that ensures that the extrapolations have unsystematic error.

Methods for extrapolating human judgments need to be improved. We introduce one such improved method: Rather than inferring values from the properties of semantic neighbours, we extrapolate semantic properties by regressing raw vector representations on the human judgment to be predicted. This methodology is based on the reasoning that if a vector representation captures semantic properties of words, then its dimensions should be usable to accurately predict semantic properties of words. We describe our methodology for extrapolating human judgments of word valence, arousal, dominance, and concreteness from vector representations of word meaning. We also show that this approach improves on previous attempts to extrapolate human judgments.

We start by providing a brief introduction to co-occurrence models of semantics and explain why we choose to work with a particular one (the skip-gram model using negative sampling).

Co-occurrence models of semantics

The basic observation motivating co-occurrence models of semantics is that words with similar meanings occur within similar contexts. The specific details of what constitutes context and how context should be represented are where variations in models exist. The two classic co-occurrence models are latent semantic analysis (LSA; Landauer & Dumais, 1997) and the hyperspace analogue to language (HAL; Lund & Burgess, 1996) and its many derivations (Durda & Buchanan, 2008; Hofmann, Kuchinke, Biemann, Tamm, & Jacobs, 2011; Jones & Mewhort, 2007; Rhode, Gonnerman, & Plaut, 2007; Shaoul & Westbury, 2006, 2010a, 2011). Within LSA-style models, context is defined as a document of text. In HAL-style models, context is defined as a finite window of words preceding and following a target word. Regardless of these differences, as the contexts of two words become more similar, their semantic similarity is also assumed to become more similar.

Within co-occurrence models, a word's meaning is represented as a context vector. Consider the sentences, *the boy ate the cookie* and *the dog ate the bone*. Within a HAL-style model, the words *boy* and *dog* would each have a vector representing its unique meaning. Within this vector, there would be dimensions for the various contexts in which the words could occur. One dimension would correspond to how often the word occurred with *the*, another for occurrences with *ate*, a third with *cookie*, a fourth with *bone*, and many more dimensions for the variety of other contexts that these words might occur within. In this particular limited case, the vector for *boy* would be [2, 1, 1, 0] whereas the vector for *dog* would be [2, 1, 0, 1]. The similarity of those vectors specifies the similarity in meaning of the words.

Such representations tend to be long (there are many unique contexts in which words can occur) and sparse (in practice, words tend to be used only in specific contexts). To simplify representations, models will usually use dimensional reduction techniques. At its simplest, this may simply involve truncation of vectors to the *n* most variable dimensions (where *n* might be a couple of thousand; Lund & Burgess, 1996). In other cases, statistical tools like singular value decomposition are used for dimensional reduction of vectors (e.g., Landauer & Dumais, 1997). Often these models will require large corpora of documents (millions or billions of words) on which to construct word vectors. Such data are readily available on the internet through websites like Wikipedia. Co-occurrence models prove to be an effective tool for modelling semantics. For instance, models can make similarity judgments between words that allow them to pass tests of English as a foreign language with basic competence (e.g., Landauer & Dumais, 1997).

Another type of co-occurrence model has recently been developed in the natural language processing literature. This type relies on machine learning tools to predict the contexts in which words occur. The

most prominent of this class is Google's skip-gram model with negative sampling (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Rather than representing word meanings as raw co-occurrence values (e.g., HAL-types), these models use a neural network to find relationships that map each word to its context of use. The skip-gram model with negative sampling (sometimes referred to as word2vec, after Google's free software that implements this algorithm) trains a neural network to predict the words preceding and following a target word across large corpora of text. Errors are back-propagated to a vector representation of the input word. The vector's values are adjusted to minimize error on future encounters with that word–context pair.

HAL-type models represent word meaning as the literal contexts in which they occur. In contrast, the skip-gram model's representations of word meaning drive predictions of contextual details. The skip-gram model is not encoding *context*; it is encoding *abstractions that predict context*. Thus, it is more akin to models like LSA, which employ dimensionality reduction techniques that extract patterns of co-occurrence relationships across documents. Also like LSA, the skip-gram model uses short vector representations (vector lengths of tens or hundreds, rather than lengths of thousands). The consequence is that the skip-gram model's (and LSA's) vectors are informationally more dense than vectors from HAL-type models. One of the critical points of divergence between LSA and the skip-gram model is scalability: The computational complexity of constructing an LSA model increases supralinearly with the number of documents over which vectors are constructed, whereas the skip-gram model only increases linearly. Consequently, the skip-gram model can be trained on substantially larger corpora, resulting in higher quality word vectors.

The technical documents describing the skip-gram model are opaque on issues of why certain design decisions were made, and why they work as well as they do (see Goldberg & Levy, 2014, for discussion, as well as our final discussion below). The motivation for most of the design decisions is probably pragmatic: The skip-gram model is designed the way it is, because that is the way that worked the best. It is hard to argue with a response like this when you compare the skip-gram model's performance to alternate models; it has unprecedented success on a range of semantic relatedness judgments (Mikolov, Chen, et al., 2013). However, given that the skip-gram model is so good at modelling semantic relationships, there are strong motivations to try and understand why.

On the technical end, Levy and Goldberg (2014) have proven that the skip-gram model converges on a factorization of the word-context matrix that constitutes its training data. The question of *why* matrix factorization results in human-plausible semantic representations is a separate topic that still does not have a satisfying answer (Levy & Goldberg, 2014). We have made preliminary steps towards addressing this question: The skip-gram model embeds word meaning within a multidimensional space whose primary dimensions of variability organize words along affective axes (Hollis & Westbury, in press). This is consistent with observations that humans organize meaning primarily along affective dimensions (e.g., Osgood, Suci, & Tannenbaum, 1957). Possibly the skip-gram model is converging on a psychologically plausible means of extracting meaning from context that is omitted from other models. For this reason, the skip-gram model presents itself as a promising model on which to extrapolate human semantic judgments from raw vector representations.

Experiment 1

Method

Word vectors

Google has previously released a dataset of vector representations for the three million most common words (e.g., *river*) and phrases (e.g., *Yangtze_River*) in a six billion word subset of the Google News corpus. These vectors are free for download and use, and can be found at the word2vec code repository (Word2vec, 2013). This dataset represents word meaning as vectors of length 300.

We use a subset of these vectors to extrapolate human judgments. This subset we use corresponds to the 60,453 words that occurred in both (a) the three million Google news vector dataset and (b) in any of the following norm sets of human judgments or responses (Balota et al., 2007; Brysbaert et al., 2014; Keuleers et al., 2012; Kuperman et al., 2012; Warriner et al., 2013).

Data for model training

Very large norm sets of human judgments have recently been collected through use of Amazon's

Mechanical Turk. The norm sets contain human judgments for valence, arousal, and dominance (n = 13,915; Warriner et al., 2013), concreteness (n = 37,058; Brysbaert et al., 2014), and age of acquisition (n = 30,121; Kuperman et al., 2012). We build and validate models that predict corresponding human judgments in the first two of the three norm sets (valence, arousal, dominance, and concreteness). We have built a model of age of acquisition, but omit it from our current work. We found it difficult to produce an adequate model for human judgments of age of acquisition, despite trying numerous transformations of both the predictors and the target variable. Specifically, words estimated as being acquired before approximately age 7 were not well fitted by any model we tried. Further work is required to find an appropriate transformation to put age of acquisition on a scale that can be effectively predicted from the skip-gram model's vectors (but see discussion in Westbury, 2013, on the formal problems with age of acquisition as a predictor of lexical access).

Model training

Separate models for each of our four human judgments were trained. In each case, models were constructed following the same sequence of steps. First, the words for which human judgments were available were randomly split into two halves: a training set, and a validation set. A linear regression model was created to predict human judgments from the skip-gram model's dimensions. This model was developed on the training data. The model initially included all 300 skip-gram model dimensions in the regression equation. Backwards stepwise regression was then performed to sequentially eliminate all dimensions that did not reduce information loss according to the Akaike information criterion (AIC). The model was then reconstructed using forward stepwise regression on only the previously chosen terms. The final model included only terms that reduced information loss (AIC) in the forward direction.

As noted above, the skip-gram model is setting vector weights by implicitly performing matrix factorization (Goldberg & Levy, 2014), and thus the 300 dimensions that represent word meaning are all near-orthogonal. We therefore have no reason to be concerned with covariation among predictors.

Model validation

We use two criteria to ensure the validity of extrapolated judgments. The first criterion is prediction of human judgments over validation subsets created from data in Warriner et al. (2013) and Brysbaert et al. (2014). We additionally check the predictive validity of our estimates by comparing them to the separate norms of Bradley and Lang (1999) and Coltheart (1981). The former shall be referred to as the ANEW affect norms, and the latter MRC (Medical Research Council) concreteness norms. This step ensures the predictive validity of extrapolated judgments.

Our second validation criterion is prediction against lexical and semantic variables. The second validation step ensures that predicted human judgments share the same statistical structure with lexical and semantic variables as actual human judgments. We validate against affective measures taken from Warriner et al. (2013; valence, arousal, dominance), concreteness measures taken from Brysbaert et al. (2014), lexical (log word frequency, word length, orthographic n) and lexical access measures (lexical decision time, word naming time) taken from the ELP (Balota et al., 2007), and a lexical access measure (lexical decision time) taken from the BLP (Keuleers et al., 2012). The English Lexicon Project provides multiple measures of word frequency. We use frequencies calculated from movie subtitles, as they have the highest predictive validity of lexical access times.

If the final forward stepwise model showed no sign of overfitting based on our validation criterion, its term weights were recalculated on the combined training and primary validation data. Human judgments are then extrapolated for the 78,286 unique words that (a) we have vector representations for and (b) are included in any of the above-mentioned norm sets, plus those used in Westbury et al. (2015).

Data transformation

Concreteness ratings and all three of the affect measures were reported on scales with restricted ranges (1–5 for concreteness, 1–9 for affect). This poses potential conceptual problems for linear regression, as predictions might be made outside of the plausible range of response (e.g., the model might make a nonsense prediction of 10 out of 9 for valence on a particular word). In response to this, we considered two alternate approaches to model construction. First, we tried constructing models on the logit-transform of the four human judgments: log $[y_s/(1-y_s)]$ where y_s is the value of y, scaled to be between 0 and 1. This transformation maps the judgments onto a range of negative to positive infinity. Second, we also tried linear regression where the

output of the model was clamped at the ends of each scale. In all cases, we found clamping of model output to provide better fits to human judgments than logit-transforming the variable being predicted. Thus, we only report results for construction of output-clamped models.

Results

Model construction

Model results for predicting human judgments are displayed in Table 1. In each case, the model validated well. The largest discrepancy between training set and validation set performance is for dominance judgments (training $r = .715$, validation $r = .685$). According to the Fisher r-to-z transformation for comparing two correlation coefficients, this is a reliable difference ($z = 3.35$, $p = .0008$). When the model's term weights are recalculated on the validation data, $r = .704$, the recalculated fit is not reliably different from the training set performance ($z = 1.65$, $p = .10$). The smallest discrepancy is for concreteness judgments (training $r = .835$; validation $r = .828$). This, too, is a reliable difference ($z = 2.14$, $p = .03$). However, when the model's term weights were recalculated on the validation data ($r = .833$) there was no reliable difference from performance on the training data ($z = 0.53$, $p = .59$). Similar patterns of results are seen for valence and arousal. These results provide evidence that the model did not overfit to the training data.

Since signs of overfitting to the training set were not present, the model's term weights were recalculated by pooling both the training and validation set together. The performance of the recalculated model is shown in Table 1. The best performance was in predicting concreteness judgments ($r = .833$), and worst performance in predicting arousal judgments ($r = .620$). Low performance on predicting arousal judgments is consistent with previous observations that arousal judgments are difficult to model (Mandera et al., 2015; Recchia & Louwerse, 2015; Westbury et al., 2015), in part because independent human judgments of arousal are not as well correlated as other human judgments (Westbury et al., 2015; Westbury et al., 2013).

Predictive validity of affective judgments

It is informative to compare model performance at predicting human judgments to measures of inter-norm reliability and split-half reliability. We start with the affective measures of valence, arousal, and dominance. Reported split-half reliabilities for valence, arousal, and dominance judgments are .914, .689, and .770, respectively (Warriner et al., 2013). Warriner et al. (2013) additionally split their affect judgments by demographic features and correlate scores between demographic groups. Specifically, they looked at male versus female, young versus old, and high and low education groups. While model predictions are overall lower than split-half reliabilities of the full dataset, they are comparable to correlations between demographic groups for valence [$r = .799$ vs. .79 (split by gender), .82 (split by age), .83 (split by education)], and superior for arousal [.620 vs. .52 (gender), .50 (age), .41 (education)] and for dominance [.704 vs. .59 (gender), .59 (age), .61 (education)].

Examining inter-norm reliabilities is another way to assess the predictive validity of model performance. There is a strong relationship between Warriner et al. (2013) norms and the ($n = 1034$) ANEW norms: $r = .953$ for valence; $r = .759$ for arousal; and $r = .795$ for dominance. These values are comparable to the split-half reliabilities reported by Warriner et al.

When our model predictions are correlated with the ANEW norms, we see lower correlations for valence ($r = .872$), arousal ($r = .704$), and dominance ($r = .723$). Our model is not accounting for all of the variance shared between norm sets, but it is accounting for a very large proportion of it. To estimate the proportion of psychologically relevant variation that our extrapolated norms are accounting for, we take the ratio of two r^2 values: the r^2 between model extrapolations and the ANEW norms, divided by the r^2 between the ANEW and Warriner et al. (2013) norms. This provides a metric of the amount of variance that our model *is* accounting for in the ANEW norms, compared to the amount of variance that our model *could* account for in the ANEW norms, as estimated by a separate norm set. This produces estimates of 95.65% for valence, 96.31% for arousal, and 95.36% for dominance. Our models are accounting for nearly all of the psychologically plausible variance in ANEW norms.

Finally, we compare the performance of our model to that of previously reported models. For affective judgments, Recchia and Louwerse (2015) is currently the best performing model. Like ours, the Recchia and Louwerse models are developed on the Warriner et al. (2013) norms. Our models perform reliably better than the Recchia and Louwerse models at predicting human judgments of valence ($r = .799$ vs. .765), $z = 7.30$, $p = 2.2e\text{-}16$, arousal

Table 1. Final model performance on both the training and validation sets.

	Training model				Combined model		Recchia norms		Westbury norms	
Target	Training *r*	*n*	Validation *r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*
Valence	.807	6962	.786	6961	.799	13,923	.765	13,777	.704	13,783
Arousal	.623	6962	.611	6961	.620	13,923	.575	13,777	.489	13,783
Dominance	.715	6962	.685	6961	.704	13,923	.654	13,777	—	—
Concreteness	.835	19,977	.829	19,977	.833	39,954	—	—	—	—

Note: After model development, term weights were recalculated by running the model on the pooled sets of data. Model performance from previously reported norms (Recchia & Louwerse, 2015; Westbury et al., 2015) are also included for comparison. Our models better predict valence, arousal, and dominance judgments than previously reported models.

($r = .620$ vs. $r = .575$), $z = 5.83$, $p = 2.2e\text{-}16$, and dominance ($r = .704$ vs. $r = .654$), $z = 7.73$, $p = 2.2e\text{-}16$. Our models also outperform those of Recchia and Louwerse when validated on the ANEW norms. There are differences in correlation strength for valence ($r = .872$ vs. $r = .800$), $z = 5.5$, $p < 2.2e\text{-}16$, arousal ($r = .704$ vs. $r = .620$), $z = 3.4$, $p = .007$, and dominance ($r = .723$ vs. $r = .660$), $z = 2.74$, $p = .0061$, judgments. Our models of valence, arousal, and dominance have reliably higher predictive validity than the next-best model.

Predictive validity of concreteness judgments

Model estimates of concreteness correlate with the Brysbaert et al. (2014) data at $r = .833$. Brysbaert et al. do not report a split-half reliability score for their concreteness norms, so we are unable to directly compare model performance to this value. We further compare our extrapolated judgments to concreteness estimates from the MRC norms. Over the $n = 3{,}937$ words that the Brysbaert et al. norm set shares with the MRC database, our estimates correlate with these concreteness norms at $r = .835$. The Brysbaert et al. norms correlate with the MRC norms at $r = .918$. Taking the ratio of r^2 values, our estimates account for 95.37% of the variance that could be accounted for in MRC concreteness judgments, based on a theoretical maximum estimated by the correlation between MRC concreteness judgments and the Brysbaert et al. concreteness norms. As with the affective measures, this is an indication that our concreteness estimates are honing in on a large portion of the relevant variation that forms human concreteness judgments.

The only other model of concreteness judgments we are aware of is reported by Mandera et al. (2015). They report that their model cross-validates with $r = .796$ on a quarter subset of the Brysbaert et al. (2014) data. Our model cross-validated to half of the Brysbaert et al. (2014) data with $r = .829$. This difference in correlations is reliable, $z = 7.94$, $p = 2.2e\text{-}16$.

Validation against lexical and semantic measures

We now verify that our predicted judgments share the same statistical structure with lexical and semantic measures as do actual human judgments. To do this, we correlate (a) actual human judgments and (b) estimated human judgments with three lexical variables (log frequency, word length, orthographic neighbourhood size), three measures of lexical access (lexical decision time and lexical naming time taken from the ELP and lexical decision time taken from the BLP), and four semantic measures (valence, arousal, dominance, and concreteness). Statistical relationships are displayed in Table 2.

Overall, actual human judgments and our estimated human judgments relate to lexical variables in similar ways. We observe no differences in statistical structure shared between human judgments or our estimates, and lexical measures.

We do observe discrepancies for concreteness and arousal on measures of lexical access. Our arousal estimates predict lexical decision times slightly better than human judgments of arousal ($r = .072$ vs. $.047$), $z = 2.0$, $p = .045$. We discount this difference based on its reliability and the number of comparisons that are being made.

Concreteness estimates show discrepancies for both word naming times and lexical decision times. However, these differences are marginal. Given that the differences are for measures of lexical access times but not lexical variables, we interpret the effect as indicating that our estimates are uncontaminated by confounding lexical factors, but have not accounted for all of the psychologically relevant variation present in human judgments of concreteness as it pertains to lexical access.

Table 2. Correlation strength between human judgments and estimations for three lexical variables, three measures of lexical access, and four semantic variables.

	Valence		Arousal		Dominance		Concreteness	
Predicted variables	Human	Model	Human	Model	Human	Model	Human	Model
Lexical variables (ELP)	$n = 12{,}707$		$n = 12{,}707$		$n = 12{,}707$		$n = 23{,}391$	
Log frequency	.188	.178	.031	.024	.170	.186	.137	.124
Word length	−.029	−.037	.109	.094	−.040	−.038	−.341	−.337
Orthographic *n*	.015	.009	−.094	−.081	.034	.034	.208	.201
Lexical access (ELP)	$n = 12{,}707$		$n = 12{,}707$		$n = 12{,}707$		$n = 23{,}391$	
Lexical decision times	−.176	−.174	.047	.072*	−.179	−.184	−.241	−.211***
Naming times	−.122	−.117	.053	.061	−.134	−.130	−.238	−.217*
Lexical access (BLP)	$n = 7{,}819$		$n = 7{,}819$		$n = 7{,}819$		$n = 12{,}868$	
Lexical decision times	−.189	−.177	−.048	−.046	−.177	−.187	−.067	−.013*
Semantic variables	$n = 13{,}793$		$n = 13{,}793$		$n = 13{,}793$		$n = 33{,}973$	
Valence	—	—	−.181	−.274***	.718	.733*	.089	.103
Arousal	−.181	−.212*	—	—	−.176	−.25***	−.167	−.177
Dominance	.718	.652***	−.176	−.279***	—	—	.011	.031
Concreteness	.089	.11	−.167	−.173	.011	.032	—	—

Notes: Cases where human judgments and estimations differ in their strength of relationship to lexical variables are marked. BLP, British Lexicon Project; ELP, English Lexicon Project.
*$p < .05$. **$p < .005$. ***$p < .0005$.

Our extrapolated judgments do show evidence of semantic contamination. Model estimates of valence are more strongly related to human judgments of arousal and less strongly related to human judgments of dominance than actual human judgments of valence are. Likewise, model estimates of both arousal and dominance are more strongly related to human judgments of affect than expected.

To our knowledge, the only other researchers performing this type of validation procedure for estimates of human judgments are Mandera et al. (2015). They find that their estimates of human judgments are highly contaminated by lexical variables. Although we do find some evidence of contamination (particularly with affective variables), that contamination is not nearly as pronounced as what is reported by Mandera and colleagues. In the case of concreteness judgments, our estimates show no indication of contamination from either lexical of semantic characteristics.

To further assess the relative quality of our extrapolated norms, we downloaded the norms of Recchia and Louwerse (2015) and Westbury et al. (2015) and applied the above validation procedure to each set. Results can be found in Table 3. Variables from both norm sets show reliable contamination from both lexical and semantic variables. Word frequency, concreteness, and dominance judgments appear to be

Table 3. Correlation strength between human judgments and their estimations (Recchia & Louwerse, 2015; Westbury et al., 2015) for three lexical variables, three measures of lexical access, and four semantic variables.

	Recchia and Louwerse (2015) norms						Westbury et al. (2015) norms			
	Valence		Arousal		Dominance		Valence		Arousal	
Predicted variables	Human	Model	Human	Model	Human	Model	Human	Model	Human	Model
Lexical variables (ELP)	$n = 12{,}703$		$n = 12{,}703$		$n = 12{,}703$		$n = 12{,}691$		$n = 12{,}691$	
Log frequency	.188	.235***	.031	−.033***	.170	.279***	.188	.100***	.030	.040
Word length	−.029	−.026	.109	.226***	−.040	−.053	−.030	−.021	.109	.072**
Orthographic n	.015	1.038	−.094	−.162***	.034	.068*	.016	−.003	−.094	−.063*
Lexical access (ELP)	$n = 12{,}703$		$n = 12{,}703$		$n = 12{,}703$		$n = 12{,}691$		$n = 12{,}691$	
Lexical decision times	−.176	−.181	.047	.124***	−.179	−.221**	−.177	−.126***	.048	.063
Naming times	−.122	−.138	.053	.128***	−.134	−.170***	−.123	−.080***	.053	.051
Lexical access (BLP)	$n = 7{,}827$		$n = 7{,}827$		$n = 7{,}827$		$n = 7{,}819$		$n = 7{,}8197$	
Lexical decision times	−.189	−.191	−.048	.005***	−.177	−.237***	−.19	−.104***	−.048	.019***
Semantic variables	$n = 13{,}777$		$n = 13{,}777$		$n = 13{,}777$		$n = 13{,}783$		$n = 13{,}783$	
Valence	—	—	−.181	−.289***	.718	.707	—	—	−.183	−.278***
Arousal	−.181	−.194	—	—	−.176	−.224***	−.183	−.182	—	—
Dominance	.718	.613***	−.176	−.274***	—	—	.718	.592***	−.178	−.267***
Concreteness	.089	.123**	−.167	−.322***	.011	.092***	.095	.065**	−.167	−.308***

Notes: Cases where human judgments and estimations differ in their strength of relationship to lexical variables are marked. BLP, British Lexicon Project; ELP, English Lexicon Project.
*$p < .05$. **$p < .005$. ***$p < .0005$.

the most consistently influential source of contamination. We conclude that our attempts at estimating human judgments is a marked improvement over previous attempts. This is evidenced both by a reduction in contamination from extraneous variables and as an improvement in the predictive validity of human judgments.

Combining models

We were curious whether our norms are strictly better than previous reported norm sets, or whether different extrapolated norm sets are honing in on different aspects of human judgments. We built two regression models to predict human judgments for each of valence, arousal, and dominance: one using just our extrapolated values, and one using both our extrapolated values and those provided by Recchia and Louwerse (2015). In each case, the model using both extrapolated norms better predicted human judgments than the model using just our extrapolated norms [for all F (1,13757), $p = 2.2e\text{-}16$]. Combined predictions correlated with valence at $r = .836$, arousal $r = .655$, and dominance $r = .731$. Including the Westbury et al. (2015) norms did not substantially improve predictions of affective measures.

Our extrapolated norms of human judgments are improvements to previously reported extrapolated norms. However, they are not strictly superior. Combining estimation methods and/or identifying new methodologies is a direction for future research.

Adding lexical predictors

We additionally constructed models including lexical variables (word length, log frequency, orthographic neighbourhood size) as predictors for human judgments. The resulting lexical-based extrapolations better predicted human judgments than the skip-gram-only models by a marginal amount: valence $r = .801$; arousal $r = .631$; dominance $r = .708$; concreteness $r = .861$ (compare to Table 1). However, in all cases, this introduced contamination by said lexical variables. As an example, for valence, lexical-based extrapolations correlated with log frequency at $r = .258$, word length at $r = -.086$, and orthographic neighbourhood size at $r = .049$. These lexical-based extrapolated values more strongly correlate with lexical properties than actual human judgments (Table 2). We do not include lexical predictors in our final extrapolated values.

Discussion

Stimulus norms are of important and varied use to psychological research. However the investment of time, labour, and money required to compile norm sets makes it challenging to engage in these activities on a large scale. Crowd-sourcing data collection can go a long way to ease the burdens of compiling norms. However, even brute-force approaches are fundamentally ill equipped to the problem considering the scope and variety of language use.

Extrapolation from computational models of semantics may help address the tractability problem inherent in brute-force approaches to compiling psychological norm sets. This is a particularly promising possibility, given the major recent advances that the field of semantic modelling is seeing. We add to this body of literature by presenting a new methodology for extrapolating human judgments that provides estimates with high predictive validity and low contamination from extraneous variables. We release an extrapolated norm set with entries for 78,286 words. We also provide software that enables users to extrapolate human judgments of valence, arousal, dominance, and concreteness for up to three million words and short phrases.

Experiment 2

A fair question is, what is the point of having such massively large norm sets in the first place? Psychology has done just fine with norms of a few thousand words so far. Even if that were true, larger norm sets means better granularity and choice when controlling stimulus properties. That choice enables psychological research to expand its scope and precision of operation.

Another argument is that large norm sets open up qualitatively new lines of research within psychology. For example, the emerging trend of megastudies in psycholinguistic research has proven to be illuminating about numerous factors that affect lexical access (Balota et al., 2007; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers et al., 2012). However, this approach is predicated on using regression over large quantities of data points (tens of thousands). Producing large norm sets helps cultivate possibilities for large-scale modelling studies of lexical access.

Large norm sets also allow language research to begin grappling with more complex textual units. Psycholinguistic research tends to treat the word unit as if

it is a privileged level of analysis from which principles about linguistic acts more generally can be inferred. However, it has been pointed out that pervasive effects seen in word reading, like effects of word frequency, may not exist for connected text reading (Wallot, Hollis, & van Rooij, 2013), and that connected text reading may be functioning in a qualitatively different way than we infer from standard research paradigms within psycholinguistics (Wallot, 2014; Wallot et al., 2013). These discrepancies motivate an involved study of more realistic reading acts than are traditionally seen in the laboratory. But that necessarily introduces more variability in the linguistic environment and word use. Very large norm sets will be essential for beginning to quantify the semantic properties of larger textual units like sentences, articles, or books.

Large norm sets also allow psychologists to begin grappling with applied problems of language use and comprehension. For instance, there is a growing demand from companies to be able to use user-generated content (e.g., product reviews, social media posts) to infer customer feelings and responses to products. To be able to process customer feelings towards products at volume, automated methods need to be developed that enable researchers to identify the types and strengths of reactions that consumers have towards products.

Such research problems fall within the domain of sentiment analysis (see Liu, 2015, for introduction). Sentiment analysis is concerned with recognizing the presence of personal feelings (typically within text) and accurately labelling those feelings. Sentiment analysis is challenging for numerous reasons. First, it often has to infer affective content from complex, error-prone, open-ended texts. For this reason, large dictionaries that map words to affective properties are often valuable. Second, it needs to identify the presence of sentiment. This is referred to as the subjective/objective problem. Compare "three people killed in fire" to "three people were killed in a fire today". Although both passages contain similar affect, the first is expressing a fact whereas the second is expressing a sentiment.

The problem of distinguishing subjective from objective passages is much harder than the problem of recognizing the type of affect that is present (Pak & Paroubek, 2010). For this reason, the majority of research on sentiment analysis has so far focused on the simpler problem of inferring the affective component of real-world texts that express a sentiment. The ANEW norms are a popular resource for aiding in the categorization of sentiment affect (Staiano & Guerini, 2014). However, because of its small size, the ANEW norm set is of limited use. The ANEW norms become particularly limiting when dealing with short, informal messages (e.g., social media posts; Go, Bhayani, & Huang, 2009), which may not contain any words referenced in the ANEW norms due to the brevity of message, or because of the presence of speech patterns that are idiosyncratic to particular social media platforms.

Large sets of extrapolated human judgments may therefore have an important role to play in sentiment analysis. We compare three norm sets (ANEW; Warriner et al., 2013, henceforth Warriner et al.; and the norms introduced here, henceforth Hollis et al.) on their ability to be used to categorize sentiment affect on two benchmark corpora: one containing 1000 high-valence sentiment and 1000 low-valence sentiment movie reviews and another containing over 1.6 million Twitter posts of both high- and low-valence sentiment.

Method

Sentiment corpora

The movie reviews corpus (introduced in Pang & Lee, 2004) contains 2000 entries harvested from internet movie review sites. Half contain positive reviews (accompanying "thumbs up" on review), and the other half contain negative reviews (accompanying "thumbs down" on review). Each review contains one or more paragraphs of text. The reviews were preprocessed to strip out nonlinguistic characters and to convert all text to lower case.

The Twitter corpus (introduced in Go et al., 2009) contains a 1.6 million tweet training set and a 359 tweet validation set. The training set was constructed and labelled by harvesting twitter posts that contained emoticons. Emoticons are a marker for the presence of subjective sentiment rather than objective fact. Twitter posts were categorized as high or low valence based on the type of emoticon in the post. For example, :), :-), and :-D would all result in a categorization of high-valence sentiment, whereas :(, :-(, and >:(would result in a categorization of low-valence sentiment. The Twitter corpus was preprocessed to strip out all hashtags, emoticons, and URLs. Posts were also converted to lower-case text.

Emoticons are noisy labels for the presence and type of sentiment. To get an accurate measure of the quality of a classifier trained on such data, its performance would need to be assessed on a separate dataset with less noisy labelling of sentiment. Go et al. (2009) provide a validation set of 359 twitter posts (split between low/high-valence sentiment) that were annotated by hand. All reported measures of model performance pertaining to the Twitter sentiment corpus are in reference to this validation set.

Model construction

For each corpus, logistic regression classifiers were constructed to group data into categories of high-valence sentiment and low-valence sentiment. Classifiers were built using one of the three affective norm sets noted above: the ANEW norms, the Warriner et al. norms, and our newly extrapolated Hollis et al. norms. Each classifier used only a single predictor variable. For the movie reviews corpus, the predictor was a measure of average valence (for all words available in each norm set) across adjectives and adverbs within the review. Previous research has demonstrated that sentiment is best recognized for long texts when only adjectives and adverbs are considered (Liu, 2015). Part of speech was marked using Python's natural language toolkit (nltk) package. The length restrictions on tweets prevented the use of such an approach; tweet valence was estimated as an average over all words within the tweet that also had an entry in the norm set being used.

Results

Classification accuracies using the ANEW norms, the Warriner et al. norms, and the Hollis et al. norms, respectively, were 56.17%, 65.05%, and 68.85% for movie reviews, and 52.92%, 72.70%, and 77.43% for tweets. The norms we have introduced here were superior to the ANEW and Warriner et al. norms at providing sentiment classifications for both movie reviews and tweets (all chi-squared $p < .05$).

A likely reason why our norms are superior to the Warriner and ANEW norms at providing sentiment classifications is the fact that our norms have entries for a larger set of words. When dealing with real-world text, this breadth provides a larger coverage of the text to be modelled. Another way to assess the quality of norms is by first reducing each norm set to the set of words that overlap between all norm sets. There are $n = 1026$ words that are present in all three norm sets. There are $n = 13{,}793$ words that are present in both the Warriner et al. norms and the Hollis et al. norms. We developed a new set of models from our three norm sets, restricting the models to only using entries for these two subsets of words. Data are presented in Figures 1 and 2 for the movie reviews and Twitter corpus, respectively.

There are no reliable statistical differences in classification performance between the best and worst performing model using the $n = 1{,}026$ subsets (Hollis et al. 56.62% vs. Warriner et al. 54.85%), $\chi^2 = 1.17$, $p = .28$, to classify movie reviews. There is no reliable difference between the Hollis et al. and Warriner et al. norms when using the $n = 13{,}793$ subsets (Hollis et al. 67.70% vs. Warriner et al. 65.05%), $\chi^2 = 3.03$, $p = .08$, to classify movie reviews. Likewise, there is no difference between the best and worst performing model using the $n = 1{,}026$ subsets (ANEW 52.92% vs. Hollis et al. 49.30%), $\chi^2 = 0.80$, $p = .37$, to classify tweets, nor is there a reliable difference between the Hollis et al. and Warriner et al. norms when using the $n = 13{,}793$ subsets (Warriner et al. 72.70% vs. Hollis et al. 70.19%), $\chi^2 = 0.43$, $p = .51$, to classify tweets. For the purposes of sentiment classification, each of the three norm sets are providing equal-quality affect estimates for words. However, due to their larger coverage of the English language, our norms provide overall superior classification accuracy for text sentiment.

Discussion

We demonstrate that, for the purposes of sentiment classification, our reported norms provide estimates of valence that are comparable in quality to both the ANEW and Warriner et al. norms. However, due to its larger size, our norm set allows for more complete coverage of real-world text and consequently produces higher accuracy classifications of the sentiment component of movie reviews and tweets.

General discussion

Our reported research has two main contributions. First, we demonstrate that the difficulties inherent in

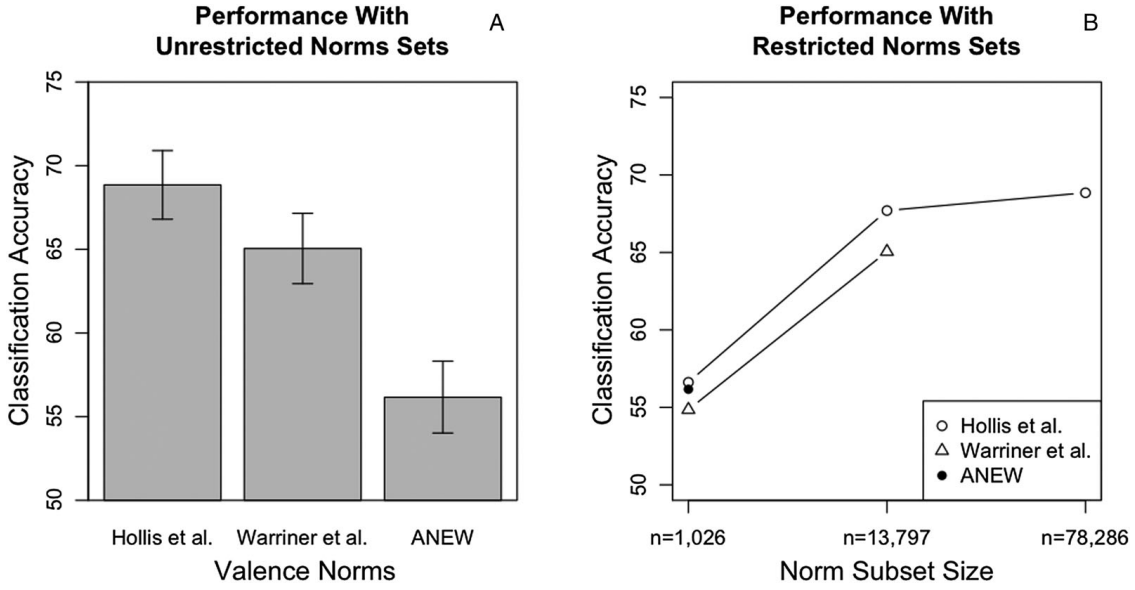

Figure 1. Accuracy for classifying movie reviews as "thumbs up" or "thumbs down" using valence estimates of words derived from the ANEW, Warriner, Kuperman, and Brysbaert (2013), or Hollis, Westbury, and Lefsrud affective norm sets. Classifiers were built using logistic regression with a single input variable of aggregate document valence. (A) Classification accuracies for each norm set, using all entries contained within that norm set. Error bars are 95% confidence intervals. (B) Classification accuracies for each norm set, restricted to only include words that have overlap with the other norm sets.

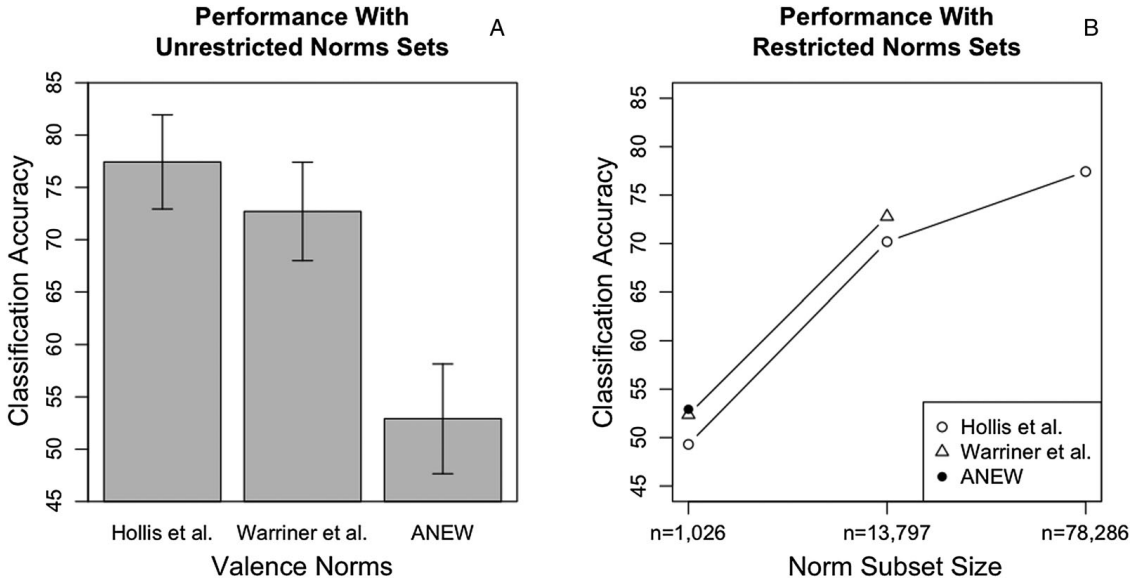

Figure 2. Accuracy for classifying tweets as containing positive or negative sentiment using valence estimates of words derived from the ANEW, Warriner, Kuperman, and Brysbaert (2013), or Hollis, Westbury, and Lefsrud affective norm sets. Classifiers were built using logistic regression with a single input variable of aggregate tweet valence. (A) Classification accuracies for each norm set, using all entries contained within that norm set. Error bars are 95% confidence intervals. (B) Classification accuracies for each norm set, restricted to only include words that have overlap with the other norm sets.

extrapolating human semantic judgments from co-occurrence models can be mitigated with new methodologies. Although the error component of our affect extrapolations do still show evidence of slight contamination from other affect measures, our reported estimates of human judgments are an improvement in terms of both predictive validity and error composition over other attempts at extrapolating human judgments using similar validation procedures. Our extrapolations of concreteness judgments appear to

be free of contamination from both semantic and lexical properties.

Second, we provide a very large norm set based on our extrapolations ($n = 78{,}286$). Not only is this the largest extrapolated norm set to date, it is also the highest quality extrapolated norm sets as assessed by multiple validation criteria. We further provide software for interested parties to generate their own estimates for up to three million unique words and short phrases. In Experiment 2, we demonstrate the value of large extrapolated norm sets for researchers interested in the classification of text sentiment.

In Experiment 1, we observe increased predictive validity of human judgments when our extrapolations are combined with those of Recchia and Louwerse (2015). This motivates a search for improved or combined methodologies for extrapolating human semantic judgments.

Another area where future work can be directed is with the study of human processing of larger textual units. Recent work has demonstrated that the way people process connected texts is different from the way they process isolated words; variables known to exert a large influence in single word recognition may have little or no influence during connected text reading (Wallot, 2014; Wallot et al., 2013). This motivates a comparison of findings observed in paradigms using the presentation of single words (e.g., lexical decision, naming experiments) to findings observed using the presentation of larger bodies of coherent text (e.g., connected text reading). However, in order to conduct such research, methods need to be available for quantifying properties of larger units of text. The results of Experiment 2 support the claim that our present work is a step towards quantifying semantic properties of large units of connected text.

There are probably multiple converging reasons as to why our extrapolated norms are superior to previously reported norm sets. There are also considerations for constructing high-quality extrapolated norms that go beyond our current analysis. Here we discuss some of these considerations.

Previous extrapolated norm sets have used lexical properties like frequency and length for extrapolation of human judgments (e.g., Mandera et al., 2015; Recchia & Louwerse, 2015). Intuitively, extrapolated judgments will become contaminated by lexical properties if lexical properties are used during extrapolation. Since we do not extrapolate based off lexical properties, our norms avoid this source of contamination. Employing validation criteria like those laid out here and in Mandera et al. (2015) will help guard against such contamination.

While we use a different methodology for extrapolating human judgments from that of previous researchers, our materials also differ from those of previous researchers: We rely on word vectors trained by a different learning algorithm as well as a different training corpus for learning word vectors. Either of these factors may have contributed to, or accounted for, the advancements we observe. The corpus our word vectors were built on contained a selection of news articles on world events. It contains content that general persons have some degree of topical knowledge on. We would probably have different results if our word vectors were trained on, for instance, literary texts, textbooks, or texts from internet discussion groups. These contexts have variations in word use and, consequently, pick up on different aspects of semantics. We reran our analysis using word vectors constructed from a 900-million word Wikipedia snapshot downloaded in 2010 (Shaoul & Westbury, 2010b). We observed decrements in predictive validity of approximately 3% variance accounted for, for each of our extrapolated judgments. Corpus quality is a consideration for training co-occurrence models of semantics. Corpus quality is probably influenced by a combination of size and register—that is, the nature of its content and its relevance to human knowledge and experience (see Brysbaert et al., 2011; Brysbaert & New, 2009).

It is possible that learning algorithms used to construct word vectors produce representations that have differential applicability to extrapolating human judgments. In particular, it may be the case that the skip-gram architecture is superior to co-occurrence architectures because the skip-gram model's focus on prediction forces it to make fine-grained discriminations that are ignored by the purely associationist co-occurrence models (see Rescorla, 1988). By their very nature, co-occurrence models record all co-occurrences, whether those co-occurrences are discriminatively relevant to a target word's context (and, by extrapolation, that word's meaning) or not. For example, the word "red" may co-occur with the word "ball" but since the word "red" occurs with many other words, and (more to the point) the word "ball" is modified by many other adjectives (including many that are more discriminative of the word's context/meaning, like "tennis", "bouncy", or "soccer"), the word "red" is not a probable guess for the

context of the word "ball". Since the skip-gram architecture is designed to predict each target word's context, it must by its very nature discriminate between words that are likely to be discriminative of that context ("tennis", "bouncy", or "soccer") and "noise" words that co-occur with that target word without being likely to be predicted by it (in this example, "red"). In sum, in associationist co-occurrence models, some recorded co-occurrences are unavoidable noise that are not useful for discriminating the target word, while the discriminative skip-gram model is explicitly designed to minimize that noise.

We were unable to test for the effect of learning algorithm in this research; the vectors we used are freely available online, but they were trained on a proprietary corpus. Thus, we are unable use the same corpus with a different learning algorithm to construct word vectors. Comparing the relative effectiveness of different learning algorithms for extrapolating human judgments should be addressed in future research. Such comparisons additionally provide a good context for testing the relative psychological merits of the various co-occurrence models of semantics (e.g., Durda & Buchanan, 2008; Hofmann et al., 2011; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Chen, et al., 2013; Rhode et al., 2007; Shaoul & Westbury, 2006, 2010a, 2011).

There are various ways to extrapolate human judgments. One of the contributions of this work is the introduction of one such method: regressing dimensions from word vector representations onto human judgments. Most research up to this point has instead relied on techniques one step removed from actual semantic representations in co-occurrence models—similarities to seed words (e.g., Westbury et al., 2015) or inferring based on properties of neighbours (e.g., Bestgen & Vincze, 2012; Mandera et al., 2015; Recchia & Louwerse, 2015). A more thorough comparison of these methods is warranted. It is possible that there is one best method, or that separate methods are complementary and can be combined.

A final consideration on the quality of extrapolated judgments has to do with the size of the training set for making extrapolations. Mandera et al. (2015) demonstrate that larger training sets improve the quality of extrapolated judgments for most extrapolation methods they test. Although it is unlikely that training set size contributes to the improvements seen in our extrapolated values relative to other recent attempts (we use comparably smaller training sets than Mandera et al., 2015; Recchia & Louwerse, 2015; Westbury et al., 2015), training set size still poses itself as a parameter consideration for future research on extrapolating human judgments. Extrapolated judgments will be most useful when only small datasets of human judgments are available. Developing methodologies that allow for high-quality extrapolation from thousands, rather than tens of thousands, of actual human judgments would be a useful step forward.

We are able to assert that our extrapolated judgments are an improvement over previously reported extrapolations. However, we are not able to say with certainty why. Improvements could be due to our extrapolation methodology, the corpus that word vectors were constructed from, the learning algorithm for constructing vectors, or interactions between these sources of difference from previous research. Further research will need to be conducted to better understand best practices for extrapolating human judgments. Such research is useful, as it can save researchers both time and money from having to collect large norm sets for human judgments of semantic variables.

We believe that extrapolated norms have the potential to be useful proxies for actual human judgments in "megastudies" approaches to language (e.g., Balota et al., 2007) and experimental work more generally. However, we note there are reasons why further validation of our norms may be necessary before such applications occur regularly. First, we point out that our three affective estimates all show slight contamination with other affective measures. Second, all three of our extrapolated affective measures correlate as strongly with measures of lexical access as actual human judgments but have lower inter-norm reliabilities with the ANEW or MRC norms than the training data. We presume this lower inter-norm reliability for extrapolated values is due to the fact that our norms are not fully capturing all of the systematic variation having to do with human judgments of affect (Experiment 1 results suggest that our norms are accounting for 95% of the relevant variation). But if this is the case, we should expect our norms to have slightly weaker correlations with measures of lexical access than actual human judgments. One interpretation is that the mild contamination from other affective measures increases the strength of relationship between our norms and measures of lexical access, cancelling the expected difference in relationship strength with measures of

lexical access. An alternate interpretation is that there are systematicities shared between separate norm sets of human judgments that have no bearing on human semantics or lexical access (e.g., due to biases in judgment introduced by rating scale response formats). The alternative interpretation is supported by results of Experiment 2; extrapolated norms performed equally well as actual human judgments for estimating text sentiment when vocabulary size was held constant.

Ultimately we believe that the first interpretation is the more prudent stance to take when considering these norms for use in experimental research. However, any contamination that does exist is probably mild and may be acceptable for research contexts where actual human judgments do not exist (e.g., proper nouns). Table 2 suggests these concerns apply more to our extrapolated measures of valence, arousal, and dominance than they do to our extrapolations of concreteness.

We note that there are reasons to doubt that human judgments must be considered the ideal gold standard for lexical research. Since human judgments are often unreliable, subject to many moderating variables (such as age, gender, or education), often correlated with many lexical variables, and totally opaque, we do not necessarily believe they serve as a good gold standard. In keeping with classical psychometric theory (as outlined in, e.g., Cronbach & Meehl, 1955), an argument can be made that psycholinguistic research is better off using predictors that are unambiguously defined than using opaque human judgments as predictors. Using human judgments as predictors amounts to correlating one unknown (the thing we are trying to explain) with another unknown (human judgments). The point of scientific explanation is surely not to simply correlate unknowns, but rather to map unknowns on to empirically accessible, unambiguously defined measures.

The accurate extrapolation of human judgments should not be considered an end-goal for research on lexical semantics. The more productive action would be to replace human judgments with unambiguous, empirically accessible measures. However, demonstrating that (a) human judgments can be accurately extrapolated, and (b) extrapolations can effectively stand in for human judgments when modelling other human decisions are both useful steps for learning what types of empirically accessible measures could replace such judgments. Until such time as a more empirically grounded framework of understanding human semantics is posed, extrapolation becomes a useful pragmatic tool for researchers currently relying on human judgments as proxy for semantic measures. The work on extrapolating human judgments broadly construed points to co-occurrence models as being a useful framework for empirically grounding the study of semantics.

Acknowledgements

We thank two anonymous reviewers for helpful advice on an earlier draft of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCiD

Geoff Hollis http://orcid.org/0000-0002-8922-1613

References

Abercrombie, H. C., Kalin, N. H., Thurow, M. E., Rosenkranz, M. A., & Davidson, R. J. (2003). Cortisol variation in humans affects memory for emotionally laden and neutral information. *Behavioral Neuroscience*, *117*(3), 505–516.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1037–1053.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*, 445–459.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1–45). Technical report C-1, the center for research in psychophysiology, University of Florida.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Durda, K., & Buchanan, L. (2008). WINDSORS: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, *40*(3), 705–712.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*, 12.

Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Hamann, S., & Mao, H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *Neuroreport*, *13*(1), 15–19.

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. doi:10.3389/fpsyg.2011.00252

Hollis, G., & Westbury, C. (in press). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1053-2

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 174. doi:10.3389/fpsyg.2010.00174

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.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

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*(2), 211–240.

Larsen, R. J., Mercer, K. A., & Balota, D. (2006). Lexical characteristics of words used in emotion Stroop studies. *Emotion*, *6*, 62–72.

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. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2177–2185). Cambridge, MA: MIT Press.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. New York, NY: Cambridge University Press.

Lodge, M., & Taber, C. S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, *26*(3), 455–482.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

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

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).

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

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, pp. 1320–1326).

Pang, B., & Lee, L. (2004, July). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Rhode, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2007). *An improved method for deriving word meaning from lexical co-occurrence*. Unpublished manuscript. Cambridge, MA: Massachusetts Institute of Technology. Retrieved April 20, 2007, from http://tedlab.mit.edu/~dr/

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Shaoul, C., & Westbury, C. (2010a). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413.

Shaoul, C., & Westbury, C. (2010b). *The Westbury lab wikipedia corpus*. Edmonton, AB: University of Alberta.

Shaoul, C., & Westbury, C. (2011). HiDEx: The high dimensional explorer. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 230–246). Hershey, PA: IGI Global.

Staiano, J., & Guerini, M. (2014). DepecheMood: A Lexicon for emotion analysis from crowd-annotated news. arXiv preprint arXiv:1405.1605.

Võ, M. L., 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*(2), 534–538.

Wallot, S. (2014). From "cracking the orthographic code" to "playing with language": Toward a usage-based foundation of the reading process. *Frontiers in Psychology*, *5*, 891. doi:10.3389/fpsyg.2014.00891

Wallot, S., Hollis, G., & van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. *PloS one*, *8*(8), e71914. Advance online publication. doi:10.1371/journal.pone.0071914

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Westbury, C. (2013). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, *43*(5), 631–649.

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. *The Quarterly Journal of Experimental Psychology*, *68*(8), 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, & the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*, 991. Advance online publication. doi:10.3389/fpsyg.2013.00991

Word2vec: Tool for computing continuous distributed representations of words. (2013). Retrieved October 18, 2015, from https://code.google.com/p/word2vec/