A Couch Potato is not a Potato on a Couch: Prompting Strategies, Image Generation, and Compositionality Prediction for Noun Compounds

Anonymous ACL submission

Abstract

Predicting the compositionality of English noun–noun compounds, such as *climate change* and *couch potato*, has traditionally relied on text-based methods. We explore a novel image-based approach, believing that images convey rich information beyond what text can capture, and that visual context may provide valuable insights. We generate images for compounds and their constituents using variants of text prompts, then encode these images with Vision Transformers, and assess the depicted meaning relatedness through cosine similarity. Evaluated against human compositionality ratings, the image-based approach performs en par with text-based methods for concrete compounds, while challenges in image acquisition and the misalignment between visual and semantic similarity negatively affect the results for abstract compounds.

1 Introduction

Compositionality is a core concept in linguistics (Partee, 1984); it posits that the meaning of complex expressions, such as noun–noun compounds, can be derived from their constituent meanings. The degree of compositionality however varies; e.g., compounds like *climate change* are highly compositional, while others like *couch potato* are less so. Accurately predicting compositionality is crucial for natural language understanding tasks such as summarization and machine translation, as misinterpretation may have significant impact.

State-of-the-art models of compositionality prediction primarily leverage text-based numerical representations. In contrast, our study suggests that visual cues, such as colors and spatial relationships, could reveal aspects of meaning not captured by the text modality. We thus examine the potential of using images of compounds and their constituents for compositionality prediction in English noun–noun compounds, and contrast our approach with traditional text-based methods.

Image acquisition is however particularly challenging for our task, as standard image search returns false positives for non-compositional compounds, e.g., a *couch potato* is depicted as a potato (instead of a lazy person) sitting on a couch (see Figure 1). Automatically generating images of (non-compositional) compounds offers a promising solution to obtain adequate images; it requires, however, carefully designed prompts to ensure that the image generation model captures the compounds' figurative meanings. We contribute (i) prompting strategies for image generation with increasing contextual description levels to address this challenge; (ii) a vision-based approach using Vision Transformer to predict compositionality; (iii) and analyses of modeling effects regarding aspects of non-compositionality, including the compounds' abstractness vs. concreteness as well as prototypicality of constituent meanings.

2 Related Work

Traditionally, most computational approaches to automatically predict the compositionality of noun compounds have been realized as text-based vector space models (Reddy et al., 2011; Salehi et al., 2015; Schulte im Walde et al., 2016; Cordeiro et al., 2019; Miletić and Schulte im Walde, 2023, i.a.). Few studies addressed compound meaning using multimodal information, such as Bruni et al. (2014) to identify figurative uses of color terms in adjective–noun phrases, Pezzelle et al. (2016) and Günther et al. (2020) to predict compound representations, and Köper and Schulte im Walde (2017) to predict the compositionality of German compounds. On a more general level, multimodality has been employed to identify (Shutova et al., 2016) and interpret (Kalarani and Bhattacharyya, 2024) metaphors, culminating into approaches of metaphor visualization through text-to-image synthesis (Chakrabarty et al., 2023; Zhang et al., 2024).

3 Methodology

Given a compound (e.g., *couch potato*), we assess how related the compound meaning is in relation to the constituent meanings. Our task is to predict its degree of compositionality in relation to the modifier (*couch*) and its head (*potato*). The image-based approach follows three main steps:

1. **Image Acquisition**: Acquire images for the compound and its constituents.
2. **Target Representation**: Create vector representations for compounds and constituents.
3. **Prediction**: Estimate the similarity between the vector representation of the compound and each constituent. These similarities serve to rank the compounds; then ranks are compared to gold compound–constituent ratings.

3.1 Image-Based Experimental Pipeline

Image Acquisition To reliably capture a word's meaning, images should accurately represent its concept while covering a range of different visual realizations and scenarios for greater robustness. We experiment with two automatic methods: (i) We employ a standard strategy to download images from Bing[1]. (ii) Given that the image search results turn out as not reliable for non-compositional compounds (e.g., a *couch potato* is depicted as a potato (instead of a lazy person) sitting on a couch) we suggest a promising alternative that we expect to be also highly valuable for the acquisition of images of figurative expressions in general: We generate images with a text-to-image model, for which we select the Diffusion Transformer PixArtSigma[2], after testing various diffusion models, and we explore four distinct prompting strategies to guide image generation[3]:

Word: Prompts consist solely of the target word, without any context or modifications.

Sentence: Prompts are sentences containing the target word, sourced from the ENCOW16AX web corpus (Schäfer and Bildhauer, 2012).

Definition: Prompts are definitions of the target word, generated by ChatGPT.

Context: Prompts are diverse, descriptive scenarios involving the target word, generated by ChatGPT.

[1] https://www.bing.com/images

[2] https://huggingface.co/PixArt-alpha/PixArt-Sigma-XL-2-1024-MS

[3] See Appendix A for examples and ChatGPT instructions.

Figure 1: `Bing` (left) and `Context` (right) images of *couch potato*.

We download 10 images per target word from `Bing`. For `Word`, we generate 10 images with different seeds. For `Sentence`, we extract 10 sentences per target, generating one image per sentence. For `Definition`, we create 3 definition prompts, generating one image each, while for `Context`, we generate 25 context prompts, with one image per prompt. Downloaded images are resized to 1024×1024, while generated images are created directly at this resolution.

Target Representation We extract feature vectors from images using a Vision Transformer model[4]. We create a single visual representation for each target word by mean-pooling the feature vectors of multiple images of the same word.

Prediction We assess the meaning relatedness between a compound and its constituents using cosine similarity, where a higher similarity corresponds to a higher degree of compositionality.

3.2 Experimental Setup

Gold Data Reddy et al. (2011) compiled a compositionality dataset with human ratings for 90 noun–noun compounds, collected via Amazon Mechanical Turk. It contains compounds with varying degrees of compositionality, including compounds where both constituents are literal (e.g., *swimming pool*), only one is literal (e.g., *flea market*), or neither is literal (e.g., *cloud nine*). Ratings range from 0 (non-compositional) to 5 (highly compositional) and include both compound–whole and compound–constituent ratings; for this work, we only consider the latter. We use 88 compounds; we excluded two (*number crunching* and *pecking order*) due to frequency limitations.

Reference We compare the proposed image-based approach to a widely used text-based compositionality prediction method, where target repre-

[4] https://pytorch.org/vision/main/models/generated/torchvision.models.vit_h_14.html

		Modifier	Head
	Bing	.345	.232
PixArtSigma	**Word**	-.005	.043
	Sentence	**.506**	.096
	Definition	.414	.288
	Context	.457	**.440**
	Skip-gram	.565	.574

Table 1: Spearman's ρ for predictions across various image acquisition strategies and `Skip-gram`.

sentations are derived from text using a Word2Vec `Skip-gram` model (Mikolov et al., 2013) trained on the ENCOW16AX web corpus (Schäfer and Bildhauer, 2012) with a window size of 20, minimum count of 5, and 300 dimensions.

Evaluation Our approach predicts two compound–constituent ratings for each target compound, one for the compound–modifier rating, and one for the compound–head rating. To assess prediction quality, we measure Spearman's rank-order correlation coefficient (ρ) between the predicted scores and the gold standard ratings provided by Reddy et al. (2011).

3.3 Results

Table 1 shows the results on image-based compositionality prediction, i.e., the modifier and head correlations for `Bing`, `Word`, `Sentence`, `Definition` and `Context`, as well as the text-based `Skip-gram` for comparison.

The image-based approach shows promising results when images are downloaded via `Bing`, with correlations of .345 and .232 for modifiers and heads, respectively. When using generated images, performance is highly influenced by the prompting strategy. `Word` yields weak correlations of -.005 and .043. A substantial improvement is seen with `Sentence`, particularly for modifiers, where the correlation rises to .506—noticeably higher than for `Bing`, though the head correlation remains weak at .096. `Definition` improves the head correlation to .288, but the modifier correlation drops slightly to .414. Finally, `Context` produces the best overall results, with correlations of .457 for modifiers and .440 for heads. Although some of these results are promising, `Skip-gram` still outperforms every variant of the image-based approach.

4 Analysis

We conduct a detailed analysis of the image-based approach, focusing on the images and predictions generated by the highest-performing candidate, `Context`, with `Skip-gram` included as a point of comparison. Section 4.1 compares prediction quality between concrete versus abstract compounds. Section 4.2 examines one compound with accurate predictions from `Context` against one with poor predictions.

	Context		**Skip-gram**	
	Mod	Head	Mod	Head
Concrete	**.448**	.174	.439	.220
Abstract	.299	.400	**.471**	.430

Table 2: Spearman's ρ for `Context` and `Skip-gram` predictions for concrete versus abstract compounds.

4.1 Concrete versus Abstract Compounds

We analyze the differences in predictions for concrete, easily perceivable, against abstract and less perceivable compounds, expecting differences in the benefit of visual perception features. As a first step, we collect human annotations, where participants rated each compound on a scale from 0 (abstract) to 5 (concrete), following previous work regarding the instructions (Brysbaert et al., 2014; Muraki et al., 2023). The 30 compounds with the highest mean ratings are categorized as concrete, and the 30 with the lowest as abstract (see Table 3).

Table 2 shows the modifier and head correlations. For concrete compounds, `Context` and `Skip-gram` perform similarly. `Context` achieves correlations of .448 for modifiers and .174 for heads, while `Skip-gram` reaches .439 and .220, respectively. In contrast, for abstract compounds, `Skip-gram` performs noticeably better, with correlations of .471 for modifiers and .430 for heads, compared to `Context`'s .299 (modifier) and .400 (head).

These results align with our expectations: the image-based approach works better for compounds with clear, recognizable features, such as concrete nouns, which are easier to capture and represent in images. In contrast, abstract compounds, which are harder to visually represent (Pezzelle et al., 2021; Tater et al., 2024), lead to poorer predictions, and the text-based approach outperforms the image-based one.

4.2 Analysis of Individual Compounds

To assess prediction quality for individual compounds, we rely on Rank Differences (RDs), which compare each compound's predicted rank to its rank within the gold ratings, thereby calculating the

Figure 2: Images of *graveyard shift*, *graveyard* and *shift*.

Figure 3: Images of *engine room*, *engine* and *room*.

absolute difference. RDs are computed separately for modifiers and heads, with the average providing an overall RD for each compound (see Table 4). To illustrate, the compound *couch potato* has a low RD of 0.5, indicating a close alignment between the `Context`-predicted ratings and the gold ratings.

Graveyard Shift For the compound *graveyard shift*, which refers to *"a work shift taking place from late night to early morning"*, `Context` performs well, achieving an RD of 2.5, while `Skip-gram` has a much higher RD of 22.5.

Figure 2 presents the underlying images. Those of *graveyard* (second row) show graveyards with tombstones, mostly in daylight. In contrast, *shift* (third row) is more abstract and harder to represent; still, the images capture the concept fairly accurately, by depicting people working in various contexts, such as bakers and construction workers. Finally, the images of *graveyard shift* (first row) closely resemble those of *shift*, as they also depict workers in various settings, but with the key distinction of always occurring at night, differentiating them from the daytime scenes associated with *shift*.

The computed visual similarities for *graveyard shift* are .243 for *graveyard* and .753 for *shift*, which aligns well with the underlying images. The corresponding gold ratings are .38 for *graveyard* and 4.50 for *shift*, showing that the visual similarities accurately reflect the semantic contributions of each constituent. This alignment between visual and semantic relationships results in strong predictions for the compound.

Engine Room *Engine room* is a compound for which `Context` predicts poor compositionality ratings with an RD of 46, while `Skip-gram` performs slightly better but still poorly, with an RD of 34.5.

Figure 3 presents the underlying images. Those of *room* (third row) are high-quality and accurately depict a variety of types of rooms (e.g., living rooms and conference rooms). In contrast, the images of *engine room* (first row) depict a mix of diverse types of engine rooms with trains and cars.

The visual cosine similarity is .45, while the gold compositionality rating is 5, the maximum value. The captured visual similarity seems reasonable, as images of *engine room* and *room* should intuitively share some features but also exhibit significant differences, because a prototypical *room* is rather a living or conference than an engine room. Unfortunately, the predicted visual similarity does not align with the compositionality rating.

We observe that the image-based approach, which relies solely on visual similarity, performs well when shared visual features align with the semantic contributions of constituents to the compound's meaning. However, it struggles in cases where visual similarity does not accurately capture these contributions, thus highlighting the limitations of using visual features alone when predicting compositionality.

5 Conclusion

This study explored the potential of an image-based approach to predict the compositionality of 88 English noun–noun compounds, relying on prompt strategies in interaction with image generation to acquire and compare adequate images, a promising novel strategy for visual representations of figurative language. Results show that image quality, diversity, and alignment with semantic meaning are crucial factors; generated images improved predictions, especially for concrete, literal compounds, though abstract compounds remained challenging. Although the text-based approach is more effective, our findings suggest that a multimodal approach combining text and image features could further enhance compositionality prediction.

Limitations

The image-based approach relies heavily on the quality and availability of relevant, accurate images for the compounds. While image generation can address some of these challenges, it comes with significant resource demands (GPU) and can be time-consuming, which may hinder scalability, especially when generating large numbers of images for many compounds. Additionally, while the approach performs well for concrete compounds, it struggles with abstract compounds and those that are difficult to visualize.

Ethics Statement

We see no ethical issues related to this work. All experiments involving human participants were voluntary, with fair compensation (12 Euros per hour), and participants were fully informed about data usage. We did not collect any information that can link the participants to the data. All modeling experiments were conducted using open-source libraries, which received proper citations. All relevant information (including created artifacts, used packages, information for reproducibility, etc.) can be found in (PLACEHOLDER for GitHub repository, will be added upon paper acceptance).

References

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

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 64:904–911.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Fritz Günther, Marco Alessandro Petillia, and Marco Marelli. 2020. Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, 112.

Abisek Rajakumar Kalarani and Pushpak Bhattacharyya. 2024. Multimodal captioning and figurative language understanding: A survey.

Maximilian Köper and Sabine Schulte im Walde. 2017. 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.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia.

Emiko J. Muraki, Summer Abdalla, Marc Brysbaert, and Penny M. Pexman. 2023. Concreteness ratings for 62,000 English multiword expressions. *Behavior Research Methods*, 5:2522–2531.

Barbara H. Partee. 1984. Compositionality. In Fred Landman and Frank Veltman, editors, *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium*, pages 281–311. Foris Publications.

Sandro Pezzelle, Ravi Shekhar, and Raffaella Bernardi. 2016. Building a bagpipe with a bag and a pipe: Exploring conceptual combination in vision. In *Proceedings of the 5th Workshop on Vision and Language*, pages 60–64, Berlin, Germany. Association for Computational Linguistics.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, pages 977–983, Denver, Colorado, USA.

Roland Schäfer and Felix Bildhauer. 2012. 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.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Tarun Tater, Sabine Schulte im Walde, and Diego Frassinelli. 2024. Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21581–21597, Miami, Floria, USA.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. 2024. GOME: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18500–18510, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt Generation Using ChatGPT

This appendix describes the procedure for generating prompts for text-to-image models using ChatGPT. The process consists of three main phases, repeated separately for each prompting strategy (`Definition`, `Context`):

- **Preparation Phase:** ChatGPT is introduced to the task, including the goal of generating prompts that accurately reflect the meanings of compounds and their constituents. Prompts are described as detailed descriptions of the intended image, formatted in CSV without headers or numbering for easy copying.
- **Instruction Phase:** ChatGPT receives guidelines for each strategy. For `Definition`, it creates three prompts based directly on the noun definitions. For `Context`, it generates 25 diverse prompts capturing real-world scenarios related to the target word, ensuring broad representation.
- **Query Phase:** ChatGPT generates prompts for each target compound in sequence (modifier, head, compound) based on the provided instructions, outputting them in CSV format for further use.

A.1 Definition Prompt Examples

***Couch Potato*:**

1. *"A person who spends a significant amount of time sitting or lying down, typically watching television or engaging in sedentary activities."*
2. *"A term describing someone who leads a sedentary lifestyle, preferring indoor activities such as watching TV or playing video games."*
3. *"An informal term for a person who is inactive or lazy, often spending leisure time on a couch or sofa."*

***Couch*:**

1. *"A piece of furniture designed for seating two or more people, typically with a back and armrests."*
2. *"A long upholstered piece of furniture for reclining or sitting, often found in living rooms or lounges."*
3. *"A sofa or settee, usually with cushions and upholstered arms and back, used for relaxation or casual seating."*

***Potato*:**

1. *"An edible tuber that is a staple food in many cultures, typically underground and harvested from the Solanum tuberosum plant."*
2. *"A starchy vegetable with a variety of culinary uses, such as boiling, baking, frying, or mashing."*
3. *"The plant itself, Solanum tuberosum, which belongs to the nightshade family and produces tubers that vary in size, shape, and color."*

A.2 Context Prompt Examples

***Couch Potato*:**

1. *"A couch potato binge-watching their favorite TV series, surrounded by cushions and blankets."*
2. *"A person on the couch, flipping through a photo album or scrapbook."*
3. *"A person lounging on a couch with a bowl of popcorn, absorbed in a movie marathon."*

***Couch*:**

1. *"A vintage leather couch with tufted upholstery, adding a touch of elegance to a study."*
2. *"A cozy reading nook with a couch by the window, bathed in natural sunlight."*
3. *"A modular couch with interchangeable pieces, allowing for easy customization and rearrangement."*

***Potato*:**

1. *"A beautifully plated baked potato topped with melting butter and dollops of sour cream."*
2. *"A farmer harvesting potatoes in a sunlit field, with rows of potato plants in the background."*
3. *"A close-up of potato peelings on a kitchen countertop, with a peeler and scattered peels."*

B Compound Subsets

C Rank Differences

Compound	Concreteness	Compound	Concreteness
car park	5.0	crash course	2.5
human being	4.9	couch potato	2.5
swimming pool	4.9	snake oil	2.5
credit card	4.7	climate change	2.4
parking lot	4.7	night owl	2.4
polo shirt	4.7	sitting duck	2.4
ground floor	4.6	sacred cow	2.4
call centre	4.6	game plan	2.4
brick wall	4.6	eye candy	2.3
cocktail dress	4.6	rock bottom	2.3
application form	4.4	monkey business	2.3
zebra crossing	4.4	face value	2.2
health insurance	4.4	role model	2.2
video game	4.3	meltin gpot	2.2
law firm	4.3	agony aunt	2.2
bank account	4.2	graveyard shift	2.2
engine room	4.1	cash cow	2.2
radio station	4.1	guilt trip	2.1
grandfather clock	4.1	memory lane	2.1
balance sheet	4.1	shrinking violet	2.1
head teacher	4.1	gravy train	2.1
speed limit	4.0	kangaroo court	2.0
gold mine	3.9	lip service	2.0
graduate student	3.9	ivory tower	2.0
brass ring	3.9	blame game	2.0
lotus position	3.9	rat run	2.0
panda car	3.8	swan song	2.0
search engine	3.7	rat race	1.9
china clay	3.6	crocodile tear	1.9
research project	3.6	cloud nine	1.9

Table 3: Top 30 (left) and bottom 30 (right) compounds ranked by (mean) concreteness, based on human-judgements. Scale: 0 (abstract) to 5 (concrete).

Compound	Context	Skip-gram	Compound	Context	Skip-gram
couch potato	0.5	7.5	mailing list	16.2	13.2
parking lot	1.8	32.8	memory lane	16.8	19.8
guilt trip	2.0	12.5	cocktail dress	17.2	13.2
graveyard shift	2.5	22.5	snail mail	18.8	16.0
rat run	3.5	24.8	swimming pool	18.8	3.0
grandfather clock	3.8	27.2	blame game	19.5	9.0
case study	5.5	8.0	diamond wedding	20.0	32.5
graduate student	6.8	7.8	end user	20.0	28.8
think tank	7.0	29.0	web site	21.0	33.0
rush hour	7.8	13.0	brass ring	21.5	5.5
crash course	8.0	8.0	sitting duck	21.8	13.8
research project	8.0	10.5	fine line	23.5	16.5
front runner	8.0	30.8	silver spoon	23.8	29.5
zebra crossing	8.0	19.5	video game	23.8	6.8
balance sheet	8.2	32.8	cash cow	24.0	14.5
rock bottom	8.5	6.5	agony aunt	25.5	20.5
nest egg	8.8	5.8	call centre	26.0	32.8
human being	8.8	13.2	bank account	26.0	7.5
spelling bee	9.0	17.5	public service	26.5	7.0
game plan	9.2	24.2	face value	27.0	19.8
melting pot	10.5	9.0	silver bullet	27.5	17.0
gravy train	10.5	25.0	chain reaction	28.2	22.0
radio station	10.5	11.8	fashion plate	29.5	13.0
eye candy	11.2	26.8	ground floor	31.2	30.2
polo shirt	11.8	18.2	rat race	31.5	22.0
credit card	12.0	9.0	brick wall	33.0	37.5
search engine	12.5	14.0	kangaroo court	33.5	20.0
cheat sheet	12.5	5.8	gold mine	33.5	40.5
interest rate	12.8	13.5	lotus position	34.5	53.0
flea market	12.8	30.2	car park	35.0	30.2
ivory tower	12.8	3.5	smoking jacket	35.2	11.2
head teacher	12.8	25.2	monkey business	35.5	39.0
spinning jenny	13.2	22.0	application form	35.8	35.2
climate change	13.2	20.8	lip service	36.0	29.5
health insurance	13.5	6.8	shrinking violet	37.2	16.5
snake oil	13.5	12.8	cloud nine	37.8	25.2
role model	13.5	23.0	rocket science	38.5	8.5
firing line	14.5	7.2	speed limit	44.8	25.2
china clay	15.0	4.8	acid test	45.0	10.0
cutting edge	15.0	10.5	engine room	46.0	34.5
silver screen	15.0	16.8	night owl	46.2	15.2
smoking gun	15.2	12.0	sacred cow	48.5	16.5
law firm	15.5	31.5	panda car	57.0	1.0
swan song	16.2	23.0	crocodile tears	62.5	17.0

Table 4: RDs between `Context`/`Skip-gram` predictions and the gold ratings, sorted by increasing `Context` RDs.