The Face of a Character Called Gmork

Marco Bragoni (marcobragoni19@gmail.com)
Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2
5037AB Tilburg, The Netherlands

Giovanni Cassani (g.cassani@tilburguniversity.edu)
Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2
5037AB Tilburg, The Netherlands

Abstract

We used cross-modal generative AI models, which rely on the Contrastive Language-Image Pretraining (CLIP) encoder, to generate portraits of fictional characters based on their names. We then studied to what extent image generation captures names' gender and age connotations when information from linguistic distribution is rich and informative (talking names, e.g., *Bolt*), present but possibly uninformative (real names, e.g., *John*), and absent (made-up names, e.g., *Arobynn*). Three pre-trained Computer Vision classifiers for each attribute exhibit reliable agreement in classifying generated images, also for made-up names. We further show a robust correlation between the classifiers' confidence in detecting an attribute and the ratings provided by participants in an online survey about how suitable each name is for characters bearing a certain attribute. These models and their learning strategies can shed light on mechanisms that support human learning of non-arbitrary form-meaning mappings.

Keywords: Form-Meaning mappings; Generative AI; Large Language Models; Multi-modal Semantics.

Introduction

Generative AI models have recently shown impressive abilities in many tasks (e.g., text-based image generation, text generation, question answering) and several studies have started to probe the representations these models learn. For example, Bianchi et al. (2023) found that text-based image generation models amplify racial and gender stereotypes when asked to generate images from prompts describing occupations, beyond gender disparities in the workforce. Furthermore, Cai, Haslett, Duan, Wang, and Pickering (2023) found that when given new names constructed to resemble male or female names, language generation systems chose the gender pronoun that aligned with the phonological composition of the name, suggesting that these models pick up subtle sublexical regularities in the training data.

In this work, we use cross-modal generative AI models to investigate systematic form-meaning mappings such as the *bouba-kiki* effect (Köhler, 1929), where participants consistently pair pseudo-words to images showing a certain semantic attribute. We investigate names of fictional characters and study whether such models capture relevant semantic properties like perceived gender and age given the name alone. Crucially, we investigate three different naming devices, which differ in the degree to which they leverage established semantic connotations (Joosse, Kuscu, & Cassani, in press). On one end of the spectrum, *talking* names[1], such as *Blackberry*, leverage existing words to possibly convey information about a character. On the other end of the spectrum, *made-up* names, such as *Gmork*, cannot be interpreted by relying on experience with the name, since it was never found in context and does not have a specific meaning[2]. Yet, studies on the interpretation of (made-up) names show that people have consistent intuitions based on the sounds and letters in the name (Elsen, 2017; Pitcher, Mesoudi, & McElligott, 2013). A third type of names, *real* given names such as *Adelaide*, sits in-between the two poles. These names do have lexical distributions which can sometimes establish connotations through antonomasia, like *Karen* for obnoxious, and have distinctive usage patterns on a relevant dimension like gender, with certain names consistently used for men and other for women (Cassidy, Kelly, & Sharoni, 1999). At the same time, however, these names lack a precise meaning, and yet their phonological make-up affects semantic judgments in participants (Sidhu, Deschamps, Bourdage, & Pexman, 2019; Sidhu, Pexman, & Saint-Aubin, 2016). Therefore, when interpreting such names, people may combine experiences and sensitivity to sound patterns. By feeding these different names to cross-modal generative AI models which can handle out-of-vocabulary words, we aim to explore the models' ability to reflect the associations people exhibit.

Our study further draws inspiration from Davis, Morrow, and Lupyan (2019), where participants were asked to draw pictures given a set of pseudo-words, and then to guess which pseudo-word inspired a specific drawing. Participants were consistent in depicting certain pseudo-words: for example, drawings of *horgous* were consistently large while drawings of *keex* were tiny. Moreover, when guessing, participants performed more accurately than expected under a random baseline, showing that even in the lack of conventional semantic

[1] We choose to use the label *talking* names for this class of names to emphasise that these names, the name itself typically communicates certain aspects of the character.

[2] The names we consider were first collected and analyzed by Joosse et al. (in press). In this study, names of fictional characters were selected to first study whether the intuition people had about names on different attributes aligned with the choices made by the authors who picked each name for a specific character with specific attributes. For this reason, made-up names coming from literary works were preferred to names made-up for the specific purpose of the study. We address the possible confounds this choice brings in the discussion.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

connotations, people consistently mapped strings to semantic attributes across modalities (language and vision, in this case). We submit a similar task to cross-modal generative AI models, using the names of fictional characters as input, and investigate the degree to which generated pictures encode perceived gender and age (Bianchi et al., 2023). Following on evidence from studies on sound symbolism (Ramachandran & Hubbard, 2001; Lockwood & Dingemanse, 2015; Ćwiek et al., 2022) and work which started investigating the semantic intuitions elicited by pseudo-words (Sabbatino, Troiano, Schweitzer, & Klinger, 2022; Cassani, Chuang, & Baayen, 2020), we hypothesize that the representations the AI models derive from names capture the target attributes, even in the lack of specific cross-modal distributions for the lexical form (as expected for made-up names). Using the same names, Joosse et al. (in press) showed that the embedding space derived from a representative corpus of English using a Distributional Semantic Model with sub-word information (Bojanowski, Grave, Joulin, & Mikolov, 2017) reflects systematic form-meaning mappings for perceived age and gender for all three name types. Building on these results, we investigate here whether the same pattern is found in cross-modal semantic models, under the hypothesis that when asked to rate a name, participants relied on cross-modal correspondences between (sub-)lexical distributions and visual attributes, which influenced their rating of how fit a name is for a character having a certain attribute. To this end, we rely on generative models which leverage the Contrastive Language-Image Pretraining (CLIP) encoder (Radford et al., 2021), which is trained on a vast dataset of image-caption pairs, and learns to encode strings and images in the same representational space, making it possible to quantify the semantic relatedness between a text and an image.

Computational models such as CLIP thus act as a possible theory of how people form cross-modal associations that generalize to completely novel strings (Sidhu & Pexman, 2018). If we can explain people's intuitions using associations derived from a computational model which only sees image-caption pairs, it would mean that systematic form-meaning mappings can be learned from situated language data alone (Sidhu & Pexman, 2018). If this generalizes to entirely novel names, we can further probe to what extent an account grounded solely in the presence of informative co-occurrences between (sub-)lexical patterns and visual features in the environment can explain people's intuitions on relevant semantic dimensions.

Crucially, after generating images, we use pre-trained Computer Vision (CV) classifiers to tag images on perceived gender and age. In this way, we test whether generative models reliably capture semantic attributes such that generated images reflect those attributes reliably enough that independent classifiers trained to recognize the same attributes in real images can recognize them in generated images. We argue that this set-up offers a more stringent test than asking new participants to rate generated images, since participants may rely on very different cues to detect perceived gender and age, and adjust to the task more flexibly. CV classifiers, on the contrary, are expected to rely on how the attributes are encoded in the training set, which consists of human faces. Evidence of CV classifiers' success in detecting the same attributes in images generated from names (including *made-up* ones) would suggest that the underlying generative model does encode those semantic attributes reliably. Finally, by controlling for language-based associations derived from distributional semantic models (Joosse et al., in press), we assess whether cross-modal associations play a unique role when inferring semantic attributes from names alone.

Materials & Methods

Fictional Characters Dataset

We use 179 characters' names from fan fiction, children, and young-adult literature collected and manually tagged by Joosse et al. (in press) based on the character's gender in the original story[3]. 119 names were also tagged according to whether the character was portrayed as young or old[4]. Names were balanced according to attribute combinations, such that there are approximately the same number of young male, young female, old male, and old female names for talking, real, and made-up names to ensure the class distribution in the input names is balanced for both attributes. The names were presented to 300 participants through an online survey (protocol approved by the Ethics Board of Tilburg University, protocol ID: 2020.203), asking them to drag a slider bar (anchored between -50 and 50) to indicate how suitable each name would be for a male/female or young/old character. Raters were thus asked to provide a continuous rating about the suitability of a name for a character with certain attributes. No specific instructions were provided on what to consider old or young to avoid biasing ratings.

Image Generation

We generated 20 images for each character name and each image generator (either Stable Diffusion[5] or VQGAN-CLIP (Crowson et al., 2022)), using the prompt "a face of a character called *<name>*". Both image generators leverage CLIP Radford et al. (2021) to encode textual inputs and guide generation. Using two generators sharing the same encoder ensures that any observed pattern does not depend on the specific generator. Figure 1 shows six generated images.

CLIP consists of an image (Dosovitskiy et al., 2021) and a text encoder (Vaswani et al., 2017), trained jointly on 400M matching sentence-image pairs. During training, CLIP learns

[3]See the original study by Joosse et al. (in press) for further details about data collection and tagging. Data and code to reproduce our analyses are available here: `https://github.com/Braga19/ClipSoundSymbolism`

[4]Age could not be determined for some characters due to indeterminacy in the characterization, since in many cases the authors did not specify the age of a character. We also only considered characters who do not age during the story, to ensure names were chosen to characterize characters at a specific time of their life.

[5]`https://github.com/CompVis/stable-diffusion`

Figure 1: Generated images using Stable Diffusion (SD) and VQGAN+CLIP (VC) given three target names, one real, one made-up, one talking.

a multi-modal latent embedding representation where matching sentence-image pairs have a higher cosine similarity. The text encoder pre-processes the input sentences using byte-pair encoding (BPE) tokenization, enabling the model to encode out-of-vocabulary (OOV) words by dividing them in subword tokens. This feature is crucial to handle infrequent and novel words in input sentences (Sennrich, Haddow, & Birch, 2015), but also our made-up names (e.g., *Gmork* is tokenized as "g", "mor", and "k").

VQGAN+CLIP (VC) uses CLIP to generate images leveraging VQGAN (Vector Quantized Generative Adversarial Network, (Esser, Rombach, & Ommer, 2021)), a generative model that combines the transformer architecture with a Generative Adversarial Network (GAN) and leverages vector quantization (VQ). Starting from a random mask, at each iteration CLIP embeds the generated image and assesses how similar it is to the text query in its latent space. Using gradient ascent, the image generator is prompted to generate an image which falls closer to the text query (Crowson et al., 2022). We leverage the VQGAN generator trained on the *faceshq* dataset and use CLIP *ViT-B/32*, saving images after 200 iterations.

Stable Diffusion (SD) consists of three major components: CLIP *ViT-L/14*, U-Net (Ronneberger, Fischer, & Brox, 2015), and an image decoder. Through an iterative process, U-Net deconstructs the noise starting from a randomly initialized image and generates a new latent array that more closely represents the input text as embedded by CLIP (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2021). Thus, CLIP conditions the final output which is generated by the image decoder. We use Stable Diffusion *v1-4*.

Image Classification

We rely on a classification task to probe the presence of systematic cross-modal form-meaning mappings in CLIP's latent space. We input each generated image to pre-trained, off-the-shelves CV classifiers trained to detect perceived gender and age in faces using different datasets and architectures. From the HuggingFace *Transformers* library (Wolf et al., 2020), we selected *rizvan* (`https://huggingface.co/rizvandwiki/gender-classification`), *crangana-gen* (`https://huggingface.co/crangana/trained-gender`), and *leilab* (`https://huggingface.co/Leilab/gender_class`) to predict perceived gender: each model outputs the probability of an image being classified as male or female. To predict age, we chose *nateraw* (`https://huggingface.co/nateraw/vit-age-classifier`), *ibombSwin* (`https://huggingface.co/ibombonato/swin-age-classifier`), and *crangana-age* (`https://huggingface.co/crangana/trained-age`). These classifiers predict the probability of the image fitting different age bins, which we aggregate as follows: 0-30 for YOUNG and 30+ for OLD, by summing probabilities allotted to each original bin. The threshold was chosen to harmonize the predictions of different classifiers.

Data analysis

Inter-Annotator Agreement Since we are interested in the robustness of the encoded patterns rather than in which classifier performs best, we compute the inter-annotator agreement (IAA) for each target attribute, considering the three CV classifiers for each attribute (perceived gender: *rizvan*, *crangana-gen*, *leilab*; age: *nateraw*, *ibombSwin*, *crangana-age*) as the annotators. We use the Fleiss'κ coefficient (Fleiss, 1971), which yields a number between -1 and 1, where a high positive κ indicates that different classifiers consistently recognize the target attribute in generated images.

Regression models

First, we obtained the output probability of each attribute (MALE/FEMALE, OLD/YOUNG) for each generated image (generated using VC or SD) outputted by each classifier (perceived gender: *rizvan*, *crangana-gen*, *leilab*; age: *nateraw*, *ibombSwin*, *crangana-age*). Then, we averaged the probability of the positive class (FEMALE and OLD, those paired with positive values in the slider bar from the behavioral task, where a value of 50 meant a name was deemed very suitable for a female character on the gender semantic differential or for an old character on the age semantic differential) for each name. It is worth stressing that human raters did not rated generated images, but rather names alone. We filtered individual behavioral ratings higher than the 9th decile and lower than the 1st decile for each name, and averaged (see Joosse et al. (in press) for additional details on data pre-processing).

We then fitted a baseline linear model predicting the average behavioral rating per name as a function of the name type (real, talking, made-up) and a measure of semantic polarization derived from a custom FastText model (Bojanowski et al., 2017) trained with 2- to 5-grams on the Corpus of American English (CoCA, Davies (2010)). This measure modifies the Word Embedding Association Test (Caliskan, Bryson, & Narayanan, 2017) and was introduced by Joosse et al. (in press). First, we embedded each name using the *fastText* model. Then, we computed the average cosine similarity be-

Table 1: The Fleiss κ coefficient across attributes (Gender; Age), generator (Stable Diffusion (SD); VQGAN+CLIP (VC)), and name type (Real; Made-up; Talking).

	Gender		Age	
	SD	VC	SD	VC
Real	0.72	0.76	0.46	0.44
Made-up	0.59	0.62	0.45	0.41
Talking	0.52	0.51	0.5	0.38

tween each name's *fastText* embedding and a pool of words A encoding attribute X (e.g., *male, masculine, man, boy, men, he* for MALE) and the average cosine similarity between the same name and a pool of words B encoding attribute Y opposite to X (e.g., *female, feminine, woman, girl, women, she* for FEMALE), to reproduce the semantic differentials on which participants in the behavioral experiment rated names[6]. Finally, we took the difference between the average cosine similarities between each name and attributes in B and in A: positive values indicate the name is closer in *fastText*'s representational space to words pertaining to FEMALE and OLD. This text-based metric reliably predicted behavioral ratings in a previous experiment (Joosse et al., in press): names that are embedded relatively closer to words encoding the attribute *female*, were rated to better fit a female character in the behavioral experiment. We thus aim to establish whether cross-modal patterns predict behavioral ratings beyond text-based associations. To this end, we added the average probability of images generated from a name using VC or SD being tagged as female (p_{female}) or old (p_{old}) by the different CV classifiers to the baseline statistical models, comparing models using AIC and assessing effect sizes using η^2.

To sum up, we start from a baseline statistical model which predicts behavioral ratings as a function of name type and text-based associations extracted for all name types using a custom *fastText* model: we fit one baseline statistical model to predict behavioral ratings on gender, and one to predict behavioral ratings on age. To each baseline model, we then add the average probability each CV classifier assigns to the 20 images generated for each name to depict a female/old character respectively. We fit separate statistical models for each combination of target attribute (age, perceived gender), image generator (SD, VC), and CV classifier (perceived gender: *rizvan*, *crangana-gen*, *leilab*; age: *nateraw*, *ibombSwin*, *crangana-age*), for a total of 12 regression models. Importantly, the focus is not on determining the best one, but on establishing whether they all exhibit coherent patterns.

Results

Table 1 provides the IAA results. The Fleiss κ is consistent across the two image generators. The highest agreement among classifiers is reported for perceived gender: predictably, real names show the highest consistency ($\kappa_{VC} = 0.76$ and $\kappa_{SD} = 0.72$), but the score is robust for made-up and talking names as well (κ always higher than 0.5). Interestingly, this pattern reflects the consistency in behavioral ratings, were participants agreed the most on the likely gender of a character bearing a real name, but also shown consistent patterns for other attributes and name types (see Joosse et al. (in press) for detailed patterns). The same pattern is reported for age, although the score is consistently lower, with κ between 0.38 and 0.5. Thus, we report moderate to strong IAA across three classifiers per attribute (perceived gender: *rizvan*, *crangana-gen*, *leilab*; age: *nateraw*, *ibombSwin*, *crangana-age*) on perceived gender and age, for all name types.

Table 2: Linear regressions fitted to predict behavioral ratings, fitted separately for each combination of attribute (gender; age), generator (Stable Diffusion, SD; VQGAN+CLIP, VC), and classifier. ΔAIC: change in fit between a baseline model considering only name type and text-based associations, and a model also including cross-modal information as the probability each classifier attributes to images generated for a name of being female/old (larger scores indicate the more complex model improves more over the simpler model); η^2_t: partial effect size of the text-based predictor; η^2_{cm}: partial effect size of the cross-modal predictor (the average probability that images generated from a name are tagged as female or old); r^2: model fit (adjusted coefficient).

Generator	Classifier	ΔAIC	η^2_t	η^2_{cm}	r^2
Gender					
SD	leilab	60.817	0.75	0.30	0.77
	rizvan	70.728	0.76	0.33	0.79
	crangana-gen	54.611	0.75	0.27	0.77
VC	leilab	26.245	0.72	0.15	0.73
	rizvan	24.738	0.71	0.14	0.72
	crangana-gen	19.383	0.71	0.11	0.72
Age					
SD	ibombSwin	39.674	0.18	0.30	0.40
	crangana-age	39.254	0.18	0.29	0.40
	nateraw	35.522	0.17	0.27	0.38
VC	ibombSwin	8.208	0.14	0.08	0.22
	crangana-age	19.736	0.15	0.17	0.30
	nateraw	34.660	0.17	0.27	0.38

Table 2 summarizes the AIC improvement over the baseline statistical model[7]. In the baseline statistical model, we predicted participants' ratings about the fit of a name for a character with a certain attribute using name type and the text-based semantic differential (Joosse et al., in press). This baseline statistical model was compared with other models which included the average probability that images generated for a name depicted a character with a certain attribute: p_{female} in-

[6]For the attribute *young*, we opted for *young, youth, child, boy, girl, baby*; for the attribute *old*, we chose *old, elderly, grandparent, grandfather, grandmother, adult*

[7]No interaction term improved over a model simply including a linear combination of the predictors.

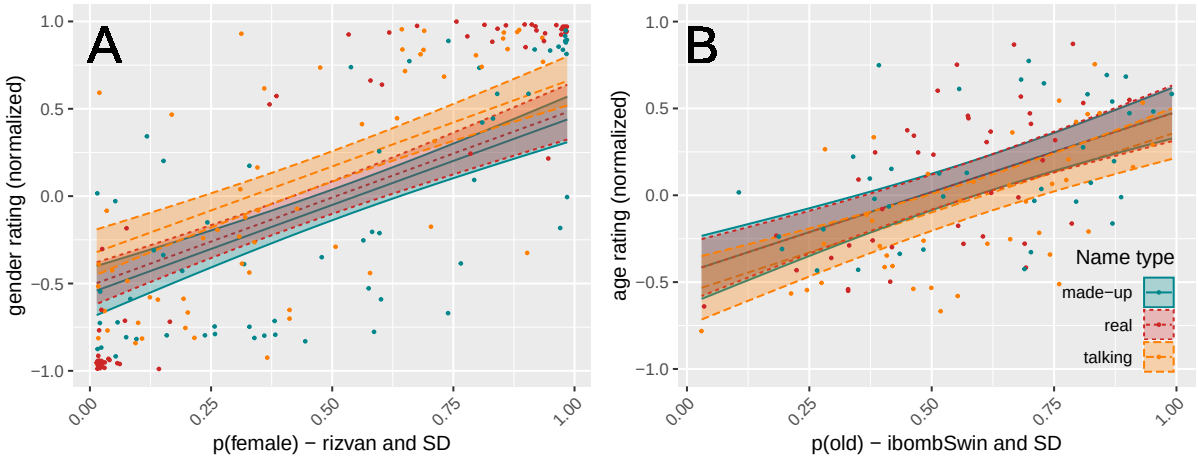

Figure 2: Effect of the average probability of images generated from a name of being tagged as female on gender behavioral ratings (A) and of being tagged as old on age behavioral ratings (B), while partialing out the effect of text-based associations on each attribute. The effect is plotted separately by name type. The plot shows the model fitted on images generated using Stable Diffusion and probabilities provided by the rizvan classifier (A) and the ibombSwin classifier (B). Points show actual data, while lines show predicted values (shaded areas indicate 95% Confidence Intervals).

dicates the average probability that the 20 images generated from a name feature a face tagged as female by the chosen CV classifiers; p_{old} indicates the average probability that the 20 images generated from a name feature a face tagged as old by the chosen CV classifiers (a probability is computed separately for each classifier). Positive AIC scores indicate that the more complex model including the probability that generated images as tagged as featuring a female/old character improves over the simpler model which only includes name type and the text-based differential. We further provide the η^2 of the text-based differential and cross-modal predictors and the adjusted r^2 as a measure of fit. For example, the first row in Table 2 indicates that, when predicting participants ratings of how well a name fits a female character the inclusion of the average probability that the *leilab* gender classifier assigns to the images generated by Stable Diffusion (SD) of depicting a female face, the model fit improves by 60.817 AIC units over a model which predicts the same dependent variable only considering name type (real, made-up, talking) and the text.based differential derived from *fastText*. The more complex model accounts for 0.77% of the variance. Finally, the text-based differential has an η^2 of 0.75 while $p(female)$ derived from the *leilab* classifier fed with images generated by Stable Diffusion (SD), i.e., the cross-modal predictor, has an η^2 of 0.3. The remaining rows summarize the comparison between the appropriate baseline model and each model including the cross-modal predictor derived from each combination of image generator (VC or SD) and CV classifier.

We see that including the cross-modal predictor improves model fit in predicting behavioral gender ratings across the board (ΔAIC generally above 20). The lack of a robust interaction with model type further suggests that this predictor reliably improves model fit for all names, regardless of the availability of rich co-occurrence information. We further see that the model fits the data well, with an r^2 consistently above 0.7. Finally, we see in Figure 2A that the effect is positive: names for which images generated with the SD generator yield a higher p_{female} according to the *rizvan* classifier tend to be rated as better fitting for female characters, and this applies equally to all name types (patterns for other classifiers and generators are similar).

The same general pattern is observed when considering age, despite some notable differences. First, r^2 values are consistently lower (the best adjusted r^2 for *ibombSwin* classifying images generated with SD is at 0.4), suggesting that predicting age ratings is a more challenging task. This is not surprising considering that age is a more graded and abstract attribute than perceived gender. Moreover, we observe that the effect size of p_{old} when using Stable Diffusion tends to be consistently stronger than that of the text-based predictor. Figure 2B further confirms that names for which images generated with the SD generator yield a higher p_{old} according to the *ibombSwin* classifier tend to be rated as better fitting for older characters in the behavioral task.

Discussion

In this study we fed names of fictional characters to two cross-modal generative AI models, Stable Diffusion (SD) and VQGAN+CLIP (VC), to generate images of a character from its name alone. Images were then fed to pre-trained, Computer Vision (CV) classifiers for perceived gender and age to detect whether they featured a male/female, young/old character. We computed Inter-Annotator Agreement (IAA) among classifiers for a same target attribute to quantify the robustness with which perceived gender and age were encoded by two generative models, Stable Diffusion (SD) and

VQGAN+CLIP (VC). Finally, we predicted participants' ratings about how fit each name is for a character having a certain attribute using each CV classifier's output probabilities, while controlling for name type and text-based associations between names and target attributes, to assess the unique contribution of cross-modal relations. Crucially, pre-trained CV classifiers were used to detect the target attributes in generated images: this ensures that the target attributes are construed in an extrinsically valid way.

We report moderate to strong IAA for both attributes on all name types, suggesting that CLIP - the encoder model at the core of both SD and VC - encodes perceived gender and age sufficiently well for different CV classifiers to come to similar conclusions about the attributes of the character portrayed. That it was easier to classify perceived gender than age fits with evidence that gender is a prominent semantic dimension in semantic models (Hollis & Westbury, 2016) and in names (Cassidy et al., 1999), as well as robustly connected to sublexical cues (Monaghan & Fletcher, 2019; Westbury, Hollis, Sidhu, & Pexman, 2018). The reported agreement for *made-up* names further suggests that CLIP encodes systematic form-meaning mappings for sublexical patterns. Moreover, since CLIP is only trained on image-caption pairs, our evidence suggests that the training data are replete with systematic associations between (sub-)lexical patterns and visual features, confirming that cross-modal systematic form-meaning mappings can be learned from situated language experience (Sidhu & Pexman, 2018).

We further observe a reliable relation between the strength with which classifiers detect an attribute in the images generated from a name and the perceived fit between a character name and an attribute in participants' ratings. This relation holds after controlling for text-based associations (Caliskan et al., 2017) derived from a sublexical distributional model which leverages n-grams and can thus embed OOV words (Bojanowski et al., 2017; Sabbatino et al., 2022; Joosse et al., in press), suggesting a unique contribution of cross-modal correspondences. We do observe differences across attributes though. Perceived gender tends to be predicted more strongly by text-based associations, in line with the important role of gender in language, even in a language which lacks explicit morphological markers for it. Age, on the contrary, shows a more prominent role of cross-modal cues. Overall, the observation that we can better predict participants' ratings when we consider how sublexical patterns relate with visual attributes suggests that their rating may have been at least partially influenced by cross-modal correspondences between (sub-)lexical patterns and visual attributes. This fits with embodied cognition theories (Barsalou, 1999) arguing that language understanding is mediated by multi-modal, embodied simulations. We add to this body of literature that such simulations may also apply to entirely novel strings and be mediated by systematic cross-modal mappings between language and the visual world (Davis et al., 2019).

A possible concern is that names were not made-up to the participants who rated them and to model. Certain names, like *Gmork*, do come from relatively well known books (*The neverending story* in the case of *Gmork*). However, the distribution of ratings suggests that the names were generally made-up to the participants: if names were rated based on a specific character, we would expect very consistent ratings for a name, whereas ratings for made-up names are only moderately correlated across participants (Joosse et al., in press). Moreover, excluding extreme ratings that deviate from the general distribution (higher than the 9th decile and lower than the 1st decile), should further avoid the risk of considering ratings which pertain to the character rather than the name, possibly due to a specific rater recognizing a name and rating it based on a character it refers to. As far as the underlying computational model is concerned, we cannot exclude the possibility that made-up names were known to CLIP since its training dataset is not available. However, made-up names are on average tokenized using more part words (2.18 ± 0.62) than real (1.28 ± 0.52) and talking (1.25 ± 0.47) names, despite a similar length in characters (made-up: 5.92 ± 1.48; real: 5.69 ± 1.39; talking: 5.3 ± 1.75). 7 made-up names were tokenized as a single unit, suggesting they were encountered as such in training: we re-ran statistical analyses excluding them and observed comparable patterns, suggesting that reported evidence was not driven by made-up names being recognized by the model. Moreover, made-up names did not feature in the training data for the *fastText* model, and despite this, the text-based semantic differential shows a robust relation with participants' ratings, strengthening the position that the chosen computational models did not leverage lexical co-occurrences but indeed relied on sublexical correspondences in the case of made-up names. Still, the issue of what counts as a pseudo-word for pre-trained neural models needs to be studied further, but we argue that our results show how the boundary between words and pseudo-words is more blurred than typically assumed (Gatti, Marelli, & Rinaldi, 2022).

In conclusion, our results show that cross-modal generative models encode biases beyond lexical items and social dimensions (Bianchi et al., 2023): even made-up names encode attributes like gender (Cai et al., 2023) and age in a sufficiently robust way for models and people to recognize them. Moreover, a model trained solely on image-caption pairs can support generation of images from made-up names that reflect the attributes people associate with the same names, suggesting that the input is replete with systematic correspondences between sublexical patterns and visual features. Beyond psycholinguistics, our findings are relevant for the social sciences, where vignette studies use names to analyze inequalities in hiring (Johfre, 2020), and marketing, where semantic congruence between brand names and logos may improve marketing strategies, also for made-up names (Klink, 2000).

Acknowledgments

We thank Federico Bianchi, Peter Hendrix, and Paul Schreiber for their help throughout the work.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 637–660.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 acm conference on fairness, accountability, and transparency* (p. 1493–1504). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3593013.3594095

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi: 10.1162/tacl_a_00051

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). *Does chatgpt resemble humans in language use?*

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases [Journal Article]. *Science*, *356*(6334), 183-186. doi: 10.1126/science.aal4230

Cassani, G., Chuang, Y. Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical category [Journal Article]. *Journal of Experimental Psychology-Learning Memory and Cognition*, *46*(4), 621-637. doi: 10.1037/xlm0000747

Cassidy, K. W., Kelly, M. H., & Sharoni, L. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology: General*, *128*, 362-381. doi: 10.1037/0096-3445.128.3.362

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., & Raff, E. (2022, 4). Vqgan-clip: Open domain image generation and editing with natural language guidance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13697 LNCS*, 88-105. doi: 10.1007/978-3-031-19836-6\textunderscore6

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, *25*(4), 447–464.

Davis, C. P., Morrow, H. M., & Lupyan, G. (2019). What does a horgous look like? nonsense words elicit meaningful drawings [Journal Article]. *Cognitive Science*, *43*(10), e12791. doi: 10.1111/cogs.12791

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Elsen, H. (2017). The two meanings of sound symbolism. *Open Linguistics*, *3*(1). doi: 10.1515/opli-2017-0024

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In (p. 12873-12883).

Fleiss, J. L. (1971, 11). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382. doi: 10.1037/H0031619

Gatti, D., Marelli, M., & Rinaldi, L. (2022). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0001304

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics [Journal Article]. *Psychonomic Bulletin and Review*, *23*(6), 1744-1756. doi: 10.3758/s13423-016-1053-2

Johfre, S. S. (2020). What age is in a name? [Journal Article]. *Sociological Science*, *7*, 367-390.

Joosse, A. Y., Kuscu, G., & Cassani, G. (in press). You sound like an evil young man: A distributional semantic analysis of systematic form-meaning associations for polarity, gender, and age in fictional characters' names [Journal Article]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0001345

Klink, R. R. (2000). Creating brand names with meaning: The use of sound symbolism. *Marketing Letters*, *11*(1), 5–20. doi: 10.1023/a:1008184423824

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, *6*, 145602.

Monaghan, P., & Fletcher, M. (2019). Do sound symbolism effects for written words relate to individual phonemes or to phoneme features? *Language and Cognition*, *11*, 235–255.

Pitcher, B. J., Mesoudi, A., & McElligott, A. G. (2013). Sex-biased sound symbolism in english-language first names. *PLoS ONE*, *8*(6), e64825. doi: 10.1371/journal.pone.0064825

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.*

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia - a window into perception, thought and language [Journal Article]. *Journal of Consciousness Studies*, *8*(12), 3-34.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021, 12). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 10674-10685. doi: 10.1109/CVPR52688.2022.01042

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234-241. doi: 10.1007/978-3-319-24574-4\textunderscore28/COVER

Sabbatino, V., Troiano, E., Schweitzer, A., & Klinger, R.

(2022). “splink” is happy and “phrouth” is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th workshop on computational approaches to subjectivity, sentiment & social media analysis.* Association for Computational Linguistics. doi: 10.18653/v1/2022.wassa-1.4

Sennrich, R., Haddow, B., & Birch, A. (2015, 8). Neural machine translation of rare words with subword units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1715-1725. doi: 10.18653/v1/p16-1162

Sidhu, D. M., Deschamps, K., Bourdage, J. S., & Pexman, P. M. (2019). Does the name say it all? investigating phoneme-personality sound symbolism in first names. *Journal of Experimental Psychology: General*, *148*(9), 1595–1614. doi: 10.1037/xge0000662

Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association [Journal Article]. *Psychonomic Bulletin and Review*, *25*(5), 1619-1643. doi: 10.3758/s13423-017-1361-1

Sidhu, D. M., Pexman, P. M., & Saint-Aubin, J. (2016). From the bob/kirk effect to the benoit/éric effect: Testing the mechanism of name sound symbolism in two languages. *Acta Psychologica*, *169*, 88–99.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017, 6). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999-6009.

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language*, *99*, 122-150. doi: https://doi.org/10.1016/j.jml.2017.09.006

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). *Huggingface's transformers: State-of-the-art natural language processing.*

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., … Lippus, P. (2022). The bouba/kiki effect is robust across cultures and writing systems [Journal Article]. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200390.