UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Modeling the Emergence of Letter Shapes

Permalink
https://escholarship.org/uc/item/1011w6cg

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Hein, Alice
Diepold, Klaus

Publication Date
2024

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Modeling the Emergence of Letter Shapes

Alice Hein (alice.hein@tum.de)
Department of Computer Engineering, School of Computation, Information and Technology
Technical University of Munich, Germany

Klaus Diepold (kldi@tum.de)
Department of Computer Engineering, School of Computation, Information and Technology
Technical University of Munich, Germany

Abstract

Graphic codes across times and cultures consistently share certain visual characteristics. According to the ecological hypothesis, this is because glyphs reflect the input statistics to which our visual system has adapted. We computationally model this hypothesis by employing a drawing-based signaling game involving two AI models to explore factors that impact empirical regularities in the surface form of artificially evolved glyphs and their similarity to human visual signs. In our first experiment, we investigate the role of the models' perception system on glyph line orientation and symmetry. We find that these characteristics are impacted by the input statistics of data used to pre-train models and, to a lesser extent, canvas shape and architectural model properties. Our second experiment analyzes the grapho-phonemic mapping that emerges when we integrate representations learned by a deep learning model trained for speech conversion into our setup.

Keywords: Cultural Evolution; Neural Recycling; Graphical Communication; Artificial Intelligence

Introduction

Writing is an ancient cognitive technology, which has been invented independently several times in the course of human history (Morin, 2022b). Even before fully-fledged writing systems, humans have produced geometric signs since at least the Paleolithic (Dutkiewicz et al., 2020). Curiously, graphic codes across times and cultures consistently share certain characteristics. Specifically, glyphs appear to reflect the input statistics to which our visual system has adapted. Letters tend to display a disproportionate rate of vertical symmetry, which is a feature of, e.g., human faces or standing bodies (Morin, 2018) and they extensively comprise topological signatures found in natural scenes (Changizi et al., 2006; Testolin et al., 2017). Furthermore, vertical and horizontal strokes are over-represented compared to obliques (Morin, 2018). This cardinality preference has been attributed to our visual acuity being better for vertical and horizontal lines than other orientations (Appelle, 1972). Finally, there is a tendency towards simplicity, as complex characters are more effortful to read and produce (Y.-C. Lin et al., 2019; Miton & Morin, 2021). These findings support an ecological hypothesis that signs have evolved to accommodate human visual perception.

Several studies have investigated the emergence of graphical conventions using signaling games where participants communicate through drawing (Galantucci, 2005; Garrod et al., 2007; Fay et al., 2010; Bergmann et al., 2013; Roberts et al., 2015; Fay et al., 2018; Fan et al., 2020; Hawkins et al., 2023). Their primary focus has been on the role of social context in the construction of sign systems and the trade-off between iconicity and abstraction. A smaller number of works develop AI models for generating sketches. Most of these models are trained to convert images into simplified drawings (Ha & Eck, 2018; Muhammad et al., 2018; Song et al., 2018; Cao et al., 2019; Mihai & Hare, 2021; Vinker et al., 2022; Qiu et al., 2022) or to play collaborative or Pictionary-type games (Fan et al., 2019; Bhunia et al., 2020).

Similar to previous work, we employ a drawing-based signaling game involving two AI models. However, the sketches produced by the models represent pre-defined, abstract classes. Thus, our goal differs in that we do not focus on iconicity nor on the evolution of language. Instead, we are interested in how different design choices impact empirical regularities in the surface form of artificially evolved glyphs and their similarity to human visual signs. We explore this question using two experiments. The first experiment investigates aspects of the receiver model's perception system that impact glyph stroke orientation and symmetry in abstract graphic codes. The second experiment takes a step towards modeling orthography by introducing an aural dimension and a notion of sender-side motor effort minimization.

Approach

Our setup consists of a sender and a receiver. The receiver is a visual model – specifically, a five-layer convolutional neural network (CNN). An architectural overview can be found in Table 1. We use a kernel size of 4×4, batch normalization, no bias, and a leaky ReLU activation with negative slope 0.2 for all convolutional layers. The sender is a simple linear model that generates a graphic code with a pre-defined number of glyphs. The criteria these glyphs should fulfill vary by experiment. The sender can place three lines per glyph on a 64×64 canvas. Three is the average number of strokes per character across many writing systems (Changizi et al., 2006). Each line is a quadratic Bézier curve defined by the x and y coordinates of its start, control, and endpoint. The sender thus has to optimize six parameters per glyph and stroke.

These parameters are optimized using the covariance matrix adaptation evolution strategy (CMA-ES) (Hansen & Ostermeier, 2001). CMA-ES is a stochastic numerical optimization method that has been found empirically to outperform other black box optimizations in a range of applica-

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.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Table 1: Summary of the sender model's architecture. C is the number of classes, which varies by dataset.

Layer	Type	Input Shape	Output Shape
1	Conv2d	$1 \times 64 \times 64$	$64 \times 32 \times 32$
2	Conv2d	$64 \times 32 \times 32$	$128 \times 16 \times 16$
3	Conv2d	$128 \times 16 \times 16$	$256 \times 8 \times 8$
4	Conv2d	$256 \times 8 \times 8$	$128 \times 4 \times 4$
5	Conv2d	$128 \times 4 \times 4$	$C \times 1 \times 1$
6	Flatten	$C \times 1 \times 1$	C
7	LogSoftmax	C	C

tions (Hansen et al., 2010), including activation maximization in CNNs (Wang & Ponce, 2022), which is closely related to our experimental setup. It also has the benefit of being quasi parameter-free and allowing us to define non-differentiable loss functions, which is not the case for gradient estimation methods such as backpropagation.

Our approach is inspired by Park (2020), who explore a similar setup of a CNN receiver and a sender drawing a set of abstract glyphs with Bezier curves. However, we use different model architectures, loss functions, and optimization algorithms. Our work also diverges in scope in that Park (2020) present a technical proof-of-concept mainly focused on aesthetics. We significantly expand on their proposal by systematically analyzing the generated codes, contextualizing them in cultural evolution research on human writing systems, and, in experiment 2, introducing an aural dimension.

Experiment 1

In our first experiment, we build a computational model of the hypothesis that letters evolved to reflect the statistics of natural visual inputs. We pre-train receivers on different image datasets and measure the effect on the cardinality and symmetry of glyphs produced by the sender.

Methods

Datasets Inspired by Changizi et al. (2006), we train one model on "natural" images (henceforth NAT) and one on images of urban landscapes and human-made objects (henceforth H-M). We also include a randomly initialized, untrained CNN for comparison. The composition of the NAT and H-M datasets can be found in Tables 2 and 3, respectively. We resize the shortest side of each image and apply centered crops

Table 2: Composition of the NAT dataset.

#	class	source
5,666	natural landscape	15-Scene (Lazebnik et al., 2006): `coast`, `forest`, `mountain`, `open country`
		Flickr (Chen et al., 2018)
5,500	face	CelebA 64x64 (Liu et al., 2015)
5,500	plant	ImageCLEF 2013 (Goëau et al., 2013)
5,399	animal	Animal Image (Banerjee, 2023)

Table 3: Composition of the H-M dataset.

#	class	source
2,200	urban landscape	15-Scene (Lazebnik et al., 2006): `street`, `suburb`, `living room`, `office`, `industry`, `building`, `inside city`, `highway`
2,200	motorcycle	COCO 2017 (T.-Y. Lin et al., 2014)
2,200	airplane	COCO 2017 (T.-Y. Lin et al., 2014)
2,200	wine glass	COCO 2017 (T.-Y. Lin et al., 2014)
2,200	bowl	COCO 2017 (T.-Y. Lin et al., 2014)

to obtain 64×64 inputs, which we gray-scale and normalize. Any images containing text were removed to prevent exposure to human writing systems. We use 80% of the data for training and 20% for validation.

We also create a dataset of human-made scripts for comparison. This dataset is based on the collection of 116 writing systems analyzed by Morin (2018). We generate one image per glyph using a consistent font (Noto Sans). Loma, Woleai, Kpelle, and Afak scripts were omitted as they are not yet part of the Unicode codespace.

Receiver Model All receiver models' architectures are identical (shown in Table 1), except for their output dimension C. C is four for the NAT and random models and five for the H-M dataset. NAT and H-M models were trained for 200 epochs on image classification, using the Adam optimizer (Kingma & Ba, 2015), negative log likelihood loss, and a batch size of 64. The final validation accuracy after early stopping was 86% for both models. Note that receiver models are not updated further during their interaction with the sender model to avoid overfitting to the produced glyphs.

Sender Model The sender is tasked with developing a graphic code with 25 glyphs, which should be perceptually distinct to the receiver. More specifically, it maximizes the distance between the activations elicited in the receiver by the different glyphs. The sender thus optimizes for what Qiu et al. (2022) term *symbolicity*, i.e., consistent separability in high-level visual embedding space. Let A represent the activations in the receiver's convolutional layers. We use embeddings from the receiver's last layer for most experiments. Each activation vector a corresponds to a different glyph. The sender's loss function aims to maximize the minimum L2 norm between each activation and its closest neighbor:

$$\text{Loss} = \frac{1}{|A|} \sum_{a_i \in A} \min_{\substack{a_j \in A \\ a_j \neq a_i}} \|a_i - a_j\|_2 \tag{1}$$

For the CMA-ES optimization of sender models, we use a population size of 32, uniform random solution initialization, and an initial standard deviation of 0.05. We let models run for 1300 iterations and train ten models per setting. We report averages across these ten models.

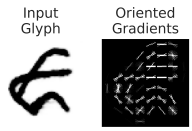

Figure 1: Histogram of oriented gradients applied to a glyph.

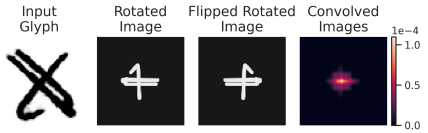

Figure 2: Illustration of the steps involved in our symmetry measure. The example shown is for an angle of 137°.

Metrics To measure the orientation of glyph strokes, we use histogram of oriented gradients (HOG) (Dalal & Triggs, 2005). HOG is a computer vision feature descriptor that splits an image into a grid of cells. For each pixel in the cell, intensity gradients, i.e., edge directions, are computed, binned into orientations, and counted to obtain a histogram. Contrast normalization may be applied block-wise for better invariance to lighting changes. An example is shown in Fig. 1. HOG is traditionally used for tasks such as object detection. We here re-purpose it as an automated alternative to the manual coding by which letter cardinality has previously been analyzed (Morin, 2018). We use 12 orientations, cells of size 16×16, three cells per block, and L2 normalization. Before applying HOG, we resize images to 128×128 and apply a Gaussian blur of size 2 to avoid square pixelation artifacts that would artificially increase cardinal dominance.

To measure glyph symmetry, conceptually, we place an axis through an image at each angle between 0°and 179°, mirror it along that axis, and record the overlap for each angle. In practice, we rotate the glyphs in SVG space by each angle between 0°and 179°. We pad images to avoid parts of the glyph rotating out of the picture at certain angles. We then flip the rotated image vertically, sum-normalize the rotated and flipped rotated images, and convolve the two via the fast Fourier transform method. Intuitively, this corresponds to moving the flipped rotated image over the rotated image and computing the overlap at each point. We use the maximum value of the convolution as a measure of the highest overlap, i.e., symmetry, at a given angle. Fig. 2 shows an example of the process outlined above.

Note that our metric is a more continuous measure of symmetry than used by Morin (2018). In their work, symmetry was coded manually, and only wholly (vertically or horizontally) symmetric letters were considered. Our Bézier curve-based glyphs are more akin to handwriting than standardized letters and contain a higher degree of noise, e.g., small shifts or rotations. We propose our automated measure as a way to still capture symmetric regularities in such cases.

Results

Stroke Orientation Fig. 3 shows that vertical and horizontal orientations are most common in both the NAT and H-M dataset, consistent with previous analyses (Coppola et al., 1998; Changizi et al., 2006; Girshick et al., 2011). Pre-training on this data promotes a preference for cardinality: Particularly, gradients near 90°occur with above-average frequency in glyphs evolved for the pre-trained receivers. In

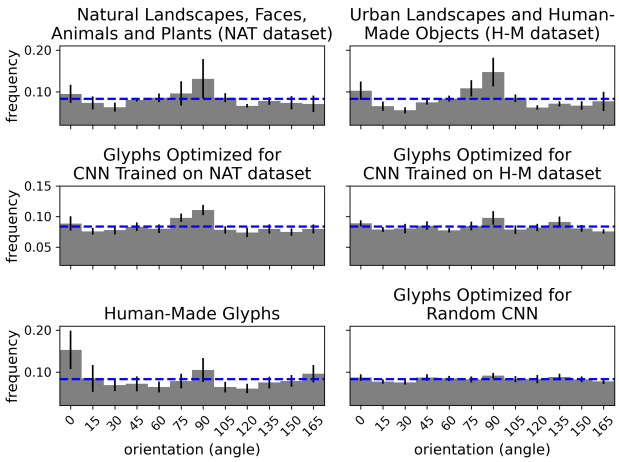

Figure 3: Histograms of oriented gradients for the pre-training datasets, artificially evolved graphic codes, and human writing systems. Blue dotted line marks expected frequency in uniform random case.

contrast, glyphs optimized for a random CNN show a nearly uniform gradient distribution. The correlation between orientation statistics of the pre-training dataset and optimized glyphs is stronger for NAT ($R = 0.92$, $p < 2 \times 10^{-5}$) than for H-M ($R = 0.68$, $p < 2 \times 10^{-2}$), likely because NAT models were exposed to more training data.

Although the tendency towards cardinality is less pronounced than in human-made letters, we find a moderate correlation between orientation characteristics of evolved and human-made glyphs ($R = 0.44$, $p < 0.15$). Overall, the results support the notion that optimizing for a visual system that has been exposed to natural input statistics can give rise to the preferred line orientations observed in human writing systems. Analogous to the mechanisms thought to have shaped human letters, CNN units will be more attuned to common orientations in their training set (Henderson & Serences, 2021), which the sender may, in turn, exploit to optimize discriminability.

Symmetry We now turn our attention to another aspect of anisotropy: Glyph symmetry. Fig. 4 shows that, consistent with human preferences, there is an above-average tendency

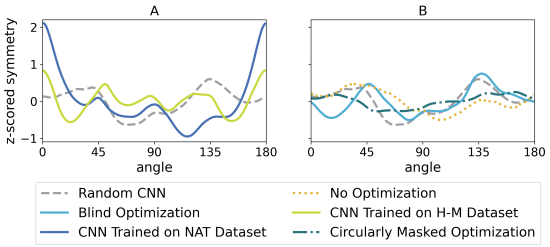

Figure 4: Z-scored symmetry distribution of artificially evolved graphic codes.

(a) (b)

Figure 5: Exemplary graphic codes evolved for the NAT (a) and H-M (b) receiver's different convolutional layers.

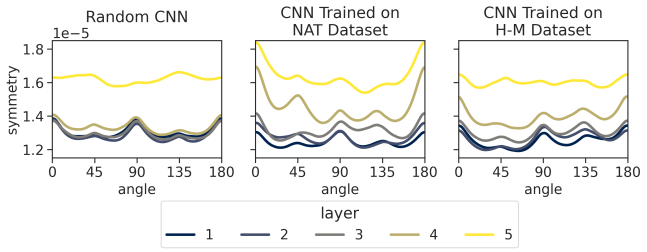

Figure 6: Comparison of glyph symmetry for codes evolved for different receiver layers.

towards vertical symmetry in our evolved glyphs, particularly for the NAT setup. This result is perhaps to be expected from our HOG analysis, as cardinal lines tend to be symmetrical. However, interestingly, even glyphs evolved for random CNNs show a slight above-chance symmetry along 45°and 135°angles despite not having been exposed to any training that could explain this preference.

We considered three possible explanations for this phenomenon: 1) a bias introduced by the CMA-ES algorithm, 2) a bias introduced by the canvas shape, and 3) an inductive bias of the CNN architecture. To test each of these options, we plotted results without any optimization (create glyphs via uniform random sampling), optimization with a circular mask (resample any time CMA-ES suggests a solution containing points outside the circle), and blind optimization (set loss to constant 0). Fig. 4B shows that, without any optimization, there are still peaks at 45°and 135°angles. This speaks against option 1. Blind optimization closely resembles that for the untrained receiver, suggesting that the random CNN's feedback, rather than containing some hidden preference, is basically arbitrary. This contradicts option 3.

However, the symmetry preferences disappear when using circularly masked optimization, confirming option 2. Considering the square canvas is uniformly sampled, a slight over-representation of points in the four corners will implicitly promote symmetry at 45°and 135°angles as measured by our "overlap" metric. This result relates to the role of physical constraints imposed by writing materials on the evolution of human scripts. E.g., rectangular canvases have been considered as a potential cause of cardinal dominance in paintings (Miller, 2007), and Indian and Southeast Asian scripts are thought to be less angular because they were written on flexible leaves (Watt, 1994).

Besides pre-training and canvas shape, we identified an additional factor influencing symmetry: the layer of the CNN receiver used to calculate equation 1. Fig. 5 shows examples of glyphs optimized for different layers. When comparing the symmetry of the produced glyphs (Fig. 6), it is interesting to see that, even in glyphs optimized for random CNNs, there is a higher level of symmetry at 0°, 90°, 180°, and to a lesser degree at 45 °and 135°angles. The tendency is less pronounced for the highest layer. We hypothesize that this symmetry emerges due to the square nature of the receiver's convolutional filters, which partition the input image into overlapping patches. From the sender's perspective, each patch essentially represents a separate noisy channel (Shannon, 1948).

Given that the sender is limited to contiguous strokes in utilizing these channels, it will tend to connect grid neighbors, resulting in increased cardinal and oblique symmetry for a grid of squares. Symmetry is less pronounced in the last layer because this layer's receptive field spans the whole image (see Table 1), limiting the sender to a single, global channel. To further illustrate the described effect, we create a setup that mimics the mechanism of CNN filters in a simplified way. We split canvases into 4×4 patches and define the signal communicated to the sender as a 16-dimensional vector with binary entries: 0 if a patch contains no ink in a square and 1 if it does. We then compare the effect of using a square and a hexagonal grid, the latter implemented as proposed by Steppa and Holch (2019).

As shown in Fig. 7, the symmetry distribution for the square grid setup highly correlates with that of the random CNN ($R = 0.91$, $p < 8 \times 10^{-72}$). In the hexagonal case, the angles of a cell's neighbors change, increasing symmetry at 30°and 150°. Thus, although CNNs carry the added complexity of overlapping filters and multiple filter channels, the results from our simplified setups lend credence to the proposed explanation of the symmetry behavior for lower layers.

Experiment 2

The glyphs produced in experiment 1 can be seen as emblems, i.e., graphic codes that do not encode a language and lack the productivity of specialized codes such as musical notation (Morin et al., 2020). A graphic code is only considered writing if it encodes words, morphemes, or phonemes, thus vicariously acquiring generality (Morin, 2022b). In our second experiment, we take a step towards modeling the development of such a glottographic code representing language.

Our setup differs from experiment 1 in two regards: The first is that the semantics of the evolved glyphs change from abstract classes to representing linguistic information. The second is that we add constraints designed to mimic the evolutionary pressure towards reduced effort in producing and processing writing.

Methodology

Speech Model Linguistic information is incorporated using a deep learning model trained for speech conversion, i.e., transforming source speech to a target voice without changing the content. This task resembles writing in that the goal is to communicate content in a standardized form, abstracting away speaker-specific acoustic features. The specific model we use was proposed and implemented by van Niekerk et al. (2022). It works by extracting features from HuBERT (Hsu et al., 2021), a widely used transformer-based speech model. HuBERT is pre-trained in a self-supervised manner on LibriSpeech-960 (Panayotov et al., 2015).

Van Niekerk et al. (2022) apply k-means ($k = 100$) to the intermediate representations of HuBERT's 7th layer and train an acoustic model to decode the resulting clusters into output speech. However, we ignore this decoder here and simply use the 100 clusters, which we will henceforth refer to as *units*. Note that these units are not explicitly trained to map to phonemes, morphemes, or syllables. They are simply representations learned by the model to optimally fulfill its speech conversion task. Note also that the model has only been trained to predict speech in spectrogram space, not to transcribe it. Its representations have thus not been shaped by exposure to English orthography.

Sender Model As in experiment 1, the sender model must optimize a graphic code, this time containing 100 glyphs representing the speech model units. However, instead of creating maximally distinguishable glyphs, it is tasked with using as few strokes as possible while still ensuring what Qiu et al. (2022) term *semanticity*, i.e., visually preserving the topology of the speech model's latent space. These constraints reflect two competing pressures on writing systems: transmission efficiency and referential efficiency (Morin, 2016, 2018; Kelly et al., 2021). We model them with a frequency penalty and a similarity constraint, respectively.

We allow the sender to draw up to $L_{\max} = 3$ strokes. To model transmission efficiency, we add a penalty P. P calculates how many strokes l were used for a glyph, multiplied by the relative frequency f with which the unit it represents occurs in natural speech. This term reflects the finding that, across writing systems, more frequent characters consistently have a lower degree of complexity (Miton & Morin, 2019; Koshevoy et al., 2023). We collect frequency statistics by applying the speech conversion model to the Flickr 8k Audio Caption Corpus (Harwath & Glass, 2015) and recording how many times each speech unit occurs. The penalty is weighted by a factor α, here set to $\frac{1}{2}$.

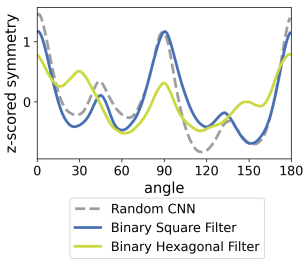

Figure 7: Symmetry of graphic codes evolved in binary grid set-ups.

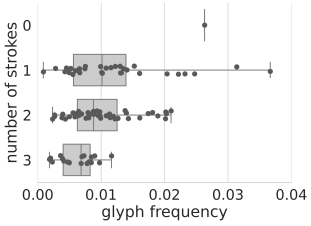

Figure 8: Glyph frequency vs. complexity for speech-based graphic code.

$$P_i = \alpha \times \left(1 - \frac{l_i}{L_{\max}}\right) \times f_i$$

For the similarity constraint, we calculate the pairwise L2 distance d between the centers of the $N = 100$ units in the speech model's activation space A. We do the same for the images of the 100 glyphs in the visual receiver model's embedding space. We normalize each row in the distance matrices and subtract it from 1 to obtain a measure of similarity s:

$$s_{ij} = 1 - \frac{d_{ij}}{\max_i d_{ij}}, \quad d_{ij} = \|A_i - A_j\|_2$$

We then minimize the mean absolute distance between the two similarity matrices S^{vis} and S^{speech}. The reasoning behind this is that characters that look similar tend to have similar canonical pronunciations across orthographies (Jee et al., 2022). The combined loss function reads as follows:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}|S_{ij}^{\text{vis}} - S_{ij}^{\text{speech}}| \times (1 + P_i)\right)$$

Given the added complexity of the problem, we increase our CMA-ES population size to 64 and the number of iterations to 30,000. We initialize solutions to 0.5 with a standard deviation of 0.1 and use a NAT CNN from experiment 1 as our receiver, with activations taken from layer 4. To mimic the variability inherent to handwriting, we add Gaussian noise to the solutions with a standard deviation of $0.005^{\frac{1}{2}}$.

Association Rule Mining In this experiment, we are interested not only in the surface form of the glyphs but also in the kind of grapho-phonemic mapping the models produce. We, therefore, analyze the co-occurrences of glyphs and speech units using TIMIT (Garofolo et al., 1993), a standard dataset used to evaluate automatic speech recognition systems. TIMIT is designed for broad phoneme coverage and includes rich word and phoneme-level annotations.

For each sentence in the corpus, we collect the phoneme annotation and the units produced by the speech model during processing. Having collected co-occurrences, we apply the apriori algorithm (Agrawal & Srikant, 1994) to identify

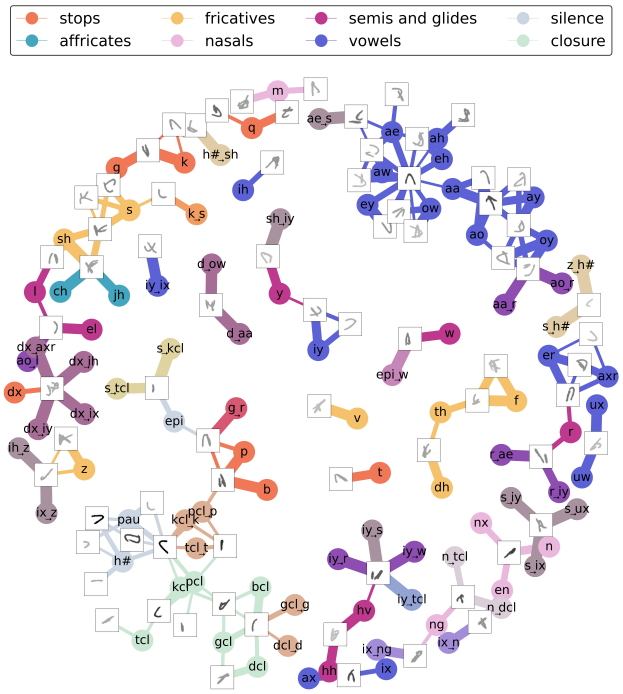

Figure 9: Association rules for glyphs and phonemes, encoded using TIMIT notation. For nodes representing transitions (→), colors corresponding to the individual phonemes' categories were combined. Edge width represents rule confidence. Glyph opacity represents frequency of occurrence.

association rules. We generate all rules with a minimum support of 0.0001, minimum confidence of 0.2, and minimum lift of 3. Support here refers to the relative frequency of a unit-phoneme pairing, confidence refers to the conditional probability of a unit given a phoneme or another preceding unit, and lift refers to the ratio between confidence and support (Agrawal & Srikant, 1994).

Results

We visualize co-occurrences between phonemes and glyphs, representing HuBERT speech units, in Fig. 9. The graph shows that the similarity constraint of our loss function works as intended: phonetic similarities are reflected visually. E.g., glyphs correlated with the phonemes s (sea), sh (she), and z (zone) look alike. Similar phonemes often share a glyph, see, e.g., y (yacht) and iy (beet), or b (bee) and p (pea).

Interestingly, some glyphs co-occur with common phoneme combinations, e.g., "ix ng/ix n", as used in the English present participle form, or the word "sh iy" (she). This mapping has some correspondence with human scripts in that no writing system consistently follows a single organizing principle (Morin, 2022b). I.e., while some scripts may be predominantly syllabic, alphabetic, or logographic, no system falls into purely one category (Mattingly, 1992).

In addition to preserving phonetic similarity, the evolved glyphs successfully reflect transmission effort. Frequent glyphs, such as those representing closures, pauses, or the sentence marker h#, tend to be simple, often consisting of a single line. Fig. 8 illustrates this in more detail. The bulk of the glyphs evolve to contain between one and two strokes, with only a few low-frequency glyphs encoded using three. One high-frequency glyph is effectively a space, i.e., it has zero strokes. Note that this glyph is not contained in Fig. 9 as we only show rules with a minimum confidence of 0.2. Consistent with "Zipf's law of meaning" (Zipf, 1949), the high-frequency glyph in question co-occurs with many different phonemes, diluting its association rules' confidence.

Discussion

Our work is informed by two fields: Machine learning and cultural evolution. From a machine learning perspective, it relates to the interpretability technique of activation maximization for CNNs (Yosinski et al., 2015) and probing studies of self-supervised speech models (Ji et al., 2022). Similar to the way that writing provides insights into human cognitive constraints, the evolved graphic codes can be seen as a window into these artificial models of visual and speech perception. From a cultural evolution perspective, we offer a concrete implementation of how features of "cultural attractors" in writing systems (Kelly et al., 2021) can emerge without pre-supposing universal, innate aesthetic preferences.

Conclusion

In summary, we find that, as predicted by the ecological hypothesis of letter shapes, glyphs evolved for models pre-trained on images reflect the statistics of their input data and display anisotropy consistent with human-made glyphs. We also observe that the square nature of the canvas and the receiver's convolutional filters impact glyph symmetry, albeit to a smaller extent. We then integrate representations learned by a pre-trained speech model as well as efficiency pressures into our setup. The resulting code yields a hybrid orthography and shows a Zipfian effect for glyph complexity.

In future work, our experimental setups could, e.g., be used to investigate the trade-off between referential and transmission efficiency more in-depth by varying the penalty factor α or code size N. Furthermore, HuBERT could easily be replaced by speech models pre-trained on other languages or augmented with visual input of speakers' mouth areas (Shi et al., 2022) to test how this influences the organizing principles of artificially evolved graphic codes.

The scope of this study could also be expanded by moving from a discrete code of predefined size to a continuous, perhaps even open-ended, meaning space. Finally, our setup is asynchronous by design, meaning agents have no shared context. However, it involves instant feedback via a loss signal, a feature of synchronous, face-to-face communication. In reality, this is an unlikely combination – a factor that has been proposed as an explanation for why the evolution of writing is such a rare and slow process (Morin et al., 2020; Morin, 2022a). Our experimental setup could thus be adapted to reflect more realistic social mechanisms.

References

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the International Conference on Very Large Databases, (VLDB)* (pp. 487–499).

Appelle, S. (1972). Perception and Discrimination as a Function of Stimulus Orientation: The "Oblique Effect" in Man and Animals. *Psychological Bulletin*, *78*(4), 266. doi: 10.1037/h0033117

Banerjee, S. (2023). *Animal Image Dataset.* Retrieved 2024-01-16, from `https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals/data`

Bergmann, T., Dale, R., & Lupyan, G. (2013). The Impact of Communicative Constraints on the Emergence of a Graphical Communication System. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci).* Berlin, Germany.

Bhunia, A. K., Das, A., Muhammad, U. R., Yang, Y., Hospedales, T. M., Xiang, T., … Song, Y. (2020). Pixelor: A Competitive Sketching AI Agent. So You Think You Can Sketch? *ACM Transactions on Graphics (TOG)*, *39*(6), 166:1–166:15. doi: 10.1145/3414685.3417840

Cao, N., Yan, X., Shi, Y., & Chen, C. (2019). AI-Sketcher: A Deep Generative Model for Producing High-Quality Sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 2564–2571).

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The Structures of Letters and Symbols throughout Human History Are Selected to Match Those Found in Objects in Natural Scenes. *The American Naturalist*, *167*(5), E117–E139. doi: 10.1086/502806

Chen, Y., Lai, Y.-K., & Liu, Y.-J. (2018). CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9465–9474). Salt Lake City, USA.

Coppola, D. M., Purves, H. R., McCoy, A. N., & Purves, D. (1998). The Distribution of Oriented Contours in the Real World. *Proceedings of the National Academy of Sciences*, *95*(7), 4002–4006. doi: 10.1073/PNAS.95.7.4002

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, p. 886-893). San Diego, USA. doi: 10.1109/CVPR.2005.177

Dutkiewicz, E., Russo, G., Lee, S., & Bentz, C. (2020). SignBase, a Collection of Geometric Signs on Mobile Objects in the Paleolithic. *Scientific Data*, *7*(1), 364. doi: 10.1038/s41597-020-00704-x

Fan, J. E., Dinculescu, M., & Ha, D. (2019). Collabdraw: An Environment for Collaborative Sketching with an Artificial Agent. In *Proceedings of the ACM SIGCHI Conference on Creativity and Cognition (C&C)* (pp. 556–561). San Diego, USA. doi: 10.1145/3325480.3326578

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic Inference and Visual Abstraction Enable Contextual Flexibility during Visual Communication. *Computational Brain & Behavior*, *3*, 86–101. doi: 10.1007/s42113-019-00058-7

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The Interactive Evolution of Human Communication Systems. *Cognitive Science*, *34*(3), 351–386. doi: 10.1111/J.1551-6709.2009.01090.X

Fay, N., Walker, B., Swoboda, N., & Garrod, S. (2018). How to Create Shared Symbols. *Cognitive Science*, *42*, 241–269. doi: 10.1111/COGS.12600

Galantucci, B. (2005). An Experimental Study of the Emergence of Human Communication Systems. *Cognitive Science*, *29*(5), 737–767. doi: 10.1207/S15516709COG0000_34

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. *NASA STI/Recon Technical Report*, *93*, 27403.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of Representation: Where Might Graphical Symbol Systems Come From? *Cognitive Science*, *31*(6), 961–987. doi: 10.1080/03640210701703659

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal Rules: Visual Orientation Perception Reflects Knowledge of Environmental Statistics. *Nature Neuroscience*, *14*(7), 926–932. doi: 10.1038/NN.2831

Goëau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N., & Molino, J.-F. (2013). The ImageCLEF Plant Identification Task 2013. In *Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data* (pp. 23–28). Barcelona, Spain.

Ha, D., & Eck, D. (2018). A Neural Representation of Sketch Drawings. In *Proceedings of the International Conference on Learning Representations (ICLR).* Vancouver, Canada.

Hansen, N., Auger, A., Ros, R., Finck, S., & Pošík, P. (2010). Comparing Results of 31 Algorithms from the Black-Box Optimization Benchmarking BBOB-2009. In *Proceedings of the Annual Conference Companion on Genetic and Evolutionary Computation (GECCO)* (pp. 1689–1696). Portland, USA.

Hansen, N., & Ostermeier, A. (2001). Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, *9*(2), 159–195.

Harwath, D., & Glass, J. (2015). Deep Multimodal Semantic Embeddings for Speech and Images. In *Ieee Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 237–244).

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual Resemblance and Interaction History Jointly Constrain Pictorial Meaning. *Nature Communications*, *14*(1), 2199. doi: 10.1038/s41467-023-37737-w

Henderson, M., & Serences, J. T. (2021). Biased Orientation Representations can be Explained by Experience with

Nonuniform Training Set Statistics. *Journal of Vision*, *21*(8), 1–22. doi: 10.1167/JOV.21.8.10

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460.

Jee, H., Tamariz, M., & Shillcock, R. (2022). Systematicity in Language and the Fast and Slow Creation of Writing Systems: Understanding Two Types of Non-arbitrary Relations between Orthographic Characters and their Canonical Pronunciation. *Cognition*, *226*, 105197. doi: 10.1016/J.COGNITION.2022.105197

Ji, H., Patel, T., & Scharenborg, O. (2022). Predicting Within and Across Language Phoneme Recognition Performance of Self-supervised Learning Speech Pre-trained Models. *CoRR*. doi: 10.48550/ARXIV.2206.12489

Kelly, P., Winters, J., Miton, H., & Morin, O. (2021). The Predictable Evolution of Letter Shapes: An Emergent Script of West Africa Recapitulates Historical Change in Writing Systems. *Current Anthropology*, *62*(6), 669–691. doi: 10.1086/717779

Kingma, D. P., & Ba, J. (2015). Adam: A method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego, USA.

Koshevoy, A., Miton, H., & Morin, O. (2023). Zipf's Law of Abbreviation Holds for Individual Characters Across a Broad Range of Writing Systems. *Cognition*, *238*, 105527. doi: doi.org/10.1016/J.COGNITION.2023.105527

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2, pp. 2169–2178). New York, USA.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 740–755). Zurich, Switzerland.

Lin, Y.-C., Chao, Y.-L., Hsu, C.-H., Hsu, H.-M., Chen, P.-T., & Kuo, L.-C. (2019). The Effect of Task Complexity on Handwriting Kinetics. *Canadian Journal of Occupational Therapy*, *86*(2), 158–168. doi: 10.1177/0008417419832327

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 3730–3738). Santiago, Chile.

Mattingly, I. G. (1992). Linguistic Awareness and Orthographic Form. In *Advances in Psychology* (Vol. 94, pp. 11–26). doi: 10.1016/S0166-4115(08)62786-7

Mihai, D., & Hare, J. S. (2021). Learning to Draw: Emergent Communication through Sketching. In *Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 7153–7166). Virtual.

Miller, R. (2007). Another Slant on the Oblique Effect in Drawings and Paintings. *Empirical Studies of the Arts*, *25*(1), 41–61. doi: 10.2190/E737-V374-0640-6766

Miton, H., & Morin, O. (2019). When Iconicity Stands in the Way of Abbreviation: No Zipfian Effect for Figurative Signals. *PLoS One*, *14*(8), e0220793. doi: 10.1371/JOURNAL.PONE.0220793

Miton, H., & Morin, O. (2021). Graphic Complexity in Writing Systems. *Cognition*, *214*, 104771. doi: 10.1016/J.COGNITION.2021.104771

Morin, O. (2016). The Spontaneous Emergence of Functional Complexity in Writing Systems: The Case of Cardinal Lines. *The Idea of Writing*, 13.

Morin, O. (2018). Spontaneous Emergence of Legibility in Writing Systems: The Case of Orientation Anisotropy. *Cognitive Science*, *42*(2), 664–677. doi: 10.1111/COGS.12550

Morin, O. (2022a). The Piecemeal Evolution of Writing. *Lingue e Linguaggio*, *21*(2), 217–237. doi: 10.1418/105963

Morin, O. (2022b). The Puzzle of Ideography. *Behavioral and Brain Sciences*, 1–69. doi: 10.1017/S0140525X22002801

Morin, O., Kelly, P., & Winters, J. (2020). Writing, Graphic Codes, and Asynchronous Communication. *Topics in Cognitive Science*, *12*(2), 727–743. doi: 10.1111/TOPS.12386

Muhammad, U. R., Yang, Y., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2018). Learning Deep Sketch Abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8014–8023). Salt Lake City, USA.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210).

Park, S. W. (2020). Generating Novel Glyph without Human Data by Learning to Communicate. In *Machine Learning for Creativity and Design Workshop @ NeurIPS*. Virtual.

Qiu, S., Xie, S., Fan, L., Gao, T., Joo, J., Zhu, S., & Zhu, Y. (2022). Emergent Graphical Conventions in a Visual Communication Game. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, USA.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How Communication Changes When We Cannot Mime the World: Experimental Evidence for the Effect of Iconicity on Combinatoriality. *Cognition*, *141*, 52–66. doi: 10.1016/J.COGNITION.2015.04.001

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379–423. doi: 10.1002/J.1538-7305.1948.tb01338.X

Shi, B., Mohamed, A., & Hsu, W. (2022). Learning lip-based audio-visual speaker embeddings with av-hubert. In *23rd proceedings of the annual conference of the international*

speech communication association (INTERSPEECH) (pp. 4785–4789). Incheon, Korea: ISCA. doi: 10.21437/INTERSPEECH.2022-885

Song, J., Pang, K., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2018). Learning to Sketch with Shortcut Cycle Consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 801–810). Salt Lake City, USA.

Steppa, C., & Holch, T. L. (2019). HexagDLy—Processing Hexagonally Sampled Data with CNNs in PyTorch. *SoftwareX*, *9*, 193 - 198. doi: 10.1016/J.SOFTX.2019.02.010

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter Perception Emerges from Unsupervised Deep Learning and Recycling of Natural Image Features. *Nature Human Behaviour*, *1*(9), 657–664. doi: 10.1038/s41562-017-0186-2

van Niekerk, B., Carbonneau, M.-A., Zaïdi, J., Baas, M., Seuté, H., & Kamper, H. (2022). A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6562–6566). Singapore.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., … Shamir, A. (2022). CLIPasso: Semantically-Aware Object Sketching. *ACM Transactions on Graphics (TOG)*, *41*(4), 86:1–86:11. doi: 10.1145/3528223.3530068

Wang, B., & Ponce, C. R. (2022). High-Performance Evolutionary Algorithms for Online Neuron Control. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)* (pp. 1308–1316). Boston, USA: ACM. doi: 10.1145/3512290.3528725

Watt, W. (1994). Curves as Angles. In *Writing Systems and Cognition: Perspectives from Psychology, Physiology, Linguistics, and Semiotics* (pp. 215–246). Springer.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural Networks through Deep Visualization. In *Deep Learning Workshop @ ICML.*

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.