Reducing the Local Alphabet Size in Tiling Systems for Picture Languages *

Stefano Crespi Reghizzi[1], Antonio Restivo[2], and Pierluigi San Pietro[1]

[1] Politecnico di Milano - DEIB
[2] Dipartimento di Matematica e Informatica, Università di Palermo

Abstract. The family of recognizable picture (or 2D) languages is defined as the projection of a local picture language defined by a set of two-by-two tiles, i.e., by a 2-strictly-locally-testable (2-SLT) 2D language, formalized in terms of a tiling system. A basic complexity measure of such 2D languages is the size of the 2-SLT alphabet, more precisely the so-called alphabetic ratio of sizes: 2-SLT-alphabet / picture-alphabet. Such 2D language family can also be defined by the projection of larger k-by-k tiles, $k > 2$, i.e., by a k-SLT 2D language. Studying how the alphabetic ratio changes moving from $k = 2$ to larger values, we obtain the following: any recognizable picture language over an alphabet of size n is the projection of an SLT language over an alphabet of size $2n$. Moreover, two is the minimal alphabetic ratio possible in general. It is suggestive to rephrase the result in the case of black-and-white pictures: any such picture is the projection of an SLT picture that uses just four colors. This result lifts into 2D a known similar property (called Extended Medvedev Theorem) of regular word languages, concerning the minimal alphabetic ratio needed to define a language by means of a projection of an SLT word language. The proof uses 2D comma-free codes, which have the property of synchronizability.

1 Introduction and preliminaries

Formal language families can be defined by different approaches: automata, generative rules, logical formulas, and by mapping one language family into another one. Thus, the original definition of the regular language family by means of finite automata and regular expressions was later supplemented with other definitions, in particular by means of the homomorphic image of the simpler language family known as *local* languages. The latter definition is also known as Medvedev's theorem [10, 11], for short *MT*. Each local language is characterized by the local testability property, stating that a word is valid iff its 2-factors, i.e., its substrings of length two, are included in a given set. Then, MT says that for each regular language R over an alphabet Σ there exists a local language L over a *local* alphabet Λ and a letter-to-letter morphism $h : \Lambda^* \to \Sigma^*$ s.t. $R = h(L)$.

The present work deals with languages whose elements, called *pictures*, are not words but rectangular arrays of cells each one containing a letter. Clearly, a word (or *1D*)

 The full version of this extended abstract is in [3].

language is also a picture (or *2D*) language that only comprises rectangles with unitary (say) height. The well-known family of 2D languages, named *tiling recognizable* REC [7], has its primary natural definition through the analog in 2D of MT. The definition relies on a *tiling system* (TS), consisting of a local picture language and of a morphism from the local alphabet Λ to the picture alphabet Σ. More precisely, the local language is specified by a set of tiles that are pictures of size two-by-two and play the role of the 2-factors of the local word languages. Notice other definitions of the REC family by means of automata and logical formulas [7], are less convenient than the corresponding definitions for 1D languages.

We recall the MT definition of regular 1D languages. The letters of the local alphabet Λ correspond to the elements of the product $Q \times \Sigma$, where Q is the state set of a finite automaton (FA) recognizing the regular language. Hence, the *ratio* $|\Lambda|/|\Sigma|$ of the alphabet sizes is a meaningful measure of the complexity of the FA, inducing an infinite hierarchy under set inclusion for certain series of regular languages. For REC the situation is similar, but less investigated. We mention that any limitation of the alphabetic ratio of tiling systems may restrict the language family.

Local 1D languages are located at the lowest level of an infinite language hierarchy (under set inclusion) called k-*strictly locally testable* (SLT) [9], each level $k \geq 2$ being characterized by the use of a sliding window of width k, used to scan the input. A similar hierarchy exists also for 2D languages, see [6], using k-tiles instead of k-factors. The corresponding 2D language family is here called k-*strictly locally testable* (k-SLT); or simply SLT when the (finite) value of parameter k is left unspecified.

The first basic question one can address is: if in the tiling system definition of REC we allow k-SLT languages with $k > 2$, do we obtain a language family larger than REC? The answer is known to be negative [6, 8]. Then, the next question is interesting: using k-tiles with $k > 2$, can we reduce the alphabetic ratio, and how much?

The answer to the latter question is known for 1D languages [4], and we refer to it as the *Extended Medvedev's Theorem* (EMT): the minimal alphabetic ratio is 2, and is obtained by assigning to the SLT parameter k a value logarithmic in the FA recognizer size. We hint to its proof, which is the starting point of the present development for pictures. Given an FA, the construction in the proof samples in each computation the sub-sequences made by k state-transitions. Then the construction encodes all such sub-sequences by means of a binary code of length k. To prevent mistakes, the proof resorts to codes that can be decoded without synchronization, using a $2k$-SLT DFA as decoder. The family of *comma-free* codes [2] has such a property, and is the one we use also here, but in 2D.

Moving from words to pictures often complicates matters, and we had to examine and discard several possible ways of encoding the tiling process operated by the TS, before founding a successful one. The result extends EMT from regular 1D languages to REC and says that any picture language is the projection, by means of a letter-to-letter morphism, of an SLT 2D language over an alphabet with size double of the original alphabet size; the alphabetic ratio 2 is the minimal possible. It is suggestive to rephrase the result in the case of black and white pictures: any such picture is the projection of a strictly locally testable picture that uses just four colors.

All the alphabets to be considered are finite. The following concepts and notations for picture languages follow mostly [7]. A *picture* is a rectangular array of letters over an alphabet. Given a picture p, $|p|_{row}$ and $|p|_{col}$ denote the number of rows and columns, respectively; $|p| = (|p|_{row}, |p|_{col})$ denotes the *picture size*. Pictures of identical size are called *isometric*. The set of all pictures over Σ of size (m, n) is denoted by $\Sigma^{m,n}$. The set of all non-empty pictures over Σ is denoted by Σ^{++}. This notation is naturally extended from an alphabet to a finite set X of isometric pictures, by writing X^{++}. A 2D (or picture) language over Σ is a subset of Σ^{++}. For brevity, the term "language" will always denote a 2D language.

Let $p, q \in \Sigma^{++}$. The *horizontal* (or *column*) *concatenation* $p \oplus q$ is defined when $|p|_{row} = |q|_{row}$ as: $\boxed{p\ q}$. The *vertical* (or *row*) *concatenation* $p \ominus q$ is defined when $|p|_{col} = |q|_{col}$ as: $\boxed{\begin{array}{l} p \\ q \end{array}}$. Concatenations are extended to languages in the obvious way.

Since the pixels on the boundary of a picture play often a special role for recognition, it is convenient to surround a picture p by a frame of width one comprising only the special symbol $\# \notin \Sigma$. Such *bordered picture* is denoted by $\widehat{p}$ and has size $(|p|_{row} + 2, |p|_{col} + 2)$ and domain $\{0, 1, \cdots, |p|_{row} + 1\} \times \{0, 1, \cdots, |p|_{col} + 1\}$.

We denote by $B_{k,k}(p)$, $k \geq 2$, the set of all subpictures, named *k-tiles*, of picture p having size (k, k). If one or both dimensions of p are smaller than k, let $B_{k,k}(p) = \emptyset$. $B_{k,k}(L) = \bigcup_{p \in L} B_{k,k}(\widehat{p})$ is the set of subpictures of size (k, k) of a language L.

Definition 1 (picture morphism). *Given two alphabets Γ, Λ, a (picture) morphism is a mapping $\varphi : \Gamma^{++} \to \Lambda^{++}$ s.t., for all $p, q \in \Gamma^{++}$:* $\begin{cases} i) \ \ \varphi(p \oplus q) = \varphi(p) \oplus \varphi(q) \\ ii) \ \varphi(p \ominus q) = \varphi(p) \ominus \varphi(q) \end{cases}$

Since $\oplus$ is a partial operation, to satisfy $i)$ we need that for all $p, q \in \Gamma^{++}$, $p \oplus q$ is satisfied iff $\varphi(p) \oplus \varphi(q)$; and similarly for $\ominus$ to satisfy $ii)$. This implies that the images by φ of the elements of alphabet Γ are isometric, i.e., for any $x, y \in \Gamma$, $|\varphi(x)|_{row} = |\varphi(y)|_{row}$ and $|\varphi(x)|_{col} = |\varphi(y)|_{col}$.

2 Strictly locally testable 2D languages and tiling recognition

Definition 2. *Given $k \geq 2$, a language $L \subseteq \Sigma^{++}$ is k-*strictly-locally-testable *(k-SLT) if there exists a finite set T_k of (allowed) k-*tiles *in $(\Sigma \cup \{\#\})^{k,k}$ s.t. $L = \{p \in \Sigma^{++} \mid B_{k,k}(\widehat{p}) \subseteq T_k\}$; we also write $L = L(T_k)$. A language L is called strictly-locally-testable (SLT) if it is k-SLT for some $k \geq 2$.*

In other words, a k-SLT picture language can be recognized by looking at its subpictures of size (k, k) and checking their inclusion in a given finite set. Local languages correspond to the special case $k = 2$. This definition ignores of with size less than (k, k), which anyway amount to a finite language. In the following, when comparing picture languages defined in terms of k-SLT languages, we ignore those finite sets.

Tiling recognition. Let Γ and Σ be alphabets; given a mapping $\pi : \Gamma \to \Sigma$, to be termed *projection*, we extend π to isometric pictures $p' \in \Gamma^{++}$, $p \in \Sigma^{++}$ by: $p = \pi(p')$ s.t. $p_{i,j} = \pi(p'_{i,j})$ for all $(i, j) \in \mathrm{dom}(p')$. Then, p' is called the *pre-image* of p.

Definition 3 (tiling system). *A* tiling system *(TS) is a quadruple* $\mathcal{T} = (\Sigma, \Gamma, T, \pi)$ *where* Σ *and* Γ *are alphabets,* T *is a 2-tile set over* $\Gamma \cup \{\#\}$*, and* $\pi : \Gamma \to \Sigma$ *is a projection. A language* $L \subseteq \Sigma^{++}$ *is recognized by such a TS if* $L = \pi(L(T))$*. We also write* $L = L(\mathcal{T})$*. The family of all* tiling recognizable *languages is denoted by* REC.

Since k-SLT picture languages include as a special case k-SLT 1D languages, the following proposition derives immediately from known properties.

Proposition 1. *The family of* k*-SLT languages for* $k \geq 2$ *is strictly included in the family of* $(k+1)$*-SLT languages, when ignoring pictures of size less than* $(k+1, k+1)$.

If we apply a projection to k-SLT languages, the hierarchy of Proposition 1 collapses (see [6, 8]). We state it to prepare the concepts needed in later developments.

Theorem 1. *Given a* k*-SLT language* $L \subseteq \Sigma^{++}$ *defined by a set of* k*-tiles* T_k *(i,e,* $L = L(T_k)$*), there exists an alphabet* Γ*, a local language* $L' \subseteq \Gamma^{++}$ *and a projection* $\pi : \Gamma \to \Sigma$ *s.t.* $L = \pi(L')$.

It easily follows that the use of larger tiles in a TS does not enlarge the language family.

Corollary 1. *The family of SLT languages is strictly included in* REC [7]*. The family of languages obtained by projections of SLT languages coincides with the family* REC *of tiling recognizable languages.*

Thus any REC language over Σ can be obtained both as projection of a local language over the alphabet Γ_2, and as a projection of a k-SLT language (with $k > 2$) over an alphabet Γ_k. However, if we use 2-tiles instead of k-tiles, we need an alphabet Γ_2 which can be larger than Γ_k. The trade-off is represented by the ratio $\frac{|\Gamma_2|}{|\Gamma_k|}$. As shown in [6, 8], this ratio is proportional to the area k^2 of the k-tiles: $\frac{|\Gamma_2|}{|\Gamma_k|} = \Theta(k^2)$.
Next, we state the unsurprising fact that the family REC constitutes an infinite hierarchy with respect to the size of the local alphabet.

Proposition 2. *For every* $\ell \geq 1$*, let* REC_ℓ *be the family of languages recognized by tiling systems with a local alphabet of cardinality at most* ℓ*. Then,* $\mathrm{REC}_\ell \subsetneq \mathrm{REC}_{\ell+1}$.

Example 1. The language $R \subseteq \{a\}^{++}$ s.t. for any $p \in R$, $|p|_{col} = 2 \cdot |p|_{row} - 1$, is defined by the TS consisting of the 2-tiles $T_2 \subseteq {\Gamma_3}^{2,2}$ with $\Gamma_3 = \{b, \searrow, \nearrow\}$ which are in the pre-image–see column 1 in the figure below, where the obvious projection is: for all $c \in \Gamma_3$, $\pi(c) = a$.

Pre-image with 3 symbols

$\searrow$	b	b	b	b	b	$\nearrow$
b	$\searrow$	b	b	b	$\nearrow$	b
b	b	$\searrow$	b	$\nearrow$	b	b
b	b	b	$\searrow$	b	b	b

Pre-image with 2 symbols

$\rightarrow$	b	b	b	b	b	$\rightarrow$
b	$\rightarrow$	b	b	b	$\rightarrow$	b
b	b	$\rightarrow$	b	$\rightarrow$	b	b
b	b	b	$\rightarrow$	b	b	b

Pre-image of illegal picture

$\rightarrow$	b	b	b	b	b	b	b	$\rightarrow$
b	b	b	b	b	b	b	b	b
b	b	b	b	$\rightarrow$	b	b	b	b
b	b	b	b	b	b	b	b	b

Merging the letters $\searrow$ and $\nearrow$ by reducing the local alphabet to $\Gamma_2 = \{b, \rightarrow\}$, the corresponding pre-image is shown in column 2; let T'_2 be its tiles. Now, illegal pictures, e.g., the one having column 3 as pre-image, can still be tiled using T'_2, hence $\pi(L(T'_2)) \supset R$. Yet, alphabet Γ_2 suffices to eliminate the illegal picture in column 3 if, instead of a 2-TS, we use a 3-TS having the 3-tiles occurring in column 2.

Comma-free codes in 2D. Our proofs are based on 2D comma-free (cf) codes composed of isometric square pictures (called code-pictures), defined analogously to 1D cf codes. A cf code is s.t. if a picture is paved with code-pictures then it is impossible to overlay a code-picture in a position where it overlaps other code-pictures. We introduce the notion of 2D code by means of a picture morphism, then we define the c.f. codes, we state a known result on the cardinality of non-overlapping codes, and we finish with the statement that the set of pictures paved with a cf code is an SLT language.

Definition 4 (code picture). *Given alphabets Γ, Λ and a one-to-one morphism φ : $\Gamma^{++} \to \Lambda^{++}$, the set $X = \varphi(\Gamma) \subseteq \Lambda^{++}$ is called a (uniform)* picture code *and its elements are called* code-pictures. *The morphism "φ" is denoted as $[\![-]\!]_X : \Gamma^{++} \to \Lambda^{++}$. For all $\gamma \in \Gamma^{++}$, the unique picture $[\![\gamma]\!]_X$ in X^{++} is called the* encoding *of γ.*

Let p be a picture of size (r, c); an *internal factor* of p is a subpicture $p_{(i,j;\,n,m)}$, s.t. $1 < i \leq j < r$ and $1 < n \leq m < c$. Given a set $X \subseteq \Lambda^{k,k}$, consider $X^{2,2}$, i.e., the set of all pictures p of size $(2k, 2k)$ of the form $(p \oplus q) \ominus (r \oplus s)$, $p, q, r, s \in X$.

Definition 5 (comma-free code). *Let Λ be an alphabet and let $k \geq 2$. A (finite) set $X \subseteq \Lambda^{k,k}$ is a* comma-free *picture* code *("cf code" for short) if, for all pictures $p \in X^{2,2}$, there is no internal factor $q \in \Lambda^{k,k}$ of p s.t. $q \in X$.*

Although the exact cardinality of a cf code $X \subseteq \Lambda^{++}$ is unknown, the following result from [1] states a lower bound for a family of binary codes that are non-overlapping, hence they are also cf codes. This bound is essential for the proof of the main result.

Theorem 2. *For all $k \geq 4$ there exist cf codes $X \subseteq \{0, 1\}^{k,k}$ of cardinality $|X| \geq \left(2^{k-2} - 1\right)^{k-2} \cdot 2^{k-3}$.*

Next, we consider a local language defined by a set of 2-tiles, and we encode each letter using a cf code, obtaining a language with the SLT property. As it is the case for Theorem 2, this is a fundamental step for the proof of the main result.

Lemma 1. *Let $T \subseteq \Gamma^{2,2}$ be a set of 2-tiles defining a local language $L(T)$ and let $X \subseteq \Lambda^{k,k}$ be a cf code s.t. $|X| = |\Gamma|$. The language $[\![L(T)]\!]_X$ is $2k$-SLT.*

3 Extended Medvedev's theorem for picture languages

Before presenting the main result, we notice that for some language in REC an alphabetic ratio < 2 does not suffice.

Theorem 3. *There exists a tiling recognizable language R over an alphabet Σ s.t. for every alphabet Γ and SLT language $L \subseteq \Gamma^{++}$, if R is the image of L under a projection, then the alphabetic ratio is $\frac{|\Gamma|}{|\Sigma|} \geq 2$.*

The proof is based on language R defined as follows. For a letter a, let R_a be the language composed of all square pictures over $\{a\}$, of size at least $(2, 2)$. R_a can only be recognized by TSs having a local alphabet Γ of cardinality at least 2, otherwise, if $|\Gamma| = 1$, a non-square (rectangular) picture and a square picture can be covered by the same set of 2-tiles. Let $\Sigma = \{b, c\}$; the language R is $R_b \cup R_c$.
The next theorem states the alphabetic ratio two is sufficient for all REC languages.

Theorem 4. *For any $R \subseteq \Sigma^{++}$ in* REC*, there exist a SLT language L over an alphabet Λ with $|\Lambda| = 2|\Sigma|$, and a projection $\rho : \Lambda \to \Sigma$ s.t. $R = \rho(L)$.*

The statement is first proved, using Th. 2 and Lm. 1, for a padding language $R^{(k)} \subseteq (\Sigma \cup \{\$\})^{++}$, where $\$ \notin \Sigma$, $k \geq 2$, instead of R, and it is then extended to R by a suitable deletion of padding symbols from the tile set. $R^{(k)}$ is obtained by padding the east and the south sides of each picture $p \in R$ with two strips of letters \$ of thickness from 0 to $k-1$, in such a way that the minimal picture of size (m, n) is obtained, where m and n are multiple of k.

Conclusions and future work. Our main result (Th. 4) can be placed next to the similar ones pertaining to regular 1D languages ([4] discussed in the Introduction) and to tree languages [5]; the latter says that every regular tree language is the letter-to-letter homomorphic image of an SLT tree language, with alphabetic ratio 2. This gives evidence that, for a quite significant sample of formal language families, an Extended Medvedev's theorem, holds: the alphabetic ratio 2 is sufficient and necessary to characterize a language as a morphic image a SLT language. What is common among the three cases considered, is the prerequisite that a (non-extended) Medvedev's theorem exists, which is based on a notion of locality, respectively, for words, for rectangular arrays, and for tree graphs. In the future, it would be interesting to see if any family endowed with the basic Medvedev's theorem necessarily possesses the extended form of the theorem with alphabetic ratio two.

References

1. M. Anselmo, D. Giammarresi, and M. Madonia. Non-expandable non-overlapping sets of pictures. *Theor. Comput. Sci.*, 657:127–136, 2017.
2. J. Berstel, D. Perrin, and C. Reutenauer. *Codes and Automata*, volume 129 of *Encyclopedia of Mathematics and its Applications*. CUP, 2009.
3. S. Crespi-Reghizzi, A. Restivo, and P. San Pietro. Reducing local alphabet size in recognizable picture languages. In N. Moreira and R. Reis, editors, *Developments in Language Theory, DLT 2021*, LNCS, pages 1–12, 2021. to appear.
4. S. Crespi-Reghizzi and P. San Pietro. From regular to strictly locally testable languages. *Int. J. Found. Comput. Sci.*, 23(8):1711–1728, 2012.
5. S. Crespi-Reghizzi and P. San Pietro. Homomorphic characterization of tree languages based on comma-free encoding. In *LATA 2021*, LNCS 12638, pages 241–254. Springer, 2021.
6. D. Giammarresi and A. Restivo. Recognizable picture languages. *Int. Journ. Pattern Recognition and Artificial Intelligence*, 6(2-3):241–256, 1992.
7. D. Giammarresi and A. Restivo. Two-dimensional languages. In G. Rozenberg and A. Salomaa, editors, *Handbook of formal languages, vol. 3*, pages 215–267. Springer, 1997.
8. D. Giammarresi, A. Restivo, S. Seibert, and W. Thomas. Monadic second-order logic over rectangular pictures and recognizability by tiling systems. *Inf. Comput.*, 125(1):32–45, 1996.
9. R. McNaughton and S. Papert. *Counter-free Automata*. MIT Press, Cambridge, USA, 1971.
10. Y. T. Medvedev. On the class of events representable in a finite automaton. In E. F. Moore, editor, *Sequential machines – Selected papers*, pages 215–227. Addison-Wesley, 1964.
11. S. Eilenberg. *Automata, Languages, and Machines*, volume A. Academic Press, 1974.