An Information-Theoretic Foundation for the Subregular Hierarchy

Hung Mai[1,2], **Khanh Nguyen**[2,3], **Luong Doan**[2], **Ngan Duong**[1], **Thao Duong**[4], **Tuan Do**[2,3]
[1]National Economics University, Ha Noi, Vietnam
[2]B0Labs, N2TP Technology Solutions JSC, Ha Noi, Vietnam
[3]Phenikaa University, Ha Noi, Vietnam
[4]East Asia University of Technology, Ha Noi, Vietnam
Correspondence: pqhung.mai@n2tp.com, tuando7758@gmail.com

Abstract

The Subregular Hypothesis posits that phonological patterns in natural languages occupy a restricted region of the formal language hierarchy, yet the cognitive basis for this restriction remains unclear. We propose an information-theoretic characterization: Strictly Local languages, when formalized as shifts of finite type, are exactly those admitting stationary Markov sources, which exhibit zero conditional mutual information between distant positions given intervening symbols. We prove that certain non-subregular patterns such as first-last assimilation admit no such Markov realization, explaining their unlearnability. Empirical validation on English phonotactics versus Finnish, Turkish, and Hungarian vowel harmony confirms that MI profiles statistically distinguish SL-like from TSL-like patterns ($p < 0.001$, $r = 0.84$). This work bridges formal language theory and information theory, offering a unified framework for understanding computational restrictions on natural language phonology.

Keywords: subregular languages, mutual information, phonotactic learning, Markov sources, vowel harmony

1 Introduction

A fundamental discovery in computational phonology is that natural language phonological patterns occupy a highly restricted region of the Chomsky hierarchy. While the class of regular languages provides an upper bound on phonological complexity, the vast majority of attested patterns fall within much weaker subregular classes: Strictly Local (SL), Strictly Piecewise (SP), and Tier-based Strictly Local (TSL) languages (Heinz, 2010; Rogers et al., 2010; Graf, 2022).

This paper argues for a single big-picture connection between these formal classes and information theory. At a high level, we show that local phonological patterns are precisely those compatible with bounded statistical dependence, while unattested patterns such as first-last assimilation require dependence that cannot be sustained by finite-memory generative processes. The paper proceeds in three stages: Section 2 introduces the formal and probabilistic ingredients, Sections 3–4 establish the theoretical characterization and its limits, and Sections 5–6 test the predictions on synthetic and cross-linguistic phonotactic data and discuss their linguistic implications.

This restriction is not merely descriptive. Experimental evidence demonstrates that humans fail to learn patterns outside these subregular classes even when such patterns are computationally simple in other respects. The canonical example is first-last assimilation: a pattern requiring the first and last segments of a word to agree. Despite involving only two salient positions, this pattern is unattested in any natural language and unlearnable in laboratory settings (Heinz and Idsardi, 2013; Avcu and Hestvik, 2020).

Why do these specific computational boundaries constrain natural language? Previous explanations appeal to working memory limitations or perceptual salience, but these fail to distinguish subregular classes from one another. We propose that the answer lies in information-theoretic structure. Specifically, subregular classes correspond to constraints on how mutual information (MI) is distributed across positions in strings.

Dai and Futrell (2020) initiated the connection between subregular languages and information theory by analyzing statistical complexity measures. Our work extends this direction with three key contributions: (1) we prove a complete SFT-Markov equivalence theorem with an explicit counterexample showing its limits; (2) we prove that first-last assimilation admits no Markov realization with exact support; and (3) we provide cross-linguistic empirical validation on natural phonotactic data.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46417–46429

Our contributions are fourfold. First, we prove that any stationary finite-order Markov source exhibits finite-range conditional dependence: conditional MI between positions vanishes beyond a fixed distance (Theorem 7). Second, we prove that first-last assimilation admits no stationary finite-order Markov source with exact support (Theorem 11), explaining why such patterns are fundamentally different. Third, we establish that for factorial and extendable languages (shifts of finite type), the existence of a stationary Markov realization is equivalent to SL membership (Theorem 15), and demonstrate via counterexample that this equivalence fails for general prefix-closed languages. Fourth, we validate these predictions empirically: English phonotactics (SL-like) shows significantly lower endpoint MI than vowel harmony languages (TSL-like), with $p < 0.001$ and effect size $r = 0.84$.

A critical distinction underlies our framework. A formal language L is a set of strings; mutual information is a property of probability distributions over strings. All our theorems are stated with explicit distributional assumptions. The key insight is that when a language admits a natural generative process (e.g., a stationary finite-order Markov source with exact support), its MI structure reflects its subregular complexity.

2 Background

2.1 Notation and Subregular Languages

This subsection fixes notation and briefly motivates the linguistic classes used throughout the paper. We make explicit the substring and subsequence operators that appear in later definitions, and we introduce vowel harmony because it is the main empirical case in which long-distance dependencies are local only on a projected tier.

A central result in computational phonology is that many phonological patterns found in natural languages occupy a much smaller region than the full class of regular languages. Rather than requiring arbitrary finite-state machinery, many attested patterns can be characterized by highly restricted dependencies, captured by the subregular classes Strictly Local (SL), Strictly Piecewise (SP), and Tier-based Strictly Local (TSL). These classes were introduced because they align well with recurring empirical generalizations in phonotactics and harmony systems: some constraints depend only on adjacent material, some on subsequences, and some on locality after ignoring irrelevant segments. This paper uses these classes not as abstract formal artifacts, but as linguistically motivated models of different dependency types.

Let Σ be a finite alphabet. We denote by Σ^* the set of all finite strings over Σ, and by Σ^n the set of strings of length exactly n. For a string $w = w_1 w_2 \cdots w_n$, we write $|w| = n$ for its length and w_i for the symbol at position i. A k-factor of w is a contiguous substring of length k; a k-subsequence is a (not necessarily contiguous) subsequence of length k.

We write $\mathrm{fac}_k(w)$ for the set of all length-k factors of w, and $\mathrm{subseq}_k(w)$ for the set of all length-k subsequences of w.

Strictly Local languages capture patterns determined by contiguous windows. Intuitively, membership can be decided by scanning the string with a bounded-size window and checking that no forbidden local configuration appears. This is well suited to phonotactic restrictions such as local segment cooccurrence restrictions or boundary-sensitive constraints. For example, the prohibition against certain word-initial clusters, or against a particular adjacent bigram, is naturally SL.

Definition 1 (Strictly Local)**.** A language $L \subseteq \Sigma^*$ is Strictly k-Local (SL_k) if there exists a finite set $G \subseteq (\Sigma \cup \{\rtimes, \ltimes\})^k$ of forbidden k-factors such that $L = \{w \in \Sigma^* : \mathrm{fac}_k(\rtimes w \ltimes) \cap G = \emptyset\}$, where $\rtimes$ and $\ltimes$ are boundary markers.

For example, if G contains the factor $\rtimes$ ng, then all strings beginning with word-initial /[N]/ are excluded. This illustrates how SL grammars can encode local phonotactic restrictions, including boundary-sensitive ones.

Strictly Piecewise languages capture a different notion of simplicity: constraints on subsequences rather than adjacent substrings. These are useful for long-distance restrictions where intervening material is irrelevant, but where no distinguished tier is assumed.

Definition 2 (Strictly Piecewise)**.** A language L is Strictly k-Piecewise (SP_k) if there exists a finite set $G \subseteq \Sigma^k$ of forbidden k-subsequences such that $L = \{w \in \Sigma^* : \mathrm{subseq}_k(w) \cap G = \emptyset\}$.

For instance, forbidding the subsequence ab excludes any string in which a precedes b, regardless of how many symbols intervene. SP therefore captures a non-local but still highly restricted dependency type.

Tier-based Strictly Local languages were proposed to model patterns such as vowel harmony and consonant harmony, where the relevant dependency is long-distance on the surface but local once irrelevant symbols are ignored.

Vowel harmony is the canonical example: vowels within a word tend to agree in features such as backness or rounding, while intervening consonants are typically irrelevant. In Turkish, for instance, suffix vowels alternate so that the vowel sequence in a word obeys a harmony pattern even when consonants intervene. This is exactly the kind of dependency that motivates TSL. The key idea is to project the string onto a linguistically meaningful tier (for example, vowels only), and then impose an SL constraint on that projected string.

Definition 3 (Tier-based Strictly Local)**.** A language L is Tier-based Strictly k-Local (TSL_k) with tier alphabet $T \subseteq \Sigma$ if there exists an SL_k language L' over T such that $L = \{w \in \Sigma^* : \pi_T(w) \in L'\}$, where π_T erases all symbols not in T.

A canonical example is Turkish vowel harmony. On the surface, harmonizing vowels may be separated by consonants, so the dependency is not contiguous. But if one projects a word onto its vowel tier, harmony can often be described by a local constraint on adjacent vowels on that tier. TSL was introduced precisely to capture such patterns.

These classes form strict inclusions: SL $\subsetneq$ TSL $\subsetneq$ Regular, with SL and SP incomparable. The linguistic importance of this hierarchy is that many attested phonological patterns fall into SL or TSL, while certain logically simple but unattested patterns, such as first-last agreement, fall outside these classes. This observation motivates the search for a deeper explanation of why natural phonology occupies this restricted region.

2.2 Distributional Framework

The key move in this subsection is to shift from categorical well-formedness to probabilistic generation. This lets us ask not only whether a pattern is possible, but whether it can arise from a stationary bounded-memory source whose finite-length distributions have exactly the desired support.

The subregular classes above are properties of formal languages, that is, sets of strings. Our goal, however, is to connect these classes to information-theoretic quantities, which are defined over probability distributions. We therefore introduce a probabilistic generative framework. The guiding question is: when can a language be generated by a process with only bounded memory?

Definition 4 (Stationary Markov Source)**.** A stationary m-order Markov source is a probability measure p on $\Sigma^{\mathbb{N}}$, where $\mathbb{N}$ denotes the positive integers indexing an infinite sequence, such that for all $t \geq m$:

$$\Pr_p(W_{t+1} = x \mid W_{1:t}) = \Pr_p(W_{t+1} = x \mid W_{t-m+1:t}) \quad (1)$$

We write p_n for the length-n marginal of p on Σ^n. We say p has support L if for every $n \geq 1$: $p_n(w) > 0 \iff w \in L \cap \Sigma^n$.

An m-order Markov source remembers only the last m symbols when generating the next one. In this sense, it formalizes a bounded-memory process. This notion is particularly natural for SL-type patterns, whose membership is also determined by bounded local context.

The support condition implies L must be prefix-closed: if $w \in L$, every prefix of w is in L. Many SL languages with boundary-sensitive constraints are not prefix-closed and cannot admit such sources. Consequently, our Markov-based results apply to prefix-closed subregular languages, a genuine restriction we address in Section 4.

2.3 Information-Theoretic Definitions

To connect bounded-memory generation with subregular structure, we study how much information different positions in a string share. If distant positions remain dependent even after the intervening material is known, then the pattern requires information to be transmitted across the string. If that dependence vanishes beyond a fixed range, the pattern is compatible with bounded-memory generation.

For a string W drawn from distribution p, let W_i denote the random variable at position i. We write p_i and p_j for the single-position marginals at positions i and j, and p_{ij} for their joint marginal. Throughout, entropies and mutual informations are measured in bits, so $\log$ denotes logarithm base 2.

Definition 5 (Mutual Information)**.** The mutual information between positions i and j is:

$$I_p(W_i; W_j) = \sum_{\sigma,\tau \in \Sigma} p_{ij}(\sigma, \tau) \log_2 \frac{p_{ij}(\sigma, \tau)}{p_i(\sigma) p_j(\tau)} \quad (2)$$

This quantity measures how much knowing the symbol at position i reduces uncertainty about the symbol at position j.

Definition 6 (Conditional MI)**.** The conditional mutual information given intervening positions is:

$$I_p(W_i; W_j \mid W_{i+1:j-1}) = H_p(W_i \mid W_{i+1:j-1}) - H_p(W_i \mid W_{i+1:j}) \tag{3}$$

The conditional MI measures the residual dependence between positions i and j after accounting for all symbols between them. This quantity is central to our characterization: for bounded-memory processes, dependence should disappear beyond the relevant memory range, whereas for patterns that require unbounded tracking, such dependence can persist even at arbitrarily large distances.

3 Theoretical Results

This section develops the theoretical core of the paper. We first establish the bounded-dependence property that every finite-order Markov source must satisfy, then contrast it with first-last assimilation, and finally show how tier structure lets TSL patterns preserve long-distance dependence in a controlled way.

3.1 Finite-Range Conditional Dependence

We begin with a standard conditional independence property of finite-order Markov sources, included here for self-containment because it provides the information-theoretic baseline for our later results. Intuitively, once the intervening context contains the last m symbols relevant to an m-order Markov process, more distant positions become conditionally independent.

Theorem 7 (Finite-Range Conditional Independence)**.** *Let p be a stationary m-order Markov source on $\Sigma^{\mathbb{N}}$. Then for all positions $i < j$ with $j - i > m$:*

$$I_p(W_i; W_j \mid W_{i+1:j-1}) = 0 \tag{4}$$

Proof sketch. By the Markov property, $\Pr_p(W_j = x \mid W_1, \ldots, W_{j-1}) = \Pr_p(W_j = x \mid W_{j-m:j-1})$. When $j - i > m$, we have $j - m \geq i + 1$, so $W_{j-m:j-1} \subseteq W_{i+1:j-1}$. Thus $\Pr_p(W_j = x \mid W_i, W_{i+1:j-1}) = \Pr_p(W_j = x \mid W_{i+1:j-1})$, implying $W_j \perp W_i \mid W_{i+1:j-1}$. The full proof appears in Appendix C.1.

Corollary 8. *If $L \in SL_k$ admits a stationary m-order Markov source with $m \geq k - 1$, then conditional MI vanishes beyond distance m.*

Theorem 7 is not itself a novelty claim; rather, it serves as the baseline fact against which our main results are contrasted. The substantive contribution begins in showing that certain unattested non-subregular patterns, such as first-last assimilation, admit no such Markov realization, and in identifying the precise conditions under which an SFT/SL-style characterization by Markov support does hold.

3.2 First-Last Assimilation Has No Markov Realization

We now analyze the canonical non-subregular pattern, showing it has fundamentally different information-theoretic properties.

Theorem 9 (Conditional MI for First-Last)**.** *Let $L_{FL} = \{w \in \Sigma^* : w_1 = w_{|w|}\}$. For any distribution p supported on $L_{FL} \cap \Sigma^n$:*

$$I_p(W_1; W_n \mid W_{2:n-1}) = H_p(W_1 \mid W_{2:n-1}) \tag{5}$$

Proof sketch. By constraint $w_1 = w_n$, we have $H_p(W_1 \mid W_{2:n-1}, W_n) = 0$ since W_n determines W_1. The result follows from the definition of conditional MI.

Corollary 10. *Under the uniform distribution $\mathcal{U}_n(L_{FL})$:*

$$I_p(W_1; W_n \mid W_{2:n-1}) = \log |\Sigma| \tag{6}$$

which is constant in n and strictly positive for $|\Sigma| \geq 2$.

Theorem 11 (No Markov Realization)**.** *For $|\Sigma| \geq 2$, there is no stationary m-order Markov source whose marginals satisfy $p_n(w) > 0 \iff w \in L_{FL} \cap \Sigma^n$ for all $n \geq 2$.*

Proof sketch. Assume such a source exists. For strings of length $m + 2$, valid strings have form xcx where $x \in \Sigma$ and $c \in \Sigma^m$. The Markov factorization gives $p_{m+2}(x, c, x) = p_{m+1}(x, c) \cdot q_c(x)$ where q_c is the context-dependent transition kernel depending only on c. For all (x, c, x) to have positive probability, we need $q_c(x) > 0$ for all x and c. But then $p_{m+2}(x, c, y) = p_{m+1}(x, c) \cdot q_c(y) > 0$ for $x \neq y$, contradicting that mismatched-endpoint strings must have zero probability. The full proof appears in Appendix C.2.

This result explains why first-last assimilation is fundamentally different from SL patterns: it requires tracking unbounded information across the

entire string, which no finite-order Markov process can accomplish while maintaining exact support.

3.3 Tier-Based Strictly Local Languages

For TSL languages, the MI structure manifests on the tier rather than at the surface level.

Theorem 12 (Finite-Range Dependence on Tier)**.** *Let $L \in TSL_k$ with tier T. If the distribution p induces an m-order Markov structure on the projected string $Z = \pi_T(W)$, then for tier positions $r < s$ with $s - r > m$:*

$$I_p(Z_r; Z_s \mid Z_{r+1:s-1}) = 0 \tag{7}$$

This follows directly from applying Theorem 7 to the tier projection. The implication is that TSL patterns exhibit finite-range conditional dependence on the tier, but may show elevated MI at the surface level when tier symbols are separated by intervening non-tier material.

4 The SFT-Markov Equivalence and Its Limits

We now sharpen the main theorem by identifying exactly when SL-style locality coincides with Markov realizability. The positive result requires moving to the boundary-free symbolic-dynamical setting of shifts of finite type (SFTs), while the counterexample shows why unrestricted boundary-sensitive SL languages fall outside that equivalence.

4.1 A Counterexample

We initially conjectured that among prefix-closed regular languages, SL membership is equivalent to admitting a stationary Markov source. This is false. The counterexample is useful conceptually because it isolates the role of boundary markers: the problem is not locality per se, but the incompatibility between one-sided boundary conditions and stationarity.

Proposition 13 (Counterexample)**.** *Let $L_{start\text{-}a} = \{\varepsilon\} \cup \{aw : w \in \{a, b\}^*\}$. Then: (1) $L_{start\text{-}a}$ is prefix-closed and SL_2; (2) under uniform sampling, $I(W_i; W_j \mid W_{i+1:j-1}) = 0$ for all $i < j$; (3) yet $L_{start\text{-}a}$ admits no stationary Markov source with exact support.*

Proof sketch. For (3): if such a source exists, stationarity implies $\Pr(W_t = b) = \Pr(W_1 = b) = 0$ for all t. Thus b never occurs, but exact support requires strings like "ab" to have positive probability.

The failure arises from boundary asymmetry: SL can enforce constraints involving the left boundary marker $\rtimes$, but stationary processes are shift-invariant.

4.2 Repaired Characterization

The repaired statement replaces boundary-sensitive SL with SFTs. An SFT (shift of finite type) is a set of infinite sequences defined by forbidding finitely many contiguous blocks; equivalently, membership depends only on the absence of finitely many local configurations. In our setting, SFTs provide the boundary-free symbolic-dynamics counterpart of SL languages and are therefore the right object for exact comparison with stationary Markov sources.

Definition 14. A language L is factorial if every factor of $w \in L$ is in L; extendable if every $w \in L$ extends on both sides within L.

Theorem 15 (SFT-Markov Equivalence)**.** *Let L be factorial and extendable. The following are equivalent: (1) L is an m-step shift of finite type (SFT); (2) L admits a stationary m-order Markov source with exact support; (3) under any m-order Markov distribution, $I(W_i; W_j \mid W_{i+1:j-1}) = 0$ for $j - i > m$.*

The proof uses the classical symbolic dynamics result that m-step SFTs are exactly block languages of m-step Markov shifts. The boundary-free fragment of SL corresponds precisely to factorial, extendable SFTs.

5 Experiments

We validate our framework through synthetic and cross-linguistic experiments. Table 1 summarizes the data sources. The NorthEuraLex database (Dellert et al., 2020) provides standardized IPA transcriptions for the harmony languages.

The experimental goal is twofold: first, to verify on controlled synthetic languages that the MI diagnostics recover the expected formal distinctions; and second, to test whether the same diagnostics separate English phonotactics from vowel-harmony systems in real lexicons. Because the harmony lexicons are modest in size, we treat the multilingual results as evidence of a stable trend rather than as an exhaustive characterization of all TSL-like systems.

5.1 Synthetic Validation

We begin with synthetic languages because their formal structure is known in advance. This lets us

Language	Source	Words	Phonemes	Pattern Type
English	CMU Dict v0.7b (Rudnicky, 2014)	125,763	39 (ARPAbet)	SL-like
Finnish	NorthEuraLex v0.9 (Dellert et al., 2020)	1,188	29 (IPA)	TSL (harmony)
Turkish	NorthEuraLex v0.9 (Dellert et al., 2020)	1,249	28 (IPA)	TSL (harmony)
Hungarian	NorthEuraLex v0.9 (Dellert et al., 2020)	1,234	33 (IPA)	TSL (harmony)

Table 1: Data sources and characteristics. The harmony lexicons are modest in size, so the multilingual analysis should be interpreted as a targeted proof of concept rather than a typologically exhaustive sample of TSL-like patterns.

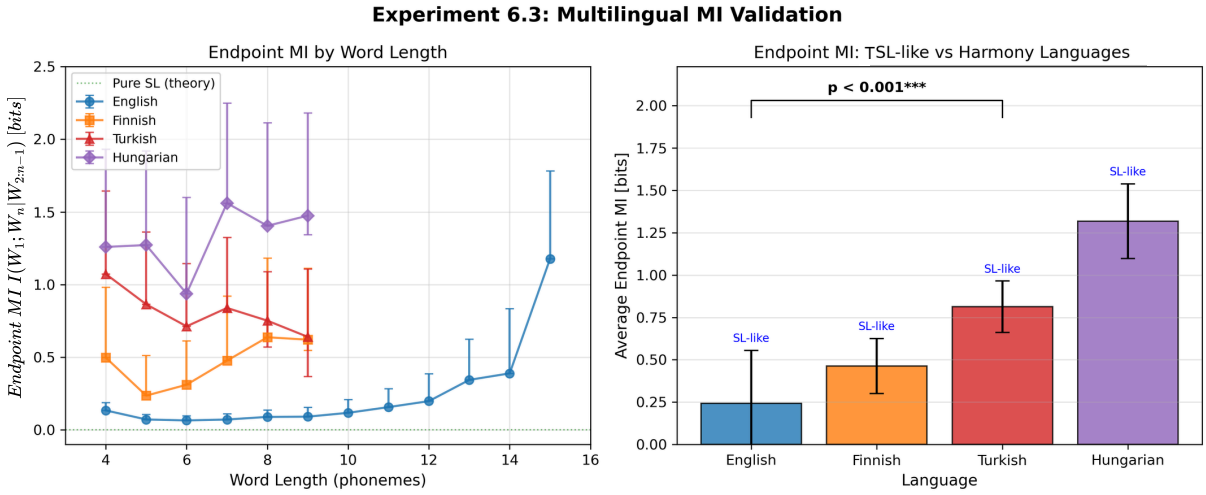

Figure 1: Endpoint MI by word length (left) and average by language (right) with significance bracket. The left panel shows all analyzed data points: one point per attested word length per language after excluding lengths with insufficient data for stable estimation.

confirm that the estimated MI quantities behave as the theorems predict before turning to naturally occurring lexical data.

To validate the theoretical predictions of our framework under controlled conditions, we conduct experiments on a suite of synthetic languages whose structural properties are known by construction. Specifically, we tested six synthetic languages: L_1 (SL_2, no "aa"), L_2 (SL_3, no "aba"), L_3 (SP_2, no "ab" subsequence), L_4 (TSL_2, sibilant harmony), L_5 (first-last), and L_6 (parity). We generated 10,000 strings per language at lengths $n \in \{10, 20, 30, 40\}$ via rejection sampling and computed endpoint conditional MI $I(W_1; W_n \mid W_{2:n-1})$.

The results, summarized in Table 2, strongly corroborate our theoretical characterization. The SL languages show endpoint MI near zero across all lengths. First-last shows constant endpoint MI $\approx \log_2(3) = 1.585$ bits, exactly matching Corollary 10.

Language	Class	n=10	n=20	n=30	n=40
L_1	SL_2	0.000	0.000	0.000	0.000
L_2	SL_3	0.000	0.000	0.000	0.000
L_4	TSL_2	0.085	0.084	0.086	0.083
L_5	First–Last	1.585	1.585	1.585	1.585

Table 2: Endpoint conditional MI for synthetic languages. Each cell is estimated from 10,000 sampled strings for that language and length.

5.2 Multilingual Phonotactic Validation

We next test whether the same information-theoretic contrast appears in natural lexical data. The comparison is intentionally simple: English serves as an SL-like baseline with predominantly local phonotactics, while Finnish, Turkish, and Hungarian provide TSL-like cases because vowel harmony induces structured long-distance dependencies on the vowel tier.

We computed endpoint MI $I(W_1; W_n \mid W_{2:n-1})$ for words of length 4–14 phonemes. English represents SL-like local phonotactics; Finnish, Turkish, and Hungarian represent TSL-like vowel harmony

Test	Statistic	p-value	Effect Size
Mann-Whitney U	17.0	6.38×10^{-5}	$r = 0.843$
Welch's t	-4.77	2.77×10^{-5}	–

Group	n	Mean	SD
SL-like	12	0.242	0.313
TSL-like	18	0.865	0.399

Table 3: Statistical comparison of endpoint MI. Sample sizes are the number of language-by-length data points contributing to the comparison: 12 for the SL-like group and 18 for the TSL-like group.

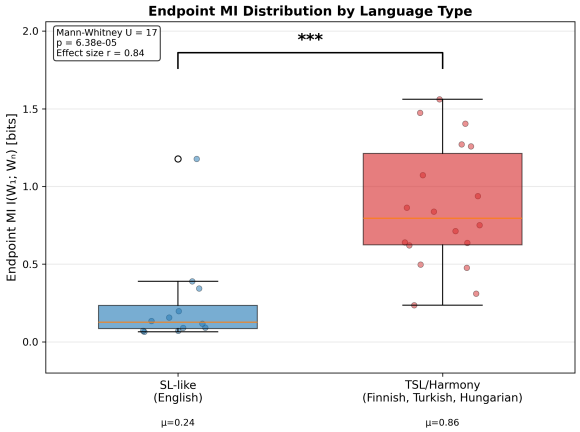

Figure 2: Endpoint MI distributions by language type with individual data points.

(harmony rates: 60%, 84%, 81% respectively). After excluding lengths with too few word types for stable estimation, this yields 12 English data points and 18 harmony-language data points in total, which are exactly the points shown in Figure 1 and summarized in Table 3.

Figure 1 presents the core empirical finding: endpoint MI systematically distinguishes pattern types. English maintains low endpoint MI (0.07–0.39 bits) across lengths, while harmony languages show elevated values: Finnish (0.24–0.64 bits), Turkish (0.64–1.07 bits), Hungarian (0.94–1.56 bits).

Figure 2 shows clear separation between SL-like (median ≈ 0.09 bits) and TSL-like (median ≈ 0.79 bits) distributions. A Mann-Whitney U test confirms significance: $U = 17.0$, $p = 6.4 \times 10^{-5}$, with effect size $r = 0.84$. Table 3 reports detailed statistics.

The mean endpoint MI for SL-like phonotactics (0.24 bits) is approximately 3.6 times lower than for TSL-like patterns (0.87 bits), supporting our theoretical prediction that MI profiles distinguish subregular language classes.

6 Discussion

The discussion makes the paper's broader claim explicit. The main takeaway is that the formal hierarchy, the probabilistic realizability results, and the empirical MI profiles all point to the same organizing principle: natural phonological patterns are constrained by how far information must be tracked.

6.1 Interpretation of Results

Our theoretical and empirical results support the hypothesis that the subregular hierarchy corresponds to constraints on mutual information structure. Theorem 7 establishes that finite-order Markov sources exhibit zero conditional MI beyond their memory parameter, a property that SL languages can satisfy but first-last assimilation cannot (Theorem 11). The SFT-Markov equivalence (Theorem 15) provides a complete characterization for the boundary-free fragment of SL, while Proposition 13 shows this fails when boundary constraints are involved.

The empirical validation demonstrates that these theoretical distinctions manifest in real phonotactic data. English, with predominantly local phonotactic constraints, shows endpoint MI near zero across all tested word lengths. Vowel harmony languages show elevated endpoint MI because harmony creates long-range dependencies between vowels that persist even when conditioned on all intervening consonants. The gradient observed across harmony languages (Finnish $<$ Turkish $<$ Hungarian) correlates with harmony strength in each language, suggesting MI profiles capture fine-grained phonological differences, not merely coarse typological categories.

6.2 Cognitive Implications

Our framework offers a potential mechanistic explanation for why certain patterns are unlearnable. If human phonological learning relies on tracking statistical dependencies with bounded memory, patterns requiring unbounded tracking (constant high MI regardless of string length) cannot be acquired. First-last assimilation exemplifies this: its endpoint MI remains at $\log |\Sigma|$ bits regardless of word length. Learnable patterns, by contrast, have MI that either decays with distance (SL) or concentrates on a tier

(TSL).

This connects to the artificial grammar learning literature. Artificial language learning experiments have shown that humans generalize based on phonological features (Finley and Badecker, 2009), and Avcu and Hestvik (2020) showed that humans fail to learn first-last patterns despite extensive training, while readily acquiring sibilant harmony (TSL). Our MI metric correctly predicts this: sibilant harmony has bounded tier-projected MI, while first-last has constant high endpoint MI. The MI profile thus serves as a quantitative predictor of learnability, formalizing the intuition that "simple" patterns are those with bounded statistical complexity.

6.3 Connections to Neural Language Models

Our framework relates to ongoing work on what formal languages neural networks can learn (Bhattamishra et al., 2020). Torres and Futrell (2023) showed that sparse LSTMs prefer subregular languages over regular languages, with the preference increasing as sparsity pressure increases. Transformers learn positional dependencies through attention mechanisms; our MI framework predicts which dependencies are "easy" (finite-range, decaying) versus "hard" (unbounded, constant). This suggests MI profiles may predict neural network learnability in addition to human learnability.

The Information Bottleneck principle (Tishby and Zaslavsky, 2015) provides a broader theoretical framework: optimal representations compress input while preserving relevant information. Our characterization can be viewed through this lens, with finite-range conditional dependence corresponding to efficient compression of positional information.

6.4 The Role of Distribution

A recurring theme in our analysis is the importance of distributional assumptions. Theorem 7 applies to Markov sources; Corollary 10 relies on the uniform distribution. The choice of distribution matters: different distributions over the same language can yield different MI profiles. Our empirical results use the natural distribution induced by lexicon frequencies, which may approximate cognitive exposure statistics. Investigating which distributions yield the cleanest characterizations remains an important direction for future work.

7 Related Work

7.1 Subregular Linguistics

The subregular hierarchy emerged from computational analyses of phonological patterns. McNaughton and Papert (1971) introduced Locally Testable languages; Rogers and Pullum (2011) and Rogers et al. (2010) systematically characterized the Strictly Local and Strictly Piecewise classes. Heinz (2010) demonstrated that long-distance phonotactic patterns fall within SP, while Heinz et al. (2011) introduced Tier-based Strictly Local languages to capture harmony systems.

Graf (2017, 2022) provided comprehensive surveys connecting subregular complexity to linguistic typology. Rawski (2019) gave a geometric characterization using tensor product representations, showing subregular classes are linearly separable. Lambert (2023) extended these characterizations with relativized adjacency relations. Lambert et al. (2021) derived subregular classes from the learning perspective. Our work complements these automata-theoretic and geometric approaches with information-theoretic characterization.

7.2 Information Theory and Subregular Languages

Dai and Futrell (2020) initiated the connection between subregular languages and statistical complexity theory, analyzing measures like excess entropy and statistical complexity for characterizing subregular classes. Their work used probabilistic finite-state automata to connect formal language theory with information theory. Our work extends this direction by: (1) proving a complete SFT-Markov equivalence (Theorem 15) with explicit counterexample showing its limits; (2) proving impossibility results for first-last assimilation (Theorem 11); and (3) providing cross-linguistic empirical validation.

More broadly, information-theoretic approaches to language have a long history (Sloane and Wyner, 2009; Plotkin and Nowak, 2000). Piantadosi et al. (2011) showed word length correlates with information content, and Futrell et al. (2015) analyzed dependency length minimization across languages. Hahn et al. (2020) proposed that word order reflects memory-surprisal tradeoffs. Futrell and Hahn (2022) argued information theory bridges language function and form. Our work differs by targeting formal language classification rather than production pressures.

7.3 Learnability and Cognitive Constraints

Gold (1967) established fundamental limits on language learning. Heinz et al. (2016) provide comprehensive treatment of grammatical inference. Heinz and Idsardi (2013) argued subregular boundaries reflect cognitive constraints. Avcu and Hestvik (2020) provided ERP evidence that humans fail to learn non-subregular patterns.

Rawski and Heinz (2019) formalized the "learning bias" hypothesis. Torres and Futrell (2023) showed that sparse neural networks prefer subregular languages, suggesting simplicity biases parallel human cognitive biases. The Information Bottleneck framework (Tishby and Zaslavsky, 2015) provides broader context for understanding how bounded-capacity systems learn. Our MI framework instantiates similar principles: patterns with finite-range conditional dependence require only bounded statistical tracking.

8 Conclusion

We have established an information-theoretic foundation for the subregular hierarchy. Our main results show that finite-order Markov sources exhibit finite-range conditional dependence (Theorem 7), that non-subregular patterns like first-last assimilation admit no such Markov realization (Theorem 11), and that for shifts of finite type, the existence of a Markov source is equivalent to SL membership (Theorem 15). The counterexample $L_{\text{start-}a}$ delineates exactly where this characterization applies.

Empirical validation on English versus Finnish, Turkish, and Hungarian demonstrates that MI profiles reliably distinguish SL-like from TSL-like phonological patterns ($p < 0.001$, $r = 0.84$). The broader vision is that computational constraints on natural language phonology reflect fundamental limits on information-theoretic tracking capacity: the subregular hierarchy is not merely formal curiosity but a reflection of how statistical dependencies must be structured to be humanly learnable.

Future work should extend the characterization to include boundary-sensitive SL patterns, develop MI-based characterizations of SP languages, and investigate whether MI profiles predict learnability in neural architectures.

Limitations

Several limitations qualify our results. First, the SFT-Markov equivalence (Theorem 15) applies only to factorial and extendable languages, excluding boundary-sensitive SL patterns. The counterexample $L_{\text{start-}a}$ shows that prefix-closure alone is insufficient; extending the characterization to the full SL class with boundary markers remains open.

Second, our empirical results are distribution-dependent. The uniform distribution does not always yield clean MI characterizations, and the "natural" distributions appropriate for phonotactic analysis require further investigation.

Third, the multilingual validation uses lexicon data (word types) rather than online processing measures. The harmony language samples are relatively small (∼1,200 words each), which may not capture the full spectrum of TSL-like phonotactics across lexical strata or dialects, though bootstrap confidence intervals and large effect sizes suggest that the coarse SL-like/TSL-like contrast is stable in the present sample.

Fourth, our TSL results assume the tier is known; automatic tier identification from MI profiles is an open problem we do not address.

Fifth, we do not provide formal learnability guarantees for the MI-Profile Learner. The algorithm is heuristic, and sample complexity bounds for MI estimation (Paninski, 2003) do not translate directly to classification accuracy.

Finally, our framework currently addresses classification of phonological patterns but does not directly model human processing or neural network learning. The cognitive implications we discuss are plausible but remain to be validated experimentally.

References

Enes Avcu and Arild Hestvik. 2020. Unlearnable phonotactics. In *Glossa: a journal of general linguistics*, volume 5. ISSN: 2397-1835 Issue: 1.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020. On the Ability and Limitations of Transformers to Recognize Formal Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, Online. Association for Computational Linguistics.

Huteng Dai and Richard Futrell. 2020. Work in progress: Information-theoretic characterization of the subregular hierarchy. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 94–97, New York, New York. Association for Computational Linguistics.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga,

Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54(1):273–301.

Sara Finley and William Badecker. 2009. Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3):423–437.

Richard Futrell and Michael Hahn. 2022. Information Theory as a Bridge Between Language Function and Language Form. In *Frontiers in Communication*, volume 7, page 657725. ISSN: 2297-900X Journal Abbreviation: Front. Commun.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

E Mark Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

Thomas Graf. 2017. The power of locality domains in phonology. *Phonology*, 34(2):385–405.

Thomas Graf. 2022. Subregular linguistics: bridging theoretical linguistics and formal grammar. *Theoretical Linguistics*, 48(3-4):145–184.

Michael Hahn, Judith Degen, and Richard Futrell. 2020. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. Repository: PsyArXiv.

Jeffrey Heinz. 2010. Learning Long-Distance Phonotactics. *Linguistic Inquiry*, 41(4):623–661.

Jeffrey Heinz, Colin De La Higuera, and Menno Van Zaanen. 2016. *Grammatical Inference for Computational Linguistics*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Jeffrey Heinz and William Idsardi. 2013. What Complexity Differences Reveal About Domains in Language*. *Topics in Cognitive Science*, 5(1):111–131.

Jeffrey Heinz, Chetan Rawal, and H. Tanner. 2011. Tier-based Strictly Local Constraints for Phonology.

Dakotah Lambert. 2023. Relativized Adjacency. *Journal of Logic, Language and Information*, 32(4):707–731.

Dakotah Jay Lambert, Jonathan Rawski, and Jeffrey Heinz. 2021. Typology emerges from simplicity in representations and learning. *Journal of Language Modelling*, 9(1).

Robert McNaughton and Seymour A. Papert. 1971. *Counter-Free Automata*. MIT Press, Cambridge, MA, USA.

Liam Paninski. 2003. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Joshua B Plotkin and Martin A Nowak. 2000. Language Evolution and Information Theory. *Journal of Theoretical Biology*, 205(1):147–159.

Jonathan Rawski. 2019. Tensor Product Representations of Subregular Formal Languages. *ArXiv*.

Jonathan Rawski and Jeffrey Heinz. 2019. No free lunch in linguistics or machine learning: Response to Pater. *Language*.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2010. On Languages Piecewise Testable in the Strict Sense. volume 6149, pages 255–265, Berlin, Heidelberg. Springer Berlin Heidelberg. Book Title: The Mathematics of Language Series Title: Lecture Notes in Computer Science.

James Rogers and Geoffrey K. Pullum. 2011. Aural Pattern Recognition Experiments and the Subregular Hierarchy. *Journal of Logic, Language and Information*, 20(3):329–342.

Alex Rudnicky. 2014. The CMU pronouncing dictionary, release 0.7b.

N. J. A. Sloane and Aaron D. Wyner. 2009. Prediction and Entropy of Printed English. IEEE. Book Title: Claude E. Shannon.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. Conference Name: 2015 IEEE Information Theory Workshop (ITW) ISBN: 9781479955244 9781479955268 Place: Jerusalem, Israel Publisher: IEEE.

Charles Torres and Richard Futrell. 2023. Simpler neural networks prefer subregular languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1651–1661, Singapore. Association for Computational Linguistics.

A MI-Profile Learner

Algorithm 1 MI-Profile Learner

Require: Sample $S = \{w^{(1)}, \ldots, w^{(m)}\}$ from target language L
Require: Locality parameter k, distance threshold $d_{\max}$
Ensure: Hypothesis class (SL, SP, TSL, or UNKNOWN)
for $d = 1$ to $d_{\max}$ **do**
  Compute $\hat{I}_{\text{uncond}}(d) =$ averaged MI(W_i, W_{i+d})
end for
for each symbol pair (σ, τ) **do**
  Compute $\hat{I}_{\text{sym}}(\sigma, \tau) = I(X_\sigma; X_\tau)$
end for
Compute $\hat{I}_{\text{endpoint}} = \hat{I}(W_1; W_n \mid W_{2:n-1})$
if $\hat{I}_{\text{uncond}}(d) \approx 0$ for $d > k$ for small k **then**
  return SL$_k$
end if
if $\hat{I}_{\text{endpoint}}$ significantly positive and constant **then**
  return UNKNOWN (potentially non-subregular)
end if
for each candidate tier T **do**
  Compute projected MI on tier
  if projected MI decays rapidly **then**
    return TSL with tier T
  end if
end for
return UNKNOWN

This algorithm is heuristic and does not carry formal guarantees. Estimating conditional MI requires distributions over exponentially many conditioning contexts, limiting practical applicability to short-range estimates.

B Detailed Experimental Results

Table 4: Unconditional MI decay profiles at n=40.

Distance	L_1 (SL$_2$)	L_2 (SL$_3$)	L_3 (SP$_2$)	L_4 (TSL$_2$)	L_5 (FL)
1	0.083	0.010	0.214	0.088	0.000
2	0.004	0.005	0.190	0.087	0.000
3	0.000	0.000	0.171	0.087	0.000
5	0.000	0.000	0.141	0.088	0.000
10	0.000	0.000	0.088	0.089	0.000
15	0.000	0.000	0.055	0.087	0.000

SL languages show rapid decay consistent with finite-range dependence. SP shows slower decay. TSL maintains constant MI across all distances (tier symbols remain correlated regardless of separation). First-last shows zero short-range MI (endpoints independent of middle).

The k-gram learner succeeds on SL but not SP; the k-subsequence learner succeeds on SP but not SL. Neither succeeds on first-last, confirming that bounded-memory learners cannot acquire patterns requiring unbounded tracking.

Table 5: MI-bounded learner results.

Language	Class	Learner	k	Acc. (seen)	Acc. (longer)
L_1	SL_2	k-gram	2	1.000	1.000
L_1	SL_2	k-subseq	2	0.500	0.500
L_3	SP_2	k-gram	2	0.941	0.983
L_3	SP_2	k-subseq	2	1.000	1.000
L_5	Non-subreg	k-gram	2	0.500	0.500
L_5	Non-subreg	k-subseq	2	0.500	0.500

C Full Proofs

C.1 Proof of Theorem 7

Theorem (Finite-Range Conditional Independence for Markov Sources)**.** Let p be a stationary m-order Markov source on $\Sigma^{\mathbb{N}}$. Then for all positions $i < j$ with $j - i > m$:

$$I_p(W_i; W_j \mid W_{i+1:j-1}) = 0 \tag{8}$$

Proof. By the m-order Markov property:

$$\begin{aligned} &\Pr_p(W_j = x \mid W_{1:j-1}) \\ &\quad = \Pr_p(W_j = x \mid W_{j-m:j-1}) \end{aligned} \tag{9}$$

Consider positions $i < j$ with $j - i > m$. This means $i \leq j - m - 1$, so W_i is among the "remote past" $W_{1:j-m-1}$.

Given the full conditioning set $W_{i+1:j-1}$, we know the last m symbols $W_{j-m:j-1}$ (since $j - m \geq i + 1$ when $j - i > m$).

Therefore:

$$\begin{aligned} &\Pr_p(W_j = x \mid W_i, W_{i+1:j-1}) \\ &\quad = \Pr_p(W_j = x \mid W_{j-m:j-1}) \\ &\quad = \Pr_p(W_j = x \mid W_{i+1:j-1}) \end{aligned} \tag{10}$$

The first equality uses the Markov property (only the last m symbols matter). The second equality holds because $W_{j-m:j-1} \subseteq W_{i+1:j-1}$.

This shows $W_j \perp W_i \mid W_{i+1:j-1}$, which implies $I_p(W_i; W_j \mid W_{i+1:j-1}) = 0$. □

C.2 Proof of Theorem 11

Theorem (No Finite-Order Markov Realization of First-Last)**.** For $|\Sigma| \geq 2$, there is no stationary m-order Markov source p on $\Sigma^{\mathbb{N}}$ whose finite-length marginals satisfy:

$$p_n(w) > 0 \iff w \in L_{FL} \cap \Sigma^n \quad \text{for all } n \geq 2 \tag{11}$$

Proof. Assume for contradiction that such a Markov source p exists for some $m \geq 1$.

Step 1: Structure of valid strings. Consider strings of length $n = m + 2$. Any string in $L_{FL} \cap \Sigma^{m+2}$ has the form $w = x\, c_1 c_2 \cdots c_m\, x$ where the first and last symbol are both $x \in \Sigma$, and the middle block $c = c_1 \cdots c_m \in \Sigma^m$ is arbitrary.

Step 2: Support requirement (positive direction). Since p has support exactly L_{FL}, we require:

$$p_{m+2}(x, c, x) > 0 \quad \text{for every } x \in \Sigma \text{ and } c \in \Sigma^m \tag{12}$$

Step 3: Markov factorization. By the m-order Markov property:

$$p_{m+2}(x, c, x) = p_{m+1}(x, c) \cdot q_c(x) \tag{13}$$

where we define the context-dependent transition kernel $q_c(y) := \Pr_p(W_{m+2} = y \mid W_{2:m+1} = c)$. Note that $q_c(\cdot)$ depends only on the context c, not on the first symbol.

Step 4: Implication for transition kernel. From Steps 2 and 3: for each context $c \in \Sigma^m$ and each symbol $x \in \Sigma$, if $p_{m+1}(x, c) > 0$, then we must have $q_c(x) > 0$.

Step 5: Support requirement (negative direction). Since p has support exactly L_{FL}, strings with mismatched endpoints must have zero probability:

$$p_{m+2}(x, c, y) = 0 \quad \text{whenever } x \neq y \tag{14}$$

Step 6: Contradiction. Fix any context $c \in \Sigma^m$. From Step 2, for each $x \in \Sigma$, the string (x, c, x) must have positive probability. This requires $p_{m+1}(x, c) > 0$ for all $x \in \Sigma$.

By Step 4, this means $q_c(x) > 0$ for all $x \in \Sigma$.

Now consider any two distinct symbols $x \neq y$ in Σ. We have $p_{m+1}(x, c) > 0$ and $q_c(y) > 0$. Therefore:

$$p_{m+2}(x, c, y) = p_{m+1}(x, c) \cdot q_c(y) > 0 \tag{15}$$

But this contradicts Step 5, which requires $p_{m+2}(x, c, y) = 0$ for $x \neq y$.

Since $|\Sigma| \geq 2$, such distinct x, y exist, completing the contradiction. □

C.3 Proof of Proposition 13

Proposition (Counterexample)**.** Let $\Sigma = \{a, b\}$ and define $L_{\text{start-}a} = \{\varepsilon\} \cup \{aw : w \in \{a, b\}^*\}$. Then:

1. $L_{\text{start-}a}$ is prefix-closed and regular
2. $L_{\text{start-}a} \in \text{SL}_2$ (forbid $\rtimes b$)
3. Under $\mathcal{U}_n(L_{\text{start-}a})$, we have $I(W_i; W_j \mid W_{i+1:j-1}) = 0$ for all $i < j$
4. $L_{\text{start-}a}$ admits no stationary finite-order Markov source with exact support

Proof. **(1)–(2):** Any prefix of a string starting with a also starts with a (or is ε). The forbidden 2-factor $\rtimes b$ captures exactly "strings not starting with a."

(3): Under $\mathcal{U}_n(L_{\text{start-}a})$ for $n \geq 1$: $W_1 = a$ deterministically, and $W_2, \ldots, W_n$ are i.i.d. uniform on $\{a, b\}$. Thus all position pairs are independent (or one is constant), giving $I(W_i; W_j \mid W_{i+1:j-1}) = 0$.

(4): Suppose a stationary process p has exact support $L_{\text{start-}a}$. For $n = 1$, only "a" is allowed, so $p_1(a) = 1$ and $p_1(b) = 0$. By stationarity, the marginal distribution of W_t is identical for all t:

$$\Pr(W_t = b) = \Pr(W_1 = b) = 0 \quad \forall t \tag{16}$$

Thus b never occurs, and the process is supported only on a^∞. But exact support requires strings like "ab", "aab", etc. (which are in $L_{\text{start-}a}$) to have positive probability, a contradiction. □