From Formal Language Theory to Statistical Learning: Finite Observability of Subregular Languages

Katsuhiko Hayashi
Language and Information Sciences
The University of Tokyo
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Hidetaka Kamigaito
Information Science
Nara Institute of Science and Technology
kamigaito.h@is.naist.jp

Abstract

We prove that all standard subregular language classes are linearly separable when represented by their deciding predicates. This establishes finite observability and guarantees learnability with simple linear models. Synthetic experiments confirm perfect separability under noise-free conditions, while real-data experiments on English morphology show that learned features align with well-known linguistic constraints. These results demonstrate that the subregular hierarchy provides a rigorous and interpretable foundation for modeling natural language structure. Our code used in real-data experiments is available at https://github.com/UTokyo-HayashiLab/subregular.

1 Introduction

The relationship between formal language theory and natural language has long been a subject of investigation, dating back to the pioneering work of Chomsky and Schützenberger (1963), which connected context-free languages to algebraic structures. This tradition has demonstrated that the formal properties of natural language are not arbitrary but are tightly constrained, often residing within computationally restricted subclasses of the regular languages.

In recent decades, research has increasingly focused on the *subregular hierarchy*, a set of language classes strictly contained within the regular languages but capable of capturing a wide range of phonological, morphological, and even syntactic phenomena (Heinz, 2010; Rogers et al., 2013; Heinz et al., 2011; Graf, 2017; Graf and Mayer, 2018; Graf, 2022b,a; Torres et al., 2023; Hanson, 2023). This line of work has revealed that linguistic constraints often fall within these low-complexity classes, explaining their learnability and robustness in acquisition (Heinz and Idsardi, 2011; Saffran et al., 1996). For example, phonotactic restrictions such as vowel harmony, consonant dissimilation, and tonal patterns have been successfully modeled with subregular languages (Rogers et al., 2013; Heinz, 2010; Jardine and Heinz, 2016). Morphological and syntactic dependencies have likewise been shown to align with subregular complexity (Graf, 2017; Chandlee, 2014).

While the algebraic and logical characterizations of subregular classes are well established (Straubing, 1994; Thomas, 1997; Pin, 1995), there has been limited work providing a *geometric* perspective on why these classes matter for learnability. At the same time, statistical learning theory has established powerful results connecting *linear separability* to generalization guarantees via the VC dimension, perceptron mistake bounds, and margin-based analyses (Vapnik, 1998; Blumer et al., 1989; Novikoff, 1962; Cortes and Vapnik, 1995; Shalev-Shwartz and Ben-David, 2014). These results suggest that if subregular classes can be shown to be linearly separable in appropriate finite feature spaces, then we can directly apply statistical learning theory to natural language phenomena.

In this paper, we introduce the notion of *finite observability* and prove that all standard subregular classes are finitely observable, which entails that they are linearly separable under suitable embeddings. This provides a novel *geometric characterization* of subregular languages, complementing existing algebraic and automata-theoretic accounts (McNaughton and Papert, 1971; Schützenberger, 1965; Eilenberg, 1974). By embedding strings into finite-dimensional Boolean spaces defined by primitive predicates, we unify disparate subregular classes within a single framework and open the door to direct application of statistical learning guarantees.

The contributions of this paper are threefold:

1. We define **finite observability**, a property ensuring that membership in a language depends only on finitely many primitive observations (predicates).

2. We prove that all standard subregular classes (SL, SP, LT, PT, LTT, TSL) are finitely observable (see Appendix A.4).

3. We show that finite observability entails **linear separability** in Boolean feature spaces, thereby bridging formal language theory, statistical learning theory, and linguistic applications.

This perspective enriches our theoretical understanding of why natural language patterns are learnable, while also suggesting practical methods for interpretable, efficient models of linguistic constraints.

2 Related Work

Subregular hierarchy: historical and theoretical foundations. The subregular hierarchy has been established as a family of language classes that are both *computationally weaker* than regular languages and yet *linguistically expressive* enough to capture many phonological and morphological constraints. The algebraic characterization of *star-free languages* and their correspondence with first-order logic FO[$<$] (via the classical results of McNaughton, Papert, and Schützenberger) laid much of the theoretical foundation (McNaughton and Papert, 1971; Straubing, 1994; Schützenberger, 1965; Eilenberg, 1974; Pin, 1995; Thomas, 1997). In particular, Simon's characterization of *Piecewise Testable (PT)* languages (Simon, 1975) and the algebraic and logical perspectives on *Locally Testable (LT)* languages (McNaughton and Papert, 1971; Straubing, 1994; Pin, 1995; Thomas, 1997) illustrate how language membership can be determined by *finite local observations*, thereby connecting naturally to SL_k and SP_k. Rogers et al. (2013) systematized the subregular complexity of phonological patterns, while Heinz (2010) showed how long-distance phonotactic restrictions are learnable precisely because they fall within the subregular hierarchy. Tier-based Strictly Local (TSL) languages, which localize nonlocal dependencies via tier projection, have proven to be a powerful extension (Heinz et al., 2011), with applications extending to syntax as well (Graf, 2017; Graf and Mayer, 2018). Our contribution complements this tradition by providing a *geometric* characterization of subregular classes via linear separability, offering an alternative perspective to classical algebraic and automata-theoretic accounts.

Linguistic evidence and applications (phonology, morphology, syntax). Subregular classes have been applied to model local and nonlocal constraints across phonology, morphology, and syntax. In phonology, Strictly Local languages capture cluster restrictions and syllable-structure constraints (Rogers et al., 2013), while Strictly Piecewise languages capture nonadjacent harmony processes (Heinz, 2010). TSL languages handle long-distance vowel harmony and other tier-based phenomena (Heinz et al., 2011; Jardine and Heinz, 2016), while tonal and assimilation/dissimilation processes have also been argued to fall within subregular subclasses (Graf and Mayer, 2018; Avcu et al., 2017). In morphology, affix ordering restrictions and stem–affix cooccurrence patterns have been modeled using SL, LT, and TSL languages (Rogers et al., 2013; Chandlee, 2014). In syntax, dependencies have been reanalyzed in terms of Strictly Local and Strictly Piecewise constraints (Graf, 2017). The present work connects these descriptive successes to *interpretable linear models*, providing a theoretical justification for lightweight learning and inference modules grounded in subregular structure.

Learnability, grammatical inference, and statistical learning theory. The grammatical inference literature has long studied the identifiability of language classes, beginning with Gold's identification-in-the-limit framework (Gold, 1967), Angluin's query learning paradigm (Angluin, 1987), and consolidated in de la Higuera's monograph (de la Higuera, 2010). Early work showed that certain subclasses of regular languages (e.g., k-testable in the strict sense) admit efficient inference algorithms (García and Vidal, 1990; Oncina and García, 1992). For subregular languages, numerous learnability results have been established for SL, SP, and TSL classes (Rogers et al., 2013; Heinz, 2010; Heinz et al., 2011; Graf and Mayer, 2018; Jardine and Heinz, 2016; Chandlee, 2014). Meanwhile, statistical learning theory provides a different angle: linear separability yields strong guarantees via the VC dimension (Vapnik, 1998; Blumer et al., 1989; Shalev-Shwartz and Ben-David, 2014), the perceptron mistake bound (Novikoff, 1962), and SVM margin analyses (Cortes and Vapnik, 1995). Our contribution demonstrates that subregular classes are linearly separable in appropriate finite feature spaces, thereby allowing these guarantees to apply directly to natural language classes themselves. This unifies classical grammatical inference with modern statistical learning theory.

Connections between logic, algebra, and geometry. The logical and algebraic characterizations of regular languages (FO[<], monoids, and aperiodicity (McNaughton and Papert, 1971; Straubing, 1994; Schützenberger, 1965; Pin, 1995; Thomas, 1997; Bojańczyk, 2014)) have traditionally been studied in symbolic and automata-theoretic terms. Recent work has refined the complexity and decidability results for locally testable languages, piecewise testable languages, and their relatives (Place and Zeitoun, 2014; Place et al., 2013). Our contribution provides a complementary *geometric* perspective, recasting these classes as halfspaces in the Boolean hypercube via minterm embeddings. Specifically, any class definable by a finite set of primitive predicates (substrings, subsequences, prefixes, thresholds) is linearly separable with constant margin, and the standard subregular families fall neatly into this framework; a bridge from algebra and logic to statistical learning and convex geometry.

3 Method

This section introduces the two central notions used throughout the paper: *predicates* and *finite observability*, and develops our main theorem that finite observability entails linear separability under an explicit embedding. We also explain the learning-theoretic consequences and provide practical guidance for using low-dimensional predicate features in applications. All formal proofs are deferred to §A.

3.1 Predicates and Finite Observability

Let Σ be a finite alphabet and let $\#$ be a boundary symbol not in Σ. For $K \in \mathbb{N}$ and $x \in \Sigma^*$, we write $\tilde{x} = \#^K\, x\, \#^K$ for the boundary-extended string. A *predicate* is a boolean-valued function

$$p : \Sigma^* \to \{0,1\}, \tag{1}$$

which detects a finite, well-defined property of x (e.g., presence of a particular substring, subsequence, or a thresholded count). We emphasize that predicates are the *atoms of observation*: once we know which predicates hold of x, no additional information about x is required for membership decisions in finitely observable classes.

Finite Observability. A language class $\mathcal{C} \subseteq \mathcal{P}(\Sigma^*)$ is *finitely observable* if there exists a *finite* set of predicates $P = \{p_1, \dots, p_n\}$ such that, for every $L \in \mathcal{C}$, there is a set $S_L \subseteq \{0,1\}^n$ with

$$x \in L \iff (p_1(x), \dots, p_n(x)) \in S_L \qquad \forall x \in \Sigma^*. \tag{2}$$

Thus membership in any language L depends only on the *truth vector* of finitely many primitive predicates.

Illustrative predicates (examples). Below we list the predicate types that will serve as building blocks for standard subregular families: Appendix A.4 gives formal constructions and proofs.

- **Substring predicates (SL$_k$):** for $g \in (\Sigma \cup \{\#\})^k$, $p_g(x) = \mathbf{1}[\, g \text{ occurs contiguously in } \tilde{x}\,]$. *Example:* in an SL$_3$ language forbidding 3-gram `ngt`, the constraint is captured by requiring $p_{\texttt{ngt}}(x) = 0$.
- **Subsequence predicates (SP$_k$):** for $h \in (\Sigma \cup \{\#\})^k$, $q_h(x) = \mathbf{1}[\, h \text{ occurs as a subsequence of } \tilde{x}\,]$. *Example:* in SP$_2$, $q_{\texttt{a\_i}}(x)$ indicates whether `a` precedes `i`; forbidding such pairs encodes vowel-harmony style constraints.
- **Boundary predicates (LT$_k$):** for $u, v \in (\Sigma \cup \{\#\})^{k-1}$, $\pi_u(x) = \mathbf{1}[\, \text{prefix}_{k-1}(\tilde{x}) = u\,]$, $\sigma_v(x) = \mathbf{1}[\, \text{suffix}_{k-1}(\tilde{x}) = v\,]$.
- **Thresholded count predicates (LTT$_{k,\tau}$):** for g with $|g| \leq k$ and $t \in \{1, \dots, \tau_g\}$, $c_{g,t}(x) = \mathbf{1}[\, \#\{g \text{ in } \tilde{x}\} \geq t\,]$.
- **Tiered predicates (TSL$_k$):** for a tier $T \subseteq \Sigma$ and $g \in (T \cup \{\#\})^k$, $s_{T,g}(x) = \mathbf{1}[\, g \text{ occurs in the projection } \pi_T(\tilde{x})\,]$, where π_T erases symbols outside T.

Despite their different linguistic motivations, these predicates are all *finite* in number for fixed class parameters (k, thresholds, and the finite alphabet), and they suffice to decide membership across the corresponding subregular families.

3.2 Minterm Linearization

Given predicates $P = \{p_1, \ldots, p_n\}$, define the *truth vector* $r(x) = (p_1(x), \ldots, p_n(x)) \in \{0,1\}^n$. For each minterm $a \in \{0,1\}^n$, define the *minterm indicator*

$$\mu_a(x) = \prod_{j=1}^{n} \left(p_j(x)^{a_j} \, (1-p_j(x))^{1-a_j}\right) \in \{0,1\}. \tag{3}$$

By construction, $\mu_a(x) = 1$ iff $r(x) = a$; otherwise $\mu_a(x) = 0$. Let the *minterm feature map* be

$$\Phi : \Sigma^* \to \{0,1\}^{2^n}, \qquad \Phi(x) = \big(\mu_a(x)\big)_{a \in \{0,1\}^n}. \tag{4}$$

Lemma 3.1 (Minterm Linearization)**.** For any $S \subseteq \{0,1\}^n$, there exist weights $w \in \mathbb{R}^{2^n}$ and bias $b \in \mathbb{R}$ such that

$$\langle w, \Phi(x)\rangle + b > 0 \iff r(x) \in S \qquad \text{for all } x \in \Sigma^*. \tag{5}$$

Moreover, the separating hyperplane attains functional margin at least $1/2$; the geometric margin equals $1/(2\|w\|_2)$.

Intuition. Each minterm indicator is a one-hot basis vector that uniquely identifies the truth assignment $r(x)$. Any boolean decision rule over $r(x)$, including non-monotone combinations such as XOR, is representable by selecting the corresponding minterm coordinates and thresholding. A short proof appears in Appendix A.2.

3.3 Main Theorem: Finite Observability ⇒ Linear Separability

Theorem 3.2. Let $\mathcal{C}$ be a finitely observable class with predicates $P = \{p_1, \ldots, p_n\}$. Then for every $L \in \mathcal{C}$ there exist (w, b) such that

$$x \in L \iff \langle w, \Phi(x)\rangle + b > 0 \qquad \text{for all } x \in \Sigma^*, \tag{6}$$

where Φ is the minterm feature map in (3). The margin is at least $1/2$.

Explanation. By finite observability, there exists $S_L \subseteq \{0,1\}^n$ such that (2) holds. Lemma 3.1 constructs a single hyperplane that separates minterms in S_L from those outside S_L. Composing with Φ yields the desired linear separator for L. See Appendix A.3 for a formal proof.

Learning-theoretic consequences. Each feature vector $\Phi(x)$ is one-hot with $\|\Phi(x)\|_2 = 1$, so the effective radius is $R = 1$ and the geometric margin is $\gamma \geq 1/(2\|w\|_2)$. Hence a perceptron run on *separable* data suffers at most $M \leq (1/\gamma)^2 \leq 4\|w\|_2^2 = 4|S|$ mistakes (Novikoff, 1962). Moreover, the class of halfspaces in $\mathbb{R}^{2^n}$ has VC dimension at most 2^n (Vapnik, 1998; Blumer et al., 1989), and margin-based generalization bounds apply directly (Cortes and Vapnik, 1995; Shalev-Shwartz and Ben-David, 2014). In short, finitely observable classes are *easy to learn* under standard linear methods.

3.4 From Theory to Practice: Low-dimensional Predicate Features

Although the minterm embedding is conceptually clean, it is exponential in n. In practice, the subregular families admit *direct* low-dimensional predicate features that already separate the target languages:

- **SL$_k$:** presence/absence of k-grams (with boundary padding) suffices to implement forbidden-substring constraints.
- **SP$_k$:** presence/absence of k-length subsequences suffices to implement forbidden non-adjacent pairs/tuples.
- **LT$_k$:** k-gram features augmented with $(k-1)$-length prefix/suffix indicators determine membership.
- **PT$_m$:** presence/absence of subsequences of length $\leq m$ (Simon's theorem (Simon, 1975)).
- **LTT$_{k,\boldsymbol{\tau}}$:** bounded-threshold counts for substrings of length $\leq k$.
- **TSL$_k$:** k-gram features on tier projections π_T (for $T \subseteq \Sigma$) capture nonlocal constraints through local ones on the tier (Heinz et al., 2011).

These directly usable features retain the interpretability of symbolic constraints and allow linear classifiers (logistic regression, linear SVMs, perceptrons) to achieve exact separation on the intended targets while remaining computationally lightweight.

Scope and limitations. Our results require fixing class parameters (e.g., k, m, $\boldsymbol{\tau}$) and the finite alphabet, which is standard when defining target language families. Unions over unbounded parameters (e.g., PT $= \bigcup_m \mathrm{PT}_m$) should be viewed as directed unions of finitely observable subclasses

(Appendix A.4). While the minterm embedding itself is exponential, practical predicate sets for the standard subregular families are polynomial in $|\Sigma|$ and the fixed parameters, and can be trained with sparse linear methods.

4 A Counterexample Showing the Limits of Fixed Finite Predicate Sets

Our main results hinge on *finite observability*: once a finite set of predicates $P = \{p_1, \ldots, p_n\}$ is fixed, any language definable in this framework is a union of the finitely many *predicate cells* (truth assignments) in $\{0,1\}^n$. This section shows that, for any *fixed* finite predicate set P, there exist regular languages that cannot be detected (i.e., represented as unions of predicate cells), hence cannot be linearly separated in the corresponding feature space. The proof relies on the Myhill–Nerode theorem (Sipser, 2012).

Setup. Fix any finite alphabet Σ and any finite predicate set $P = \{p_1, \ldots, p_n\}$ with $n \geq 1$. Let $r(x) = (p_1(x), \ldots, p_n(x)) \in \{0,1\}^n$. The induced equivalence relation $\equiv_P$ on Σ^* groups two strings iff they share the same truth vector: $x \equiv_P y \iff r(x) = r(y)$. There are at most 2^n such equivalence classes (*cells*). Any language expressible via P is necessarily a union of $\equiv_P$-classes.

Theorem 4.1 (Regular language not representable by a fixed P)**.** For any fixed finite predicate set P over Σ, there exists a regular language L such that L is not a union of $\equiv_P$-classes. Equivalently, no linear classifier over the fixed predicate feature space can separate L from its complement.

Proof. Let $m > 2^n$ be any integer and consider the one-letter alphabet $\Sigma = \{a\}$. Define the regular language

$$L_m = \{\, a^k \mid k \equiv 0 \pmod{m} \,\}.$$

It is well known (and follows from Myhill–Nerode) that L_m has exactly m distinct right-congruence classes, corresponding to length modulo m (Sipser, 2012). For any equivalence relation E whose index (number of classes) is strictly smaller than m, E cannot refine the Myhill–Nerode congruence of L_m; hence L_m cannot be a union of E-classes. Since $\equiv_P$ has at most 2^n classes and $m > 2^n$, $\equiv_P$ does not refine the Myhill–Nerode congruence of L_m. Therefore L_m is not a union of $\equiv_P$-classes and cannot be represented in the predicate framework determined by P. Because linear classification over P can only realize unions of $\equiv_P$-cells, no linear separator exists for L_m in that fixed feature space. ■

Implications. Theorem 4.1 formalizes an intuitive limitation: fixing a finite observation vocabulary P yields at most 2^n observable profiles, while regular languages can require arbitrarily many distinct right-congruence classes (e.g., m for L_m), even over a unary alphabet. Thus, unlike the subregular families discussed in this paper, the class of all regular languages cannot be captured (nor linearly separated) within any single finite predicate space. This clarifies the scope of our finite-observability program: it targets *families whose essential structure is visible through a bounded set of primitive observations*.

Star-Free Languages as an Intermediate Case. In addition to regular languages, it is instructive to consider *star-free* languages. They are strictly weaker than the full class of regular languages but strictly more expressive than the standard subregular families. Classical results show that locally testable and piecewise testable languages are contained within star-free, while strictly local and strictly piecewise families form even smaller subclasses (McNaughton and Papert, 1971; Brzozowski and Simon, 1973; Simon, 1975; Rogers et al., 2007). Thus we have a proper hierarchy:

$$\text{SL, SP} \subset \text{LT, PT} \subset \text{Star-Free} \subset \text{Regular}.$$

Unlike the subregular classes, the star-free family does not admit finite observability. This follows from the fact that FO[<], which characterizes star-free languages, forms strict quantifier-rank and dot-depth hierarchies (Brzozowski and Cohen, 1971; Kufleitner and Lauser, 2012; Place and Zeitoun, 2021). No finite predicate basis suffices to capture all star-free languages. Thus, star-free languages provide a useful counterexample: finite observability holds robustly for the subregular families but breaks down once we move upward in expressive power, even before reaching all regular languages.

5 Synthetic Experiments

5.1 Linear Separability

To empirically validate our theoretical results, we conducted controlled experiments on artificially

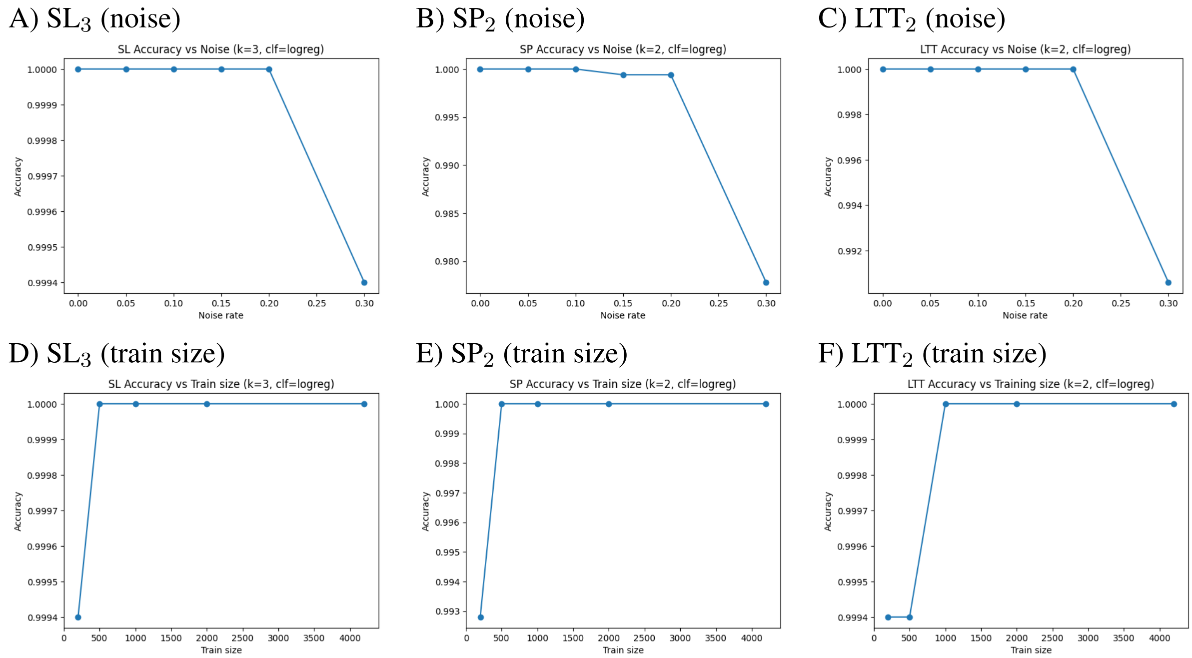

Figure 1: Accuracy of synthetic classification: (A) SL_3, (B) SP_2 and (C) LTT_2 under label noise, (D) SL_3, (E) SP_2 and (F) LTT_2 with varying training sizes.

generated languages. The goal of these experiments is not to maximize accuracy per se, but rather to visualize how linear separability manifests across different subregular classes when noise or sample size is varied.

To keep the presentation concise and focused, we concentrate on three representative classes from the subregular hierarchy: *Strictly Local (SL)*, *Strictly Piecewise (SP)*, and *Locally Threshold Testable (LTT)*. SL captures local adjacency constraints and provides the most transparent connection to phonotactic well-formedness. SP models long-distance restrictions, such as co-occurrence bans, which are common in phonology and morphology. Finally, LTT extends local tests with counting thresholds, allowing us to probe more complex phenomena and to examine how separability behaves as the constraints become richer.

For each class, we generate positive and negative examples according to the defining predicates, then train logistic regression models using the corresponding predicate features. By sweeping noise levels and training set sizes, we observe how empirical performance converges toward the theoretical guarantee of linear separability.

5.1.1 Strictly Local Languages (SL_k)

Setup. We generated synthetic datasets for SL_3 languages. A set of forbidden 3-grams was defined over a finite alphabet with explicit inclusion of boundary symbols <#>, </#>. Positive examples were sampled uniformly from sequences that avoided all forbidden 3-grams, while negative examples were constructed to guarantee the presence of at least one forbidden 3-gram (including those at boundaries). Each string length was sampled randomly within a fixed range (5–15 characters). The feature representation was the full set of boundary-aware 3-grams over the extended alphabet $\Sigma \cup$ {<#>, </#>}. This guarantees that the deciding predicates are included.

Results. Figure 1 (A) shows the accuracy under a noise sweep, where increasing proportions of training labels were randomly flipped. As predicted, performance is exactly 100% in the noise-free setting, and decreases as noise increases. Figure 1 (D) shows the size sweep, where training set size was varied. Performance converges rapidly toward 100% as more data are available, reflecting the finite observability of deciding predicates.

5.1.2 Strictly Piecewise Languages (SP_k)

Setup. For SP_2 languages, we defined a forbidden set of subsequences of length two (e.g., (a, c), (b, d)) over an alphabet of four symbols. Positive examples were strings avoiding these subsequences, while negative examples were constructed to ensure that at least one forbidden subsequence occurred. String lengths were randomly sampled in the range 6–18 characters. The feature representation was the set of all possible symbol pairs (ordered subsequences) over the alphabet. This again ensures that all deciding predicates are explicitly represented in the feature space.

Results. As with SL_3, Figure 1 (B) shows accuracy under a noise sweep, and Figure 1 (E) shows results for a size sweep. In the absence of label noise, classification achieves 100% accuracy, exactly matching theoretical expectations. When noise is introduced, performance degrades smoothly, illustrating the robustness of the separating representation.

5.1.3 Locally Threshold Testable Languages (LTT_k)

Setup. For the LTT_2 condition, we defined membership in terms of thresholded counts of substrings with boundaries. Specifically, positive examples were required to contain at least two instances of the symbol a, begin with a at the left boundary, contain at most one occurrence of the 2-gram bb, and contain at most one occurrence of c at the right boundary. Positive examples were generated by rejection sampling with targeted repairs to ensure that all threshold constraints were satisfied. Negative examples were produced by minimally breaking one constraint, for example by reducing the number of a's, inserting an additional bb, altering the prefix so it no longer begins with a, or forcing a terminal c. This guaranteed that every negative sequence violated at least one deciding predicate. Feature vectors were constructed from all substrings of length at most two, extended with boundary markers, with each feature indicating whether the count of that substring exceeded a given threshold. These features explicitly encode the deciding predicates of LTT_2.

Results. Figure 1 (C) shows the effect of label noise on classification accuracy, and Figure 1 (F) shows the effect of training size. In the noise-free setting, accuracy reached 100%, confirming the theoretical prediction of perfect separability. As noise increased, accuracy decreased, while larger training sizes quickly restored performance toward the ideal. These results provide empirical evidence that threshold-based constraints in LTT languages are finitely observable and linearly separable when represented with appropriate features.

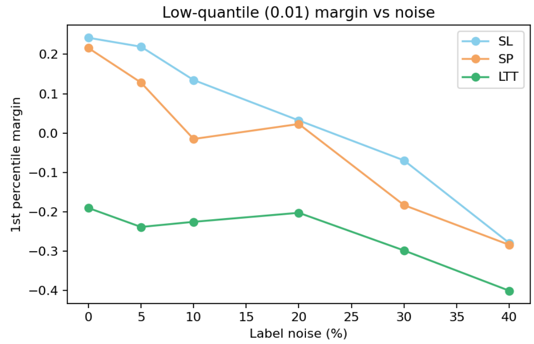

Figure 2: 1% quantile normalized margin vs. label noise for SL, SP, and LTT languages.

5.2 Low-Quantile Margins under Label Noise

Setup. To further probe the robustness of linear separability, we measured how the low-quantile normalized margins behave as label noise increases. For each subregular class: Strictly Local (SL), Strictly Piecewise (SP), and Locally Threshold Testable (LTT), we trained logistic regression classifiers on synthetic datasets of size 2,000, where a fraction of the training labels were flipped at random. Test sets of 1,000 sequences per class were kept noise-free. After training, we computed the normalized margin for each test example, defined as $m_i = y_i \cdot \frac{w^\top x_i + b}{\|w\|_2}$ and extracted the 1st percentile ($q_{0.01}$) across the distribution. This quantile analysis highlights the fragility of the worst-classified points, thus serving as a probe of separability under noise.

Results. Figure 2 shows the 1% quantile margin as a function of label noise. SL (light blue) maintains relatively high margins at low noise levels and exhibits a gradual decline, indicating strong robustness of local constraints. SP (light orange) drops more quickly with noise, reflecting the fact that long-distance constraints are easily disrupted by even a small number of mislabels. LTT (light green) displays consistently negative margins, as its threshold-based definition inherently places a significant portion of examples near the decision boundary; nevertheless, its degradation with noise is moderate and smoother than that of SP.

6 Experiments on Morphological Data

Setup. To test the applicability of our theoretical findings to natural language data, we conducted experiments on English derivational morphology using the MorphoLex-en database (Sánchez-Gutiérrez et al., 2018). MorphoLex provides a large-scale lexicon of English words annotated with morphological features, including derivational affixes. We extracted affix sequences for each word, consisting of ordered prefixes followed by suffixes. Since MorphoLex does not always provide explicit segmentation, we complemented the data with a heuristic segmenter based on a comprehensive inventory of common English affixes. This yielded a dataset in which each item was represented by a sequence of affixes.

To create a classification task aligned with the subregular framework, we constructed positive examples from well-formed affix sequences in MorphoLex, and generated negative examples by minimally perturbing them. Specifically, negatives were created by either permuting the order of affixes or substituting one affix with another from the same vocabulary. This ensured that negative sequences closely resembled real ones in length and vocabulary but violated natural morphological constraints. The resulting dataset was divided into training, development, and test sets, with splits stratified by word to avoid leakage.

We represented each affix sequence using two subregular feature classes: Piecewise Testable (PT) predicates, capturing all subsequences up to length $m = 2$, and Locally Threshold Testable (LTT) predicates, encoding boundary-aware $k = 2$. As in the synthetic experiments, we trained logistic regression classifiers on these feature vectors. Importantly, the classifiers were not constructed directly from the theoretical separating hyperplanes, but were trained from data, demonstrating that linear separability is empirically recoverable by standard learning algorithms.

Results. On the held-out test set, the classifiers achieved strong performance. Logistic regression with PT+LTT features typically reached accuracy above 85%, with F1 scores 86%. Confusion matrices in Figure 3 showed that more errors occurred in negative items that were close to positive forms in surface structure. To better understand what the classifier has learned from the real morphological data, we inspected the top-weighted predicates in the logistic regression model (Figure 4). Strikingly, the most influential features correspond to linguistically interpretable affix constraints. For instance, the model strongly weights boundary-sensitive predicates such as the presence of *-ly* or

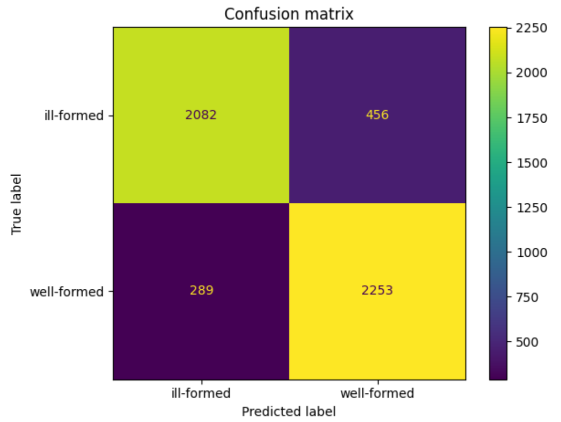

Figure 3: Confusion matrix for the morphological well-formedness classification task, showing accurate separation of well-formed (positive) and ill-formed (negative) affix sequences.

-ness at the right edge, reflecting the well-attested distributional restrictions of these suffixes. Similarly, certain prefix–suffix interactions emerge as highly discriminative, such as restrictions involving *dis-* or *re-* in combination with other affixes. These results demonstrate that the classifier is not only achieving high predictive accuracy, but is also recovering linguistically meaningful generalizations that align with theoretical expectations in morphology.

Analysis. To assess the practical linear separability of morphological patterns, we analyzed the distribution of normalized margins on the held-out test set. For each word, the margin was defined as $m_i = y_i \cdot \frac{w^\top x_i + b}{\|w\|_2}$, where $y_i \in \{+1, -1\}$ is the gold label, x_i the feature vector, w the learned weight vector, and b the intercept. A positive margin indicates that the example is correctly classified and lies on the correct side of the separating hyperplane, while the magnitude of the margin reflects the confidence of the classification.

Figure 5 shows the histogram of margins for all test examples. The distribution is clearly skewed toward positive values, with the majority of examples clustered around small but positive margins. This indicates that the classifier has successfully learned a separating hyperplane consistent with the morphological constraints. At the same time, the long tail of low and slightly negative margins reveals the presence of ambiguous or difficult cases, such as words with rare affixes or irregular formations.

The overall pattern supports our theoretical

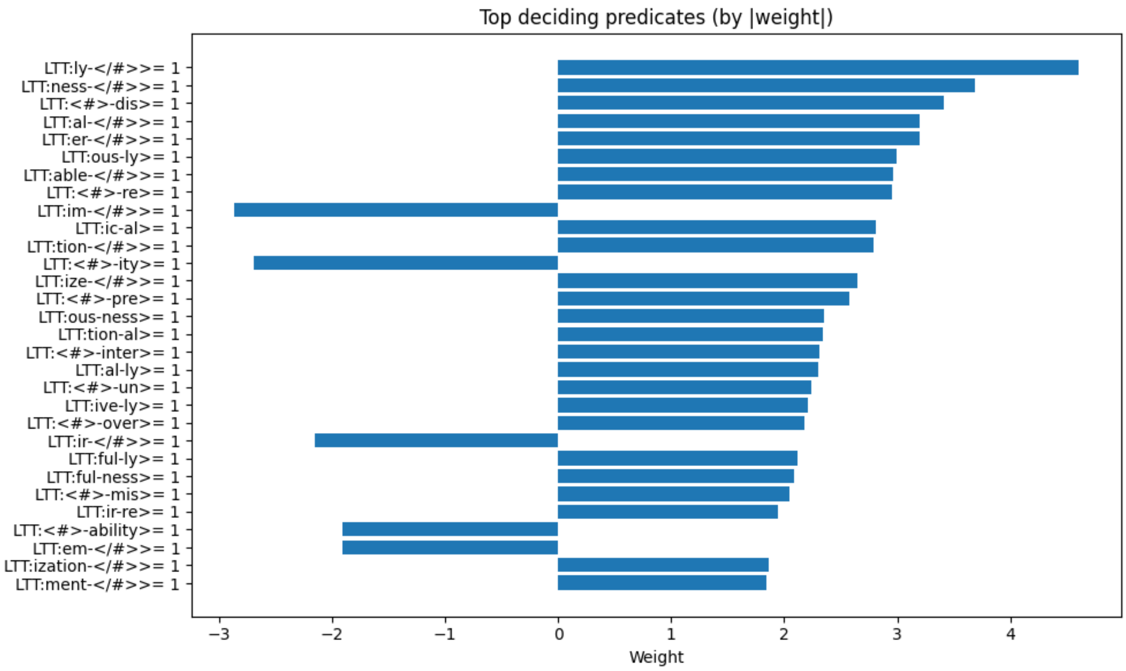

Figure 4: Top-weighted predicates learned by the classifier on the morphological dataset, illustrating linguistically meaningful affix constraints such as boundary-sensitive suffixes and prefix–suffix co-occurrence restrictions.

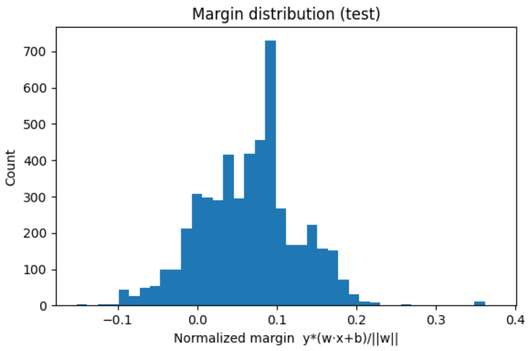

Figure 5: Histogram of normalized margins on the test set. Most examples lie at positive margins, confirming effective linear separation.

claim: although real data inevitably contains noise and irregularities, the linear predicate-based representation yields a decision boundary that cleanly separates most examples. The margin distribution therefore provides empirical evidence that subregular constraints, when instantiated on lexical data, remain both observable and learnable within a linear framework.

7 Conclusion

This work has shown that all major subregular language classes are linearly separable when expressed through their deciding predicates. This establishes finite observability across the subregular hierarchy and guarantees learnability with simple linear models. Synthetic and morphological experiments confirm these guarantees while also recovering linguistically interpretable constraints, thereby providing a geometric foundation for lightweight, interpretable models that connects naturally to earlier probabilistic approaches such as Hayes and Wilson (2008)'s Maximum Entropy model of phonotactics.

Our perspective contrasts with research on the expressive power of recurrent neural networks (RNNs), which emphasizes their maximal computational capacity, from Turing completeness under infinite precision (Siegelmann and Sontag, 1992) to pushdown-automaton simulations with LSTMs (Weiss et al., 2018). Rather than focusing on what is computationally possible in principle, we identify a restricted region of the hierarchy where natural language patterns actually reside and show that these classes are inherently simple to learn. Future work will examine whether larger families admit analogous notions of observability, and whether the effective capacity of neural models in practice aligns more closely with subregular structure than with their theoretical upper bounds.

Class	Predicate set P (finite)	Rationale
SL_k	k-gram presence in $\tilde{x}$	forbidden substrings
SP_k	k-subsequence presence in $\tilde{x}$	forbidden subsequences
LT_k	k-grams + $(k-1)$ prefix/suffix	local testability
PT_m	subsequences of length $\leq m$	Simon's theorem
$\mathrm{LTT}_{k,\boldsymbol{\tau}}$	thresholded counts for $\lvert g\rvert \leq k$	bounded counting
TSL_k	k-grams on all tiers $\pi_T(\tilde{x})$	tier projection + SL

Table 1: Summary: Finite observability of standard subregular classes.

A Appendix: Proofs and Constructions

A.1 Notation and Preliminaries

We write $\mathcal{P}(\Sigma^*)$ for the power set of Σ^*, i.e., the set of all languages over Σ. For a boolean condition φ, we use $\mathbf{1}[\varphi]$ to denote its indicator in $\{0, 1\}$. The boundary-extended string is $\tilde{x} = \#^K x \#^K$ with K chosen per construction so that all local observations are well-defined.

A.2 Proof of Lemma 3.1 (Minterm Linearization)

Let $S \subseteq \{0, 1\}^n$, and define weights $w \in \mathbb{R}^{2^n}$ and bias $b \in \mathbb{R}$ by

$$w_a = \begin{cases} 1 & \text{if } a \in S, \\ 0 & \text{otherwise,} \end{cases} \qquad b = -\tfrac{1}{2}.$$

For any x, exactly one minterm indicator equals 1, namely $\mu_{r(x)}(x)$, and all others are 0. Therefore $\sum_a w_a \mu_a(x) = \mathbf{1}[\, r(x) \in S \,]$. Adding $b = -\frac{1}{2}$ yields $+\frac{1}{2}$ on positives and $-\frac{1}{2}$ on negatives, establishing linear separability with margin at least $1/2$. ■

A.3 Proof of Theorem 3.2

By finite observability, there exists $S_L \subseteq \{0, 1\}^n$ such that $x \in L$ iff $r(x) \in S_L$. Apply Lemma 3.1 to $S = S_L$ to obtain (w, b) with $\langle w, \Phi(x)\rangle + b > 0$ iff $r(x) \in S_L$. This proves the theorem, with margin at least $1/2$ inherited from Lemma 3.1. ■

A.4 Finite Observability of Standard Subregular Classes

We sketch predicate sets witnessing finite observability for each standard family; detailed definitions match those in the main text. We summarize these results in Table 1.

SL_k. Let $P = \{p_g : g \in (\Sigma \cup \{\#\})^k\}$ where $p_g(x) = \mathbf{1}[\, g \prec \tilde{x} \,]$. A language L is specified by a finite forbidden set F of k-grams; membership is equivalent to $p_g(x) = 0$ for all $g \in F$. Thus $S_L = \{v \in \{0, 1\}^{|P|} : v_g = 0 \,\forall g \in F\}$.

SP_k. Let $P = \{q_h : h \in (\Sigma \cup \{\#\})^k\}$ where $q_h(x) = \mathbf{1}[\, h \sqsubseteq \tilde{x} \,]$ (subsequence). With forbidden subsequences F, set $S_L = \{v : v_h = 0 \;\forall h \in F\}$.

LT_k. Let P contain all k-gram indicators together with prefix/suffix indicators of length $k-1$: $p_g(x) = \mathbf{1}[\, g \prec \tilde{x} \,]$, $\pi_u(x) = \mathbf{1}[\, \mathrm{prefix}_{k-1}(\tilde{x}) = u \,]$, $\sigma_v(x) = \mathbf{1}[\, \mathrm{suffix}_{k-1}(\tilde{x}) = v \,]$. By the classical characterization of LT_k, membership depends only on this predicate vector, hence some S_L exists.

PT_m. Let $P = \{r_u : u \in \bigcup_{\ell=1}^{m} (\Sigma \cup \{\#\})^\ell\}$ with $r_u(x) = \mathbf{1}[\, u \sqsubseteq \tilde{x} \,]$. By Simon's theorem (Simon, 1975), membership depends only on the set of subsequences of length $\leq m$, hence some S_L exists.

$\mathrm{LTT}_{k,\boldsymbol{\tau}}$. Let $P = \{c_{g,t} : |g| \leq k,\; t \in \{1, \ldots, \tau_g\}\}$ with $c_{g,t}(x) = \mathbf{1}[\, \#\{g \text{ in } \tilde{x}\} \geq t \,]$, optionally augmented with thresholded prefix/suffix indicators of length $k-1$. Counts saturate at τ_g, so P is finite and determines membership.

TSL_k. Let $\mathcal{T} = \{T \subseteq \Sigma\}$ be the (finite) set of tiers. For each $T \in \mathcal{T}$ and $g \in (T \cup \{\#\})^k$, include $s_{T,g}(x) = \mathbf{1}[\, g \prec \pi_T(\tilde{x}) \,]$. If L is specified by (T, F_T) (tier T and forbidden tier k-grams), then membership is equivalent to $s_{T,g}(x) = 0$ for all $g \in F_T$, i.e., $(s_{T,g}(x)) \in S_L$ for a suitable S_L.

Unions of parameterized subclasses. Families like $\mathrm{PT} = \bigcup_m \mathrm{PT}_m$ or $\mathrm{SL} = \bigcup_k \mathrm{SL}_k$ are directed unions of finitely observable subclasses. Unless parameters are bounded, no single finite predicate set suffices for the whole union; this does not affect applications where the parameters are fixed by design.

References

Dana Angluin. 1987. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106.

Emre Avcu, Adam Brodsky, and Jeffrey Heinz. 2017. Subregular complexity and reduplication. In *Proceedings of NELS 47*.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the vapnik–chervonenkis dimension. *Journal of the ACM*, 36(4):929–965.

Mikolaj Bojańczyk. 2014. Logic and regular languages. In *Encyclopedia of Algorithms*. Springer.

Janusz A. Brzozowski and Ralph S. Cohen. 1971. Dot-depth of star-free events. *Journal of Computer and System Sciences*, 5(1):1–16.

Janusz A Brzozowski and Imre Simon. 1973. Characterizations of locally testable events. *Discrete Mathematics*, 4(3):243–271.

Jane Chandlee. 2014. *Strictly Local Phonological Processes*. Ph.D. thesis, University of Delaware.

Noam Chomsky and Marcel-Paul Schützenberger. 1963. The algebraic theory of context-free languages. In *Computer Programming and Formal Systems*, pages 118–161. North-Holland.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Samuel Eilenberg. 1974. *Automata, Languages, and Machines, Vol. A*. Academic Press, New York.

Pedro García and Enrique Vidal. 1990. Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):920–925.

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

Thomas Graf. 2017. *The power of locality: Strictly local and strictly piecewise languages in syntax*. Ph.D. thesis, University of California, Los Angeles.

Thomas Graf. 2022a. Diving deeper into subregular syntax. *Theoretical Linguistics*, 48(3-4):245–278.

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

Thomas Graf and Connor Mayer. 2018. The subregular complexity of tonal patterns. *Linguistic Inquiry*, 49(2):393–406.

Kenneth Hanson. 2023. Strict locality in syntax. *Proceedings of CLS 2023*.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Jeffrey Heinz. 2010. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623–661.

Jeffrey Heinz and William J. Idsardi. 2011. Sentence and word complexity. *Science*, 333(6040):295–297.

Jeffrey Heinz, Chetan Rawal, and Herbert Tanner. 2011. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Colin de la Higuera. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.

Adam Jardine and Jeffrey Heinz. 2016. Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics*, 4:87–98.

Manfred Kufleitner and Alexander Lauser. 2012. Around dot-depth one. *International Journal of Foundations of Computer Science*, 23(06):1323–1339.

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

Albert B. J. Novikoff. 1962. On convergence proofs for perceptrons. In *Proceedings of the Symposia in Applied Mathematics*, volume 12, pages 615–622.

José Oncina and Pedro García. 1992. Inferring regular languages in polynomial update time. In *Pattern Recognition and Image Analysis*, pages 49–61.

Jean-Éric Pin. 1995. Logic, semigroups and automata on words. *Annals of Mathematics and Artificial Intelligence*, 16(1–2):343–384.

Thomas Place, Léo van Rooijen, and Marc Zeitoun. 2013. Separating regular languages with two quantifier alternations. In *Mathematical Foundations of Computer Science (MFCS)*, pages 729–740.

Thomas Place and Marc Zeitoun. 2014. Separating regular languages with first-order logic. *Logical Methods in Computer Science*, 10(3):1–28.

Thomas Place and Marc Zeitoun. 2021. Separation for dot-depth two. *Logical Methods in Computer Science*, 17.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2007. On languages piecewise testable in the strict sense. In *Conference on Mathematics of Language*, pages 255–265. Springer.

James Rogers, Jeffrey Heinz, Margaret Feroe, et al. 2013. Subregular complexity and phonological patterns. *Lingua*, 128:72–90.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. 2018. Morpholex: A derivational morphological database for 70,000 english words. *Behavior research methods*, 50(4):1568–1580.

Marcel-Paul Schützenberger. 1965. On finite monoids having only trivial subgroups. *Information and Control*, 8(2):190–194.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Hava T Siegelmann and Eduardo D Sontag. 1992. On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 440–449.

Imre Simon. 1975. Piecewise testable events. In *Proceedings of ICALP*, pages 214–222.

Michael Sipser. 2012. *Introduction to the Theory of Computation*, 3rd edition. Cengage Learning.

Howard Straubing. 1994. *Finite Automata, Formal Logic, and Circuit Complexity*. Birkhäuser, Boston.

Wolfgang Thomas. 1997. Languages, automata, and logic. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages*, volume 3, pages 389–455. Springer.

Charles Torres, Kenneth Hanson, Thomas Graf, and Connor Mayer. 2023. Modeling island effects with probabilistic tier-based strictly local grammars over trees. *Proceedings of the Society for Computation in Linguistics*.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision rnns for language recognition. *arXiv preprint arXiv:1805.04908*.