Three Studies on Predicting Word Concreteness with Embedding Vectors

Michael Flor
Educational Testing Service
Princeton, NJ, USA
mflor@ets.org

Abstract

Human-assigned concreteness ratings for words are commonly used in psycholinguistic and computational linguistic studies. Previous research has shown that such ratings can be modeled and extrapolated by using dense word-embedding representations. However, due to rater disagreement, considerable amounts of human ratings in published datasets are not reliable. We investigate how such unreliable data influences modeling of concreteness with word embeddings. Study 1 compares fourteen embedding models over three datasets of concreteness ratings, showing that most models achieve high correlations with human ratings, and exhibit low error rates on predictions. Study 2 investigates how exclusion of the less reliable ratings influences the modeling results. It indicates that improved results can be achieved when data is cleaned. Study 3 adds additional conditions over those of study 2 and indicates that the improved results hold only for the cleaned data, and that in the general case removing the less reliable data points is not useful.

Keywords: word concreteness, word embeddings, data reduction

1. Introduction

The importance of distinction between concrete and abstract concepts has been long noted in psycholinguistics (Paivio, Yuille, & Madigan, 1968). The so called 'concreteness effect' often finds that human participants process concrete words faster and more accurately than abstract words, in a variety of tasks, such as word naming, recognition, and recall, as well as sentence comprehension (Paivio 1991). Jessen et al. (2000) conducted fMRI studies indicating that concrete nouns are processed differently in the brain than abstract nouns.

Notions of concreteness and abstractness have also been used in computational approaches, both to investigate lexical relations, and for analysis of text. Concreteness of words has been widely used for metaphor detection (Maudslay et al., 2020; Köper and Schulte im Walde, 2017; Beigman Klebanov et al., 2015; Tsvetkov et al., 2014; Turney et al., 2011). For example, when a sentence describes an abstract agent performing a concrete action, it can be a strong indication of metaphorical usage. Choi and Downie (2019) used word concreteness scores to analyze trends in popular song lyrics across several decades, finding that concreteness in songs has been decreasing before the year 1991 and began increasing since then. Hills and Adelman (2015) analyzed distributions of word concreteness in published books; they noted "*a systematic rise in concrete language in American English over the last 200 years.*" Flor and Somasundaran (2019) investigated word concreteness in narrative writing of students, finding that concreteness positively correlates with rater scores of narrative quality.

Hill et al. (2014) analyzed the associations that concrete and abstract words have in a large corpus. They found that the more concrete words have smaller sets of context words, while abstract words have larger sets of context words. Naumann et al. (2018) investigated the concreteness of the contexts of concrete and abstract English words. They found that abstract words mainly co-occur with abstract words, but for concrete words cooccurrence patterns differ by part-of-speech. Tater et al. (2022) investigated selectional preferences of English nouns and verbs, and found that strong preferences exist with respect to concreteness and abstractness of subject and direct object slot fillers for verbs.

Early work in psycholinguistics has shown that concreteness/abstractness is not dichotomous but a matter of degree, and researchers began collecting human-assigned ratings for various words and producing lexical norms (Paivio et al., 1968). Presently three large human-rated datasets of concreteness are available for English (Coltheart, 1981; Brysbaert et al., 2014; Scott et al., 2019).

In parallel with utilizing the experimental ratings, researchers have also been interested in extrapolation of concreteness ratings to other words, for which ratings are yet unavailable. Notably, the interest in using computational linguistic approaches to extrapolate human semantic judgments is not limited to concreteness ratings, Methods to extrapolate ratings for a variety of variables, such as sentiment, arousal, and dominance, have been studied (Bestgen & Vincze, 2012; Turney & Littman, 2003); for a synthesis of some approaches see Mandera et al (2015).

Many researchers have reported that utilizing dense word representations (word embeddings) from distributional semantic language models can be useful for predicting and extrapolating concreteness values. Mandera et al. (2015) used several approaches to learn to predict psycholinguistic values from corpus data. For prediction of concreteness ratings, they used the data from Brysbaert et al. (2014). Using Random Forest learning over word vectors, they achieved a correlation of .781 with original scores, and even a higher correlation of .796 when using a KNN approach. Hollis et al. used regression over word2vec vectors and achieved a correlation r=.833. Paetzold and Specia (2016) used bootstrapped regression over word2vec embedding vectors from a corpus of 7 billion words. For predicting concreteness, their best result had Pearson

The Workshop on Cognitive Aspects of the Lexicon, pages 140–150

correlation of r=.862, with human ratings. Thompson and Lupyan (2018) used multiple linear regression over word vectors and obtained correlation of r=.86 with human ratings. Ljubešić et al. (2018) utilized word embedding vectors trained on Wikipedia to predict concreteness scores with SVM regression; they reported Spearman correlation of ρ=.872 between estimated and original values.

While human ratings provide the core data for extrapolation studies, such ratings are not without problems themselves. The published ratings for each word are usually average values across several human participants, and humans often disagree in their judgments; the standard deviations of human ratings per word vary considerably. Pollock (2018) provided an in-depth critique of crowd-sourced ratings of semantic psycholinguistic variables, such as concreteness, imageability, and emotional valence. Munoz-Rubke et al. (2018) have argued against using Likert-type rating scales for rating studies such as concreteness. Computational linguists have also noted problems with words for which human ratings show high disagreement (Tater et al., 2022; Beigman Klebanov et al., 2015).

In this paper we set to investigate to what extent words that have considerable rating disagreements influence word-embedding-based modeling of concreteness ratings. The paper is structured as follows. First, we describe the three large, published datasets of word-concreteness ratings for English. In study 1 we compare twelve word-embedding models as to their ability to model the concreteness ratings in those datasets. To the best of our knowledge this is the largest such comparison to date. In study 2, we pick two models and investigate how their predictions are influenced by exclusion of words with high standard deviations of concreteness ratings. In study 3 we introduce additional conditions on exclusion of such words, which shed light on their influence in the modeling process.

2. Datasets

The MRC Psycholinguistic Database (Coltheart, 1981; Wilson 1988) is one of the earliest large compilations of linguistic and psycholinguistic values for English words. It has concreteness ratings for 4295 English words, which were derived from experimentally established sets where participants rated words for perceptual concreteness on a 1-7 rating scale. In the MRC database they are expressed on a 100-700 scale (and rescaled back to 1-7 for the current study). Notably, the MRC database does not list the per-word standard deviations of the ratings.

Brysbaert et al. (2014) published a collection of 37,057 English words (mostly lemmas) with human-provided concreteness ratings (the BWK dataset). It is the largest such collection of ratings for English. The authors noted that previous collections of human concreteness ratings tended to focus too much on visual perception, and so for their rating study they emphasized all types of experiences (not only sensory, but also actions/activities). In that study, participants (native English speakers) received word lists and had to rate each word for concreteness, on a 5-point Likert scale, where only integer values could be chosen. After careful validation and filtering, the authors retained only those words that were known by at least 85% of the raters, and each word was rated by about 25 participants. The resulting concreteness score for each word is an average of the scores it received from its raters. The authors also released the standard deviation values of the ratings for each word. The BWK and MRC sets have an overlap of 3,935 words, and Pearson correlation of concreteness ratings is r=.919, a very high level of agreement.

Scott et al. (2019) published normative ratings for 5,553 English words on nine psycholinguistic dimensions, including concreteness. The authors called this data the Glasgow Norms (hereafter the GN dataset). In that study, for any given subset of words, the same participants provided ratings across all nine dimensions, and on average each word was rated by 33 participants. For concreteness, integer ratings were assigned on a 7-point Likert scale. Average concreteness values and standard deviations of ratings for each word were released by the authors. Some of the words in that study were polysemous and were presented with a disambiguator, e.g., *blubber (cry)* and *blubber (fat)*. By excluding the 871 such entries in that data, we utilize the 4682 single words (lemmas) that have concreteness ratings. The GN dataset has an overlap of 4,445 words with the BWK dataset, and Pearson correlation of average concreteness ratings between the sets is r=.93, indicating very high agreement of ratings.

For all three datasets, the published concreteness scores are real numbers, in the respective scale ranges. It is interesting to note the distribution of concreteness scores in the three datasets. Although the scales are of different magnitudes, it can be seen in the binned distributions (Figure 1) that the BWK data is skewed towards the more abstract side, while the GN data is skewed to the concrete side of the scale. The MRC has more words on the concrete side, but the extreme bins are 'underpopulated', especially the bin for very abstract words with scores in the range of 1-2.

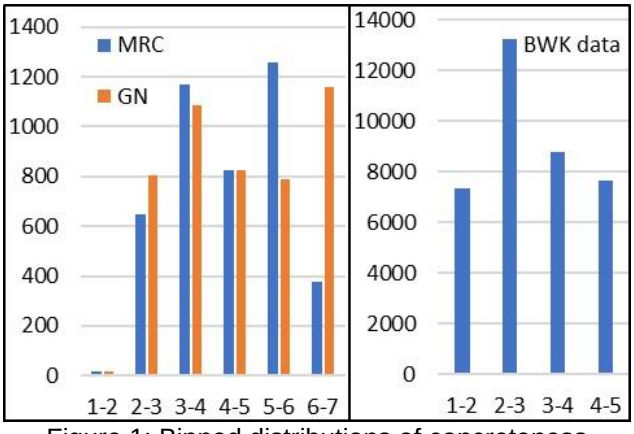

Figure 1: Binned distributions of concreteness scores in three datasets: MRC, GN, and BWK. Score bins on X-axes, word counts on Y-axes.

3. Experiments

3.1 Study 1

In Study 1 we investigate to what extent vector representations of words can be utilized for predicting word concreteness scores. Following previous studies (Thompson and Lupyan, 2018), we employ multiple linear regression as the learning method for the experiments. In such setting, the embedding vector dimensions serve as predictor variables.

We experimented with fourteen different embeddings models, as listed in Table 1. We included the widely used word2vec cbow model trained on Google News (Mikolov et al., 2013) and refer to it as *mikolov.w2v*. From Baroni et al. (2014) we adopted the word2vec cbow model trained on a window of 2 words (*baroni.w2v*), and also a vector model based on SVD of PMI word co-occurrence values (*baroni.pmi*). Two GloVe embeddings models (Pennington et al., 2014) were used – one trained on a corpus of 6 billion words (*glove.6b*) and a larger model, trained on 42 billion words (*glove.42b*). From the work of Levi and Goldberg (2014) we used two word2vec models trained on English Wikipedia data: one used a window of 5 words around a target word (*l&g.w5*), the other model used dependency parse relations (*l&g.deprel*). The *eigenwords* embeddings come from the work of Dhillon et al. (2015). *Lexsub* embeddings are a model introduced by Melamud (2015). The *ftWiki* model is a model trained on English Wikipedia, part of the Fast Text family of models (Bojanowski et al., 2017). The *paragram* model (Wietig et al., 2015) used a large database of English paraphrases to tune the word embeddings. Numberbatch embeddings (*nb17*) is a model based on the ConceptNet project data (Speer and Lowry-Duda, 2017). Two additional models use embeddings derived from Transformer architectures. We used the popular SentenceBERT library (Reimers and Gurevych, 2019) as an embedder, to produce static embeddings for the words in our experiments. The *MiniLM-L6-v2* model produces vectors of dimension 368, based on the BERT transformer model. The *distilroberta-v1* model produces vectors of dimension 768, derived from the DistilRoBERTa transformer model. In all experiments in this study, all vectors were normalized with L2 normalization.

It is worth noting that different vector models have different coverage for the words in the datasets (see Table 1). For the BWK data, among the classic models, the lowest coverage is by the *l&g.deprel* model, only 26,605 words (72% of the dataset), and the highest is by *glove.42b*, 35,491 words (96% coverage). Embeddings derived from SentenceBERT achieve full coverage, as such modern models can provide embeddings for any string. For the smaller MRC and GN datasets, the coverage was much better. Lowest coverage for MRC data was 4,140 words (96%), and for GN data: 4629 words (99%).

Experiments were performed for the MRC, BWK and GN datasets separately. All experiments involved 10-fold cross validation, with a 9:1 training:testing ratio. We used value clamping to prevent regression-based predicted values from falling outside of the original scales. Predicted values below 1 were reset to 1, and those above maxima (5 or 7) were reset to the max value.

Two evaluation measures were used to estimate the success of various models. One measure was Pearson correlation between the original published concreteness values and the predicted values. The higher the correlation, the better is the prediction. The other measure is Root Mean Square Error (RMSE), which measures the average squared difference between original and predicted scores. Lower values of RMSE indicate better prediction performance. Results (micro-averages) for all the experiments are presented in Table 1. A single model, *nb17*, achieved the best results in all datasets, on both the correlation and the RMSE measures.

Results for the BWK dataset indicate that all models show rather impressive prediction power – correlations ranging above 0.8 (except *glove.6b*), but none reaches 0.9. Across different language models, the RMSE values for BWK data range between 0.472 and 0.633. Divided by the scale range, 4, those RSMEs are at a magnitude of 12-16% of the score range.

Results for the GN dataset indicate that all models show very strong results, all correlations range above 0.8, and two models – *nb17* and *lexsub* achieve correlations above 0.9. The RMSE values for the GN data range from 0.572 to 0.872. Those values are larger than values obtained for the BWK data. However, GN data was rated on a 1-7 scale, and so higher error values should be expected. If we divide RMSE values by the scale range, we can see that the error results in the two experiments are comparable. Lowest RMSE values: for BWK data 0.472/4=0.118; for GN data 0.572/6=0.095. The highest RMSE: for BWK: 0.633/4=0.158; for GN: 0.872/6=0.145.

The results for MRC data resemble those of GN data, although each language model achieves slightly worse (lower) Pearson correlation values for MRC than for GN, but slightly better (lower) RMSE values for MRC than for GN data.

3.2 Study 2

The background for Study 2 stems from the criticism that some researchers have pointed toward the reliability of psycholinguistic ratings with Likert-type scales. Munoz-Rubke et al. (2018) have noted that when participant ratings are averaged and assigned as final word scores, for categories such as concreteness, the approach may have important limitations, as the results can be highly distorted by outliers. Specifically for concreteness values norms from the Brysbaert et al. (2014) study, Pollock (2018) has argued that the mean concreteness values for words do not reflect the judgments that actual participants made: "*this problem applies to nearly every word in the middle of the concreteness scale.*"

		BWK data			GN data			MRC data		
Model name	**dims**	**coverage**	**Pearson**	**RMSE**	**Coverage**	**Pearson**	**RMSE**	**coverage**	**Pearson**	**RMSE**
sbert Mini-LM6-v2	368	37057	0.825	*0.573*	4681	0.858	*0.736*	4295	0.812	*0.708*
sbert distilroberta-v1	768	37057	0.815	*0.588*	4681	0.797	*0.872*	4295	0.752	*0.807*
mikolov.w2v	300	33975	0.848	*0.539*	4629	0.887	*0.661*	4220	0.854	*0.627*
nb17	300	35488	**0.885**	***0.472***	4679	**0.917**	***0.572***	4292	**0.883**	***0.569***
glove.42b	300	35491	0.821	*0.579*	4682	0.855	*0.745*	4290	0.814	*0.704*
glove.6b	300	31619	0.783	*0.633*	4680	0.825	*0.811*	4261	0.784	*0.752*
l&g.deprel	300	26605	0.868	*0.507*	4651	0.891	*0.649*	4195	0.874	*0.588*
ftWiki	300	35319	0.850	*0.535*	4680	0.878	*0.686*	4294	0.849	*0.640*
Lexsub	600	28274	0.874	*0.496*	4677	0.905	*0.612*	4232	0.879	*0.577*
Eigenwords	200	28276	0.865	*0.511*	4651	0.883	*0.673*	4140	0.867	*0.601*
l&g.w5	300	27212	0.834	*0.563*	4657	0.870	*0.706*	4214	0.850	*0.636*
Paragram	300	35308	0.805	*0.601*	4682	0.811	*0.840*	4286	0.773	*0.768*
baroni.ppmi	500	30260	0.839	*0.555*	4681	0.877	*0.690*	4261	0.851	*0.638*
baroni.w2	400	30260	0.811	*0.596*	4681	0.880	*0.688*	4261	0.841	*0.656*

Table 2: Results of word-concreteness score prediction for three datasets, with 14 different vector-space models. *Dims* is the number of dimensions per vector. The columns labeled *Coverage* are counts of words that had vectors in the respective language model. For Pearson correlations, higher value means better prediction; for RMSE, lower value means better prediction.

He recommended that researchers who use such ratings pay attention to the standard deviations of ratings and use only the stimuli for which standard deviations are as low as possible. The relevance of such critique to our work is quite direct. What would happen if we excluded from our data all items (words) that are 'less reliable'? Would it improve the concreteness prediction models? On the other hand, excluding some data would make the datasets smaller; and having less data may lead to inferior learning.

Beigman Klebanov and Beigman (2014) and Jamison and Gurevych (2015) have suggested that, in supervised machine learning, the presence of difficult items in the training sets is detrimental to learning performance and that performance can be improved if systems are trained on only easy data. They define 'easy' as less controversial in human annotations. This seems exactly analogous to our current case. Words that have high standard deviations (SD) of human-rated concreteness are 'less reliable' as to their real concreteness value, they are more 'difficult' cases. Excluding them from the training data may leave just the more reliable, 'easier' data for learning and thus might lead to improved model performance.

Standard deviations of rating values for each word are available for the BWK and the Glasgow Norms datasets. To understand the potential scope of data reduction, we plot the number of words in each dataset as a function of different SD value thresholds, and also by score-bins of the ratings. Figure 2 (left panel) presents the plot for the BWK dataset. The black bars represent the data when nothing is excluded, corresponding to Figure 1. The red bars indicate the counts of remaining words when all words with SD>1.5 are excluded. Such exclusion affects mostly words in the score bins 2-3 and 3-4. The green bars indicate the counts when all words with SD>1.2 are excluded. Again, we can see that the largest data reduction occurs for words in the score bins 2-3 and 3-4. With exclusion threshold of SD>1.0 (maroon-color bars), almost all words in bins 2-3 and 3-4 get excluded. The exclusion rates are much more gradual for score bins 1-2 and 4-5, which are closer to the extremes of the concreteness rating scale.

Figure 2 (right) presents the distributions of words for the GN dataset. The black bars represent the data when nothing is excluded, corresponding to Figure 1. Data reduction thresholds for this set are somewhat different. At the exclusion threshold of SD>1.0, only the bin of scores 6-7 retains some considerable number of words, while all other bins are almost emptied. Data reduction is especially dramatic for bins of scores 3-4 and 4-5 (the middle of the rating scale).

The design for study 2 is as follows. We investigate how gradual elimination of some data from the datasets influences the quality of the learned models. For each dataset, we exclude all words that exceed a given SD threshold and train a multiple regression model with 10-fold cross-validation. This mode of exclusion is systematic. For the sake of comparison, we also check what happens if the same number of words are excluded, but chosen randomly, rather than by an SD threshold. For example, for BWK data (37,057 words), for a threshold of SD≤1.4 we exclude 7,053 words, and experiment (full 10-fold cross-

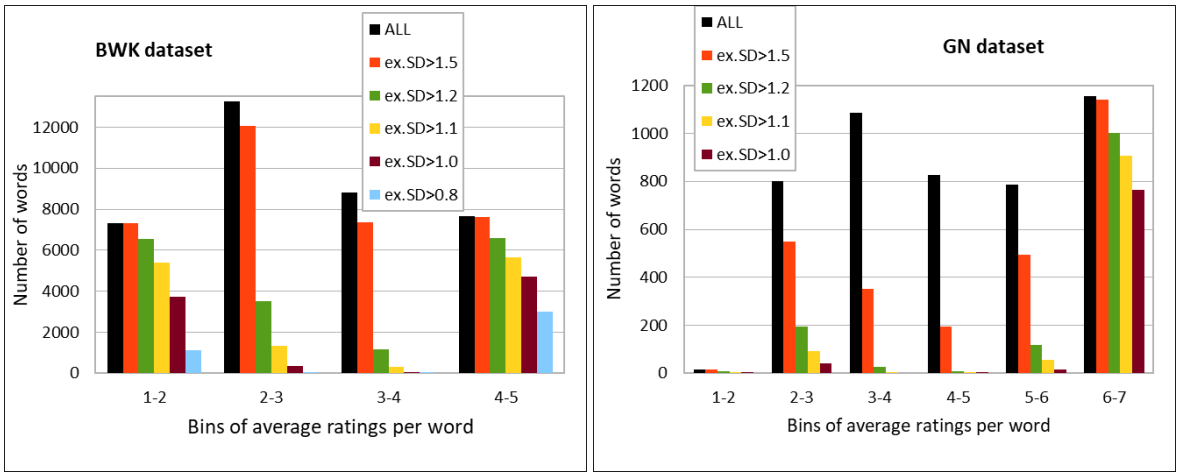

Figure 2: Binned distributions of concreteness scores, data exclusion by SD thresholds.

validation) with the remaining 30,004 words. In a matching control condition, we exclude 7,053 words randomly chosen, and run the experiment with the remaining 30K words.

For the BWK dataset, the systematic exclusion steps are from SD value of 1.0 to 0.6 with a step of 0.1 (more data is excluded on each step). For the GN data, the SD thresholds are from 2.0 to 1.0 with a step of 0.1. For each dataset we also use a condition where no words are excluded, as in study 1. Table 2 presents the counts of remaining words for each condition.

Inclusion	BWK data	GN data
All data	37057 (100%)	4682 (100%)
SD≤2.0		4599 (98%)
SD≤1.9		4485 (96%)
SD≤1.8		4260 (91%)
SD≤1.7	36942 (99%)	3901 (83%)
SD≤1.6	36345 (98%)	3383 (72%)
SD≤1.5	34375 (93%)	2748 (59%)
SD≤1.4	30004 (81%)	2198 (47%)
SD≤1.3	23916 (64%)	1738 (37%)
SD≤1.2	17814 (48%)	1357 (29%)
SD≤1.1	12681 (34%)	1069 (23%)
SD≤1.0	8847 (24%)	825 (18%)
SD≤0.9	6118 (17%)	
SD≤0.8	4147 (11%)	
SD≤0.7	2860 (7%)	
SD≤0.6	1998 (5%)	

Table 2: Number of remaining words in two datasets, by inclusion thresholds on SD values.

For Study 2 we use two embedding models from Study 1 that have good performance and also have good lexical coverage over the BWK and GN datasets – *nb17* and *sbert MiniLM-L6-v2*. Note that the number of words used in each experimental condition, as presented in Table 2, applies only to the *sbert* model, as it has full coverage of the datasets; *nb17* has lower coverage and thus the number of words used is slightly lower in each respective condition. Just as in study 1, RMSE and Pearson correlation are used as evaluation measures in study 2.

Results for the BWK dataset are presented in Figure 3. The correlation results with *sbert* and *nb17* are quite similar (Figure 3, left panels). When very little data is excluded (thresholds 1.7 and 1.6), the results of systematic or random exclusion are quite the same, and very close to those of no exclusion. However, the results begin to separate from threshold 1.5. The results from systematic exclusion become higher and higher with each successive exclusion threshold, they reach beyond correlation of .9, and for *nb17* – even beyond .95. The peak results are achieved at SD≤0.8. After that threshold, the correlation values begin decreasing, though they are still higher than for the full dataset. For the control conditions with random exclusion, the correlation values do not improve with successive exclusions, they even have a slight tendency of decreasing, and never get higher than values for the full-data condition.

RMSE results for the BWK dataset are presented in Figure 3, right-side panels. Note that for RMSE, lower error values indicate better performance. The results with *nb17* and *sbert* are quite similar. For inclusion thresholds 1.7 to 1.3, the RMSE results for systematic or random exclusion are very close to each other, and approximately the same as under the no-exclusion condition. However, as more and more data gets excluded, RMSE values for systematic exclusion begin decreasing; the decrease even accelerates (the black-color lines curve down), whereas the error levels for random exclusion (orange-colored lines) remain the same, or even increase slightly. Notably the separation of results between systematic and random conditions begins at SD≤1.2 for the *nb17* model and at SD≤1.0 for the *sbert* model.

Results for the GN dataset are presented in Figure 3. The correlation results with *nb17* and *sbert* are quite similar (Figure 3, left panels). The trends are also similar to those of the BWK dataset results. When very little data is excluded (thresholds 2.0 to 1.8), the results of systematic or random exclusion are quite similar, and very close to those of the no-exclusion

condition. For further thresholds, systematic exclusion leads to higher correlation results, until threshold levels of 1.3 or 1.2, and then the correlation results start decreasing quite sharply. The sharp decreases might be due to the sharp reduction in the size of the dataset, or due to the dramatic change in the distribution of values in the reduced corpus (see Figure 2). The best correlation results are, for *sbert*: r=.887; for *nb17*: r=.944; all when SD≤1.4. Under the random exclusion, the correlations tend to decrease.

The RMSE results for the GN data are presented in Figure 3, right-side panels. The results for systematic exclusion are similar to the RMSE results in the BWK dataset – at first the error levels are quite similar to those under the no-exclusion condition, but then the RMSE values get increasingly lower and lower (black-color lines tend to curve down). The results for random exclusion are markedly different from systematic exclusion. At first the error levels are close to those under the no-exclusion condition, but then the RMSE values begin rapidly increasing (orange lines curve up), indicating worsening performance.

The results of study 2 indicate that when the 'less reliable' data is excluded from the datasets, regression models based on word-embeddings can achieve much better results that with the full data.

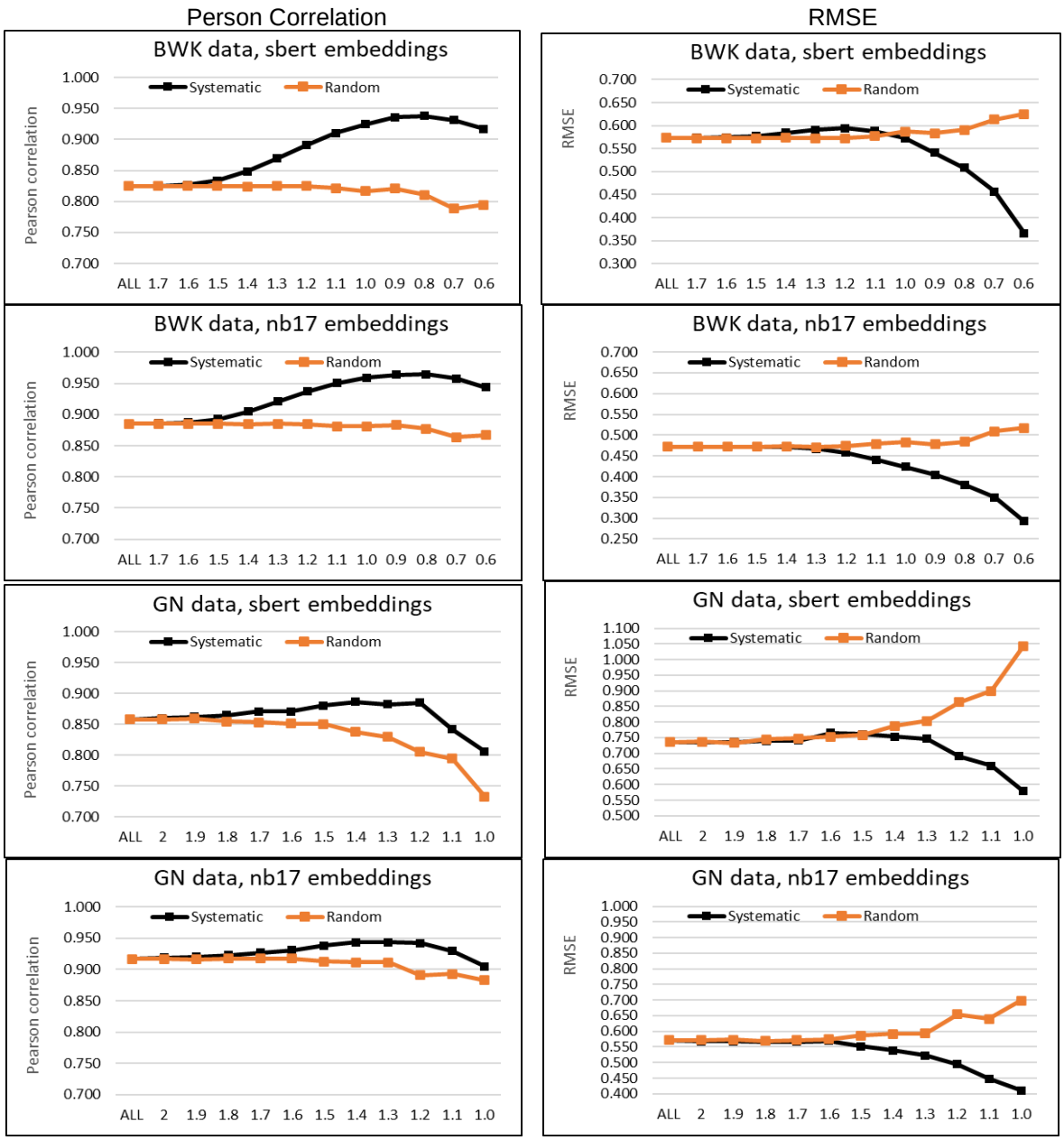

Figure 3: Pearson correlations (left) and RMSE (right) for predicting concreteness scores on two datasets, as a function of data reduction (systematic by SD thresholds, or random), using two different language models. Data points marked 'ALL' represent a condition where all available words were included.

3.3 Study 3

An important aspect in study 2 was the change in distribution of concreteness values, under the systematic exclusion condition. Do models achieve better results because they learn on increasingly 'cleaner' data, or simply because of the different distribution of values? And what about the 'less reliable' values? Should we exclude them from modeling at all? How would models trained on cleaner, reduced data perform on unfiltered data? Study 3 addresses those questions.

In study2, we tested what happens when the dataset was successively reduced. Under systematic reduction, both the training folds and the testing folds were reduced as per the SD thresholds. In the other condition, random exclusion was used for all folds. In study 3 we add two new conditions where we mix the data exclusion methods. In a condition called SR, data reduction in training folds uses systematic reduction (by SD thresholds), but in the testing folds a comparable amount of data is excluded randomly. This condition evaluates what happens when training data is systematically cleaned (and the distribution of concreteness scores changes), but the testing data is just randomly reduced, and so it keeps the same distribution as the whole dataset. Under another condition, called RS, we reverse the reduction methods. Data for the testing folds is reduced systematically (by SD thresholds), but the training folds get a proportional random reduction. Thus, the models are trained on approximately the same distribution as the whole dataset, but are tested on just the 'cleaner' data. The overall amounts of included data decrease in the same way under the new conditions, just as in study 2. For study 3 we used the same datasets and same vectors as in study 2. All experiments were run with 10-fold cross-validation. Results are presented in Figure 4. For ease of comparison, the results from study 2 are shown again, with the results of the new conditions added (yellow lines for SR and green lines for RS).

When models are trained on increasingly 'cleaner' data, their ability to predict values for 'non-cleaned' data (yellow lines) keeps up with models that do not 'clean' the data (red lines), both for correlation and RMSE. However, after certain levels of data reduction the 'clean'-trained models begin losing it – they achieve slightly lower correlations and make dramatically larger errors, as compared to models that train and test on randomly-reduced data (red lines). Comparing yellow lines to black lines (in both cases models train on cleaned data) shows that the composition of the test data makes a huge difference – when test data is also clean, the best overall results are achieved, but when the test data is unfiltered, the worst results are achieved (lowest correlations and largest errors).

Next, we consider models that test on just the clean data, but train on cleaned (black) or unfiltered data (green lines). Looking left to right on each panel (left side) in Figure 4, the green line keeps up with the black line until SD 1.2 (BWK) or SD 1.4 (GN), reduction to about 47% of the full data. After that the green lines show worse results than the black lines, but still better than the other lines. It seems that the models trained on unfiltered data retain most of the information needed to predict clean data; that is until the distributions become so different that prediction deteriorates (the respective RMSE values start rising while correlations get lower).

4. Discussion

Many previous studies used the large BWK dataset. Thompson and Lupyan (2018) reported a correlation of r=.86; Hollis et al. (2017) reported a correlation of r=.833; Mandera et al. (2015) obtained a correlation of r=.781. Ljubešić et al. (2018) reported Spearman correlation ρ=.887 on BWK data and ρ=.872 on MRC data. Paetzold and Specia (2016) reported a correlation of r=.862 on MRC data. Our results in study 1 indicate that comparable or better prediction levels can be obtained with several different language models, using ordinary multiple regression. While previous studies have used BWK and MRC datasets, the current study is first to also use the Glasgow Norms data for concreteness prediction. The results resemble those of BWK and MRC data. None of the previous studies used RMSE as an evaluation measure for concreteness ratings prediction. In study 1, RMSE results for BWK data are typically lower than for MRC and GN data, probably due to differences in scales. Beyond that, RMSE results for different embedding models are quite similar to correlation results – embeddings that get better correlations also show lower error results.

Study 2 was motivated by the notion of unreliable word concreteness ratings, which reflect considerable disagreements among human raters. In the BWK dataset, less than 50% of the words have standard-deviation values below 1.2, and only 24% have SD values below 1.0. In the GN dataset less than 29% of the words have SD values below 1.2 and just 18% have values below SD 1.0. Study 2 investigated how exclusion of unreliable data points influences regression modeling. It was found that when human raters agree more on concreteness of words, such ratings can be modeled/predicted very well with vector space models. Higher correlations and lower errors are obtained as compared to learning on the full data.

However, the distributions of concreteness scores in the BWK and the GN datasets change drastically when less reliable words are excluded – most unreliable words are in the middle of the distributions and are excluded with successive data cleaning. Study 3 investigated whether results in study 2 were due to changes in concreteness score distributions. The results showed that training on clean data does not generalize well to unfiltered data, especially with regard to magnitude of errors (RMSE). On the other hand, training on unfiltered data and testing on just the clean data reveals that the models have enough information to predict scores for clean data, especially

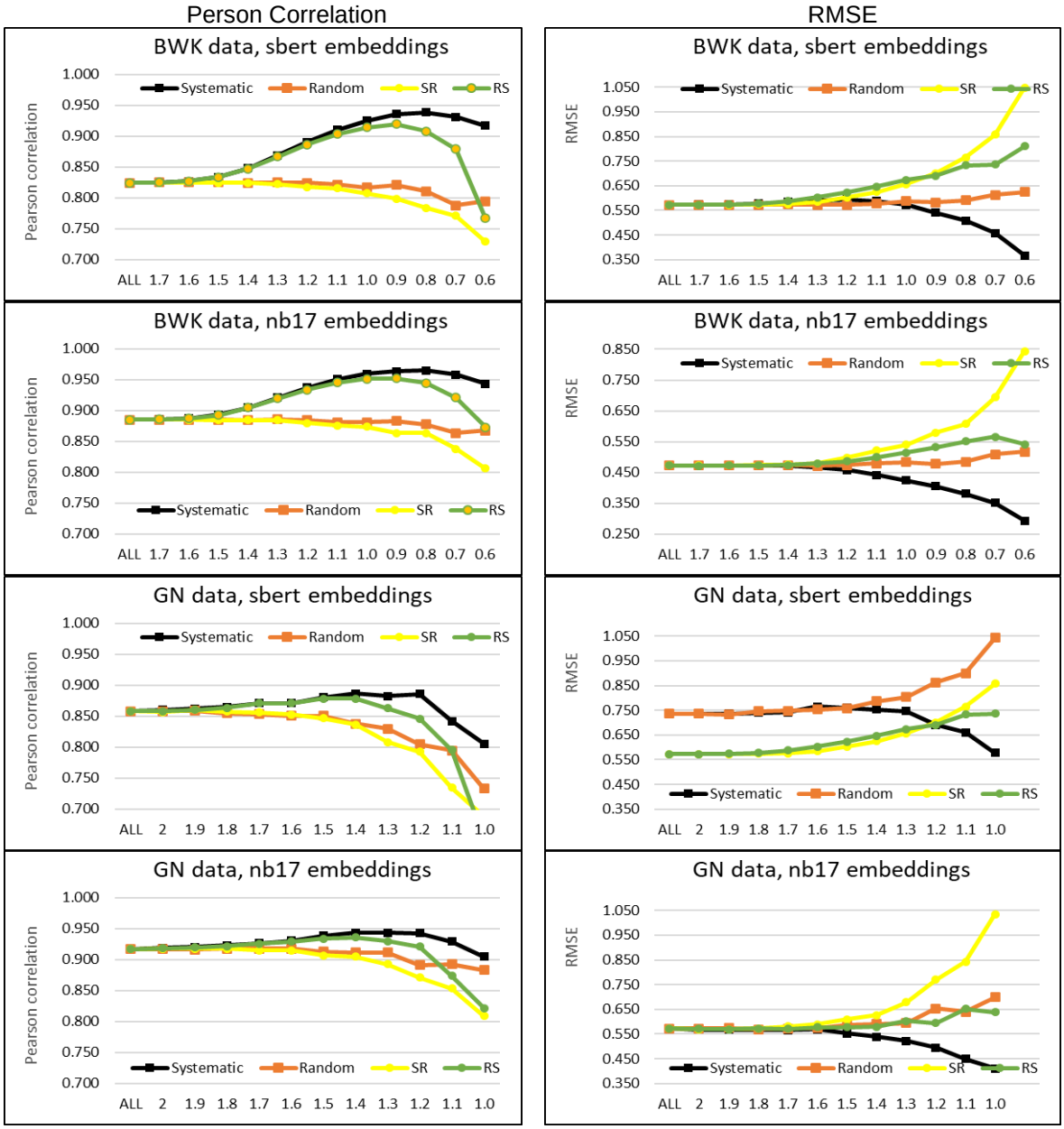

Figure 4: Pearson correlations (left) and RMSE (right) for predicting concreteness scores on two datasets, with four methods of data reduction, using two different language models.

on the correlation measure, but less so on RMSE. Thus, if we are only interested in predicting concreteness for 'reliable' words, cleaning the training data can be useful. If the potential 'reliability' of new words is unknown (as would be the case for most new words), filtering the training data is not recommended.

We continue the discussion in relation to the distinction between easy and difficult cases. Uma et al (2021) provide an extensive review on the influence of hard cases for machine learning, where difficulty arises from disagreements in human annotation. They provide a taxonomy of potential reasons for disagreement. Among the sources of disagreement, Uma et al. mention a) annotator/rater errors b) problems with the task interface, c) problems with task definition, d) situational item difficulty, e) genuine ambiguity of the data, and f) rater subjectivity.

Annotator/rater errors can be mistakes or slips made due to inattention, or other random factors. Interface issues can arise when task interface may have technical complications (e.g., selecting text spans). Problems with task definition may lead to disagreements when the task is not well defined, includes vague statements, or, in case of classification, classes that are not mutually exclusive. Item difficulty (for rating/annotation) relates to cases when the interpretation of the data is unclear. For example, for image labelling, if the image is too blurred, annotators may disagree as to what they actually see there, and thus disagree on a label. In the task of textual entailment, an item may be difficult

because the text is convoluted and a core assertion is not easy to discern. As noted by Uma et al., the problem is not that an item lacks a 'true' label, only that the 'true' label is difficult to distinguish. As they note, the conclusion either follows from the premise, or it doesn't, but not both. This contrasts with the ambiguity category, where the data items can be truly ambiguous, i.e., have different valid interpretations. Uma et al. mention that ambiguity cases have been shown to arise in annotation of anaphora and of POS tags. The final category, subjectivity of judgement, relates to cases where annotators/raters hold different opinions. The prototypical example is annotation of offensive language, where annotators may disagree on whether a given expression is offensive, and different opinions can be simultaneously valid.

Notably, the above taxonomy was developed in relation to disagreements on tasks that involve data classification, and the labels are on nominal scales (but the amount of disagreement can be expressed on continuous scales). Uma et al. (2021) presented several studies around the question on how to integrate disagreements into machine learning processes. There were few studies with data on other scales. The study by Jamison and Gurevych (2015) included a dataset on biased language, where the labels were on an ordinal scale (*no bias, some bias, very biased*), and a dataset on affect recognition for text snippets, with a scoring scale of 0-100. Loukina et al. (2018) investigated automated speech scoring (for language proficiency assessment), where spoken segments were scored on a 1-4 integer scale. In both studies the question was whether training on the easier data (with clear-cut cases and less disagreement) would be beneficial for training ML systems. The results were mixed. Jamison and Gurevych found that for data on nominal scales (classification tasks), training on easier data leads to improved performance. For affect data, training on easier cases can lead to improved results, when testing on easy cases, but only marginal or no improvement when testing on all data or just the hard cases. Loukina et al. found that training on easy data (as compared to mixed data) did not lead to better performance on test data. On the other hand, they found that the choice of data for testing the systems did matter – performance on easier testing data was always better than performance of mixed testing data. Yet, evaluation on just the easier cases should not be dismissed, as it provides an important validity indicator: making many errors on difficult cases might be tolerable, making many errors on clear-cut cases may raise serious doubts about validity of the system.

It is interesting to note how ratings of psycholinguistic variables, such as concreteness, valence, affect, etc., relate to the above taxonomy of rater disagreements. Concreteness scores from human raters are typically obtained on Likert scales. While attention and other random errors might be involved, Munoz-Rubke et al. (2018) also mention potential outlier effects. There could also be issues with reliability of raters (though responses from unreliable raters were eliminated in the Brysbaert et al. (2014) study). Task definition for rating concreteness/abstractness has also been criticized. Brysbaert et al. (2014) made special emphasis in rater instructions on concreteness in other modalities beyond visual perception, however, their results do not differ much from MRC and GN datasets, where such instructions were not explicitly presented. Attributing rater disagreement in concreteness ratings to 'situational item difficulty' is not quite plausible since ratings involved single words. A more plausible explanation for disagreement may be in the genuine ambiguity of some words, and/or the very subjective nature of concreteness ratings (Pollock, 2018).

Cases of ambiguity may arise when words have multiple senses or even just different parts of speech. For example, in the BWK dataset (scale 1-5), the word '*official*' has concreteness of 2.53 and SD of 1.43, while '*officially*' has concreteness 1.63 and much lower SD of 0.83. It might be that some raters interpreted '*official*' as a noun (and thus denoting a person), while others considered the adjective meaning (which is more abstract). The word '*officially*' is related to the same core meaning but has no such ambiguity. Perhaps concreteness ratings should be assigned per sense and not per wordform. Indeed, the Glasgow Norms (scale 1-7) have taken an early step in that direction, where 871 polysemous words were presented with a disambiguator, and thus the concreteness rating is per sense. However, even in such a disambiguated subset considerable variability of individual ratings exists – 360 entries on that list have SD>1.5, and the average SD of the disambiguated subset is 1.36. It seems raters disagreed even while rating specific senses of words.

The notion of collecting ratings per word sense is also related to predicting concreteness from word embeddings. Most of the classical word embeddings datasets (such as Google News word2vec, GloVe, etc) are not sense disambiguated, and their embeddings represent either a mix of senses or the most prevalent senses of words. For compatibility with such data, we opted to use Sentence-BERT embeddings in a similar way (i.e., per wordform). We opted to not use contextual BERT (or similar) embeddings per word and average them across multiple contexts. The issue in such case would be which contexts should be used for such averaging, and whether selection of contexts could have an influence on the senses that are implicitly modeled. However, this path that was not taken is also a path for future research. By carefully selecting contexts over which one averages contextual embeddings, a researcher might thus obtain sense-specific vectors, and potentially model sense-specific concreteness (and other psycholinguistic variables).

In sum, there seems more future work might be needed, both for collecting more reliable concreteness ratings, and for developing more sophisticated computational models of concreteness.

5. Conclusion

We investigated modeling of word-concreteness ratings with word embeddings. Study 1 demonstrated

that human-produced concreteness scores can be successfully predicted by using ordinary multiple regression with word embeddings. We compared 14 embedding models over three different datasets of human-produced concreteness scores. In all cases we obtained high Pearson correlation values (between .8 and .9) between original and estimated ratings. Using the RMSE evaluation measure, we find that all models achieve relatively low average error levels (mostly ranging from .5 to .8), which translates to 10-15% on the corresponding rating scales. Studies 2 and 3 investigated the effect of words that have 'less reliable' human-ratings. Rater disagreements for any given word result in higher standard-deviation of scores for that word. Using two datasets where standard deviation values for each word were released, we investigated how exclusion of words with high standard deviation values affects embedding-based regression models that learn to estimate the concreteness scores for words. We find that systematic exclusion of 'less reliable' words from the learning data can lead to evident improvement of results. However, study 3 indicates that such improvements stem from drastic changes in the distribution of concreteness scores when data is 'cleaned'. Training on filtered data does not generalize well to unfiltered data, whereas training on unfiltered data has enough information for modeling values for clean data.

6. Acknowledgments

We thank Beata Beigman Klebanov and two anonymous reviewers for valuable comments that helped to improve this paper.

7. Bibliographical References

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL* (Volume 1: Long Papers), pages 238–247, Baltimore, Maryland.

Beigman Klebanov, B., Leong, Ch.W., Flor, M. (2015). Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20.

Beigman Klebanov, B., and Beigman, E. 2014. Difficult Cases: From Data to Learning, and Back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Short Papers), pages 390–396.

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (20170. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 64:904-911.

Choi, K., and Downie, J.S. (2019). A Trend Analysis on Concreteness of Popular Song Lyrics. In *Proceedings of the 6th International Conference on Digital Libraries for Musicology* (DLfM '19), pp. 43–52.

M. Coltheart (1981), The MRC Psycholinguistic Database, *Quarterly Journal of Experimental Psychology*, 33A:497-505.

Dhillon, P.S., Foster, D.P., Ungar, L.H. (2015). Eigenwords: Spectral Word Embeddings. *Journal of Machine Learning Research,* 16:3035-3078.

Flor, M., and Somasundaran, S. (2019). Lexical concreteness in narrative. In *Proceedings of the Second Storytelling Workshop*, pages 75–80. Florence, Italy, August 1, 2019.

Hill, F., Korhonen, A., and Bentz, C. (2014). A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science*, 38(1):162–177.

Hills, T.T., and Adelman, J.S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, 143:87–92.

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

Jamison, E.K. and Gurevych, I. (2015). Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297.

Jessen, F., Heun, R., Erb, M., Granath, D.O., Klose, U., Papassotiropoulos, A., and Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74(1), 103–112.

Köper M., and Schulte im Walde, S. (2017). Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the First Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Levy, O., and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pages 302–308, Baltimore, Maryland.

Ljubešić, N., Fišer, D., and Peti-Stantić, A. (2018). Predicting Concreteness and Imageability of Words Within and Across Languages via Word

Embeddings. In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 217–222. Melbourne, Australia.

Loukina, A., Zechner, K., Bruno, J., and Beigman Klebanov, B. (2018). Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12, New Orleans, Louisiana.

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.

Maudslay, R. H., Pimentel, T., Cotterell, R., and Teufel, S. (2020). Metaphor Detection Using Context and Concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226.

Melamud, O., Dagan, I., and Goldberger, J. (2015). Modeling Word Meaning in Context with Substitute Vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 472–482, Denver, Colorado.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR (Workshop Poster)* 2013.

Munoz-Rubke, F., Kafadar, K., and James, K. H. (2018). A new statistical model for analyzing rating scale data pertaining to word meaning. *Psychological Research*, 82:787–805.

Naumann, D., Frassinelli, D., and Schulte im Walde, S. (2018). Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, LA, USA.

Paetzold, G.H., and Specia, L. (2016). Inferring Psycholinguistic Properties of Words. In *Proceedings of NAACL-HLT 2016*, pages 435–440, San Diego, California.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.

Paivio, A., Yuille, J.C. and Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology* Monograph Supplement, 76.

Pennington, J., Socher, R., and Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 1532–1543, Doha, Qatar.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 50:1198–1216.

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Scott, G.G., Keitel,A., Becirspahic, M., Yao, B., and Sereno S.C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3), 1258–1270.

Speer, R. and Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada.

Tater, T., Frassinelli, D., and Schulte im Walde, S. (2022). Concreteness vs. Abstractness: A Selectional Preference Perspective. In *Proceedings of the AACL-IJCNLP 2022 Student Research Workshop*, pages 92–98.

Thompson, B, and Lupyan, G. (2018). Automatic Estimation of Lexical Concreteness in 77 Languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Tsvetkov Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 248–258, Baltimore, Maryland.

Turney, P.D., and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Turney, P.D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690. Edinburgh, Scotland, UK.

Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Planck, B., and Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72, 1385-1470

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From Paraphrase Database to Compositional Paraphrase Model and Back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers,* 20:6–10.