Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns

Sai Abishek Bhaskar
Manipal Institute of Technology, India
sai.abishek@learner.manipal.edu

Maximilian Köper, Sabine Schulte Im Walde and Diego Frassinelli
University of Stuttgart, Germany
{koepermn,schulte,frassinelli}@ims.uni-stuttgart.de

Abstract

This paper explores variants of multi-modal computational models that aim to distinguish between abstract and concrete nouns. We assumed that textual vs. visual modalities might have different strengths in providing information on abstract vs. concrete words. While the overall predictions of our models were highly successful (reaching an accuracy of 96.45% in a binary classification and a Spearman correlation of 0.86 in a regression analysis), the differences between the textual, visual and combined modalities were however negligible, hence both text and images seem to provide reliable, non-complementary information to represent both abstract and concrete words.

1 Introduction

Over the years, different disciplines have been interested in exploring the contributions of contextual and perceptual information in human language acquisition and processing. From a psycholinguistic perspective, the grounding theory indicates that the mental representation of a concept is built not only through linguistic exposure but also incorporating multi-modal information extracted from real world situations, including auditory, visual, etc. stimuli (Barsalou, 1999; Shapiro, 2007; Glenberg and Kaschak, 2002). From a computational perspective, multi-modality has been shown to enhance corpus-based co-occurrence models that predict lexical information on various tasks, such as simulating word association, predicting semantic and visual similarity, determining the compositionality of multi-word expressions, and distinguishing between abstract and concrete concepts (Andrews et al., 2008; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013; Bruni et al., 2014; Hill et al., 2014; Lazaridou et al., 2015).

One focus of interest on the computational side has addressed the question of when and which perceptual information is helpful for semantic predictions in computational models, i.e., under which conditions perceptual information enhances or even outperforms textual information. For example, previous work described filters that added visual information to corpus-based information into a computational model of word meaning only in specific conditions: Kiela et al. (2014) suggested the *dispersion filter* that integrates only images that resemble each other to a certain degree. Köper and Schulte im Walde (2017) applied the same filter to a computational model of compositionality and added two more filters: the *imageability filter* integrating only images for highly imaginable words, as determined by existing imageability ratings; and the *clustering filter* only using images for a word that were similar to each other, as determined by a cluster analysis.

In this paper, we explore variants of multi-modal computational models that aim to distinguish between abstract and concrete nouns, to contribute to both psycholinguistic research by exploring differences in abstract vs. concrete concept representation and processing (Murphy, 2002; Barsalou and Wiemer-Hastings, 2005; Pecher et al., 2011), and to computational linguistic research by exploring differences between textual and visual information in modelling semantic knowledge. More specifically, we apply a binary classifier as well as a regression model to differentiate between abstract vs. concrete English nouns, as determined by concreteness ratings from Brysbaert et al. (2014). As features we compare standard textual count co-occurrences from a web corpus (Schäfer and Bildhauer, 2012), *word2vec* embeddings (Mikolov et al., 2013), *GoogLeNet* image vectors (Szegedy et al., 2015), and variants of text and image vector concatenation. As qualitative analysis, we compare classification errors across modalities, to zoom into the strengths and limits of the modalities and model parameters.

2 Material and Methods

2.1 Target Words

For our studies, we extracted nouns from the Brysbaert et al. (2014) collection of concreteness ratings for 40,000 English words. In this collection, each word was evaluated by at least 25 participants on a scale from 1 (abstract) to 5 (concrete). Given that participants were not aware of the part-of-speech (POS) of the word they were rating, we automatically assigned each word its most frequently occurring POS in our corpus (see Section 2.2). We focused our analyses on nouns because they are usually easier for humans to classify according to their concreteness compared to adjectives and verbs. In total we had 9,241 nouns that were also covered in an extensive selection of behavioural measures, such as valency scores (Warriner et al., 2013) and reaction times (Balota et al., 2007), which we aim to include in further analyses.

In one of our classification experiments (see Section 2.3), we included only a subset of the 9,241 nouns, i.e., the 1,000 most abstract and the 1,000 most concrete nouns, in order to assess a binary distinction of the most extreme noun instances regarding their abstractness vs. concreteness.

Figure 1 shows the distribution of the abstractness/concreteness scores across the 9,241 target nouns in two different ways. In the upper plot, the histogram indicates that there are considerably more concrete than abstract target nouns, which is the reason why we did not select nouns from specific ranges (e.g., 1–2 for abstract nouns vs. 4–5 for concrete nouns) for the binary classification, but rather an identical number of nouns from the two extremes (i.e., 1,000 nouns from the extreme ranges of each category). In the lower plot, the boxplots show the quartiles for the two sets of selected abstract/concrete targets vs. the remaining "middle" set of 7,241 nouns. This view once more illustrates that the 1,000 most abstract nouns cover a larger range of scores (from 1.07 for *spirituality* to 2.17 for *skill*) than the 1,000 most concrete nouns (from 4.97 for e.g. *shoe* to 5.00 for e.g. *lemon*), and also that the median of the middle part is rather high (i.e., confirming that many targets are highly concrete).

2.2 Text and Image Data

As textual modality for our target nouns, we compared standard count co-occurrence vectors and *word2vec* embedding vectors. The count vectors used *ENCOW14* (Schäfer and Bildhauer, 2012; Schäfer, 2015), currently one of the largest web corpora, to induce co-occurrence frequency matrices for the target nouns. As dimensions in the target noun vectors we compared two variants: on the one hand, we used the full set of 9,241 target nouns as co-occurring word dimensions; on the other hand, we used the reduced set of 2,000 target nouns as co-occurring word dimensions. The main reason for using the targets also as vector dimensions was to enable explorations of interdependencies between the abstractness/concreteness scores of our targets and their co-occurring words (see Frassinelli et al. (2017) for details). As window size for the co-occurrence counts we looked at two words to the left and to the right of the target words. As embedding vectors, we used the publicly available representations obtained from the *word2vec* cbow model (Mikolov et al., 2013). This model was trained on a Google-internal news corpus with 100 billion tokens.

The visual features were extracted from images downloaded from the Google search engine, following Kiela et al. (2016). We queried the search engine for up to 25 images per word, and converted all images into high-dimensional numerical representations by using the *Caffe* toolkit (Jia et al., 2014) and pre-trained models.

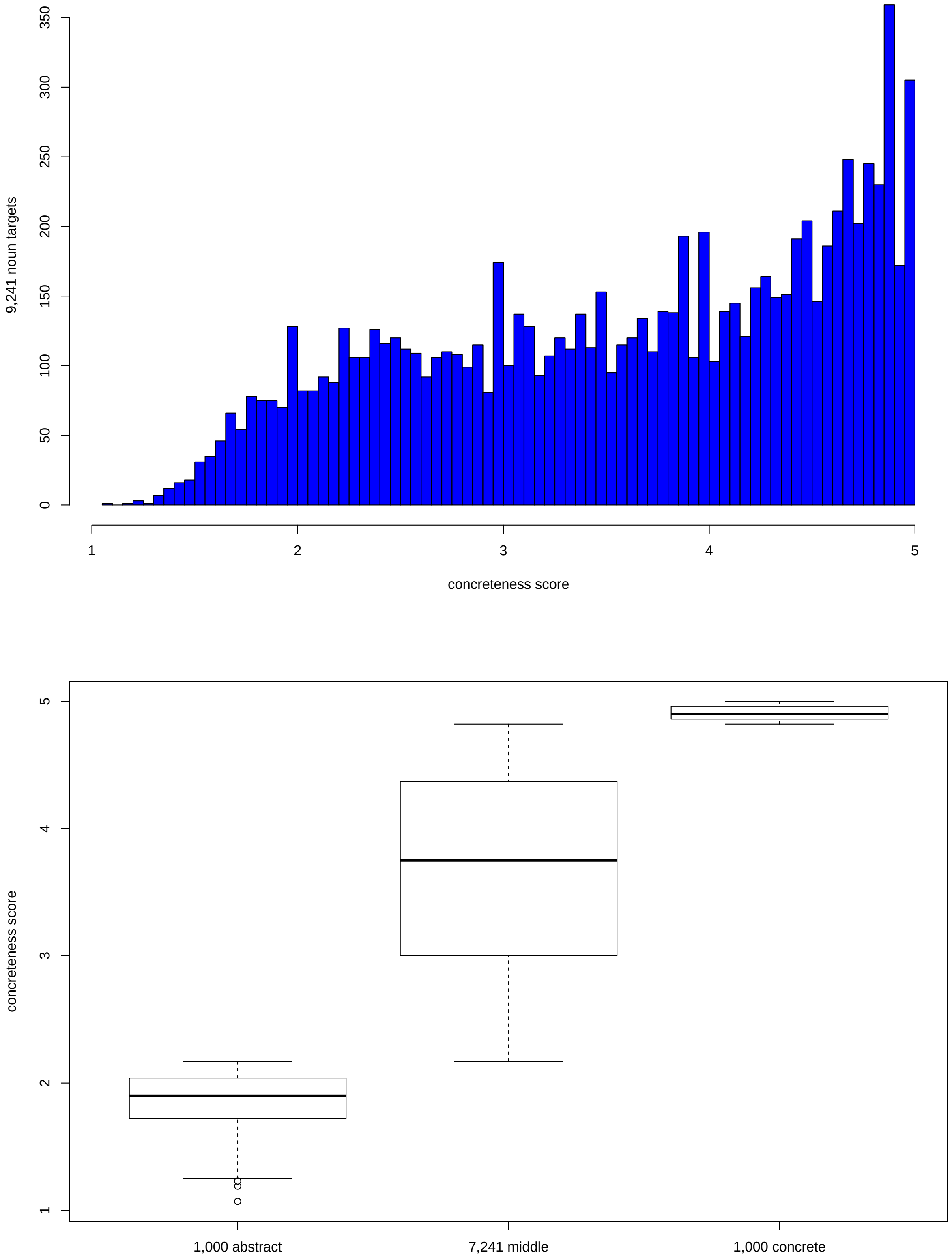

Figure 1: Abstractness/concreteness scores across our 9,241 target nouns.

For image recognition, we applied state-of-the-art convolutional neural networks:

1. BVLC *GoogLeNet* (Szegedy et al., 2015), a 22-layer deep network. We obtained vectors by outputting the value of layer pool5/7x7 s1, which is the last layer before the final softmax and contains 1,024 elements. These 1,024 elements determine the dimensionality of our vectors.
2. BVLC *AlexNet* (Krizhevsky et al., 2012), a neural network with 60 million parameters and 500,000 neurons, consisting of five convolutional layers, some of which are followed by max-pooling layers, and two globally connected layers with a final 1,000-way softmax. We output the vectors from layer fc7, a 4,096-dimensional feature vector for every image.

In all settings, a word is represented in the visual space by the mean vector of its (up to) 25 image representations.

When concatenating the textual and visual target vectors, we compared two variants:

1. CONCAT-SIMPLE: We concatenated the two vectors for each target noun without normalisation, i.e., disregarding the different dimensionalities of the vectors.
2. CONCAT-NORM: We concatenated the two vectors for each target noun after normalisation, i.e., we first calculated the proportion of each dimension in the two vectors separately, before concatenating them.

2.3 Classification

In order to evaluate the performance of our models in correctly classifying words according to their concreteness scores, we performed 10-fold cross-validation on the output of two classification tasks.

1. BINARY CLASSIFICATION: A binary classification of the most extreme abstract vs. concrete nouns should offer the best access to the relevance of the different modalities. This classification thus makes use of the subset of 2,000 target nouns (1,000 extremely abstract and 1,000 extremely concrete nouns), and applies a Random Forest Classifier. The binary classification is evaluated by computing the accuracy of the class assignments.
2. REGRESSION MODEL: In the second classification setup, we apply Gradient Boosting to predict the abstractness/concreteness scores for all 9,241 target nouns, for the subset of 2,000 target nouns, and for the “middle” set of 7,241 target nouns. The predicted scores are evaluated against the human ratings from the Brysbaert et al. (2014) collection using Spearman’s rank-order correlation coefficient ρ.

3 Results

In this section, we present the classification and regression results of our experiments, across the various sets of target nouns, their textual and visual features, and the two classification methods.

3.1 Quantitative Results

Table 1 shows a quantitative view on the results. For the binary classification, we compare the textual and visual features for the 2,000 extreme abstract/concrete target nouns. For the regression model, we compare the same features for the full set of 9,241 target nouns, the 2,000 extreme target nouns, as well as the “middle” set. Since there were only marginal differences between using 2,000 vs. 9,241 count vector dimensions as well as between using AlexNet vs. GoogLeNet image vectors, we selected only one from each of the alternatives for presentation. Regarding the concatenation of the vectors, the normalised version CONCAT-NORM consistently outperformed the non-normalised concatenation, so we focus on the former.

	Text (T)		**Images (V)**	**Text+Images (T+V)**		
Targets	9241 (count)	word2vec	GoogLeNet	2000+GoogLeNet	9241+GoogLeNet	word2vec+GoogLeNet
Binary Classification *(evaluation: accuracy)*						
2,000	94.78	92.21	92.79	95.89	**96.45**	93.79
Regression Model *(evaluation: ρ)*						
2,000	**0.78**	0.77	0.75	0.77	**0.78**	0.77
7,241	0.75	0.73	0.61	0.76	**0.79**	0.76
9,241	0.83	0.82	0.71	0.84	**0.86**	0.83

Table 1: Classification and regression results.

The differences between the textual (T), the visual (V) and the concatenated textual+visual (T+V) features are marginal. Looking at the best results per category (T/V/T+V), the textual features are slightly better than the visual features in the binary distinction between extremely abstract vs. concrete target nouns, and the combined multi-modal features are slightly better than the text-only features. For both the textual mode and the multi-modal mode, the count vectors slightly outperform the embedding vectors.

The same tendencies can be observed in the regression model, though the difference between the count and embedding vectors vanishes, and the differences between the modalities differ across the target noun sets: the predicted rankings of the concreteness scores have a better fit to the human rankings when taking all 9,241 target nouns into account, in comparison to taking subsets into account. In addition, the difference between relying on textual vs. visual information is stronger for the larger sets of 7,241 and 9,241 target nouns than for the set of 2,000 nouns.

3.2 Qualitative Results

The quantitative analysis showed that the improvements in semantic classification when using images in addition to text information were only marginal. We therefore aimed to explore differences in our abstractness/concreteness classifications across modalities by relying on a qualitative analysis.

Table 2 shows the strongest differences in abstractness/concreteness binary classification when comparing the best textual model (9241 count) and the best visual model (GoogLeNet), by calculating the disagreements between predictions and human ratings. I.e., the column 'T $>$ V' lists the 15 target nouns where the error between prediction and gold standard was much smaller across the many runs when relying on the textual vs. the visual features; vice versa for the column 'T $<$ V'. We can see that –according to our small sample– none of the differences in classification can clearly be attributed to only abstract (see 'A' and magenta font) or only concrete (see 'C' and blue font) target nouns: 7 out of the top 15 words that are classified better when relying on textual rather than visual features are concrete, as are 9 out of the top 15 words that are classified better when relying on visual rather than textual features. More or less the same applies in Table 3 when comparing the textual and the textual+visual features (9 out of 15 words for T vs. 7 out of 15 words for T+V are concrete).

4 Conclusion

In this paper, we explored variants of multi-modal computational models that aimed to distinguish between abstract and concrete nouns. While the overall predictions of many variants were highly successful (reaching an accuracy of 96.45% in a binary classification and a Spearman correlation of 0.86 using a regression model), the differences between the textual, visual and combined modalities were negligible, hence the information types in the different modalities are not complementary. A small-scale qualitative evaluation indicated that all variants of modalities do model both abstract and concrete words similarly well, i.e., both text and images seem to provide sufficient information to represent both abstract and concrete words.

T > V			**T < V**		
Target	Score	Abstract/Concrete	Target	Score	Abstract/Concrete
trickery	1.90	A	sausage	4.88	C
bamboo	4.86	C	butter	4.90	C
satire	1.96	A	firewood	4.93	C
bull	4.85	C	snout	4.89	C
extravagance	1.73	A	willingness	1.81	A
lightbulb	5.00	C	underpants	4.89	C
marshmallow	4.85	C	sociology	1.79	A
whim	1.69	A	gutter	4.85	C
regression	1.92	A	reason	1.93	A
thanks	2.15	A	bone	4.90	C
oven	4.97	C	purpose	1.52	A
pitcher	4.93	C	frustration	2.06	A
sentimentality	1.73	A	sparrow	4.85	C
bandage	4.85	C	cowardice	2.10	A
dishonesty	1.90	A	thought	1.97	C

Table 2: Most striking classification differences using textual vs. visual features.

T > T+V			**T < T+V**		
Target	Score	Abstract/Concrete	Target	Score	Abstract/Concrete
beard	4.96	C	principle	1.70	A
offensive	2.10	A	sausage	4.88	C
severity	1.76	A	purpose	1.52	A
ecstasy	2.04	A	reason	1.93	A
bowl	4.87	C	butter	4.99	C
lollipop	4.96	C	impossibility	1.52	A
walnut	4.97	C	sociology	1.79	A
breakup	2.00	A	underpants	4.89	C
fridge	4.92	C	willingness	1.81	A
complement	2.00	A	bone	4.90	C
hotel	4.93	C	gutter	4.85	C
butterfly	4.93	C	firewood	4.93	C
wish	1.77	A	frustration	2.06	A
jeans	5.00	C	weird	1.59	A
cabin	4.92	C	snout	4.89	C

Table 3: Most striking classification differences using textual vs. textual+visual features.

Acknowledgments

The first author (Sai Abishek Bhaskar) would like to thank his parents for supporting him during his time at the University of Stuttgart. The research was further supported by the DFG Collaborative Research Centre SFB 732 (Diego Frassinelli, Maximilian Köper), and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde). We also thank the two anonymous reviewers for their comments.

References

Mark Andrews, David Vinson, and Gabriella Vigliocco. Inferring a Probabilistic Model of Semantic Memory from Word Association Norms. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

David A. Balota, Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459, 2007.

Lawrence W. Barsalou. Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22:577–660, 1999.

Lawrence W. Barsalou and Katja Wiemer-Hastings. Situating Abstract Concepts. In D. Pecher and R. Zwaan, editors, *Grounding cognition: The role of perception and action in memory, language, and thinking*, chapter 7, pages 129–163. Cambridge University Press, New York, 2005.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods*, 64:904–911, 2014.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte im Walde. Contextual Characteristics of Concrete and Abstract Words. In *Proceedings of the 12th International Conference on Computational Semantics*, Montpellier, France, 2017. To appear.

Arthur M. Glenberg and Michael P. Kaschak. Grounding Language in Action. *Psychonomic Bulletin and Review*, 9(3):558–565, 2002.

Felix Hill, Roi Reichart, and Anna Korhonen. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2(1):285–296, 2014.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 835–841, 2014.

Douwe Kiela, Anita L. Verő, and Stephen Clark. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Maximilian Köper and Sabine Schulte im Walde. Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 200–206, Valencia, Spain, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, Nevada, USA, 2013.

Gregory Murphy. *The Big Book of Concepts*. MIT Press, 2002.

Diane Pecher, Inge Boot, and Saskia Van Dantzig. Abstract Concepts. Sensory-Motor Grounding, Metaphors, and Beyond. *Psychology of Learning and Motivation – Advances in Research and Theory*, 54:217–248, 2011.

Stephen Roller and Sabine Schulte im Walde. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, WA, 2013.

Roland Schäfer. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany, 2015.

Roland Schäfer and Felix Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey, 2012.

Larry Shapiro. The Embodied Cognition Research Programme. *Philosophy Compass*, 2(2):338–346, 2007.

Carina Silberer and Mirella Lapata. Grounded Models of Semantic Representation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea, 2012.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *Computer Vision and Pattern Recognition*, 2015.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.