WHY ARE NUMERAL SYSTEMS REGULAR? THE ROLE OF LEARNABILITY AND GENERALISABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Human recursive numeral systems (e.g. counting systems such as Mandarin or English base-10 numerals) are, like many other linguistic systems, highly regular, i.e. they can be represented by a simple grammar. We are interested in modelling what pressures select for such representations over the course of linguistic evolution. Following prior work that relates cross-linguistic tendencies to biases in learning, we hypothesise that regular systems are common because regularity allows the extraction of highly-learnable, highly-compressible representations. We compile a dataset of 20,732 unique numeral systems (both human and artificial) and experiment with both reinforcement and supervised learning as high-level analogues for human learning. Experiments confirm that highly regular human(-like) systems are indeed more learnable than unattested irregular systems, but only when learnability is taken as the capacity to derive a generalisable representation of the whole numeral system from skewed, limited data. Intriguingly, our results indicate that the advantage of regularity is more significant for reinforcement learning. We also conduct exploratory analyses which show that a multitude of factors affect numeral systems' learnability, but crucially, regularity robustly remains one of the key predictors. Our results contribute to the body of work linking learnability to cross-linguistic prevalence.

1 INTRODUCTION

Numeral systems across many diverse human languages have converged on highly regular solutions for expressing number (Prasertsom et al., 2026): Mandarin for instance expresses every numeral from 1 to 99 with just ten morphemes, 一 *yī* (1) through 十 *shí* (10) and two rules of composition; e.g. 十九 *shíjiǔ* (10+9), 二十 *èrshí* ($2 * 10$) and 二十一 *èrshíyī* ($2 * 10+1$). Why are the numeral systems of natural languages so regular? We explore one explanation: *regularity* is what makes a numeral system *learnable* from the kind of skewed and sparse input learners are likely to receive. The frequency rates of different number words are wildly skewed: a learner is likely to encounter small numbers like *one* and *two* extremely frequently, but will only very rarely hear a large (non-round) number like *eighty-nine* (Dehaene & Mehler, 1992). While rote memorisation might be enough to master the frequent small numerals, the limits of human memory would make the full range impossible; mastering the rare, large numerals requires learning a compressed representation of the system which permits generalisation. Recursive numeral systems are designed to achieve such a representation; they can express any integer exactly. Thus, the domain constitutes an unusually clean testbed for the learnability advantage of regularity. In this paper, we show that regular recursive numeral systems are learnable from scarce, skewed input data. This ease of learning is expected to shape a linguistic system over time: a system that is easier to learn should be transmitted more faithfully across generations, and therefore more widespread (Smith et al., 2003).

Some evidence exists that regular numeral systems are indeed more easily learned by humans (Guerrero et al., 2020; Holt & Barner, 2025), neural network models make it possible to test the learnability of many different numeral systems: unlike human participants, they can be trained in large numbers on many languages under identical conditions (Portelance & Jasbi, 2024). Work in this vein has shown that more systematic, more human-like languages are learned more efficiently in several linguistic domains (Steinert-Threlkeld & Szymanik, 2020; Galke et al., 2024; Kallini et al., 2024; Xu et al., 2026). We apply this approach to recursive numeral systems, testing the impact of regularity (and other possible properties) on the learnability of human and artificial numeral systems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

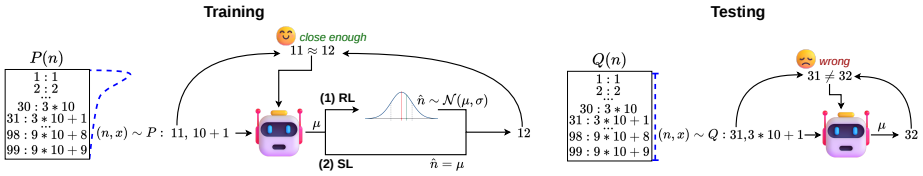


Figure 1: Framework to estimate the learnability of recursive numeral systems. An agent receives a numeral representation sampled from a training distribution, and learns to output the corresponding number. When trained via RL, it samples the prediction from a Gaussian distribution with output as the mean, whereas with SL it uses the output as the prediction directly. The agent is trained either via proximity-based reward (RL) or MSE loss (SL). Testing is done by checking exact matching between the model's prediction and the original number sampled.

We collect a dataset of 20,732 unique numeral systems, both attested and artificial, and train simple LSTM-based agents as high-level analogues for human learning, under both reinforcement learning (RL) and supervised learning (SL); Figure 1 gives an overview of the framework. Our hypothesis is that regularity is a key predictor of learnability, even when we account for other properties.

Our results highlight where the relationship between regularity and learnability is clear, and where it is not. When agents are trained *and* tested on a heavily skewed distribution of number-word frequencies, regularity barely matters, because it is enough to learn only the handful of very frequent small numerals (§4.1; cf. Smith et al., 2023). The expected relationship emerges only once learnability is taken as the capacity to derive a **generalisable representation of the whole system from skewed, limited data** (§4.2), including the rare, large numerals that recursive systems exist to express (cf. Pitt et al., 2022). This finding supports the idea that the scarcity and skew of the input are in general key factors for human language learning.

Moreover, our results pinpoint a structural difference in how RL and SL exploit regularity: where RL seems more sensitive to structure in the input, SL is efficient at learning less regular, idiosyncratic representations (§4.2; cf. Cheng et al., 2023). As both methods are heavily used in training modern large language models, this suggests that the training regime could influence the kind of representations, and ultimately the kind of languages, these models learn.

Finally, while regularity is a key predictor of learnability, we also demonstrate that multiple factors influence learnability (§4.3): research connecting linguistic features to learnability in language models may benefit from asking how those features combine, rather than which one dominates.

2 Background

Efficiency of recursive numeral systems. Recent research suggests that linguistic systems are optimised for *efficient communication* (Gibson et al., 2019), i.e. they minimise complexity while maximising informativeness. However, recursive numeral systems are not subject to this trade-off, as they are maximally informative, meaning they can precisely express any integer (Xu et al., 2020). Therefore, Denić & Szymanik (2024) argue that instead of this trade-off, recursive numeral systems optimise a trade-off between lexicon size (the number of unique morphemes) and average numeral length. Prasertsom et al. (2026) showed how this trade-off fails to take into account the regularity of numeral forms, and propose that recursive numeral systems are instead better thought of as being primarily subject to pressures to maximise *regularity* (how much structured and regular form re-use there is) and minimise *processing complexity* (how difficult it is to parse numerals), using measures based on the Minimum Description Length principle (following Brighton, 2003). This approach is able to delineate human recursive numeral systems from theoretically possible but unattested ones, with the natural systems exhibiting higher regularity and lower processing complexity.

Simplicity, learnability and cross-linguistic prevalence. Natural languages are highly diverse and yet exhibit clear shared features, a phenomenon known as *constrained variation* (Croft, 2002; Evans & Levinson, 2009). Theories accounting for constrained variation often appeal to pressures acting on language in learning (e.g. Culbertson, 2012), production and comprehension (e.g. MacDonald, 2013; Keogh, 2025), and communicative constraints (e.g. Kirby et al., 2015). Specifically *learning*

was identified by Prasertsom et al. (2026) as the potential mechanism through which recursive numeral systems are optimised for regularity. Regular systems permit the formation of compressed representations which have short description lengths, which make them easier to learn, more stable in intergenerational transmission, and therefore more widespread. There is a wealth of experimental evidence showing that simpler systems (i.e. those than can be described in a more compressed way in a given description language) are easier for humans to learn (e.g. Shepard et al., 1961; Raviv et al., 2021; Wang et al., 2025; Moreton et al., 2017; see Chater & Vitányi, 2003 for a broad review of simplicity principles in cognition), and that learners introduce regularity into irregular linguistic systems (e.g., Smith & Wonnacott, 2010; Keogh et al., 2024).

Studying the learnability of prevalent features with machine learning. Similar results also hold in machine learning. Recent work shows that more systematic, more human-like languages are learned more efficiently both by modern language models and by smaller models trained via SL (Kallini et al., 2024; Xu et al., 2026; Steinert-Threlkeld & Szymanik, 2020; Galke et al., 2024; Osmelak et al., 2026). Similarly, RL models in a two-agent setup have been used as a tool to study the emergence of communication (Foerster et al., 2016; Lazaridou et al., 2017; Mordatch & Abbeel, 2018); in these frameworks communicatively efficient languages, a common characteristic of human languages, are easier for the agents to converge to (Kågebäck et al., 2020; Carlsson et al., 2024). These results suggest that we can use neural networks to study the learnability of theoretically-possible linguistic systems in order to measure the degree to which different factors influence their learnability. We therefore hypothesise that regular numeral systems are easier for agents to learn—as quantified by how rapidly an agent learns to minimise errors given a fixed number of learning exposures—and remain agnostic as to whether SL or RL is a better proxy for human learning.

3 METHODS

To study the link between regularity and learnability, we first compile a dataset of 20,732 unique systems, both human and artificial (§3.1). We then develop RL and SL models that learn numeral-number mappings and use them to measure the learnability of different numeral systems (§3.2). These systems vary in multiple features, including regularity (§3.3), which allows us to determine to what extent a range of features including regularity affect learnability.

3.1 DATASET

We draw our sample from prior work on the efficiency of recursive numeral systems (Denić & Szymanik, 2024; Yang & Regier, 2025; Prasertsom et al., 2026). This dataset consists of attested systems and artificial systems that are generated for comparison purposes. In this prior work, artificial systems were generated using some variant of Hurford's grammar, a proposed universal template for natural language numerals (Hurford, 1975), represented with the following rules:

$$\begin{aligned} \mathit{Num} &\rightarrow D \mid \mathit{Phrase} \mid \mathit{Phrase} + \mathit{Num} \mid \mathit{Phrase} - \mathit{Num} \\ \mathit{Phrase} &\rightarrow M \mid \mathit{Num} * M \end{aligned}$$

where D stands for *digits* (monomorphemic numerals), and M for *multipliers* (bases). For example, Mandarin could be described with $D = \{1, ..., 9\}$ and $M = \{10\}$. For computational tractability and comparability, we follow all prior work in investigating only numerals in the range of 1 to 99.

Human systems. We use Denić & Szymanik's (2024) set of 128 natural language systems. As they were concerned with numeral length and lexicon size only, their annotation of the data does not take into account morpheme order and simply consistently order the D-M combination before the digit (e.g., they annotate *fifteen* as $10 + 5$ rather than $5 + 10$). As our irregularity metric is sensitive to morpheme order, we reannotated these systems to be order-sensitive[1]. Our reannotations contain 60 unique human recursive numeral systems, as some of the original 128 share the same representation.

Human-like systems. Previous work has also generated artificial systems. Some of these are "human-like," i.e., they were optimized to match human systems in some key aspects. We collect 50 optimal artificial numeral systems from Yang & Regier (2025), which are optimised for the

[1]We also identified other minor errors in annotation and fixed them to the best of our ability. We also introduced $\ominus$ as an operator in addition to $+, -, *$. This represents subtraction patterns found in some languages, where the subtrahend comes before the minuend (e.g. Kunama represents 9 as $1 \ominus 10$, 'one less than 10').

trade-off between average length and lexicon size following Denić & Szymanik (2024) but are also constrained (brute force) by using Hurford's grammar to produce systems with human features. We further add 23 human systems under Denić & Szymanik's original annotation (i.e., those that were annotated with fixed morpheme orders). These annotations were inaccurate but nevertheless are treated as *human-like* here because they match (or exceed) actual human systems in regularity.

Artificial systems. Other artificial systems are less similar to human languages. For example, we collect the 1701 unique optimal artificial systems from Denić & Szymanik (2024) which are optimised for average length and lexicon size, but are quite clearly non-human like, e.g., in their regularity. We add to those another 8,898 systems from the intermediate steps of Denić & Szymanik's optimisation algorithm, which exhibit a wide range of regularity. Finally, we also add 10,000 artificial systems sampled by Prasertsom et al. (2026). These systems contain digits and multipliers attested in human systems, but the combination of them and the numerals are randomly sampled.

3.2 Learning Framework

Figure 1 illustrates the general framework for measuring learnability. An artificial agent is trained to map numerals (the forms) to numbers (the meanings) under a fixed budget of number of interactions, either under RL or SL. Specifically, each agent is a long short-term memory network (LSTM; Hochreiter & Schmidhuber, 1997). The model processes a numeral (stripped of parentheses as natural language numerals do not contain them) by representing each morpheme within it with a learned size-5 embedding, and then processing the resulting sequence of embeddings one-by-one (hidden size of 10). From the last hidden state the network outputs a scalar μ, then multiplied by 100 and rounded to the nearest integer to convert it to a number between 1 and 99. This choice of architecture and size provides a simple neural model able to learn the task, which is quick and cheap to train.

When the agent is trained via RL, the predicted scalar mean is interpreted as the mean of a Gaussian policy $\mathcal{N}(\mu, \sigma)$. As the training proceeds, we decrease the standard deviation σ in the policy in order to encourage exploration early and exploitation later. The true prediction $\hat{n}$ is a sample from the Gaussian distribution, multiplied by 100 to achieve a number between 1 and 99. The reward function $\mathcal{R}$ is bounded, exponentially-decaying, and Gaussian-like: $\mathcal{R}(\hat{n}, n) = \mathbb{I}[0 < \text{int}(\hat{n}) < 100] \cdot \exp(-\alpha|\hat{n} - n|)$. Here int($\hat{n}$) approximates $\hat{n}$ to its nearest integer and $\alpha = 0.5$ determines the sharpness of the reward function. The first part of the reward function (the indicator function) ensures predictions outside the range 1-99 are not rewarded. The second part penalises predictions that are far away from the target. The RL agent is trained via policy-gradient objective REINFORCE (Williams, 1992) over a fixed budget of 30,000 epochs (enough for the learner to become proficient on any system from our sample), sampled with replacement from the *training distribution* $P(n)$. When the agent is trained via SL, the mean is instead directly taken as the predicted value, which is likewise multiplied by 100 to obtain a number in the range (1-99). The true numbers are used as labels to train the network via MSE loss, where distances are calculated between the predicted numbers and the labels. We also decrease the budget to 3,000 epochs as the supervised objective makes learning converge faster. Other hyperparameters can be found in Appendix A.

During evaluation, we use the network's output directly in both regimes (again, multiplied by 100 and rounded to the nearest integer), effectively taking the normal distribution's mean as prediction in RL instead of sampling. This avoids noisy guesses in the early stages of training caused by large values of σ. This setup effectively frames the problem of learning numeral systems as a regression problem, rather than classification as in prior literature (Guo et al., 2019; Carlsson et al., 2021; Silvi et al., 2025). This already provides structure to the agent's output so that it can focus on learning numerals without also having to learn the order of numbers. We run our pipeline 100 times per numeral system, randomising the network's initial weights and dataset batches.

3.3 Metrics

We use $\mathcal{N} = \{(n, x_n)\}_{n=1}^{99}$, to denote a numeral system in the range 1–99, where x_n is the underlying morphosyntactic representation of the numeral[2] referring to number n.

[2]Here, we always assume that arithmetic combinators like $+, -, *$ are present in the representation, regardless of whether they appear overtly or not; i.e. we treat the ones that are not realised as *null/zero* morphemes.

Learnability. Let $\hat{y}_t(x)$ be the number predicted by the agent for numeral x after epoch t. Then, at each epoch, we compute exact-match accuracy on the test set which we then sum together to obtain the overall area under the test-accuracy curve. Thus our learnability metric is $\mathrm{L}_{\mathcal{N}} = \sum_{t=1}^{T} \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,n)\in\mathcal{D}_{\text{test}}} \mathbb{I}\left[\hat{y}_t(x) = n\right]$, where the size of the per-epoch test set $|\mathcal{D}_{\text{test}}|$ is 99. This metric ranges between 0 (i.e. a learner that outputs wrong predictions for every sample in every epoch) and the number of epochs of training-test T (30,000 for RL and 3,000 for SL). We take the point mean for learnability across 100 runs to obtain more accurate estimates.

Irregularity. We use the same formalisation of *ir*regularity in Prasertsom et al. (2026). The measure rests on the intuition that regular systems can be represented with smaller grammars, where a numeral grammar is represented by a minimal partial deterministic finite-state automaton (DFA) that generates all and only valid numeral strings. Figure 6 in Appendix B.1 gives an example of the inferred DFA for the highly regular Mandarin numeral system. The *irregularity* is the size of the numeral grammar, measured by the number of bits required to encode the DFA: $\mathrm{Ir}_{\mathcal{N}} = |Z|\left(2 \cdot \log_2 |S| + \log_2 |\Sigma|\right) + \log_2 |S| + |S|$ where [blue] is the number of bits required to encode the transition table, where Z is the set of transitions, S the set of states and Σ the set of symbols (morphemes), [teal] and [violet] are the numbers of bits for encoding the identity of initial state and accepting states respectively. Our main prediction is that this measure should negatively correlate with learnability: more irregular systems are harder to learn.

Processing complexity. Prasertsom et al. (2026) argues that *processing complexity* also constrains natural recursive numeral systems. This measure quantifies how hard it is to parse or generate a numeral, and can be quantified by calculating the average amount of information needed to determine a unique path through the DFA for a specific numeral-number (i.e. form-meaning) pair (Brighton, 2003). We compute processing complexity following Prasertsom et al. (2026) (cf. Appendix B.2 for a full definition.) As processing complexity is concerned with difficulty in generation/parsing, we expect that it will also, at least indirectly, affect learnability.

Average numeral length. Denić & Szymanik (2024) suggest *average numeral length* as a key pressure that shapes recursive numeral systems. It is essentially a weighted average of the number of morphemes per numeral: $\overline{\mathrm{NL}}_{\mathcal{N}} = \sum_{n=1}^{99} P(n)\,\mathrm{NL}(n;\mathcal{N}) = \sum_{n=1}^{99} P(n)\,|x_n|$, with $P(n) \propto n^{-2}$. As an example, the length of English *fifty-three* is $\mathrm{NL}(53; \textit{English}) = |5 * 10 + 3| = 5$. In general, longer numerals are expected to be harder to learn simply because there is more to memorise.

Lexicon size. Denić & Szymanik (2024) argue that the pressure for shorter numerals trades off with *lexicon size*, defined as the number of unique morphemes in a given system (excluding arithmetic combinators). Intuitively, one might expect that a system with a larger lexicon size is harder to learn as there is more to memorise.

Local edit distance between numerals. While irregularity captures the size of the grammar of whole systems, one aspect it does not capture is *where* on the number line the regularity lies. In natural languages, regular rules tend to apply over consecutive ranges without gaps, resulting in similar-looking consecutive numerals (Yang & Regier, 2025). We might expect that human (and machine) learners are sensitive to this kind of local regularity, given that numbers are ordered and local regularity allows learners to use nearby numerals as cues for inference[3]. We define it as $\mathrm{LD}_{\mathcal{N}} = \frac{1}{5}\sum_{w=5}^{9}\sum_{k=1}^{99-w+1} D_{k,w}$, where $D_{k,w}$ is the mean normalised Levenshtein distance among the w consecutive numerals beginning at number k, and we vary w between 5 and 9 to capture different aspects of this regularity. More details on its definition are in Appendix B.3.

PEMDAS deviation. PEMDAS[4] is a convention for precedence order of arithmetic operations. Although natural language numerals do not contain parentheses, their underlying representations do. In some systems, stripping parentheses from the numerals and applying PEMDAS rules does not recover the original number. For example, the expression in Kunama for 61, *ʃēb-kōntálle ēlláá* 'ten five one one', represented as $10 * (5 + 1) + 1$, when removing parentheses, is evaluated by PEMDAS as $10 * 5 + 1 + 1 = 52$. Such deviations constitute a different form of irregularity, namely

For example, French's morphosyntactic representation for the numeral referring to $n = 85$ (*quatre-vingt-cinq*) is $x_n = 4 * 20 + 5$, where the combinators $*$ and $+$ are expressed with null morphemes.

[3] In humans, this could in turn be due to a general bias for form-meaning correlations (*structural iconicity*, cf. Haiman, 1980), or due to the fact that numerals are learned through counting routines that help them attend to structure of consecutive numerals (Holt & Barner, 2025), or both.

[4] Short for Parentheses, Exponents, Multiplication, Division, Addition, Subtraction.

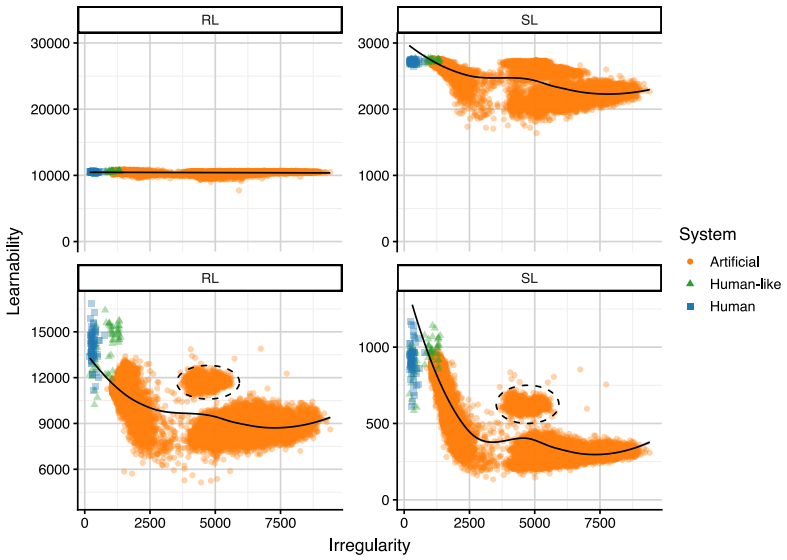


Figure 2: Irregularity and learnability of artificial, human-like and human numeral systems with skewed training and skewed (first row) or uniform (second row) test distribution under RL (left) and SL (right). The dashed-border marks the outlier cluster from §4.2. Trend lines are LOESS regression fits. The y-axis for each regime is restricted to the range where the learnability measures lie.

nonconformity to a general precedence convention: the extent to which a numeral's intended compositional structure cannot be recovered by applying PEMDAS to its unparenthesized representation. While this in principle does not imply that the numeral system must be internally inconsistent as it may simply follow a different order of operations, in our case Hurford's grammar systematically has a PEMDAS bias (see details in Appendix B.4) and deviation from PEMDAS thus implies inconsistency. We define this property as $\mathrm{DP}_{\mathcal{N}} = \sum_{n=1}^{99} \mathbb{I}\left[E_{\text{PEMDAS}}\left(\tilde{x}_n\right) \neq n\right]$, where $E_{\text{PEMDAS}}\left(\tilde{x}_n\right)$ evaluates the parenthesis-stripped numeral $\tilde{x}_n$ under PEMDAS rules.

4 Experiments

A possibility is that regularity and learnability should correlate regardless of training regime. Experiment 1 (§4.1) shows that this is not the case. Importantly, this turns out to be directly related to how learnability is measured: in Experiment 1 we train and test learners on the numerals drawn from a frequency-skewed distribution; in Experiment 2 we train them on this distribution but test them on generalisation to the whole system (§4.2). In the latter case, the expected relationship is clear and robust. We then ask how regularity fares against other properties that affect learnability (§4.3).

4.1 Experiment 1: Learnability and frequency of numbers

Prior work (Xu et al., 2020; Silvi et al., 2025) has assumed power law distributions for both training, $P(n) \propto n^{-2}$, and test distributions $Q(n) \propto n^{-2}$, following cross-linguistic frequencies of number words (Dehaene & Mehler, 1992). However, in Experiment 1, we find that there is virtually no difference in learnability among our sample of numeral systems under RL, and only a very weak trend under SL (Figure 2, first row and Appendix C for more details). Under this highly-skewed test distribution, about 75% of test samples will consist of numbers 1 and 2 alone, and about 90% will consist of numbers in the range 1-6. This means that, at test, the agent only has to perform well on the lower ranges of numerals in order for that system to score high on learnability; those same numerals are also highly frequent in the training samples. Learnability in this setting will thus primarily measure the *learnability of frequently-used signals*. As Prasertsom et al. (2026) point out, what is characteristic of recursive numeral systems is that they can precisely express *any* integer, including large, rare numbers in the tail of the distribution (cf. Pitt et al., 2022).

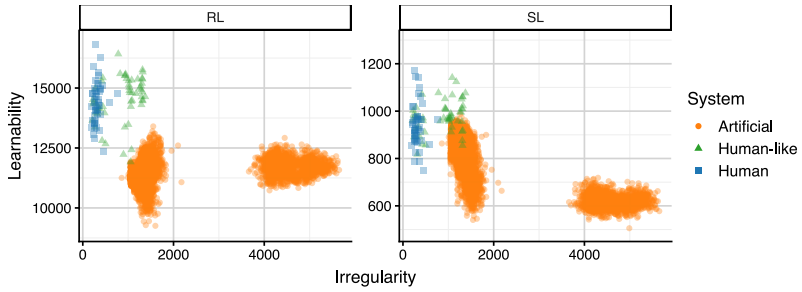


Figure 3: Irregularity and learnability under the two training regimes for systems with zero PEMDAS deviation. The cluster analysed in §4.2, point (2) is to the right of both panels. The figures suggest that regularity explains high learnability above and beyond PEMDAS deviation.

4.2 Experiment 2: Learnability as generalisability

In Experiment 2 we therefore retain the empirically-motivated assumption that the training distribution $P(n)$ is a power-law based on the frequency of use, as an analogue to human learning, but assume that $Q(n)$ is flat, i.e., all numerals are tested exactly once per iteration, testing generalisation to rarer numerals. Figure 2 (second row) shows the learnability under RL and SL with this uniform test distribution. Importantly, unlike in Experiment 1, there are clear (non-linear) trends between regularity and learnability: more regular systems allow for more successful generalisation, broadly confirming a refined version of our prediction. Three features of these results stand out.

1) The relationship between regularity and learnability is non-linear. Regularity has the clearest effect in the high-regularity region, but as systems grow increasingly irregular, regularity has less effect on learnability. This is likely because, for any regular pattern to be useful, there must be enough evidence for that regularity to be identified and exploited. When a system is highly irregular, an increase in regularity does not help much with the learning (cf. also Raviv et al., 2021).

2) Conformity to PEMDAS and monomorphemicity in low numerals (partially) explain learnability. Regularity alone does not explain all the results. Specifically, there is a cluster of artificial systems that do not follow the general learnability curve for either RL or SL (marked by dashed ellipses in Figure 2, second row). These are systems drawn from the random baseline systems sampled in Prasertsom et al. (2026), and their relatively high learnability is likely due to two features: first, these systems are **consistent** in that they all evaluate correctly under PEMDAS for all numbers, as shown in Figure 10 in Appendix D. Secondly, like many human languages, these systems consistently have monomorphemic numerals for low numbers (1-9), unlike other artificial systems with the same level of regularity but lower learnability. With monomorphemic low numerals (and complex high numerals), an agent can learn primitives more easily as these are also frequently occurring in the training data, and then capitalise on this to learn higher numerals composed of such primitives. The opposite makes a system harder to learn. Crucially, however, neither property alone explains the higher learnability of human recursive numeral systems. While PEMDAS deviation might overall correlate well with learnability (Figure 9 in Appendix D), Figure 3 shows irregularity and learnability for the $9{,}211$ systems with zero PEMDAS deviation. Here, the human(-like) systems remain sharply separated from the artificial ones: under RL, their median learnability is $14{,}506$, above the maximum of $13{,}400$ attained by any of the $9{,}108$ artificial systems in this subset. Moreover, 86% of these artificial systems fall below even the least learnable human(-like) system. What separates human numeral systems is **regularity**: within this PEMDAS-conforming subset the human(-like) systems have a median irregularity of 377, against $1{,}302$ for the artificial ones.

3) RL and SL differ in the level of efficiency at which they learn different systems. The separation in learnability between the high-regularity human and human-like systems is greater when the artificial agents are trained via RL rather than SL ($\approx 10\%$ of the human(-like) systems overlapped with the artificial ones for RL vs. $\approx 38\%$ for SL; see Figure 4). This might be because SL is not as sensitive to regular internal structure as RL, which would be consistent with previous findings in emergent communication that RL agents are more sensitive to differences in – and introduce selection for – compositionality (Cheng et al., 2023). Moreover, Appendix E shows that SL agents tend to learn highly unnatural, large-lexicon systems as efficiently or even more efficiently than human systems;

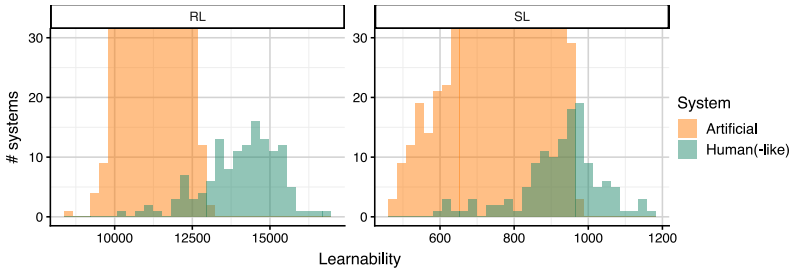
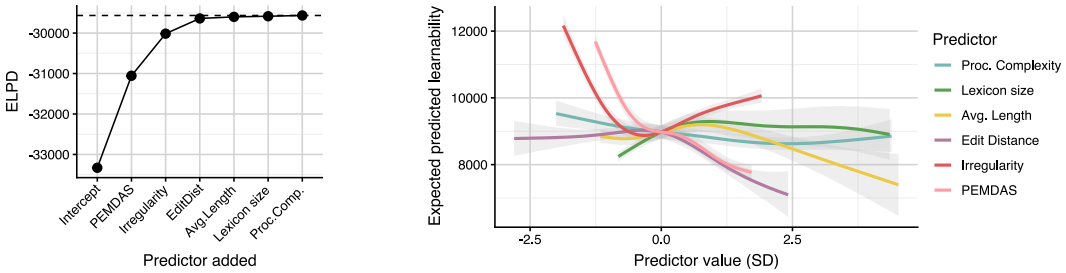


Figure 4: Learnability of artificial and human(-like) systems under RL and SL. The upper bound of the y-axis is limited to 30 to show the distribution of the human(-like) systems clearly.



(a) (b)

Figure 5: **(a)** Expected log pointwise predictive density of submodels of growing sizes under RL. Dots correspond to the best-performing submodels at specific model sizes, starting with an intercept-only model on the left, moving to the best one-predictor model (PEMDAS), the best two-predictor model (PEMDAS and irregularity), and so on up to the full six-predictor model. The dashed line indicates the ceiling performance (i.e., that of the full model). **(b)** Expected predicted learnability under RL for different predictors over standardised scales. The ranges of standardised predictor values are different for different predictors due to different probability density concentrations.

RL in contrast learns these less efficiently than most human(-like) systems. This supports the idea that highly idiosyncratic systems are easy to learn for SL, while RL is better at exploiting structure in the data, perhaps indicating RL as a better analogue to human learning in our setting.

4.3 Multivariate regression

The results above suggest that learnability is potentially determined by multiple factors, including regularity, PEMDAS adherence, and the distribution of monomorphemic forms in the lexicon. One might wonder, then, to what extent each factor contributes to learnability, and how important regularity (as one of a constellation of properties of counting systems) actually is. Here, we investigate the importance of different predictors (including regularity) as defined in §3.3 in determining the learnability of recursive numeral systems. Here, we focus on RL; SL results are in Appendix H.2.

First, as our sample set of numeral systems covers the space of possible systems unevenly, we attempt to correct as best as we can for these sample biases, following the thinning process in Appendix F. We end up with 3,738 systems for our regression model. To allow for non-linearity, we fit Bayesian generalised additive models (GAMs, Wood, 2017) predicting learnability from our formal metrics defined in §3.3. We take a projection predictive approach (Pavone et al., 2023), using leave-one-out cross-validation on GAMs to identify relevant variables (Catalina et al., 2022). This is essentially model comparison: it identifies the most parsimonious model that has the same predictive performance (measured as the sum of expected log pointwise predictive densities under leave-one-out cross-validation, or ELPD (Vehtari et al., 2017)) as the model that includes all the potentially relevant predictors. This allows us to identify whether e.g. adding lexicon size to a model that already incorporates irregularity meaningfully improves predictive accuracy. This method is robust against predictor correlations and addresses the problem of inference bias induced by the selection process (McLatchie et al., 2025; Vehtari, 2017). More details of our method are in Appendix G.

Figure 5a shows the predictive performance as given by ELPD of GAM submodels of growing sizes. PEMDAS adherence, regularity and local edit distance together explain most of the trend. The variable selection procedure, however, suggests that all six predictors have a non-trivial contribution, as no submodel smaller than the full six-term model achieves a comparable performance. Figure 5b shows the effect sizes of the full model (see also Table 1 in Appendix H.1 for estimates). Irregularity and PEMDAS deviation have the largest effect sizes, especially in the high-regularity/low-PEMDAS-deviation region. Moving down from the most regular systems by one SD regularity unit reduces learnability by about 2,500 (from over 12,000 to about 9,500). There is a slight increase in learnability towards the tail of the regularity effect line, which is likely due to a combination of the idiosyncratic properties of the outlier cluster that strengthens the overall learnability towards the tail (as discussed in §4.2), and an artefact of the sample, where systems in the 2800–3500 irregularity range are sparse. PEMDAS deviation also has a similar effect size in the low-PEMDAS-deviation region. We also observed a non-linear negative effect with local edit distance and average length, where systems become less learnable once the consecutive numerals become sufficiently different within local windows, or when they become too long. The effects of other predictors are modest.

Overall, these results confirm the key findings that 1) regularity is a primary (though by no means only) predictor of learnability, 2) its effect is strongest in the high-regularity region, and 3) it contributes to learnability in conjunction with other predictors, including PEMDAS deviation and lexicon size and average numeral length, corroborating results from Denić & Szymanik (2024).

5 Discussion and Conclusion

We asked whether the cross-linguistic prevalence of regularity in recursive numeral systems could be explained by learnability, i.e., whether regularity makes numeral systems easier to learn, and therefore more stable across repeated inter-generational transmission. We found that learnability correlates with regularity, but crucially only when we evaluate the capacity of learners to generalise to infrequently-encountered forms from skewed input. Learnability is strongly affected by regularity — though less so the more irregular systems become — and by multiple other factors. Our results also indicate that SL and RL find different features easier to exploit when learning.

Input order might also influence learnability, particularly for numeral systems. Our models were trained via a random sampling process; however, a growing body of work in numeral systems acquisition suggests that counting routines play an important role in learning recursive numeral systems (Carey, 2000; Carey & Barner, 2019; Fuson, 2012), even a necessary one (Pitt et al., 2022). Given that consecutive numerals often make use of similar rules (as formalised by our local regularity measure, §3.3), counting routines, which require learners to go through consecutive numerals, could additionally strengthen the regularity advantage in learning. Future work could investigate this by modelling our task as a Markov Decision Process (Sutton & Barto, 2018), or by testing learning strategies like *curriculum learning* (Bengio et al., 2009), which has been shown to help learning compositional tasks more efficiently in humans and animals (Lee et al., 2024).

Beyond regularity, our results highlight the multiply-determined nature of learnability for recursive numeral systems. Recent work (Carlsson, 2024; Osmelak et al., 2026) suggests using learnability as a direct substitute for complexity measures in analysing linguistic systems. This may be reasonable if one wishes to simply quantify complexity in some general sense. However, our results suggest that this substitution approach may be of limited use if one's goal is to actually understand which *aspects of complexity* of the systems of interest make them more or less learnable.

More generally, our results replicate prior findings (Cheng et al., 2023) that SL is more efficient at learning idiosyncratic forms, whereas RL appears to be more sensitive to structure. More experimental work for potential optimization and loss-geometry confounding factors is worth exploring and left for future work. In general, research measuring the learnability of linguistic systems with neural networks should therefore explore different regimes. Furthermore, as modern language models are trained with a combination of RL and SL, NLP practitioners may find interest in exploring the kinds of languages RL and SL promote to train the optimal language models for their purposes.

AI USE STATEMENT

We used generative AI tools to assist with software engineering and presentation of the results. AI assistance was used in the process of writing portions of the code used to run and evaluate the experiments and to visualise their results, and to polish and document the released code, metrics and numeral data. AI tools were not used to develop the research questions, design the experiments or metrics, interpret the findings or write the paper. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, figures and released artifacts.

ETHICS STATEMENT

The authors feel that their paper submission does not raise questions regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

All code and data needed to reproduce our experiments are available at `https://anonymous.4open.science/r/generalisability_numerals-F212`. This includes the training code for both regimes, the metrics table and morphosyntactic forms for all 20,732 numeral systems, and the R code for the regression analysis. Section 3.2 and Appendix A specify the model and training configuration, Section 3.3 and Appendix B define every metric, and Appendix F documents the sampling correction applied before regression modelling.

ACKNOWLEDGMENTS

Removed for anonymous review.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Henry Brighton. *Simplicity as a Driving Force in Linguistic Evolution*. Phd thesis, The University of Edinburgh, 2003. URL http://hdl.handle.net/1842/23810.

Paul-Christian Bürkner. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54, 2021. doi: 10.18637/jss.v100.i05.

Susan Carey. The origin of concepts. *Journal of Cognition and Development*, 1(1):37–41, 2000.

Susan Carey and David Barner. Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, 23(10):823–835, 2019.

Emil Carlsson. *Reinforcement Learning: Efficient Communication and Sample Efficient Learning*. Doctoral dissertation, Chalmers University of Technology, Gothenburg, Sweden, 2024.

Emil Carlsson, Devdatt Dubhashi, and Fredrik D Johansson. Learning approximate and exact numeral systems via reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

Emil Carlsson, Devdatt Dubhashi, and Terry Regier. Cultural evolution via iterated learning and communication explains efficient color naming systems. *Journal of Language Evolution*, 9(1-2): 49–66, 11 2024. ISSN 2058-458X. doi: 10.1093/jole/lzae010. URL https://doi.org/10.1093/jole/lzae010.

Alejandro Catalina, Paul-Christian Bürkner, and Aki Vehtari. Projection predictive inference for generalized linear and additive multilevel models. In *International conference on artificial intelligence and statistics*, pp. 4446–4461. PMLR, 2022.

Nick Chater and Paul Vitányi. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22, 2003. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(02)00005-0. URL https://www.sciencedirect.com/science/article/pii/S1364661302000050.

Emily Cheng, Mathieu Rita, and Thierry Poibeau. On the correspondence between compositionality and imitation in emergent neural communication. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12432–12447, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.787. URL https://aclanthology.org/2023.findings-acl.787/.

William Croft. *Typology and universals*. Cambridge university press, 2002.

Jennifer Culbertson. Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, 6(5):310–329, 2012. URL https://doi.org/10.1002/lnc3.338.

Stanislas Dehaene and Jacques Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29, 1992. URL https://doi.org/10.1016/0010-0277(92)90030-L.

Milica Denić and Jakub Szymanik. Recursive numeral systems optimize the trade-off between lexicon size and average morphosyntactic complexity. *Cognitive Science*, 48(3):e13424, 2024. doi: https://doi.org/10.1111/cogs.13424.

Nicholas Evans and Stephen C Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448, 2009.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2145–2153, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Karen C Fuson. *Children's counting and concepts of number*. Springer Science & Business Media, 2012.

Lukas Galke, Yoav Ram, and Limor Raviv. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(1), December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-55158-1. URL http://dx.doi.org/10.1038/s41467-024-55158-1.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407, May 2019. ISSN 13646613. doi: 10.1016/j.tics.2019.02.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661319300580.

Diego Guerrero, Jihyun Hwang, Brynn Boutin, Tom Roeper, and Joonkoo Park. Is thirty-two three tens and two ones? the embedded structure of cardinal numbers. *Cognition*, 203:104331, 2020.

Shangmin Guo, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *CoRR*, abs/1910.05291, 2019. URL http://arxiv.org/abs/1910.05291.

John Haiman. The iconicity of grammar: Isomorphism and motivation. *Language*, 56(3):515–540, 1980. doi: 10.2307/414448.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Sebastian Holt and David Barner. Learning a novel number system: The role of compositional rules and counting procedures. *Cognitive Science*, 49(6):e70071, 2025. URL https://doi.org/10.1111/cogs.70071.

James R. Hurford. *The Linguistic Theory of Numerals*. Cambridge University Press, Cambridge, 1975.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14691–14714, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.787. URL https://aclanthology.org/2024.acl-long.787.

Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental science*, 10(3):307–321, 2007.

Aislinn Keogh. *Language Adapts to Pressures from Production: Experimental and Computational Evidence*. PhD thesis, The University of Edinburgh, 2025.

Aislinn Keogh, Simon Kirby, and Jennifer Culbertson. Predictability and variation in language are differentially affected by learning and production. *Cognitive Science*, 48(4):e13435, 2024.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015. URL https://doi.org/10.1016/j.cognition.2015.03.016.

Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, and Asad Sayeed. A reinforcement-learning approach to efficient communication. *PLOS ONE*, 15(7):1–26, 07 2020. doi: 10.1371/journal.pone.0234894. URL https://doi.org/10.1371/journal.pone.0234894.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hk8N3Sclg.

Jin Hwa Lee, Stefano Sarao Mannelli, and Andrew Saxe. Why do animals need shaping? a theory of task composition and curriculum learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Maryellen C MacDonald. How language production shapes language form and comprehension. *Frontiers in psychology*, 4:226, 2013.

Yann McLatchie, Sölvi Rögnvaldsson, Frank Weber, and Aki Vehtari. Advances in projection predictive inference. *Statistical Science*, 40(1):128–147, 2025.

Stoyan Mihov and Klaus U. Schulz. *Finite-State Techniques: Automata, Transducers and Bimachines*. Cambridge University Press, 1 edition, 2019. ISBN 978-1-108-75694-5 978-1-108-48541-8. doi: 10.1017/9781108756945.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Elliott Moreton, Joe Pater, and Katya Pertsova. Phonological concept learning. *Cognitive science*, 41(1):4–69, 2017.

Doreen Osmelak, Yang Xu, Michael Hahn, and Kate McCurdy. Systematicity between forms and meanings across languages supports efficient communication, 2026. URL https://arxiv.org/abs/2601.17181.

Federico Pavone, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Using reference models in variable selection. *Computational Statistics*, 38(1):349–371, 2023.

Juho Piironen, Markus Paasiniemi, Alejandro Catalina, Frank Weber, Osvaldo Martin, and Aki Vehtari. projpred: Projection predictive feature selection, 2025. URL https://mc-stan.org/projpred/. R package version 2.10.0.

Benjamin Pitt, Edward Gibson, and Steven T Piantadosi. Exact number concepts are limited to the verbal count range. *Psychological Science*, 33(3):371–381, 2022.

Eva Portelance and Masoud Jasbi. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001, 2024. doi: https://doi.org/10.1111/lnc3.70001. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.70001.

Ponrawee Prasertsom, Andrea Silvi, Jennifer Culbertson, Devdatt Dubhashi, Moa Johansson, and Kenny Smith. Recursive numeral systems are highly regular and easy to process. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4873–4885, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.226. URL https://aclanthology.org/2026.eacl-long.226/.

Limor Raviv, Marianne de Heer Kloots, and Antje Meyer. What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210:104620, 2021. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2021.104620. URL https://www.sciencedirect.com/science/article/pii/S0010027721000391.

Roger N Shepard, Carl I Hovland, and Herbert M Jenkins. Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1, 1961.

Andrea Silvi, Jonathan Thomas, Emil Carlsson, Devdatt Dubhashi, and Moa Johansson. Learning efficient recursive numeral systems via reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025. URL https://escholarship.org/uc/item/3cc5053z.

Kenny Smith and Elizabeth Wonnacott. Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449, 2010.

Kenny Smith, Simon Kirby, and Henry Brighton. Iterated learning: A framework for the emergence of language. *Artificial life*, 9(4):371–386, 2003.

Kenny Smith, Clem Ashton, and Helen Sims-Williams. The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023. URL https://escholarship.org/uc/item/1mz1q97f.

Shane Steinert-Threlkeld and Jakub Szymanik. Ease of learning explains semantic universals. *Cognition*, 195:104076, 2020. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2019.104076. URL https://www.sciencedirect.com/science/article/pii/S0010027719302495.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Aki Vehtari. Bayesian version of Does Model Averaging Make Sense?, 2017. URL https://users.aalto.fi/~ave/casestudies/collinear/collinear.html. Modified 2026-03-27.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

Fang Wang, Simon Kirby, and Jennifer Culbertson. The learning bias for cross-category harmony is sensitive to semantic similarity: Evidence from artificial language learning experiments. *Language*, 101(1):109–150, 2025.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36, 2011.

Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2017. doi: 10.1201/9781315370279.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. Can language models learn typologically implausible languages? *Transactions of the Association for Computational Linguistics*, 14:588–611, 2026.

Yang Xu, Emmy Liu, and Terry Regier. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, 4:57–70, 2020. URL https://doi.org/10.1162/opmi_a_00034.

David Yang and Terry Regier. Re-examining the tradeoff between lexicon size and average morphosyntactic complexity in recursive numeral systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025. URL https://escholarship.org/uc/item/5k8646v4.

A Learner configuration

Both RL and SL agents are trained with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, no weight decay), with an initial learning rate of 10^{-3} and a slow cosine annealing. For each training epoch they are exposed to 5 batches of 32 numerals. The mean μ is obtained by passing the last hidden state through a linear layer and a sigmoid, rescaled to $[-0.1, 1.1]$ so that the targets for 1 and 99 do not lie in the saturated part of the sigmoid. In RL, the standard deviation σ of the policy decays linearly from 0.15 to 0.01 over the first 10,000 epochs and then stays constant. REINFORCE uses a moving-average baseline of the batch-mean reward, $b \leftarrow 0.99\, b + 0.01\, \bar{\mathcal{R}}$ with $b_0 = 0$, which is subtracted from the reward in the policy gradient.

B Detailed metrics definitions

B.1 Irregularity

Figure 6 gives an example of an inferred DFA that generates all and only valid Mandarin numerals, a highly regular system. Note that we use the same algorithm (Mihov & Schulz, 2019) and implementation (`automata-lib` in Python) for minimal DFA inference as Prasertsom et al. (2026).

B.2 Processing complexity

We compute the processing complexity $L(N \mid G)$ of recursive numeral system $\mathcal{N}$ for a set of numbers $\{x_n\}_{n=1}^{99}$ as:

$$\sum_{n=1}^{99} \Big\{ P(n) \cdot \sum_{j=1}^{|S_n|} \big(\log_2 |Z(s_j)| + F(s_j) \big) \Big\} \tag{1}$$

where $P(n)$ is the prior for number n, S_n are the states through which parsing the numeral x_n leads, $Z(s_j)$ is the set of outgoing transitions from the jth state in the parsing of the numeral x_n, and F is an indicator function that evaluates to 1 if $s_j \in S_n$ is a final state. We use $P(n) \propto n^{-2}$, derived from cross-linguistic frequencies of number words (Dehaene & Mehler, 1992).

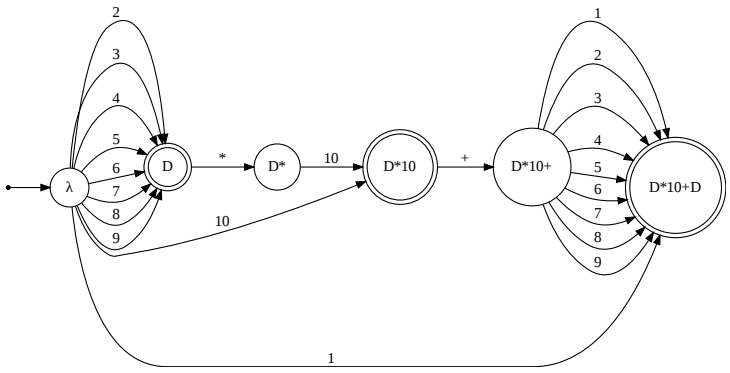


Figure 6: Inferred minimal partial DFA representing the Mandarin recursive numeral grammar. Transitions (arrows) represent output morphemes. States (circles) are labelled roughly according to the relevant grammar rules they capture. Double borders indicate accepting states. Generating the numeral for 20 ($2 * 10$) would involve going from the initial state (λ) to D (emitting 2), from D to $D*$ (emitting $*$), and from $D*$ to $D * 10$ (emitting 10) and terminate the machine there. D is a mnemonic for the category of digits (i.e. morphemes excluding the base of 10), such that $D * 10$ represents a state equivalent to $2 * 10$, $3 * 10$, $4 * 10$, ...

B.3 Local edit distance

To define local edit distance, we first define the normalised pairwise distance between two numerals of numbers i and j as

$$d(i,j) = \frac{\text{Lev}\,(x_i, x_j)}{\max\,(|x_i|, |x_j|)}, \tag{2}$$

where $\text{Lev}(x_i, x_j)$ is the Levenshtein edit distance (Levenshtein, 1966) between morpheme sequences of numerals x_i and x_j. For a window size w, let $W_{k,w} = \{k, k+1, \ldots, k+w-1\}$ be the k-th window of w consecutive numerals. We calculate the average pairwise distances for a window of size w as $D_{k,w} = \frac{2}{w(w-1)} \sum_{i,j \in W_{k,w}} d(i,j)$, with $i < j$. To smooth out the effect of specific window sizes, we average over window sizes 5 to 9. Finally, the local edit distance of a numeral system $\mathcal{N}$ is defined as $\text{LD}_{\mathcal{N}} = \frac{1}{5} \sum_{w=5}^{9} \sum_{k=1}^{99-w+1} D_{k,w}$. Summing over the windows within each size, rather than averaging over them, means that narrower windows contribute slightly more to the total, since there are more of them; this captures the higher importance of more local relations.

B.4 PEMDAS deviation

Here we provide an intuition as to why PEMDAS is the correct baseline to calculate errors against.

Recall that the grammar used to generate all artificial numeral systems is some variant of Hurford's grammar. This grammar first generates a $Phrase$, which involves multiplication ($Phrase \rightarrow Num * M$), before addition or subtraction of a digit (e.g., $Phrase + D$), following PEMDAS. Thus, except cases of complex multiplicands (e.g., $5 + 1 * 10$ for 60, viz. $(5 + 1) * 10$) and complex subtrahends (e.g., $38 - 61 - 57$ for 34, viz. $38 - (61 - 57)$), the correct way to evaluate numerals from Hurford's grammar stripped of parentheses is to follow PEMDAS. This essentially means that the grammar introduces a PEMDAS evaluation bias: the most common correct evaluation pattern that could be learned from numerals will be PEMDAS-conforming, and any system that has other patterns of evaluation mixed in will be potentially more difficult to learn. Note, too, that PEMDAS-conforming numerals result in higher penalty for multiplication errors than addition/subtraction errors (the evaluated number would tend be further from the correct number under multiplication error vs. addition cases).

This likely has a cognitive analogue for human learners. Like machine learning models, humans can extract consistent, higher-level patterns from data (i.e., overhypothesis learning; Kemp et al., 2007) and in this case are provided additional cues for order of operations from the forms (e.g., English

uses the allomorph *-ty* when using 10 as a multiplier, but *te(e)n* when it does not). In other words, this metric approximately measures the consistency of the correct evaluation pattern (which happens to be PEMDAS-conforming because of Hurford's grammar), which machines and humans alike can exploit.

C Learnability and regularity

Here we show detailed results of our experiments under the naïve assumption that training and test distributions are both the power law distribution $P(n) \propto n^{-2}$, $Q(n) \propto n^{-2}$. Figure 7 reports once more the learnability (as the area under the test accuracy curve) of different recursive numeral systems in our sample. Perhaps surprisingly, there is virtually no difference in learnability among the systems under RL, but there is a weak trend under SL, though all systems are highly learnable (learnability scores between slightly below 2000 and the maximum of 2,772).

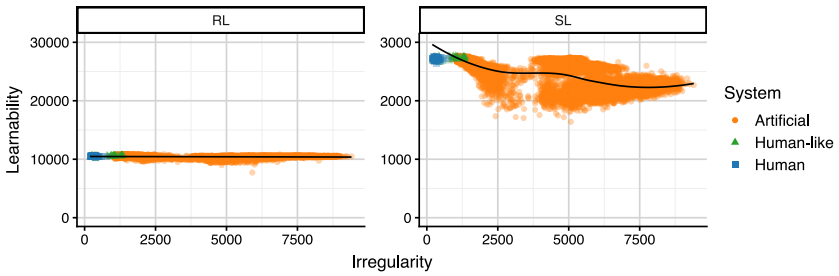


Figure 7: Irregularity and learnability (area under the test accuracy curve) of artificial, human-like and human recursive numeral systems with a power law test distribution under reinforcement learning (left) and supervised learning (right). Max. learnability is 30,000 for RL, and 3,000 for SL. Trend line is an OLS linear regression fit for RL and LOESS regression fit for SL.

However, we also observe a clear difference in the variance between RL and SL ($\hat{\sigma}_{\text{RL}} = 0.005$ vs. $\hat{\sigma}_{\text{SL}} = 0.078$ after normalising so that the max. learnability is 1 in each case). Figure 8 suggests that the variance in SL, which is absent in RL, is explained primarily by average numeral length and to some extent by lexicon size. The fact that high lexicon size confers learnability advantage may appear surprising at first, but note that these highly learnable systems are essentially systems that have dedicated monomorphemic numerals for most numbers in the range 1–99. This suggests that SL's learning here is driven by surface-level characteristics (length and lexicon size), possibly indicative of reliance on memorisation by SL: It is easier to map short, simple numerals to the correct numbers.

Overall, the results suggest that the setup here does not accurately capture the pressures that shape the regularity of human recursive numeral systems.

D Learnability and PEMDAS deviation

Figure 9 shows the overall relationship between learnability, PEMDAS deviation and irregularity (represented as the blue gradient). Figure 10 shows the positions of the outlier cluster when plotted against irregularity and PEMDAS deviation, under RL and SL. We identified the cluster by sub-setting the data from the SL learnability results under Experiment 2 to include randomly sampled systems that score over 500.

E Learnability and lexicon size

Figure 11 shows the relationship between learnability, irregularity and lexicon size (colour coded in a blue gradient). Note how high lexicon size numeral systems are generally considered fairly regular by our regularity measure, and are either almost as learnable (cf. RL results in the left panel of Figure 11), or even as learnable (cf. SL results in the right panel of Figure 11) with human and human-like numeral systems (the most regular systems in the upper left corner of both panels).

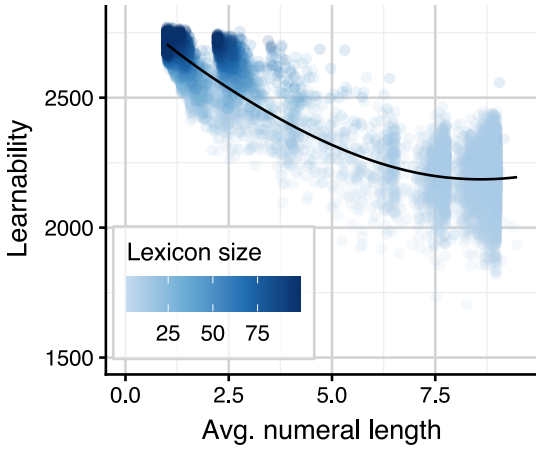


Figure 8: Average numeral length, lexicon size, and learnability (area under the test accuracy curve) of recursive numeral systems under supervised learning with a power law test distribution. Trend line is a LOESS regression fit. 37 data points with extremely high numeral lengths are excluded.

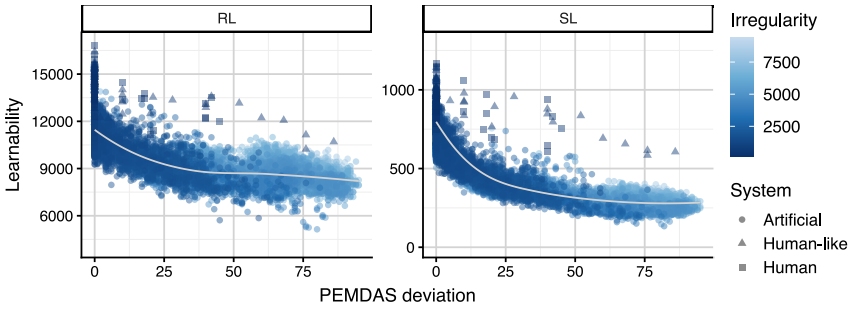


Figure 9: PEMDAS deviation, irregularity and learnability of recursive numeral systems with uniform test distribution. Trend line is a LOESS regression fit.

F Data preprocessing for regression modelling

Note that ours is a convenience sample of recursive numeral systems, which contains some known sampling bias. Crucially, this includes Denić & Szymanik's systems and our own sample of systems optimised for their lexicon-size-numeral-length trade-off objective. These are (by design) heavily biased towards low average length or low lexicon size. While we most likely will not be able to obtain the true distribution of numeral systems, we corrected for these known biases to the best of our ability. First, we recover the human(-like) systems that have been de-duplicated in the main experiments to preserve their attested linguistic frequency and ensure that the parameter uncertainty in the human-like region is not underestimated. We then thinned out the dense region among systems obtained from Denić & Szymanik (2024) by downsampling using kernel density estimates as inverse sampling weights (cf. Algorithm 1). We set the thinned sample size at 1,000. We also downsampled the 10,000 random systems to avoid them having excessive influence on parameter and uncertainty estimation. We expect these random systems to be easier to obtain than other parts of the space, so we downsampled them to a sample of a size about 2 times the rest of the space. The final sample has a total of 3,738 systems, consisting of 1000 thinned D&S systems, 2492 randomly sampled systems, 128 human (from the 60 unique systems) and 118 human-like systems.

G Regression modelling details

The approach we follow first fits a reference model that includes all the predictors (standardised to make them comparable), then projects the parameter estimates onto submodels of growing sizes by finding new estimates that minimise the KL divergence between the reference model's and the

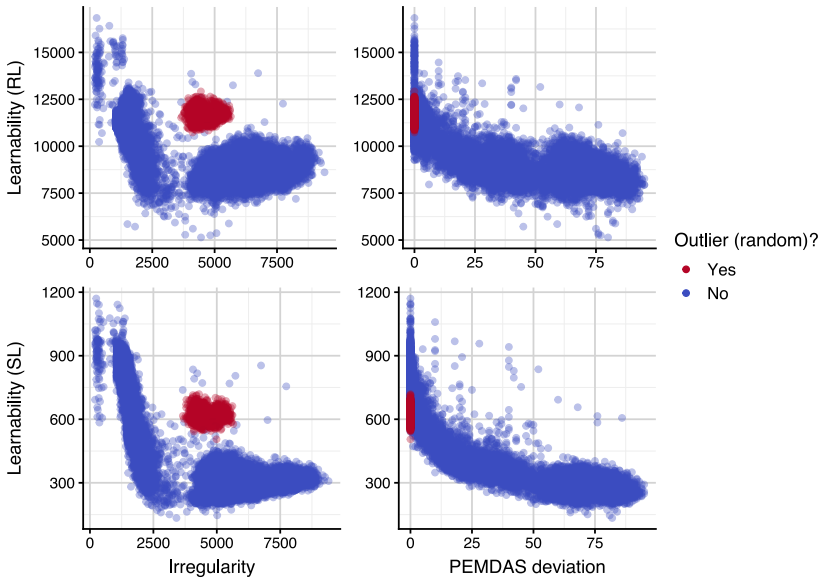


Figure 10: Learnability under reinforcement learning (top) and supervised learning (bottom) of the outlier systems within the random sample (red) vs. the rest of the space (blue). Facets show positions of the cluster along the degree of irregularity vs. PEMDAS deviation.

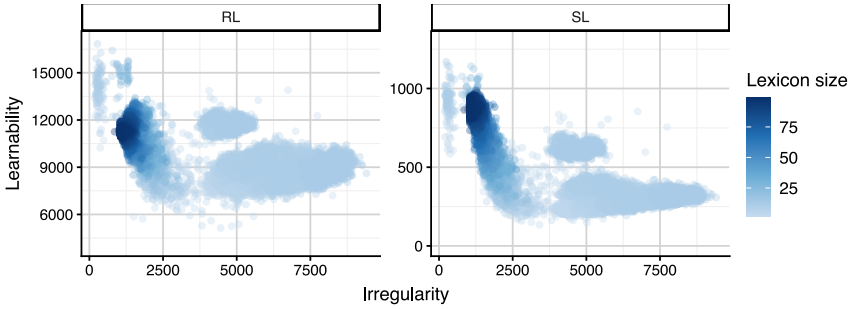


Figure 11: Lexicon size, irregularity and learnability of recursive numeral systems with uniform test distribution.

submodel's posterior predictive. We inspect the predictive accuracy, select the smallest model that performs comparably well to the full reference model, and take the variables present in the small model to be the most important ones in predicting learnability. The predictive accuracy is given by the sum of expected log pointwise predictive densities under leave-one-out cross-validation (ELPD; Vehtari et al., 2017). Predictor importance can be obtained as a ranking of predictors in terms of

Algorithm 1 Density-based thinning

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; sample size $n = 1000$; bandwidth $h = 0.05$.
▷ *x: lexicon size; y: average numeral length*

$\mathbf{Z} \leftarrow \textsc{Standardize}(\mathcal{D})$
$\hat{f} \leftarrow \textsc{GaussianKDE}(\mathbf{Z}, h)$
$w_i \leftarrow \hat{f}(\mathbf{z}_i)^{-1}$ for $i = 1, \ldots, N$
$p_i \leftarrow w_i / \sum_{j=1}^{N} w_j$ for $i = 1, \ldots, N$
$I \leftarrow \textsc{WeightedSample}(\{1, \ldots, N\}, n, \mathbf{p})$ ▷ *w/o replacement*
return $\mathcal{D}[I]$

how much ELPD is gained by their additions to the model. A submodel is taken to be as comparable as the reference if the upper bound of the 68.27% uncertainty interval over their ELPD difference is equal to or larger than zero. In other words, they are considered equivalent if their performance does not theoretically differ by more than one standard error.

After the first step, we take the best submodel and inspect the distribution of the expected posterior predictions conditional on other predictors being at their average level. To gauge effect sizes, we compare visually the effect of changing a given predictor by one standard deviation unit on learnability across the space.

We use `brms` for model fitting (Bürkner, 2021), and our own fork of the package `projpred` for projection (Piironen et al., 2025), which we minimally modify to support explicitly setting k (the number of basis functions).

We use the default priors provided by the package `brms` (Bürkner, 2021). The prior is flat for fixed effects. For the scale parameters, the prior is Student's t with d.f. = 3, the mean = 0. For the intercept, it is the same Student's t, but with the median of the data (for the intercept parameters) and the scale equal to median absolute deviation. We use the thin plate splines (`tp`), and set relatively low k (the number of basis functions) at 5 to penalise wiggliness, as we are interested in broad, general trends.

H Additional GAM results

Here, we additionally present the table of coefficient estimates in H.1, and the results under SL in H.2.

H.1 Reinforcement learning

The main results of this setup are reported in §4.3. Table 1 shows Bayesian estimates of the model from fitting.

Table 1: GAM coefficient estimates and 95% credible intervals for the model predicting learnability under RL. Smooth SDs quantify the amount of wiggliness of the smooth terms, i.e., deviation from a perfectly linear predictor, under `brms`'s parameterisation. Predictor values are standardised before fitting.

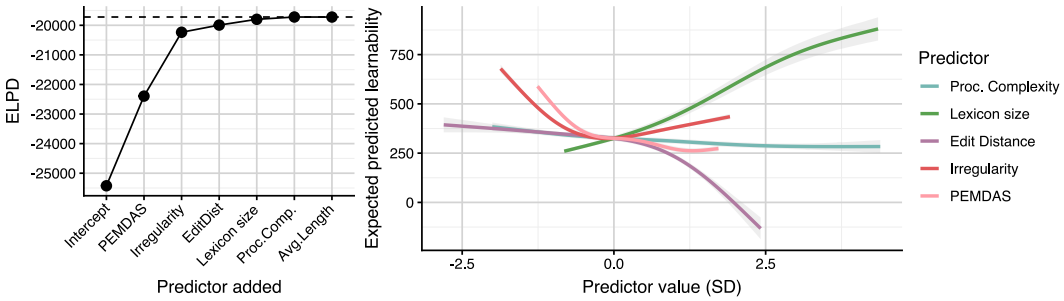
	Estimate	S.E.	l-95% CrI	u-95% CrI
Intercept				
Intercept	9745.517	10.865	9724.123	9767.084
Coefficient				
Irregularity	-9475.580	611.108	-10649.815	-8261.140
Processing Complexity	-1006.234	874.493	-2671.739	789.735
Average length	-3010.521	978.734	-4987.716	-1115.400
Lexicon size	5611.344	1374.291	2882.593	8294.130
Local edit distance	-2199.846	1270.517	-4682.052	397.334
PEMDAS deviation	-13131.115	472.281	-14065.191	-12221.834
Smooth SD				
Irregularity	3797.904	1432.506	2055.863	7437.623
Processing Complexity	1237.967	750.553	258.714	3106.858
Average length	2609.035	1088.209	1221.953	5447.070
Lexicon size	3179.330	1238.903	1536.714	6441.004
Local edit distance	2394.177	1003.379	1127.747	4903.145
PEMDAS deviation	3837.155	1453.773	2048.119	7370.163
SD				
SD	655.907	7.687	640.820	670.990

H.2 SUPERVISED LEARNING

Figure 12a shows the ELPD differences between submodels of growing sizes. The best performing submodel has a model size of 5, i.e., the full model except average length. This is expected, given that SL is much more sensitive to Denić & Szymanik's systems with high lexicon size. These systems consist of mostly monomorphemic numerals, which are also short. It therefore makes sense that the effects of average length on predictive accuracy be absorbed by lexicon size (and other predictors).

Following the methods outlined in §4.3, we inspect the effect sizes of this best performing submodel. Figure 12b shows the effect sizes derived from projected estimates, i.e., the expected predictions conditional on other predictors at their means. As expected, this largely recapitulates the effects seen in the RL results, except the fact that lexicon size has an especially strong effect, in line with our suggestion in §4.2 that SL finds high-lexicon-size systems especially easy. Table 2 shows the coefficient estimates.

Note that the coefficient estimates are not directly comparable to the RL model, as the RL model's estimates are from `brms`, which parameterises the same GAM differently from the projected SL model that is reconstructed from fits using the package `mgcv` (Wood, 2011; used by `projpred`).



(a) (b)

Figure 12: **(a)** Expected log pointwise predictive density of submodels of growing sizes under SL. Dots correspond to the best-performing submodels at specific model sizes, starting with an intercept-only model on the left, moving to the best one-predictor model (PEMDAS), the best two-predictor model (PEMDAS and irregularity), and so on up to the full six-predictor model. The dashed line indicates the ceiling performance (i.e., that of the full model). **(b)** Expected predicted learnability under SL for different predictors over standardised scales. The ranges of standardised predictor values are different for different predictors due to different probability density concentrations.

Table 2: GAM coefficient estimates and 95% credible intervals for the projected submodel predicting learnability under SL. Note that these estimates are derived from `projpred`, which internally uses `mgcv` (Wood, 2011) for fitting submodels during leave-one-out cross-validation; the parameterisation is different from the `brms` one and hence each term is not separated between a fixed (linear) component and a random (wiggliness) component as in Table 1, but instead consists of four bases that together form a smooth curve. These are not interpretable on their own, but are reported for completeness. Predictor values are standardised before fitting.

Term	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept				
Intercept	446.212	0.768	444.740	447.683
Coefficient (one row per basis)				
PEMDAS deviation (1)	137.385	5.821	126.027	148.645
PEMDAS deviation (2)	281.412	18.082	245.710	316.441
PEMDAS deviation (3)	-384.450	18.714	-420.498	-346.508
PEMDAS deviation (4)	-169.809	5.094	-179.495	-159.668
Irregularity (1)	-157.897	12.617	-183.144	-133.074
Irregularity (2)	90.038	16.285	56.866	120.948
Irregularity (3)	-334.152	23.044	-378.432	-288.252
Irregularity (4)	-140.068	6.894	-153.964	-126.817
Local edit distance (1)	69.858	9.771	50.164	89.038
Local edit distance (2)	70.191	19.805	32.506	108.779
Local edit distance (3)	265.987	40.704	187.557	346.106
Local edit distance (4)	-161.501	12.657	-186.003	-136.511
Lexicon size (1)	141.821	19.367	104.525	179.408
Lexicon size (2)	-7.930	29.241	-63.375	48.946
Lexicon size (3)	86.207	31.043	25.815	146.018
Lexicon size (4)	76.835	6.170	64.794	89.148
Processing complexity (1)	4.916	4.986	-3.887	15.260
Processing complexity (2)	-17.459	9.494	-36.702	-1.133
Processing complexity (3)	-50.031	26.213	-102.375	-2.870
Processing complexity (4)	-17.327	4.435	-26.082	-8.420
SD				
SD	47.095	0.542	46.044	48.150