Understanding the Cognitive Complexity in Language Elicited by Product Images

Yan-Ying Chen [*1] **Shabnam Hakimi** [*1] **Monica Van** [*1] **Francine Chen** [1] **Matthew Hong** [1] **Matt Klenk** [1] **Charlene Wu** [1]

Abstract

Product images (e.g., a phone) can be used to elicit a diverse set of consumer-reported features expressed through language, including surface-level perceptual attributes (e.g., "white") and more complex ones, like perceived utility (e.g., "battery"). The *cognitive complexity* of elicited language reveals the nature of cognitive processes and the context required to understand them; cognitive complexity also predicts consumers' subsequent choices. This work offers an approach for measuring and validating the cognitive complexity of human language elicited by product images, providing a tool for understanding the cognitive processes of human as well as virtual respondents simulated by Large Language Models (LLMs). We also introduce a large dataset that includes diverse descriptive labels for product images, including human-rated complexity. We demonstrate that human-rated cognitive complexity can be approximated using a set of natural language models that, combined, roughly capture the complexity construct. Moreover, this approach is minimally supervised and scalable, even in use cases with limited human assessment of complexity. The dataset will be made public[1].

1. Introduction

Language can reveal how *context* shapes the way consumers evaluate products, signaling the range of psychological processes underlying their assessments. Previous work has shown how models of language capture stimulus feature processing using properties like concreteness, or the degree to which a word refers to a perceptible entity in mind (Brysbaert et al., 2014; Charbonnier & Wartena, 2019), while other models use language to infer a person's psychological state by quantifying factors such as sentiment (Zhang & Qian, 2020), emotion (Demszky et al., 2020), and personality (Flekova & Gurevych, 2015). Such models have not yet elucidated how the complexity of thought processes influences choice. We examine cognitive complexity, or the *additional context needed outside of an eliciting visual stimulus* to understand or explain human-generated language (Hakimi et al., 2023), as a means for inferring the relationship between thoughts and actions. For example, when viewing a specific product image, such as a red car, some might focus on the dominant perceptible feature—the color red—while others may comment on more seemingly "complex" features shaped both by perception and projected utility and experiences, like "engine." Both words are concrete but involve different levels of complexity in processing visual stimuli.

In this light, a measure of cognitive complexity should make individual variability in language generation and associated underlying preferences more interpretable because it quantifies a shared, underlying structure between idiosyncratic representations. For example, given the same product image stimulus, two people might describe very different evoked memories; although the contents of these descriptions may differ greatly in other natural language measures, both sets of text would have high complexity. By leveraging shared features of the processes that generated these descriptions, we can more accurately and effectively predict subsequent behaviors (e.g., choices, language use) both for the two individuals and the population as a whole. Indeed, previous work demonstrated that cognitive complexity explains unique variance in preferences and improves choice prediction (Hakimi et al., 2023). Because cognitive complexity captures the process through which a stimulus can potentiate emergent psychological phenomena—i.e., thoughts and behaviors that are not intrinsic to the stimulus itself—it can also be used to infer whether the language was generated by a human or an LLM imitating a human. For example, an observed discrepancy between human- and LLM-generated

[*]Equal contribution [1]Toyota Research Institute, Los Altos, California. Correspondence to: Yan-Ying Chen <yanying.chen@tri.global>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]The data will be available at https://github.com/TRI-MAC/cognitive_complexity

responses in the distribution of cognitive complexity scores for certain visual stimuli can be used to filter out unqualified subjects or data (Veselovsky et al., 2023), e.g., subjects who submit machine-generated responses to a survey that requires human responses. Similarly, distributional alignment can also be used to evaluate an LLM's performance in generating language with human-like cognitive complexity.

Here, we validate a computational measure of cognitive complexity by interrogating its relationship with human ratings as well as related natural language constructs. Because cognitive complexity is a multidimensional construct (Conway III et al., 2014; Glaze et al., 2018), we evaluate the putative informational content of the "additional context" needed for understanding by testing the relationship between cognitive complexity and four related constructs—visibility, semantics, uniqueness, and concreteness—that have been reliably encoded using validated vision and language models. Further, to verify how well cognitive complexity aligns with human judgments, we first elicit user-generated text labels for product images (as (Hakimi et al., 2023)) then compare computed to human-rated measures. This work presents the first large and diverse dataset collected to understand the cognitive complexity of product image descriptions.

Our contributions include: (1) the first dataset with 4000+ product images across 14 categories and 45,609 human-generated text labels and complexity ratings for computing cognitive complexity, and (2) the use of relevant vision and language models to characterize and approximate the cognitive complexity of language elicited by product images.

2. Related Work

Images provide information that shapes the cognitive process (Cavanagh, 2011; Marr, 2010). Hence, what is visible to people may influence the cognitive complexity of their reactions. Recent work on joint text and vision embeddings has emphasized the strong ties between vision and language semantics for cognitive understanding in such diverse domains as hate speech recognition (Kiela et al., 2020) and noun property prediction (Yang et al., 2022). Such work may lead to improved models that exploit unimodal priors.

Besides visibility and semantics, constructs built on language itself such as readability and uniqueness of words can influence text simplification (Maddela & Xu, 2018). Concreteness of words that were rated by humans (Brysbaert et al., 2014; Iliev & Axelrod, 2016) or measured by learned word embeddings (Charbonnier & Wartena, 2019; Wartena, 2022) has been used to measure abstractness. (Hessel et al., 2018) explored word concreteness in four image datasets by looking at the image clustering structure. (Ramakrishnan & Deniz, 2021) found that GPT word embeddings captured all information in brain imaging data for concreteness prediction based on logistic regression models. However, the granularity of word concreteness is not grounded in a particular image. Cognitive complexity differs in that it assumes a joint vision-language representation: the complexity of a word is specific to the image it describes because reactions to different images vary. Moreover, cognitive complexity is a large theoretic construct (Conway III et al., 2014) that overlaps with multiple constructs as discussed above.

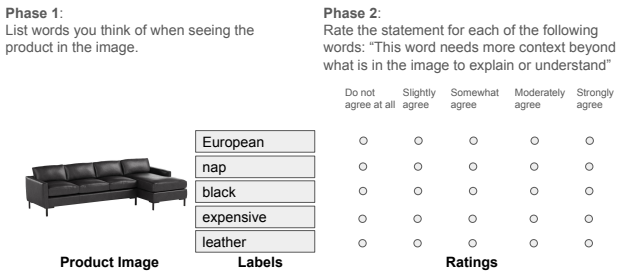


Figure 1. Data collection paradigm.

Category	images	labels	vocab.	complexity
furniture	2,174	23,210	3,278	1.75 ±0.94
decor	1,054	10,481	2,664	1.77 ±0.95
car	181	2,925	741	1.90 ±0.95
bed/bath	239	2,686	871	1.76 ±0.96
all	4,093	45,609	6,583	1.77 ±0.94

Table 1. Dataset with product image and label counts as well as vocabulary size and complexity ratings (mean±sd) for the largest four product categories. See all 14 categories in Appendix Table 5.

3. Data Elicitation

In a series of studies, language was used to assess humans' representation of and preferences for a variety of real consumer products, either cars[2] or items from the Amazon Berkeley Object (ABO) dataset (Hakimi et al., 2023; Collins et al., 2022). The complexity of these responses to product images was then evaluated by separate human raters.

In Phase 1 (Fig. 1), 3,009 crowdsourced (Amazon Mechanical Turk, 2022) labelers were shown a product image and asked to "list words you think of when seeing the product in the image." Each labeler was shown 16 images sampled from diverse product categories, and listed 5 to 20 text labels for each image. Note that the image background was subtracted and the images with text printed on products were removed to eliminate bias. Overall 45,609 text labels for 4,093 images were collected. Data statistics are reported in Table 1 and Appendix.

As in prior work (Conway III et al., 2014; Brysbaert et al., 2014), human ratings are a reasonable proxy of ground

[2]Car images are licensed by EVOX (EVOX).

truth representing human intuition and are used to verify subjective measures. Hence, in Phase 2 (Fig. 1, we asked a separate group of participants (Prolific.co, 2022) to rate the complexity of the previously elicited text labels, allowing for the estimation of ground truth that could be used to evaluate each model.

Raters were not explicitly asked to rate the "complexity" of the words, since pilot results indicated that participants were confused by the ambiguity of determining whether a word was "complex" and often shifted how they interpreted the definition of "complex". Thus, raters were asked to instead determine whether product images provided sufficient context to interpret the generated words.

For each product image and an associated set of 5-20 text labels, 1,210 raters assessed whether "this word needs more context beyond what is in the image to explain or understand" using a unipolar, 5-point Likert-style scale: "Do not agree at all" (0) to "Strongly agree" (4), which are converted to the complexity ratings (0-4). At least three (but not all) raters rated each label.

4. Approximating Cognitive Complexity

Because cognitive complexity is a broad psychological construct that likely includes multiple sub-constructs, we triangulate it using the five models described below. We hypothesized that these relevant, quantifiable constructs could be used together to approximate the complexity. We list hypotheses for each construct in italics.

Visibility as Complexity. Visual features that can be recognized at a glance, e.g., a primary color, may drive less complex cognitive processes while less visually salient features may be more complex. We therefore leveraged text used to visually describe product images as a reference corpus to represent what people can intuitively see and recognize. A reference corpus for a product image is generated by a transformer-based image-to-text generation model OFA (Wang et al., 2022). It is trained by visual description datasets such as MSCOCO (Chen et al., 2015) and generates multiple caption-like outputs using beam search to expand the most possible words with variability. Because annotators for MSCOCO were asked to describe all the important parts of the scene, and not to describe things beyond the scene, the training datasets and approach force the generated descriptions to be visually perceived intuitively. We formulate the complexity score of a text label l_m elicited by an image I_n, namely the complexity based on visibility (V) construct, as $\Theta_v(l_m, I_n) = 1 - f(l_m, D_n)$, where visibility $f(\cdot)$ means the number of sentences containing l_m divided by the total number of sentences in the reference visual descriptions D_n of I_n. Negating $f(\cdot)$ inverts its relationship with the score Θ_v, *i.e. the lower visibility, the higher complexity*.

Semantics as Complexity. Semantic intuition also influences complexity of elicited language such as the comprehension of language referring to the target product image, which is relevant to the semantic relationships between product images and elicited language. We leverage a text and image embedding space based on CLIP (Radford et al., 2021) to measure the semantic relationship. Given a product image I_n and a text label l_m, we encode the input to the joint embedding space and compute the cosine similarity. The complexity score is formulated as $\Theta_s(l_m, I_n) = 1 - \cos(E_T(l_m), E_X(I_n))$, where E_T and E_X are the text and image encoders, repectively. *A text label is assigned with a higher complexity Θ_s if it is more semantically distant to the image stimulus (less similarity)*.

Uniqueness as Complexity. In addition to word features (e.g., length), we consider corpus-based word frequency (Maddela & Xu, 2018) to reflect word complexity. We sort the text labels into multiple corpora by the product categories of their corresponding product image stimulus. We then calculate the term frequency of a text label in its corpora. *The complexity Θ_u of a text label is higher if it is less frequently used to describe product images of the same category, thus being more unique*.

Concreteness as Complexity. We use concreteness ratings collected by (Brysbaert et al., 2014) for 40,000 English lemmas as a lookup table B to derive the concreteness c_m of each text label l_m elicited by the product images. The complexity score is formulated as $\Theta_c(l_m) = b - c_m$, where b is set to the maximum score in B. The concreteness c_m has an inverse relationship with Θ_c, i.e., *the lower concreteness, the higher complexity*.

5. Experiments

Our experiments are designed to investigate (1) how well the cognitive complexity approximated by the models aligns with human judgments and (2) whether the models are complementary to each other. In Table 2, we use the dataset introduced in Sec. 3 for evaluation. We also present results for a subset of this dataset in Table 3, where product-label pairs obtained higher agreement (correlation > 0.75) among multiple human raters, to reveal the influence of agreement on the results. Due to space limitations, the results for all 14 categories are reported in Appendix.

5.1. How Well Do the Models Align with Humans?

While human ratings do not comprehensively represent the ground truth, they can be used to test the alignment of a model with human thoughts. Here, we calculate the Spearman correlation between the complexity rated by humans and as measured by models based on different constructs to verify our hypotheses in Sec. 4.

Models	furn.	deco.	car	bed	all
Θ_v	.230	**.247**	**.281**	.239	.225
Θ_s	**.247**	.219	.252	**.247**	**.228**
Θ_u	.175	.177	.207	.230	.168
Θ_c	.180	.178	.171	.208	.160
$\Theta_{v,s}$	.258	.252	.266	.259	.250
$\Theta_{v,s,u}$	.267	.270	.272	.276	.261
$\Theta_{v,s,u,c}$	**.273**	**.276**	**.316**	**.294**	**.271**

Table 2. Correlation with human ratings of complexity. Results for all 14 categories in Table 6 and 7 in Appendix.

Models	furn.	deco.	car	bed	all
Θ_v	.391	**.441**	.471	.513	**.418**
Θ_s	**.425**	.375	**.497**	**.520**	.414
Θ_u	.293	.287	.355	.466	.260
Θ_c	.277	.278	.190	.404	.267
$\Theta_{v,s}$	.436	.443	.509	.550	.433
$\Theta_{v,s,u}$	**.459**	.458	.531	.611	.456
$\Theta_{v,s,u,c}$	.454	**.462**	**.543**	**.614**	**.472**

Table 3. Correlation with human ratings of complexity (the subset with high agreement among raters). Results for all 14 categories in Table 8 and 9 in Appendix.

The results in Table 2 and 3 suggest that models of visibility, semantics, uniqueness, and concreteness are correlated with human judgments as we hypothesize (hypotheses are in italics in Sec. 4) The visibility and semantics constructs are top performers, perhaps because the two models use both text and image for the measurement, while other models measure complexity purely using a label or its metadata. This suggests that an image stimulus itself might provide meaningful information for computing cognitive complexity.

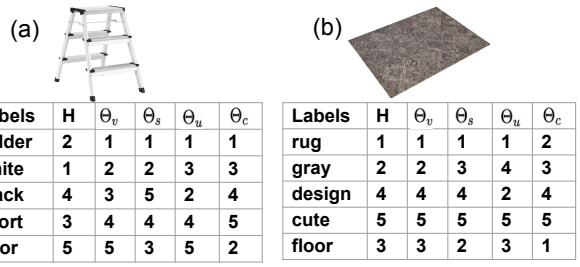
Data also suggest that the alignment between each construct and human judgment varies across different product categories. For example, the visibility construct best aligns with humans for decor while semantics alignment is better aligned for furniture. As shown in Fig. 2, the proposed constructs can help dive into different facets for efficient sense-making. A combination of multiple constructs may help generalize the measurement to various products. Note that significant individual differences (Hakimi et al., 2023) suggest variance in the driving psychological factors supporting cognitive complexity. As such, high correlations with any sub-construct are unlikely.

5.2. Are the Models Complementary?

We want to verify whether the models of constructs offer complementary information for measuring cognitive complexity. First, we compute the partial correlation between

Constructs	Θ_v	Θ_s	Θ_u	Θ_c
visibility (Θ_v)	1	-	-	-
semantics (Θ_s)	.343	1	-	-
uniqueness (Θ_u)	.352	.070	1	-
concreteness (Θ_c)	.283	.168	.040	1

Table 4. Partial Correlation between the models for measuring complexity based on different constructs



Labels	H	Θ_v	Θ_s	Θ_u	Θ_c
ladder	2	1	1	1	1
white	1	2	2	3	3
black	4	3	5	2	4
short	3	4	4	4	5
floor	5	5	3	5	2

Labels	H	Θ_v	Θ_s	Θ_u	Θ_c
rug	1	1	1	1	2
gray	2	2	3	4	3
design	4	4	4	2	4
cute	5	5	5	5	5
floor	3	3	2	3	1

Figure 2. Ranking cognitive complexity (low to high) of text labels elicited by product images, rated by human (H) and measured by the models of individual constructs

every two constructs while the rest of the constructs are the controlling variables. Table 4 shows that some constructs have very low correlations such as uniqueness vs. concreteness. Visibility Θ_v that is measured based on what people visually perceived has higher correlations with other constructs as all these models consider perceptual information in different ways. However, all pairs of constructs are far from identical and with each construct providing unique information.

The bottom part of Table 3 presents the correlations between combinations of top-performing models of constructs and human judgments. A combination takes the weighted sum of scores predicted by each construct, where the weight is empirically determined by their individual correlation with human judgment. In most cases, the listed combinations align with humans better than individual constructs. Adding more positively correlated constructs in combination usually increases the correlation with human judgments, probably because parts of the non-overlapped information are complementary. The findings are consistent with experiments based on the whole dataset in Table 2.

6. Discussion and Conclusion

Human cognition is complex and diverse. While LLMs have reached or surpassed human in several knowledge-based tasks, there remain many challenges to approximating the kind of human cognitive complexity that characterizes subjective experience and behaviors such as elicited responses to products. As a further step to triangulate complexity computationally, we propose grounding cognitive complexity

in the language elicited by product images and characterize the constructs that might be predictive of this new measure. The data elicitation for cognitive complexity is more complicated than typical image annotations. First, it requires sampling a diverse population of participants, in this case, representing consumers. Second, it increases the labeling effort from the size of language data $O(N)$ to the size of pairs of language and image data $O(N \times M)$. Hence, we foresee the cost of obtaining huge training data is nontrivial and attempt to investigate whether the foundational models, e.g., OFA (Wang et al., 2022), CLIP (Radford et al., 2021), help compute cognitive complexity and generalize the measure to domains that are essentially different from the presented elicitation framework. Future work includes fine-tuning strategies to optimize the combination of constructs for better prediction power and broader use of measuring cognitive complexity in feedback to consumer products or other visual stimuli, and in responses synthesized by LLMs.

References

Amazon Mechanical Turk. *Amazon Mechanical Turk*. Amazon.com, Inc., New Jersey, US, 2022. URL https://www.mturk.com.

Brysbaert, M., Warriner, A., and Kuperman, V. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911, 2014.

Cavanagh, P. Visual cognition. *Vision research*, 51(13): 1538–1551, 2011.

Charbonnier, J. and Wartena, C. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pp. 176–187, Gothenburg, Sweden, May 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-0415. URL https://aclanthology.org/W19-0415.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, L. C. Microsoft coco captions: Data collection and evaluation server, 2015. URL https://arxiv.org/abs/1504.00325.

Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Yago Vicente, T. F., Dideriksen, T., Arora, H., Guillaumin, M., and Malik, J. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

Conway III, L. G., Conway, K. R., Gornick, L. J., and Houck, S. C. Automated integrative complexity. *Political Psychology*, 35(5):603–24, 2014.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372.

EVOX. *EVOX Images*. URL https://www.evoximages.com/.

Flekova, L. and Gurevych, I. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1805–1816, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1208. URL https://aclanthology.org/D15-1208.

Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V., and Gold, J. I. A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, 2: 213–224, 2018.

Hakimi, S., Chen, Y.-Y., Van, M., Carter, S., Sumner, E., Bravo, N., Murakami, K., Zhang, Y., Wu, C., and Klenk, M. Machine learning-based measure of cognitive complexity explains variance in economic preferences. In *Proceedings of the 45th Annual Conference of the Cognitive Science*, 2023.

Hessel, J., Mimno, D., and Lee, L. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2194–2205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1199. URL https://aclanthology.org/N18-1199.

Iliev, R. and Axelrod, R. The paradox of abstraction: Precision versus concreteness. *J Psycholinguist Res.*, 46(3): 715–729, 2016.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 2611–2624, 2020. URL https://dl.acm.org/doi/10.5555/3495724.3495944.

Maddela, M. and Xu, W. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3749–3760, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1410. URL https://aclanthology.org/D18-1410.

Marr, D. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Prolific.co. *Prolific*. Prolific Academic Ltd, London, UK, 2022. URL https://www.prolific.co.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ramakrishnan, K. and Deniz, F. Non-complementarity of information in word-embedding and brain representations in distinguishing between concrete and abstract words. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 1–11, 2021.

Veselovsky, V., Ribeiro, M. H., Cozzolino, P., Gordon, A., Rothschild, D., and West, R. Prevalence and prevention of large language model use in crowd work, 2023.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.

Wartena, C. On the geometry of concreteness. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 204–212, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.21. URL https://aclanthology.org/2022.repl4nlp-1.21.

Yang, Y., Panagopoulou, A., Apidianaki, M., Yatskar, M., and Callison-Burch, C. Visualizing the obvious: A concreteness-based ensemble model for noun property prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 638–655, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.45.

Zhang, M. and Qian, T. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3540–3549, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.286. URL https://aclanthology.org/2020.emnlp-main.286.

A. Appendix

A.1. Dataset

Data statistics of a full list of product categories are reported in Table 5.

Category	products	labels	vocab.	complexity
furn.	2,174	23,210	3,278	1.75 ±0.94
deco.	1,054	10,481	2,664	1.77 ±0.95
car	181	2,925	741	1.90 ±0.95
bed	239	2,686	871	1.76 ±0.96
outd.	155	2,167	880	1.77 ±0.94
stor.	130	1,560	693	1.80 ±0.93
acce.	29	452	282	1.89 ±0.98
home.	23	382	278	1.78 ±0.98
heat.	21	377	184	1.69 ±1.00
pers.	23	344	227	1.93 ±0.90
spor.	23	324	161	1.74 ±0.94
elec.	17	310	193	1.90 ±0.95
kitc.	13	241	166	1.80 ±0.92
offi.	11	150	119	1.71 ±0.98
all	4,093	45,609	6,583	1.77 ±0.94

Table 5. Dataset counts for the products, labels and vocabulary size. Key: furn:furniture; deco:decor; car:car, bed:bed bath; outd:outdoor; stor:storage organization; acce:accessories; home:home improvement; heat:heating cooling; pers:personal electronic devices; spor:sports fitness; elec:home electronics; kitc:kitchen; offi:office

A.2. Experiment Results

The experiments results for a full list of product categories are presented in Table 6, 7, 8 and 9.

Models	furn.	deco.	car	bed	outd.	stor.	acce.	home	heat.
Θ_v	.230	.247	.281	.239	.166	.177	.205	.297	.301
Θ_s	.247	.219	.252	.247	.227	.176	.267	.238	.264
Θ_r	.042	.060	.131	.055	.038	.007	.027	-.109	.159
Θ_u	.175	.177	.207	.230	.174	.131	.136	.247	.290
Θ_c	.180	.178	.171	.208	.163	.121	.129	.116	.218
$\Theta_{v,s}$	.258	.252	.266	.259	.221	.185	.279	.295	.305
$\Theta_{v,s,u}$	.267	.270	.272	.276	.236	.204	.277	.376	.267
$\Theta_{v,s,u,c}$	.273	.276	.316	.294	.242	.199	.275	.294	.347

Table 6. Correlation with human ratings of complexity (the whole dataset)

Models	pers.	spor.	elec.	kitc.	offi.	all
Θ_v	.236	.139	.164	.242	.224	.225
Θ_s	.291	.128	.126	.250	.258	.228
Θ_r	.100	-.029	.041	.089	.073	.049
Θ_u	.173	.090	.061	.155	.109	.168
Θ_c	.160	-.047	.122	.170	.354	.160
$\Theta_{v,s}$	.315	.136	.127	.275	.329	.250
$\Theta_{v,s,u}$	.316	.153	.129	.278	.329	.261
$\Theta_{v,s,u,c}$	.324	.132	.152	.272	.397	.271

Table 7. (Continued with Table 6) Correlation with human ratings of complexity (the whole dataset)

Models	furn.	deco.	car	bed	outd.	stor.	acce.	home	heat.
Θ_v	.391	.441	.471	.513	.222	.381	.128	.499	.425
Θ_s	.425	.375	.497	.520	.178	.390	.471	.343	.384
Θ_r	.064	.129	.119	.031	-.032	-.135	.051	-.192	-.033
Θ_u	.293	.287	.355	.466	.164	.356	.097	.258	.234
Θ_c	.277	.278	.190	.404	.101	.337	-.040	.388	-.058
$\Theta_{v,s}$	.436	.443	.509	.550	.219	.416	.413	.425	.415
$\Theta_{v,s,u}$	.459	.458	.531	.611	.234	.494	.367	.443	.417
$\Theta_{v,s,u,c}$	.454	.462	.543	.614	.225	.514	.280	.475	.371

Table 8. Correlation with human ratings of complexity (the subset with high rater agreement, correlation > 0.75)

Models	pers.	spor.	elec.	kitc.	offi.	all
Θ_v	.447	.657	-.187	.724	.740	.418
Θ_s	.447	.515	.004	.630	.621	.414
Θ_r	.648	-.031	.220	.267	-.167	.067
Θ_u	-.787	.501	-.260	.464	.535	.260
Θ_c	.894	-.240	.017	.425	.774	.267
$\Theta_{v,s}$	.447	.581	-.25	.677	.784	.433
$\Theta_{v,s,u}$	.447	.713	-.242	.662	.784	.456
$\Theta_{v,s,u,c}$	.783	.591	-.180	.652	.827	.472

Table 9. (Continued with Table 8) Correlation with human ratings of complexity (the subset with high rater agreement, correlation > 0.75)