Journal of Universal Computer Science, vol. 1, no. 12 (1995), 790-810
submitted: 29/11/95, accepted: 7/12/95, appeared: 28/12/95©Springer Pub. Co.

Parikh prime words and GO-like territories[1]

A. Mateescu
(Academy of Finland and University of Turku
Department of Mathematics, FI-20500 Turku, Finland)

Gh. Păun
(Institute of Mathematics of the Romanian Academy
PO Box 1 – 764, RO-70700 Bucharest, Romania)

G. Rozenberg
(Leiden University, Department of Computer Science
P.O. Box 9512, NL-2300 RA Leiden, the Netherlands)

A. Salomaa
(Academy of Finland and University of Turku
Department of Mathematics, FI-20500 Turku, Finland)

Abstract: An n-dimensional vector of natural numbers is said to be prime if the greatest common divisor of its components is one. A word is said to be Parikh prime if its Parikh vector is prime. The languages of Parikh prime and of Parikh non-prime words are investigated (they are neither semilinear nor slender, hence are not context-free or D0L languages; both of them can be generated by matrix grammars with appearance checking).
Marking in the plane the points identified by prime (2-dimensional) vectors, interesting patterns of non-marked ("free") points appear (they are similar to the territories in the game of GO). The shape of such possible territories is investigated (with an exhaustive analysis of tro-, tetro-, pento- and hexominoes). Some open problems are formulated (both concerning the mentioned languages and the "GO territories theory").

Key Words: formal languages, context free languages, L-systems, Parikh mapping, word problems

Category: F.4.2, F.4.3, G.2.1

1 Introduction

The investigation here starts from the already (in)famous open problem asking whether or not the set Q of all primitive words is context-free or not. (A word is primitive if it cannot be written as a power of a different word.) Many things are known about this language and about its complement (see [6], [9] and their references), but not the answer to the mentioned question. The conjecture is that Q is not context-free. The topic is part of a more general area of interest in language theory and combinatorics on words [10]: investigate languages consisting of words containing or not containing given patterns. The history of the problem goes back to Axel Thue [15], who considered words without adjacent repeats. From pumping lemmas, a language consisting of such words cannot be context-free. For a while the problem was open whether or not the language of

[1] Research supported by the Academy of Finland, project 11281, and the ESPRIT Basic Research Working Group ASMICS II.

repetitive words (words containing adjacent repeats, that is of the form $x_1x_2x_2x_3$, with non-empty x_2) is context-free. It was conjectured in [1] that the answer is negative and, indeed, this has been confirmed in [12].

In view of the difficulty of the context-freeness problem for the language Q, it is natural to look for variants of it. If w is a non-primitive word, $w = z^k, k \geq 2$, then its Parikh vector, $\Psi(w)$, will be of the form $\Psi(w) = k \cdot \Psi(z)$. Consequently, also $\Psi(w)$ is non-primitive, in the sense that all its components are multiples of an integer greater than or equal to 2. The converse is not true: $\Psi(aabb) = (2, 2)$, but $aabb$ is primitive. Anyway, this suggests to consider the language of all words having "primitive" Parikh vectors, in the sense that the greatest common divisor of their elements is one. We call them *prime vectors* and *Parikh prime words*, respectively.

The place of the languages of Parikh prime and of Parikh non-prime words in the Chomsky hierarchy and in the L hierarchy is investigated in the following section. Then, the patterns of points identified by non-prime 2-dimensional vectors are examined. Being surrounded by points identified by prime vectors, these patterns are reminiscant of the notion of a *territory* in the celebrated game of GO, definitely one of the most interesting games ever invented. (GO is probably the oldest significant logical game, and yet considered the ultimate challenge for artificial intelligence: although spectacular progresses were recently made concerning the mathematical theory of parts of a game – of the end games in [2] – the GO programs are still far from being competitive, a quite different situation compared to that in chess. We do not present here the rules of GO, every *real* scientist knows the game...) A series of surprising results are obtained about such territories (if a pattern appears once, then it appears infinitely many times, there are arbitrarily large territories, etc.), and a complete study of territories with less than seven free points is done. However, a series of challenging problems remains unsettled.

2 Parikh primality versus Chomsky and L hierarchy

As usual, for an alphabet V we denote by V^* the set of all words over V; the empty string is denoted by λ and $V^* - \{\lambda\}$ by V^+. The length of $x \in V^*$ is $|x|$ and $|x|_a$ is the number of occurrences of the symbol $a \in V$ in the string $x \in V^*$. The left derivative (of a string x with respect to a symbol a) is denoted by $\partial_a(x)$ and it is defined by $\partial_a(x) = x'$ iff $x = ax'$.

If $V = \{a_1, \ldots, a_n\}$, then the Parikh mapping $\Psi_V : V^* \longrightarrow \mathbf{N}^n$ is defined by $\Psi_V(w) = (|w|_{a_1}, \ldots, |w|_{a_n}), w \in V^*$, and it is extended to languages $L \subseteq V^*$ by $\Psi_V(L) = \{\Psi_V(w) \mid w \in L\}$.

For (other) basic elements of formal language theory (including L systems), we refer to [13], [14]. We denote by CF, CS the families of context-free and of context-sensitive languages, respectively; by $0L, D0L, E0L, ET0L$ we denote the families of languages generated by 0L systems, D0L, E0L, and ET0L systems, respectively.

We introduce now some new notions.

A vector $(v_1, \ldots, v_n) \in \mathbf{N}^n$ is said to be *prime* if $gcd(v_1, \ldots, v_n) = 1$.

A word $w \in V^*, V = \{a_1, \ldots, a_n\}$, is called *Parikh prime* if $\Psi_V(w)$ is prime. For an alphabet V, we denote by P_V the set of all Parikh prime words over V and by $\bar{P}_V$ its complement, $\bar{P}_V = V^* - P_V$.

Remark. By definition, $\lambda \in \bar{P}_V$. Considering that $gcd(n) = n$ for all $n \geq 1$, we get $P_{\{a\}} = \{a\}$ and $\bar{P}_{\{a\}} = \{\lambda\} \cup \{a^i \mid i \geq 2\}$, hence this case is trivial. From now on, we will assume that the alphabet we work with contains at least two letters.

It is a natural question to find the place of languages $P_V, \bar{P}_V$ in the Chomsky (and in the L) hierarchy. We shall do that using the notion of *semilinearity* (and of *slenderness*).

A set $M \subseteq \mathbf{N}^n$ is called *linear* if there are $v_0, v_1, \ldots, v_p \in \mathbf{N}^n$ such that $M = \{v_0 + \sum_{i=1}^{p} v_i r_i \mid r_i \in \mathbf{N}\}$. A set $M \subseteq \mathbf{N}^n$ is *semilinear* if it is the union of a finite number of linear sets. A language $L \subseteq V^*$ is *semilinear* if $\Psi_V(L)$ is semilinear.

The well-known *Parikh theorem* says that every context-free language is semilinear. However, we have

Theorem 1. *For all V with $card(V) \geq 2$, the language P_V is not semilinear.*

Proof. Take the semilinear set $M = \{(v_1, v_2, 0, 0, \ldots, 0) \mid v_1, v_2 \in \mathbf{N}\}$. For each V with $\mathrm{card}(V) \geq 2$, if $\Psi_V(P_V)$ is semilinear, then $\Psi_V(P_V) \cap M$ is semilinear (the class of semilinear sets of vectors is closed under intersection [7]). Therefore it is enough to prove that $\Psi_V(P_V) \cap M$ is not semilinear, hence it is enough to consider the case of two-letter alphabets. We denote simply by P the language P_V for $V = \{a, b\}$ (we also write a, b instead of a_1, a_2).

Assume that $\Psi_V(P)$ is semilinear, hence $\Psi_V(P) = \cup_{j=1}^{l} T_j$, for $T_j, 1 \leq j \leq l$, linear subsets of $\mathbf{N}^2$. Assume

$$T_j = \{v_{0j} + \sum_{i=1}^{k_j} v_{ij} r_i \mid r_i \in \mathbf{N}\} \qquad (*)$$

for given vectors $v_{0j}, v_{1j}, \ldots, v_{k_j j} \in \mathbf{N}^2, 1 \leq j \leq l$.

All strings of the form $a^p b^n$ for p a prime number and $n \in \mathbf{N} - p \cdot \mathbf{N}$ are in P, hence (p, n) are in $\Psi_V(P)$. For each such (p, n) there is $j, 1 \leq j \leq l$, such that $(p, n) \in T_j$. For each prime p, the number n as above can be arbitrarily large, hence $(p, n) \in T_j$ implies that there is v_{ij} with $i > 0$ and $r_j > 0$ in the expression $(*)$ for (p, n). More precisely, for a fixed prime p and arbitrarily large n, we need a vector $v_{i_0 j} = (0, m)$ with $m > 0$ used in the writing $(p, n) = v_{0,j} + \sum_{i=1}^{k_j} v_{ij} r_i$.

This implies that all vectors of the form

$$(p, n) + r(0, m) = (p, n + rm) \qquad (**)$$

for $r \in \mathbf{N}$ are in $\Psi_V(P)$.

The set of prime numbers is infinite, the set of vectors $(0, m)$ used in writing $\Psi_V(P)$ as above is finite. Consequently, there are such p and m with $p > m$. For such p and n we have $gcd(p, m) = 1$, hence there are $s, t \in \mathbf{Z}$ such that $sp + tm = 1$. Without loss of generality we may assume that $s > 0, t < 0$ (otherwise we replace s with $s - mq$ and t with $t + pq$, with negative q, large enough in absolute value to have $s - mq > 0$ and $t + pq < 0$). Write

$$1 + |t|m = sp$$

and multiply by n:

$$n + n|t|m = nsp.$$

For $r = n|t|$ in $(**)$ we obtain

$$(p, n) + n|t|(0, m) = (p, n + n|t|m) = (p, nsp),$$

which must be in $\Psi_V(P)$, a contradiction with $gcd(p, nsp) = p$. In conclusion, $\Psi_V(P)$ cannot be semilinear. □

Also $\bar{P}_V$ is non-semilinear, and the proof for this language is much easier.

Theorem 2. *For all V with $card(V) \geq 2$, the language $\bar{P}_V$ is not semilinear.*

Proof. As above, it is enough to consider the case of V consisting of two letters. We denote by $\bar{P}$ the language $\bar{P}_V$ for $V = \{a, b\}$.

Assume that $\Psi_V(\bar{P})$ is semilinear and consider it as the union of finitely many linear sets T_j as in the previous proof. All vectors (p, pn), with prime p and $n \geq 1$, are in $\Psi_V(\bar{P})$. In order to write such vectors (p, pn) with given p and arbitrarily large n we need a vector $(0, m), m > 0$, in the writing of some set T_j. All vectors

$$(p, pn) + r(0, m) = (p, pn + rm),$$

$r \geq 0$, are in the corresponding set T_j. Because $gcd(p, pn + rm)$ must be greater than 1, we must have $gcd(p, pn + rm) = p$, which implies that m must be a multiple of p (n and r can be arbitrary). However, there are only finitely many vectors $(0, m)$ in the writing of sets T_j, they cannot contain as divisors all prime numbers, a contradiction. □

Corollary. *For all V with $card(V) \geq 2$, the languages P_V and $\bar{P}_V$ are not context-free, simple matrix, matrix of finite index, unordered vector languages.*

For definitions of the mentioned families, other than that of context-free languages, as well as for the proof that they contain only semilinear languages, the reader is referred to [4] and to its bibliography.

We recall now from [4] the definition of a matrix grammar.

A *matrix grammar* (with appearance checking) is a system

$$G = (N, T, S, M, F),$$

where N, T are disjoint alphabets (of nonterminal and terminal symbols, respectively), $S \in N$, M is a finite set of sequences $(A_1 \to u_1, \ldots, A_n \to u_n), n \geq 1$, of context-free rules over $N \cup T$ (called *matrices*), and F consists of some rules appearing in M.

For $x, y \in (N \cup T)^*$ we write $x \Longrightarrow y$ if there is $(A_1 \to u_1, \ldots, A_n \to u_n) \in M$ and $w_1, w_2, \ldots, w_{n+1} \in (N \cup T)^*$ such that $x = w_1, y = w_{n+1}$ and for each $i = 1, 2, \ldots, n$ either $w_i = w_i' A_i w_i'', w_{i+1} = w_i' u_i w_i''$ or $|w_i|_{A_i} = 0, w_i = w_{i+1}$ and $A_i \to u_i \in F$. (The rules of a matrix are used consecutively, in the order indicated, possibly skipping rules appearing in F, providing they cannot be applied to the current string.) Then $L(G) = \{x \in T^* \mid S \Longrightarrow^* x\}$.

We denote by MAT_{ac} the family of languages generated by matrix grammars as above, with λ-free rules. If $F = \emptyset$ (hence all rules must be effectively used), then we say that the grammar is without appearance checking. The corresponding family of languages is denoted by MAT (again λ-rules are not allowed).

It is known that $CF \subset MAT \subset MAT_{ac} \subset CS$, all inclusions being proper, and that MAT contains non-semilinear languages.

For saving space, in the following theorems we consider only the languages P and $\bar{P}$ (hence over $V = \{a, b\}$), but similar constructions can be obtained also for general alphabets.

Theorem 3. $P \in MAT_{ac}$.

Proof. We construct the matrix grammar

$$G = (N, \{a, b, c\}, S_0, M, F),$$

where

$$N = \{S_0, S, A, A', A'', \bar{A}, B, \bar{B}, X, Y, Z, Z', Z'', Z''', U, U', U'', V, \#\},$$

F contains all rules with the right-hand member equal with $\#$, and M contains the following matrices:

$$\begin{aligned} 1.\ & (S \to XS), \\ & (X \to X, S \to AS), \\ & (X \to X, S \to BS), \\ & (X \to Y, S \to A), \\ & (X \to Y, S \to B). \end{aligned}$$

(One produces a string $Yw, w \in \{A, B\}^+$.)

$$2.\ (Y \to Z, A \to A', A \to A').$$

(One introduces two occurrences of A'; we shall check whether or not 2 is a common divisor for $|w|_A$ and $|w|_B$. In general, the current number of A' occurrences will be checked as a possible common divisor of $|w|_A$ and $|w|_B$.)

$$\begin{aligned} 3.\ & (Z \to Z, A' \to A'', A \to \bar{A}), \\ & (Z \to Z', A' \to \#), \\ & (Z' \to Z', A'' \to A'), \\ & (Z' \to Z, A'' \to \#). \end{aligned}$$

(In the presence of Z one marks by a bar as many occurrences of A as there are occurrences of A' in the current string; the operation is iterated.)

$$4.\ (Z \to Z'', A' \to \#, A \to \#).$$

(One finishes at the same time both the A' occurrences and the A occurrences, hence $|w|_A$ is divisible by the number of A' occurrences.)

$$5.\ (Z \to Z''', A' \to A', A \to \#).$$

(The number $|w|_A$ is not a multiple of the number of A' occurrences.)

$$\begin{aligned} 6.\ & (Z''' \to Z''', A'' \to A'), \\ & (Z''' \to Z''', \bar{A} \to A), \\ & (Z''' \to Z, A'' \to \#, \bar{A} \to \#, A \to A'). \end{aligned}$$

(One returns to a string containing only symbols A' and A, with the number of A' occurrences increased by one. Matrices in groups $3-5$ are now applied for the divisibility by the new number of A' symbols.)

$$\begin{aligned} 7.\ & (Z'' \to Z'', A'' \to A'), \\ & (Z'' \to U, A'' \to \#), \\ & (U \to U, A' \to A'', B \to \bar{B}), \\ & (U \to U', A' \to \#), \\ & (U' \to U', A'' \to A'), \\ & (U' \to U, A'' \to \#). \end{aligned}$$

(Having concluded in 4 that the number of A occurrences is a multiple of the number of A' occurrences, one now checks whether the number of B occurrences is a multiple of the number of A' occurrences.)

$$8.\ (U \to U'', A' \to A', B \to \#).$$

(The number of B occurrences is not divisible by the number of A' occurrences. Only in this case we can continue, otherwise # is introduced.)

$$\begin{aligned} 9.\ & (U'' \to U'', \bar{B} \to B), \\ & (U'' \to Z''', \bar{B} \to \#). \end{aligned}$$

(One returns to group 6, for continuing the process with an increased number of occurrences of A'.)

$$\begin{aligned} 10.\ & (Z''' \to V, A'' \to \#, \bar{A} \to \#, A \to \#), \\ & (V \to V, A' \to a), \\ & (V \to V, B \to b), \\ & (V \to c). \end{aligned}$$

(When all occurrences of A were replaced by A' and still $|w|_B$ is not divisible by this number, the string is "accepted", the nonterminals are replaced by terminals.)

¿From the explanations above we have

$$P = \partial_c(L(G)) \cup b^*ab^*$$

(the grammar G produces all strings cw with $w \in P, |w|_a \geq 2$). As MAT_{ac} is closed under left derivative and union, we have $P \in MAT_{ac}$. □

Theorem 4. $\bar{P} \in MAT_{ac}$.

Proof. We construct the matrix grammar

$$G = (N, \{a, b, c\}, S_0, M, F),$$

with

$$N = \{S_0, S, A, B, A', A'', X, Y, Z, Z', U, U', V, \#\},$$

F containing all rules introducing the symbol $\#$, and M consisting of the following matrices:

$$\begin{aligned} 1.\ & (S \to XS), \\ & (X \to X, S \to AS), \\ & (X \to X, S \to BS), \\ & (X \to Y, S \to A), \\ & (X \to Y, S \to B). \end{aligned}$$

(As above, one produces a string $Yw, w \in \{A, B\}^+$.)

$$\begin{aligned} 2.\ & (Y \to Y, A \to A'), \\ & (Y \to Z, A \to A', A \to A'). \end{aligned}$$

(At least two occurrences of A are replaced by A'. The derivation will end correctly if and only if $|w|_A$ and $|w|_B$ are both divisible by the number of A' occurrences.)

$$\begin{aligned} 3.\ & (Z \to Z, A' \to A'', A \to a), \\ & (Z \to Z', A' \to \#), \\ & (Z' \to Z', A'' \to A'), \\ & (Z' \to Z, A'' \to \#). \end{aligned}$$

(The number of A occurrences is checked for divisibility with the number of A'.)

$$4.\ (Z' \to U, A'' \to \#).$$

(After introducing U, the symbols A cannot be rewritten, hence their number must have been multiple of the number of A' occurrences.)

$$\begin{aligned} 5.\ & (U \to U, A' \to A'', B \to b), \\ & (U \to U', A' \to \#), \\ & (U' \to U', A'' \to A'), \\ & (U' \to U, A'' \to \#). \end{aligned}$$

(The symbols B are terminated, in blocks of size equal to the total number of occurrences of A'.)

$$6.\ (U' \to V, A'' \to \#).$$

(No further B will be rewritten, hence their number must have been a multiple of the number of A' occurrences.)

$$\begin{aligned} 7.\ & (V \to V, A' \to a), \\ & (V \to c). \end{aligned}$$

We have

$$\bar{P} = \partial_c(L(G)) \cup a^* \cup b^*,$$

hence $\bar{P} \in MAT_{ac}$. □

Consider now the place of languages $P, \bar{P}$ in the L hierarchy.

A language $L \subseteq V^*$ is called *slender* [11] iff there is a constant k such that $\mathrm{card}(L \cap V^n) \leq k$ for all $n \in \mathbf{N}$ (the number of words in L of any given length is bounded).

In [5] it is proved that each D0L language is slender. Clearly, P and $\bar{P}$ are not slender: for any prime numbers p, q, all permutations of $a^p b^q$ are in P and all permutations of $a^p b^{2p}$ are in $\bar{P}$. Consequently, P and $\bar{P}$ are not D0L languages. Moreover, we have

Theorem 5. $P \notin 0L$.

Proof. Assume that $P = L(G)$ for some 0L system $G = (\{a,b\}, w, \sigma)$ (σ is a finite substitution, but we consider it as a set of rules of the form $c \to u$ for $u \in \sigma(c), c \in \{a,b\}$).

1. G must be propagating.

Assume that $a \to \lambda \in \sigma$. There is at least one rule $b \to x$. If $|x|_a \geq 1, |x|_b \geq 1$, then $aaabb \Longrightarrow xx$, which is not in P. If we have both $b \to a^i$ and $b \to b^j$ in σ, $i, j \geq 1$, then $abbbb \Longrightarrow b^{2i}b^{2j}$, again not in P. If we have $b \to a^i$ and $a \to b^j, i, j \geq 1$, or $b \to b^i$ and $a \to a^j, i, j \geq 1$, then $aaabb \Longrightarrow a^{2i}b^{2j}$ or $aaabb \Longrightarrow a^{2j}b^{2i}$, which are not in P. If the only b-rule is $b \to \lambda$, then we must have $a \to x$ in σ, with $|x|_a \geq 1, |x|_b \geq 1$ and then $aab \Longrightarrow xx$, not in P. In all cases (because we need rules introducing both a and b) we reach a contradiction.

Consequently, G is propagating, hence its axiom must be one of the shortest strings in P. These strings are

$$a, b, ab, ba, aab, aba, baa, abb, bab, bba.$$

Assume that the axiom is a. The case of $w = b$ is symmetric.

Then the rule $a \to b$ must be in σ (in order to obtain $b \in P$).

2. No rule $c \to d^i, c \in \{a,b\}, d \in \{a,b\}, i \geq 2$, is possible: take $e \neq c$ such that $\{e, c\} = \{a, b\}$ and a rule $e \to x$. Then $e^i c \Longrightarrow x^i d^i$, which is not in P.

3. Assume that we have the rule $a \to ab$ in σ. Then:
   - if $a \to a \in \sigma$, then $aaabb \Longrightarrow (ab)(a)(b)xx \notin P$,
   - if $b \to a \in \sigma$, then $aabbb \Longrightarrow (ab)(b)(a)(a)(a) \notin P$,
   - if $b \to b \in \sigma$, then $aaab \Longrightarrow (ab)(ab)(b)(b) \notin P$.

All cases are contradictory.

4. Examine now the case of the string aab. It cannot be produced starting from a string of length three, because, as we have seen above, we cannot use rules $a \to a, b \to a, b \to b$. Moreover, it cannot be produced from a string of length two: we need either $a \to a$ or $a \to aa$ in the case of $ab \Longrightarrow aab$, and either $b \to a$ or $b \to aa$ in the case of $ba \Longrightarrow aab$, and this is contradictory.

If $b \to aab \in \sigma$, then $aab \Longrightarrow (ab)(b)(aab) \notin P$. Consequently, we must have the rule $a \to aab$ in σ.

Consider now the case of the string abb. As above, it cannot be produced from strings of length three. If $ab \Longrightarrow abb$, then we need $b \to b$ or $b \to bb$, a contradiction. If $ba \Longrightarrow abb$, then we need $b \to ab$ and $aaab \Longrightarrow (ab)(ab)(ab)(ab) \notin P$.

They remain two cases:

$a \to abb \in \sigma$; then $a^4b^3 \Longrightarrow (a^2b)(a^2b)(ab^2)(ab^2)xxx \notin P$,

$b \to abb \in \sigma$; then $ab \Longrightarrow (a^2b)(ab^2) \notin P$.

All cases are contradictory, the assumption that $a \to ab \in \sigma$ cannot be true.

5. If $b \to ab \in \sigma$, then:

– if $a \to a \in \sigma$, then $aabbb \Longrightarrow (a)(b)(ab)(ab)(ab) \notin P$,

– if $b \to a \in \sigma$, then $aaabb \Longrightarrow (b)(b)(b)(a)(ab) \notin P$,

– if $b \to b \in \sigma$, then $abbb \Longrightarrow (b)(ab)(ab)(b) \notin P$.

All cases are contradictory.

6. Examine the possibilities to produce the string aab. We cannot obtain it from strings of length three (we cannot use rules as above) and also not from strings ab, ba (in both cases we need rules already discussed: $a \to a, a \to aa, b \to a, b \to aa$).

If $a \to aab \in \sigma$, then $aab \Longrightarrow (aab)(b)(ab) \notin P$.

If $b \to aab \in \sigma$, then $abb \Longrightarrow (b)(ab)(aab) \notin P$.

The string aab cannot be generated, a contradiction with $L(G) = P$, hence $P \notin 0L$. □

Open problems. Is $\bar{P}$ a 0L language ? Which are the relationships between $P, \bar{P}$ and $E0L, ET0L$ and MAT ?

The syntactic monoid (of course, infinite) of the languages $\bar{P}_V$ has a nice property: it is isomorphic with $\mathbf{N}^k$, for $k = \mathrm{card}(V)$.

Indeed, take $V = \{a_1, \ldots, a_k\}$. For $x, y \in V^*$ we define

$$x \approx y \ \text{ iff } \ (uxv \in \bar{P}_V \Leftrightarrow uyv \in \bar{P}_V), \text{ for all } u, v \in V^*.$$

Lemma 1. *For all $x, y \in V^*, x \approx y$ if and only if $\Psi_V(x) = \Psi_V(y)$.*

Proof. If $\Psi_V(x) = \Psi_V(y)$, then for all $u, v \in V^*$ we have $\Psi_V(uxv) = \Psi_V(uyv)$, therefore $uxv \in \bar{P}_V$ if and only if $uyv \in \bar{P}_V$.

Conversely, suppose that $x \approx y$, but $\Psi_V(x) \neq \Psi_V(y)$. Assume $\Psi_V(x) = (i_1, \ldots, i_k)$, and $\Psi_V(y) = (j_1, \ldots, j_k)$.

Without loss of generality, we may assume that there is $r, 1 \leq r \leq k$, such that $i_r < j_r$ (if necessary, we interchange x and y). Denote $D = j_r - i_r$.

Take two consecutive prime numbers p_1, p_2 such that $p_2 - p_1 > D$ and such that

$$\frac{p_1}{3} > \max\{i_h \mid 1 \leq h \leq k\}.$$

Denote $q = p_1 - i_r$. Because

$$p_1 = i_r + q < j_r + q = p_1 + (j_r - i_r) = p_1 + D < p_2,$$

it follows that $j_r + q$ is not a prime number. Therefore, $j_r + q = df$ for some numbers $d > 1, f > 1$ such that $d < \dfrac{p_1}{2}$.

Let $d_1, \ldots, d_k$ be such that $(j_h + d_h) \equiv 0 \pmod{d}$, $1 \leq h \leq k$. We can take these numbers such that $0 \leq d_h < d, 1 \leq h \leq k$.

Consider the words

$$\begin{aligned} u &= a_1^{d_1} a_2^{d_2} \ldots a_{r-1}^{d_{r-1}} a_r^q a_{r+1}^{d_{r+1}} \ldots a_k^{d_k}, \\ v &= \lambda. \end{aligned}$$

We have

$$\Psi_V(uyv) = (j_1 + d_1, \ldots, j_{r-1} + d_{r-1}, j_r + q, j_{r+1} + d_{r+1}, \ldots, j_k + d_k).$$

¿From the previous construction, d divides all components of this vector, hence $uyv \in \bar{P}_V$.

On the other hand,

$$\Psi_V(uxv) = (i_1 + d_1, \ldots, i_{r-1} + d_{r-1}, i_r + q, i_{r+1} + d_{r+1}, \ldots, i_k + d_k) = \\ = (i_1 + d_1, \ldots, i_{r-1} + d_{r-1}, p_1, i_{r+1} + d_{r+1}, \ldots, i_k + d_k).$$

For all $h, 1 \leq h \leq k, h \neq r$, we have

$$i_h + d_h < \frac{p_1}{3} + \frac{p_1}{2} < p_1.$$

Consequently, p_1 does not divide any of the components $i_h + d_h, h \neq k$, that is $gcd(i_1 + d_1, \ldots, i_{r-1} + d_{r-1}, p_1, i_{r+1} + d_{r+1}, \ldots, i_k + d_k) = 1$. This means that $uxv \notin \bar{P}_V$, a contradiction to $x \approx y$. □

Consequently, for each $x \in V^*$, the equivalence class of x, denoted by $\hat{x}$, consists of all permutations of x.

Theorem 6. *The monoid* $M = V^*/_{\approx}$ *is isomorphic with the monoid* $(\mathbf{N}^k, +, (0, \ldots, 0))$.

Proof. We define $\varphi : M \longrightarrow \mathbf{N}^k$ by $\varphi(\hat{x}) = \Psi_V(x)$ for all $x \in V^*$. From the previous lemma, if $x \approx y$, then $\Psi_V(x) = \Psi_V(y)$, hence φ is well defined.

Because for all $x, y \in V^*$ we have

$$\varphi(\hat{x}\hat{y}) = \varphi(\widehat{xy}) = \Psi_V(xy) = \Psi_V(x) + \Psi_V(y) = \varphi(\hat{x}) + \varphi(\hat{y}),$$

this is a morphism.

If $\varphi(\hat{x}) = \varphi(\hat{y})$, then $\Psi_V(x) = \Psi_V(y)$, and according to Lemma 1 we have $\hat{x} = \hat{y}$, hence φ is injective. The surjectivity is obvious: for each vector $t \in \mathbf{N}^k$ there is a word $w \in V^*$ such that $\Psi_V(w) = t$, hence $\varphi(\hat{w}) = t$. In conclusion, φ is an isomorphism of monoids. □

3 Prime vectors in the plane

Let us mark by a dot the points of coordinates (n, m) with $gcd(n, m) = 1$. We have done this in Figure 1 for n, m positive and with $n < m, 1 \leq n \leq 19, 1 \leq m \leq 23$. It is obvious that the arrangements of dots is symmetric with respect to the axes and with respect to the diagonals of the plane, hence it is enough to examine only the region of points $(n, m), n, m \in \mathbf{N}$, with $n < m$ (as in Figure 1).

If we interpret the dots as (say black) GO stones, we find in this figure a series of *territories*: regions of empty points, connected horizontally or vertically, and surrounded by occupied points.

Which shapes can these territories have ? Which is their relative distribution ? Are there territories of an arbitrarily large size (as regards the number of points) ?

These questions and other related ones can be formulated starting from Figure 1. We shall answer part of them, after introducing some more formal terminology.

Every vector $(n, m) \in \mathbf{Z}^2$ is called a *point*. Two points $(n_1, m_1), (n_2, m_2)$ are *neighbors* if $|n_1 - n_2| + |m_1 - m_2| = 1$ (two diagonally adjacent points are not considered neighbors). A point (n, m) with $gcd(n, m) = 1$ is called *marked/ occupied* (the marked points correspond to Parikh prime strings over two letters). A point (n, m) with $gcd(n, m) > 1$ is called *free* (point of territory). A sequence of points $(n_1, m_1), (n_2, m_2), \ldots, (n_k, m_k)$ is called a *path* (from (n_1, m_1) to (n_k, m_k)) if for all $j = 1, 2, \ldots, k-1$, the points $(n_j, m_j), (n_{j+1}, m_{j+1})$ are neighbors. A set $M \subseteq \mathbf{Z}^2$ is said to be *connected* if for all two points $(n_1, m_1), (n_2, m_2)$ in M there is a path from (n_1, m_1) to (n_2, m_2) using only points of M. A maximal connected set of free points is a *territory* (maximality means that no further free point can be added without losing the connectedness). Of course, every one-point set is connected.

Before examining the possible territories, let us consider their size (the number of free points).

Obviously, the points on the plane axes are all free (by convention, $gcd(0, n) = n$ for all n), hence they form together an infinite territory. This is the only infinite territory:

Theorem 7. *Outside the axes, there are arbitrarily large territories, but no one of them is unbounded.*

Proof. For a given $n \in \mathbf{N}$, take n consecutive natural numbers, $m+1, m+2, \ldots, m+n, m \geq 1$. For $q = \prod_{i=1}^{n}(m+i)$, consider all points $(m+i, q), 1 \leq i \leq n$. They are clearly in the same territory: $gcd(q, m+i) = m+i, 1 \leq i \leq n$, and $(q, m+j), (q, m+j+1)$ are adjacent for each j. We have found a territory of at least n points, thus proving the first assertion.

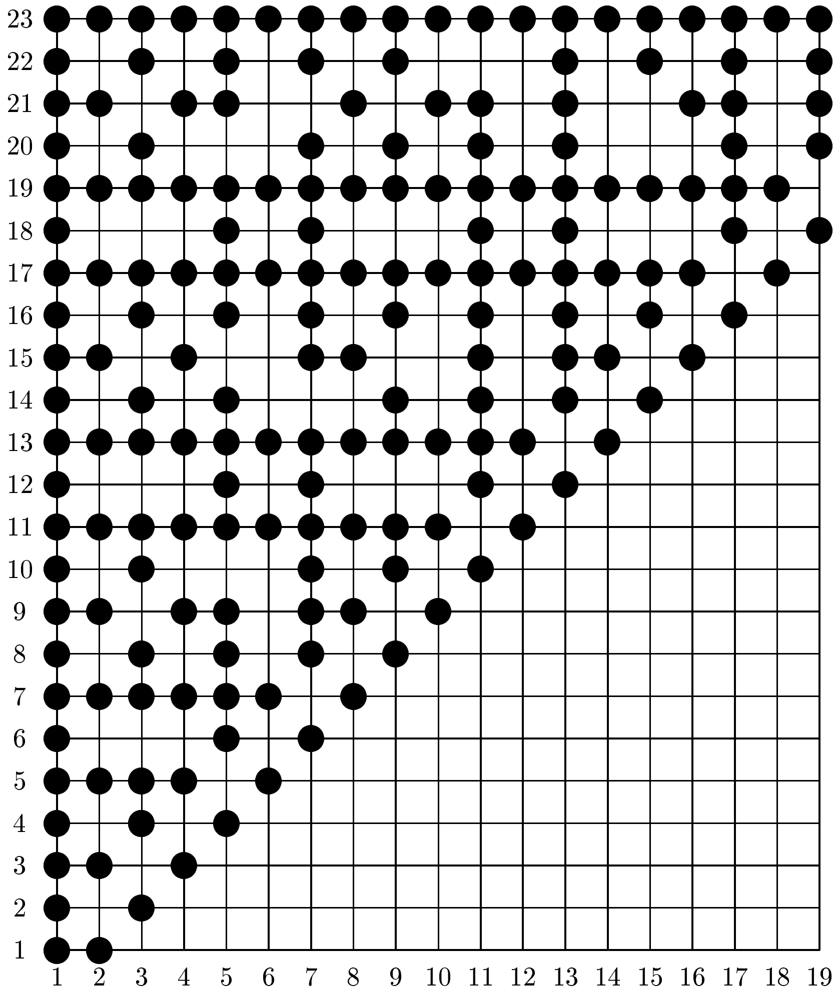

Figure 1

All points (n,m) with either $|n| = 1$ or $|m| = 1$ are marked. Moreover, all points (n,m) with $||n|-|m|| = 1$ are marked. Therefore, it is enough to prove that there is no unbounded territory in the region corresponding to Figure 1, of points (n,m) with $0 < n < m$. For every prime number p, all points $(n,p), 0 < n < p$, are marked ($gcd(n,p) = 1$). Consequently, each territory from this region is included in the horizontal stripe delimited by two consecutive primes, hence it cannot be unbounded. $\square$

In fact, a stronger result is true.

Lemma 2. *If $M \subseteq \mathbf{Z}^2$ is a finite territory and (p,q) is an arbitrary point in $\mathbf{Z}^2 - \{(0,0)\}$, then there is a translation $f : \mathbf{Z}^2 \longrightarrow \mathbf{Z}^2$ such that all points $f(n,m)$ for $(n,m) \in M$, as well as $f(p,q)$ are free.*

Proof. Assume $M = \{(n_1, m_1), \ldots, (n_k, m_k)\}$ and denote $gcd(n_i, m_i) = d_i, 1 \leq i \leq k$. For $D = \prod_{i=1}^{k} d_i$, define the translation

$$f_{t_1,t_2}(n, m) = (n + t_1 D, m + t_2 D),$$

for any given integers t_1, t_2.

For every point (n_i, m_i) of M we have $f_{t_1,t_2}(n_i, m_i) = (n_i + t_1 D, m_i + t_2 D)$ and $gcd(n_i + t_1 D, m_i + t_2 D) \geq d_i > 1$. Therefore, all points $f_{t_1,t_2}(n_i, m_i)$ are free, for all t_1, t_2.

Take now $t_1 = p, t_2 = q$. We obtain

$$f_{p,q}(p, q) = (p + pD, q + qD) = (p(D + 1), q(D + 1)).$$

Because $gcd(p(D + 1), q(D + 1)) \geq D + 1 > 1$, it follows that also $f_{p,q}(p, q)$ is a free point. $\square$

For a territory M, denote by $w(M)$ and call *the width* of M, the size (of the edge) of the largest square contained in M. (In Figure 1 we have only territories of width 1, with the only exception of the territory containing the point $(14, 20)$, whose width is 2.)

Theorem 8. *There are territories of arbitrarily large width.*

Proof. Take an arbitrary finite territory M_0 and a point (n, m), neighbor to a point in M_0 and marked. Whichever this point is, using the procedure in Lemma 2, we can find a translation $f : \mathbf{Z}^2 \longrightarrow \mathbf{Z}^2$ such that all points in $M_1' = f(M_0) \cup \{f(n, m)\}$ are free and M_1' is connected (if two points $(n_1, m_1), (n_2, m_2)$ are neighbors, then also $f(n_1, m_1), f(n_2, m_2)$ are neighbors). Denote by M_1 the territory which includes M_1' (it is possible that further marked points around M_0 are translated to free points around $f(M_0)$). We can continue this procedure arbitrarily many times, choosing the marked point (n, m) in such a way to obtain territories with larger and larger widths. $\square$

Consequently, for every territory there are arbitrarily many territories of the same shape or larger. Can we find arbitrarily many territories precisely *of the same shape* ? Surprisingly, the answer is affirmative.

We say that two territories M_1, M_2 are (strongly) *congruent* if there is a translation $f : \mathbf{Z}^2 \longrightarrow \mathbf{Z}^2$ such that $f(M_1) = M_2$. Given a territory M, we denote by $F(M)$ the *frontier* of M, that is the set of all marked points (n, m) for which there is a free point (n', m') in M such that $(n, m), (n', m')$ are neighbors.

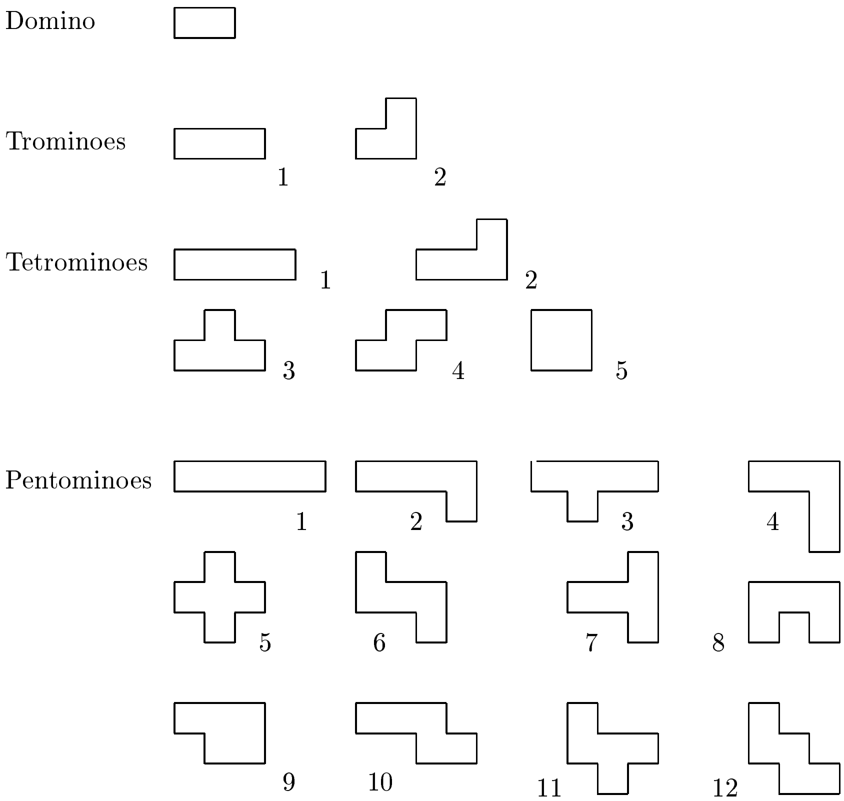

Figure 2

Theorem 9. *Given a finite territory, there are infinitely many territories congruent with it.*

Proof. Consider a territory $M = \{(n_1, m_1), \ldots, (n_k m_k)\}$, denote $d_i = gcd(n_i, m_i)$, $1 \leq i \leq k$, and $d = \prod_{i=1}^{k} d_i$. Assume that $F(M) = \{(p_1, q_1), \ldots, (p_s, q_s)\}$. Clearly, all p_i, q_i are non-zero numbers and we have $gcd(p_i, q_i) = 1$ for all $i, 1 \leq i \leq s$.

Consider the translation $f : \mathbf{Z}^2 \longrightarrow \mathbf{Z}^2$ defined by

$$f(n, m) = (n + pqd, m),$$

where

$$p = \prod_{i=1}^{s} p_i, \quad q = \prod_{i=1}^{s} q_i.$$

Because $gcd(n_i + pqd, m_i) \geq d_i > 1$, each point $f(n_i, m_i), 1 \leq i \leq k$, is a free point.

Moreover, for each $(p_i, q_i) \in F(M)$ we have

$$f(p_i, q_i) = (p_i + pqd, q_i) = (p_i(1 + \frac{p}{p_i} qd), q_i),$$

therefore $gcd(p_i + pqd, q_i) = 1$, hence $f(p_i, q_i)$ is a marked point.

Clearly, if two points $(n, m), (n', m')$ are neighbors, then also $f(n, m), f(n', m')$ are neighbors. Consequently, $M' = f(M)$ is a territory and $F(M') = f(F(M))$.

The two territories M, M' are congruent (and different). Continuing the procedure (starting now from M', then from the currently constructed territory), we can find arbitrarily many congruent territories, all congruent with M. $\square$

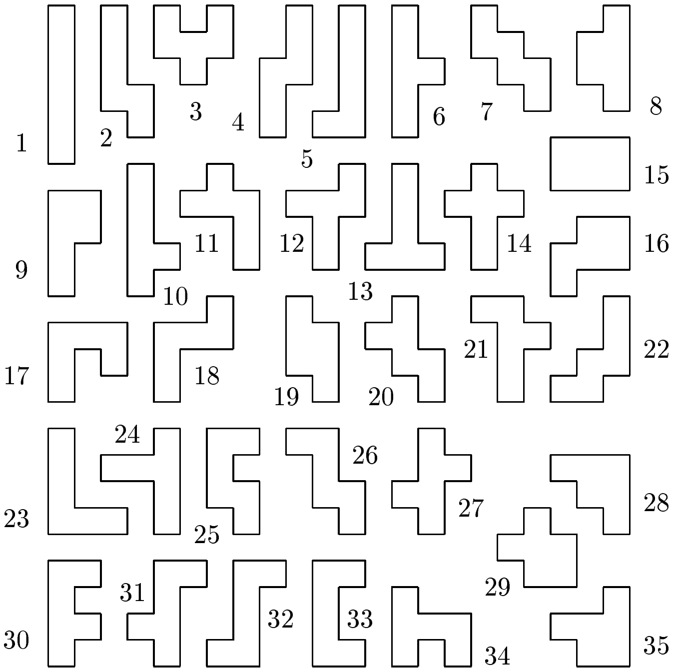

Figure 3

Observe that the previous Theorems 7 – 9 can be extended to the n-dimensional space.

Therefore, according to Theorem 9, if a pattern appears, then it appears infinitely many times. However, the question arises: which patterns actually appear ? For instance, Figure 1 contains a series of patterns, but no domino (a territory consisting of two points). As we shall see, only a few of the small polyominoes are possible. We shall complete the analysis for the domino, trominoes, tetrominoes, pentominoes (Figure 2), and hexominoes (Figure 3). We shall refer to these polyominoes with the number associated to them in these figures. (Lists of polyominoes can be found in many places; we refer here to the monumental book [3]. As usual, two polyominoes which can be obtained from one another by rotations and mirroring are considered identical.)

Before discussing particular polyominoes, we give a simple and useful general lemma.

Given a territory M, we say that a set E is a *horizontal edge* of M if

1. $E = \{(n, m), (n+1, m), \ldots, (n+k, m)\} \subseteq M$,
2. $(n-1, m)$ and $(n+k+1, m)$ are marked points,
3. either all points $(n, m-1), (n+1, m-1), \ldots, (n+k, m-1)$, or all points $(n, m+1), (n+1, m+1), \ldots, (n+k, m+1)$ are marked.

Similarly we can define a *vertical edge*.

Lemma 3. *There is no territory having an edge of even length.*

Proof. Assume that there is a territory M having a horizontal edge of length $2k, k \geq 1$; the case of vertical edges is similar. Assume that we have a situation as in Figure 4; the case when the marked neighboring points are above is similar.

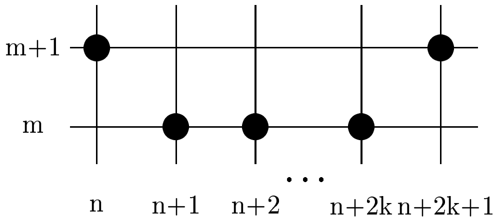

Figure 4

If m is even, because at least one of $n+1, n+2$ is even, one of the points $(n+1, m), (n+2, m)$ must belong to M, a contradiction.

If m is odd, then $m+1$ is even. One of $n, n+2k+1$ is even, hence either $(n, m+1)$ or $(n+2k+1, m+1)$ belongs to M, again a contradiction.

In conclusion, M cannot be a territory. □

Theorem 10. (i) *No straight, horizontal or vertical, line of even length can be a territory* (*hence the domino cannot appear*).

(ii) *Both trominoes appear as territories.*

(iii) *Only the tetromino 3 appears as a territory.*

Proof. All the negative assertions are proved by the previous lemma (all the mentioned patterns have an edge of even length). Both the trominoes and the tetromino 3 are present in Figure 1. □

Theorem 11. *From pentominoes, only those with numbers 1, 4 and 5 in Figure 2 appear as territories.*

Proof. The pentomino 4 appears in Figure 1. An example of territory of the form of pentomino 1 is $\{(2,30), (3,30), (4,30), (5,30), (6,30)\}$. Figure 5 indicates a place where the pentomino 5 appears. Because

$$\begin{array}{ll} 103 = \text{prime}, & 214 = 2 \cdot 107, \\ 104 = 2^3 \cdot 13, & 215 = 5 \cdot 43, \\ 105 = 3 \cdot 5 \cdot 7, & 216 = 2^3 \cdot 3^3, \\ 106 = 2 \cdot 53, & 217 = 7 \cdot 31, \\ 107 = \text{prime}, & 218 = 2 \cdot 109, \end{array}$$

we have indeed a territory of the specified shape.

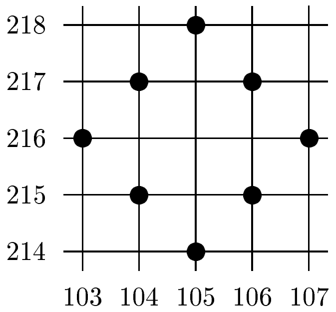

Figure 5

The pentominoes 2, 3, 6, 8, 9, 10, 11, 12 contain edges of even lengths, hence they cannot appear.

There remains the pentomino 7. It is also impossible as territory (but it cannot be refuted by Lemma 3). Consider the position in Figure 6; all other orientations can be handled in the same way.

If m is odd, then $m+1$ is even; one of $n+1$ and $n+2$ is even, hence one of the points $(n+1, m+1), (n+2, m+1)$ belongs to the territory, a contradiction.

If m is even, then also $m+2$ is even. If n is even, then $(n, m+2)$ must be in the territory, which is contradictory. If n is odd, then $n+3$ is even, hence $(n+3, m)$ must be in the territory. In conclusion, the situation in Figure 6 cannot appear. $\square$

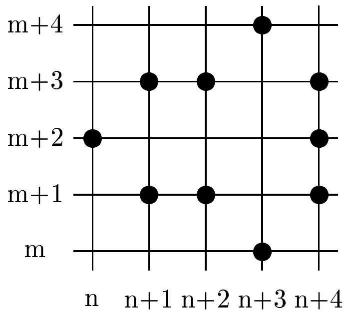

Figure 6

Theorem 12. *From hexominoes, only those with numbers 10, 11, and 28 in Figure 3 are possible.*

Proof. From Lemma 3 it immediately follows that the following hexominoes cannot appear as territories: 1, 2, 3, 5, 7, 8, 9, 12, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 29, 30, 31, 32, 33, 34, 35.

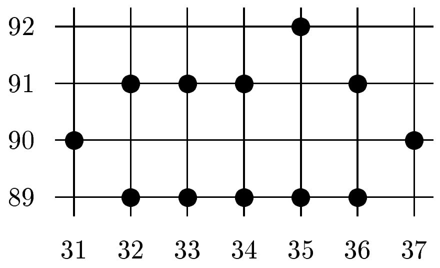

Figure 7

In Figure 1 the hexominoes 11 and 28 already appear. Also that with number 10 is possible, as indicated in Figure 7. We have

$$\begin{array}{ll} 89 = \text{prime}, & 31 = \text{prime}, \\ 90 = 2 \cdot 3^2 \cdot 5, & 32 = 2^5, \\ 91 = 7 \cdot 13, & 33 = 3 \cdot 11, \\ 92 = 2^2 \cdot 23, & 34 = 2 \cdot 17, \\ & 35 = 5 \cdot 7, \\ & 36 = 2^2 \cdot 3^2, \\ & 37 = \text{prime}, \end{array}$$

and this shows that we have, indeed, a territory of the desired form.

It remains to consider the hexominoes 4, 6, 13, 14, 19, 27. The impossibility of each of them to appear as a territory can be proved in the same way, by examining the parity of coordinates. We consider only two cases:

Figure 8 shows that the hexomino 4 cannot appear: if m is even, then one of $(n+1, m), (n+2, m)$ is free; if m is odd, then $m+1$ is even, hence one of $(n+4, m+1), (n+5, m+1)$ is free. A contradiction is obtained in each case.

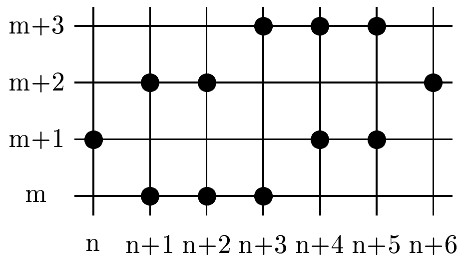

Figure 8

Figure 9 shows that the hexomino 6 is impossible, too: if m is even, then one of $(n+1, m), (n+2, m)$ must be in the territory. If m is odd, then both $m+1$ and $m+3$ are even. If n is even, then $(n, m+1)$ is free, if n is odd, then $n+3$ is even, hence $(n+3, m+3)$ is free. All cases are contradictory.

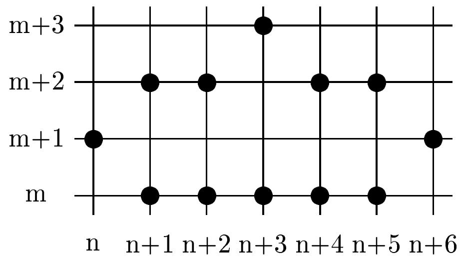

Figure 9

The reader can check in the same way the other cases. □

We can conclude that only a few polyominoes of the specified types can appear as territories. What about larger polyominoes ? What about squares, for instance ? (Of course, squares of even dimensions cannot appear.) Such problems remain to be investigated.

As a challenge for the reader, we remark that the 3×3 square appears as a territory. Indeed, consider the situation in Figure 10. We have

$$\begin{array}{ll} 103 = \text{prime}, & 6203 = \text{prime}, \\ 104 = 2^3 \cdot 13, & 6202 = 2 \cdot 7 \cdot 443, \\ 105 = 3 \cdot 5 \cdot 7, & 6201 = 3^2 \cdot 13 \cdot 53, \\ 106 = 2 \cdot 53, & 6200 = 2^3 \cdot 5^2 \cdot 31, \\ 107 = \text{prime}, & 6199 = \text{prime}. \end{array}$$

It is easy to see that the situation in Figure 10 is correct, the 3×3 square is a territory. Other 3×3 squares can be found centered around the points of coordinates (105, 150891), (105, 295581), (105, 440271), (105,584961), etc. Of course, these examples have been found using a computer. Observe the places where the squares appear; it seems that the first one is one of the closest to the origin of the plane. (It is surely the lowest one on the vertical lines 104, 105, 106.

What about larger squares ? We *conjecture* that all squares $(2k+1) \times (2k+1)$ are possible.

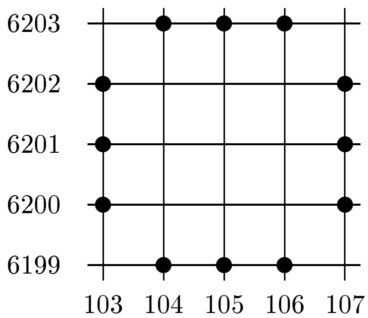

Figure 10

A series of number-theoretic (the geometry of numbers) questions can be formulated about the above free/marked points, starting with a problem similar to the famous Gauss' one (see problem F1 in [8], as well as the references of [8]): how many free/marked points there exist inside the circle with centre in the origin and radius r ? (In the case of Gauss' problem, all lattice points are counted.) What about the presumably easier problem concerning the number of free/marked points in a square centered in the origin and with a given size ? Which is the ratio of the number of free points over the number of marked points in such a square ? Is this ratio convergent ? If yes, which is the limit ?

4 References

1. J. M. Autebert, J. Beauquier, L. Boasson, M. Nivat, Quelques problèmes ouverts en théorie des langages algébriques, *RAIRO, Th. Informatics*, 13, 4 (1979), 363 – 378.
2. E. R. Berlekamp, D. Wolfe, *Mathematical GO: Chilling Gets the Last Point*, Ishi Press, 1994.
3. E. R. Berlekamp, J. H. Conway, R. K. Guy, *Winning Ways for Your Mathematical Plays*, Academic Press, New York, 1982.
4. J. Dassow, Gh. Păun, *Regulated Rewriting in Formal Language Theory*, Springer-Verlag, Berlin, Heidelberg, 1989.
5. J. Dassow, Gh. Păun, A. Salomaa, On thiness and slenderness of L languages, *Bulletin of EATCS*, 49 (1993), 152 – 158.
6. P. Dömösi, S. Horváth, M. Ito, L. Kászonyi, M. Katsura, Formal languages consisting of primitive words, *FCT Conf. 1993*, Szeged (Z. Esik, ed.), *LNCS* 710, Springer-Verlag, 1993, 194 – 203.
7. S. Ginsburg, *The Mathematical Theory of Context-Free Languages*, McGraw-Hill Book Comp., New York, 1966.
8. R. K. Guy, *Unsolved Problems in Number Theory*, Springer-Verlag, New York, 1981.
9. M. Ito, M. Katsura, Context-free languages consisting of non-primitive words, *Intern. J. Computer Math.*, 40 (1991), 157 – 167.
10. M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, Mass., 1983.
11. Gh. Păun, A. Salomaa, Thin and slender languages, *Discrete Appl. Math.*, 61 (1995), 257 – 270.
12. R. Ross, K. Winklmann, Repetitive strings are not context-free, *Report CS-81-070*, Computer Sci. Dept., Washington Univ., 1981.
13. G. Rozenberg, A. Salomaa, *The Mathematical Theory of L Systems*, Academic Press, New York, 1980.
14. A. Salomaa, *Formal Languages*, Academic Press, New York, London, 1973.
15. A. Thue, Über unendliche Zeichenreihen, *Norske Videns. Selsk. Skrifter Mat.-Nat. Kl.*, Kristiania, 7 (1906), 1 – 22.