



Row-Column Combination of Dyck Words

Stefano Crespi Reghizzi¹, Antonio Restivo², and Pierluigi San Pietro¹(✉)

¹ DEIB - Politecnico di Milano, Milan, Italy

{stefano.crespireghizzi, pierluigi.sanpietro}@polimi.it

² Dipartimento di Matematica e Informatica, Università di Palermo, Palermo, Italy

antonio.restivo@unipa.it

Abstract. We lift the notion of Dyck language from words to 2-dimensional arrays of symbols, i.e., pictures. We define the Dyck crossword language DC_k as the row-column combination of Dyck word languages, which prescribes that each column and row is a Dyck word over an alphabet of size $4k$. The standard relation between matching parentheses is represented in DC_k by an edge of the matching graph situated on the picture array. Such edges form a circuit, of path length multiple of four, where row and column matches alternate. Length-four circuits are rectangular patterns, while longer ones exhibit a large variety of patterns. DC_k languages are not recognizable by the Tiling Systems of Giammarresi and Restivo. DC_k contains pictures where circuits of unbounded length occur, and where any Dyck word occurs in a row or in a column. We prove that the only Hamiltonian circuits of the matching graph of DC_k have length four. A proper subset of DC_k , called quaternate, includes only the rectangular patterns; we define a proper subset of quaternate pictures that (unlike the general ones) preserves a characteristic property of Dyck words: availability of a cancellation rule based on a geometrical partial order relation between rectangular circuits. Open problems are mentioned.

1 Introduction

The Dyck language is a fundamental concept in formal language theory. Its alphabet $\{a_1, \dots, a_k, b_1, \dots, b_k\}$, for any $k \geq 1$, is associated with the pairs $[a_1, b_1], \dots, [a_k, b_k]$. The language is the set of all words that can be reduced to the empty word by cancellations of two coupled letters: $a_i b_i \rightarrow \varepsilon$. Dyck words represent the last-in-first-out order of events, a fundamental concept for theoretical computer science and especially for formal language and automata theory, where the Chomsky-Schützenberger theorem [1] states that any context-free language is the homomorphic image of the intersection of a Dyck language and a regular one.

Motivated by our interest in the theory of two-dimensional (2D) or picture languages (from now on simply “languages”), we investigate the possibility to transport the Dyck concept from one dimension to 2D. When moving from 1D to 2D, most formal language concepts and relationships drastically change. In particular, in 2D the Chomsky hierarchy of languages is blurred because the notions of regularity and context-freeness cannot be formulated for pictures without giving up some characteristic properties that hold for words. In fact, it is known [7] that the three equivalent definitions of

regular languages by means of finite-state recognizers, by regular expressions, and by the homomorphism of local languages, produce in 2D three distinct language families. The third one gives the family of *tiling system recognizable languages* (REC) [7], that is perhaps the best known definition for regularity in 2D.

The situation is less satisfactory for context-free languages, of which Dyck languages are a notable example, where a transposition in 2D remains problematic. None of the existing proposals of “context-free” picture grammars ([3, 5, 9–12], a survey is [2]) match the expressiveness and richness of formal properties of 1D context-free grammars. In particular, we are not aware of any existing definitions of 2D Dyck languages,¹ and we hope that the present one will open a new direction of research on (picture) languages.

It is time to describe our proposal. We consider the picture languages obtained by the *row-column combination*, also known as *crossword*, of two Dyck word languages over the same alphabet. In such a combination, all rows and all columns are Dyck words. Crosswords have been studied for regular languages (e.g., in [6, 8]) but not, to our knowledge, for context-free ones. In particular it is known [8] that the REC family coincides with the projection of the crosswords of two regular languages.

The family of Dyck crosswords over an alphabet of size $4k$, denoted by DC_k , $k \geq 1$, represents, for reasons later explained, a rather general case. It includes a spectrum of pictures where a surprising variety of complex patterns may occur. To analyze them, we introduce the *matching graph* of a picture, where the array cells are the nodes and the matching relation defines the edges. The graph is partitioned into simple (disjoint) circuits, made by alternating horizontal and vertical edges, representing a Dyck match on a row and on a column. A circuit label is a word of length multiple of 4. The edges of a circuit path may cross each other—the case of zero crossings is the length 4 circuit or rectangle. Pictures containing just such rectangular circuits may present quite evident geometrical analogies with the Dyck word case.

We prove that DC_k is not in REC and we positively answer the question whether each Dyck word can occur in DC_k pictures. We present some interesting types of Dyck crosswords that contain multiple circuits including complex ones, but much remains to be understood about the general patterns that are possible and the trade-off between circuit length and the number of circuits that cover a picture. We show that the only pictures covered by one circuit (i.e., Hamiltonian) have size 2×2 ; furthermore, we prove that for any $h \geq 0$ there exist pictures in DC_k featuring a circuit of length $4 + 8h$, i.e., the circuit length is unbounded.

As said, the structure of pictures, called *quaternate*, such that their circuits are rectangular, is intuitively similar to the structure of Dyck words since the vertexes of a rectangle delimit a subpicture much as two coupled parentheses delimit a substring. To formalize such an intuition, we introduce a further subset of Dyck crosswords characterized by a variant of the Dyck cancellation rule. First, we transform cancellation into a *neutralization* rule that maps the four vertex letters of a rectangle on a new neutral

¹ We just know of a particular example, the *Chinese box language* in [3], that intuitively consists of embedded or concatenated rectangles, and was proposed to illustrate the expressiveness of the grammars there introduced. But that language is not a satisfactory proposal, since it is in the family REC, hence “regular” rather than “context-free”.

(i.e., non-coupled) letter N . Then a quaternate picture is *neutralizable* if it reduces to a picture over alphabet $\{N\}$ by applying neutralization steps. We prove that neutralizable pictures are a subset of quaternate ones. The analogy between Dyck words and neutralizable pictures is thus substantiated by the fact that both use neutralization rules for recognition, but there is a difference. The partial order of neutralization is a tree order for words, while for pictures it is a directed acyclic graph that represents the geometric relation of partial containment between rectangles.

Section 2 lists basic concepts of picture languages and Dyck word languages. Section 3 introduces the DC_k languages, exemplifies the variety of circuits they may contain, proves formal properties, and defines the quaternate subclass. Section 4 studies the neutralizable case. Section 5 mentions open problems.

2 Preliminaries

All the alphabets considered are finite. The concepts and notations for picture languages follow mostly [7]. We assume some familiarity with the basic theory of the family REC of tiling system languages, defined as the projection of a local 2D language; the relevant properties of REC will be reminded when needed. A *picture* is a rectangular array of letters over an alphabet. The set of all non-empty pictures over Σ is denoted by Σ^{++} .

A *domain* d of a picture p is a quadruple (i, j, i', j') , with $1 \leq i \leq i' \leq |p|_{row}$, and $1 \leq j \leq j' \leq |p|_{col}$, where $|p|_{row}$ and $|p|_{col}$ denote the number of rows and columns, respectively. The *subpicture* of p with domain $d = (i, j, i', j')$, denoted by $subp(p, d)$ is the (rectangular) portion of p defined by the top-left coordinates (i, j) and by the bottom-right coordinates (i', j') .

Let $p, q \in \Sigma^{++}$. The *horizontal concatenation* of p and q is denoted as $p \oplus q$ and it is defined when $|p|_{row} = |q|_{row}$. Similarly, the *vertical concatenation* $p \ominus q$ is defined when $|p|_{col} = |q|_{col}$. We also use the power operations $p^{\ominus k}$ and $p^{\oplus k}$, $k \geq 1$, their closures $p^{\ominus+}$, $p^{\oplus+}$ and the closure under both concatenations $p^{\ominus+, \oplus+}$; concatenations and closures are extended to languages in the obvious way.

The notation $N^{m,n}$, where N is a symbol and $m, n > 0$, stands for a homogeneous picture in N^{++} of size m, n . For later convenience, we extend this notation to the case where either m or n are 0, to introduce identity elements for vertical and horizontal concatenations: given a picture p of size (m, n) , by definition $p \oplus N^{m,0} = N^{m,0} \oplus p = p$ and $p \ominus N^{0,n} = N^{0,n} \ominus p = p$.

Dyck Alphabet and Language. The definition and properties of Dyck word languages are basic concepts in formal language theory. Let Γ_k , $k \geq 1$, be an alphabet of cardinality $2k$. Γ_k is called a Dyck alphabet if it is associated with a partition into two sets Γ', Γ'' of cardinality k and with a one-to-one total mapping, called *coupling*, from Γ' into Γ'' . If the pair $[a, b]$ is in the coupling, $a \in \Gamma'$, $b \in \Gamma''$, then it is called *coupled pair* and the *coupled letters* a, b are called, respectively, *open* and *closed*. The Dyck language D_k over alphabet Γ_k is the set of words congruent to ε , via the *cancellation rule* $a_i b_i \rightarrow \varepsilon$ that erases two adjacent coupled letters. A pair of coupled letters occurring in a word is called *matching* if it is erased by the same cancellation rule application. Notice that the number of letters between the two letters of a matching pair is always even.

3 Row-Column Combination of Dyck Languages

In this section we define the languages, called simple Dyck Crosswords (DC), such that their pictures have Dyck words in rows and in columns. They may be viewed as analogous in 2D of Dyck 1D languages. Following [7] we introduce the row-column combination operation that takes two word languages and produces a picture language.

Definition 1 (row-column combination a.k.a. crossword). *Let $S', S'' \subseteq \Sigma^*$ be two word languages, resp. called row and column component languages. The row-column combination or crossword of S' and S'' is the language L such that a picture $p \in \Sigma^{++}$ belongs to L if, and only if, the words corresponding to each row (in left-to-right order) and to each column (in top-down order) of p belong to S' and S'' , respectively.*

The crossword of regular languages has received attention in the past since its alphabetic projection coincide with the REC family [7]; some complexity issues for this case are recently addressed in [6] where the crosswords are called “regex crosswords”.

Remark 1. Given two regular languages S', S'' , it is undecidable to establish whether their crossword is empty. This implies that in general there are crosswords that do not saturate their components, i.e., such that the set of all rows (or the set of all columns) occurring in pictures of the crossword is a proper subset of the row component language (or of the column component language).

We investigate the properties of the crossword of a fundamental type of context-free, non-regular languages, the Dyck ones. First, we discuss the alphabet size and couplings.

Theorem 1 (Alphabet size of crosswords of Dyck languages). *Let D', D'' be two Dyck languages over the same alphabet Δ (with possibly distinct couplings over Δ).*

- i) If Δ has fewer than four letters, then the crossword of D', D'' is empty.*
- ii) If Δ has four letters, then (up to isomorphism) there is one and only one coupling for D' and for D'' such that the crossword of D', D'' is not empty.*
- iii) If the number of letters of Δ is a multiple of four, then there is a coupling for D' and for D'' such that the crossword of D', D'' is not empty.*

Proof. Part (i): a Dyck alphabet has an even number of letters, hence the only relevant case is the binary alphabet, e.g., $\{a, b\}$. If the coupling for the row language is, say, $[a, b]$, then $[a, b]$ or $[b, a]$ is the coupling for the columns. Given a picture with an occurrence of a , say, in the leftmost column, a letter b must occur in the same column, which would require a coupling $[b, a]$ for rows, a contradiction.

Part (ii): let $\{a, b, c, d\}$ be a Dyck alphabet of four letters. As in Part (i), in the top left corner of any picture there is a letter, say, a , which is an open letter for both rows and columns. Hence, the row language has a coupled pair, say, $[a, b]$ and the column language has a coupled pair, say, $[a, c]$ —we proved above that the couplings $[a, b]$ or $[b, a]$ for columns would lead to an empty language. The letter b is thus on the first row, hence it is an open letter for the column language, therefore the latter must include the coupled pair $[b, d]$ and similarly the row language must include the coupled pair $[c, d]$: there is no other letter left and any other choice than d for the closed letter in either case

would again lead to the empty language. The corresponding crossword is not empty since, among others, it includes all pictures of the form: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{\oplus+, \ominus+}$.

Part (iii): The cardinality of Δ is $4k$, for some $k \geq 1$. It is enough to partition Δ in k subsets of four elements and then for each subset use the same coupling of Part (ii). \square

In particular, the alphabet used in the proof of Part (ii) of Theorem 1 can be denoted as $\Delta_1 = \{a, b, c, d\}$, with the coupling $\{[a, b][c, d]\}$ for the rows and $\{[a, c][b, d]\}$ for the columns. The corresponding (unique) crossword is denoted as DC_1 . A simple example of a picture in DC_1 is in Fig. 1.

We now generalize the definition of DC_1 to alphabets of any cardinality multiple of 4 (as in the proof of Part (iii) of Theorem 1).

Definition 2 (Dyck crossword alphabet and language). *The Dyck crossword alphabet Δ_k is a set of quadruplets, namely $\{a_i, b_i, c_i, d_i \mid 1 \leq i \leq k\}$, together with the following couplings of the Dyck row alphabet Δ_k^{Row} for the row component language D_k^{Row} , and of the column alphabet Δ_k^{Col} for the column component language D_k^{Col} :*

$$\begin{cases} \text{for } \Delta_k^{Row} : \{[a_i, b_i] \mid 1 \leq i \leq k\} \cup \{[c_i, d_i] \mid 1 \leq i \leq k\} \\ \text{for } \Delta_k^{Col} : \{[a_i, c_i] \mid 1 \leq i \leq k\} \cup \{[b_i, d_i] \mid 1 \leq i \leq k\} \end{cases} \quad (1)$$

The simple² Dyck crossword DC_k is the row-column combination of D_k^{Row} and D_k^{Col} .

For brevity, we later drop “simple” when referring to Dyck crosswords.

It is easy to notice that, for every $k \geq 1$, the language DC_k is closed under horizontal and vertical concatenation and their closures, and that for every $n, m \geq 1$ there exist pictures of DC_k of size $(2n, 2m)$.

We prove that DC_k is not recognizable by a tiling system, hence it is not in REC.

Theorem 2 (Comparison with REC). *For every $k \geq 1$, the language DC_k is not in the REC family.*

Proof. By contradiction, assume that DC_k is in REC. Without loss of generality, we consider only the case $k = 1$. Consider the following picture p in DC_1 : $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. From closure properties of REC, the language $p^{\oplus+}$ is in REC, hence also the language:

$$R = (a^{\oplus+} \oplus b^{\oplus+}) \ominus ((a \ominus c) \oplus p^{\oplus+} \oplus (b \ominus d)) \ominus (c^{\oplus+} \oplus d^{\oplus+}).$$

A picture in R has a^+b^+ in the top row and c^+d^+ in the bottom row. Let $T = DC_1 \cap R^{\ominus+}$. By closure properties of REC, both T and $T^{\ominus+}$ are in REC. The first row of every picture in $T^{\ominus+}$ has the form $a^n b^n$, since it is the intersection of Dyck word language over $\{a, b\}$ with the regular language a^+b^+ . By applying the Horizontal Iteration Lemma of [7] (Lemma 9.1) to $T^{\ominus+}$, there exists a (suitably large) picture t in $T^{\ominus+}$ which can be written as the horizontal concatenation of the three (non empty) pictures x, q, y , i.e., $t = x \oplus q \oplus y$, such that $x \oplus q^{i\oplus} \oplus y$ is also in $T^{\ominus+}$, thus contradicting the fact that the top row of the pictures in $T^{\ominus+}$ must have the form $a^n b^n$. \square

² More general definitions of Dyck crosswords are possible if the component languages have different alphabets.

A question, related to Remark 1, to be positively answered, is whether the row and column languages of DC_k , respectively, saturate the row and column components D_k^{Row}, D_k^{Col} . Let $P \subseteq \Delta^{++}$ be a language over an alphabet Δ ; the *row language* of P is: $ROW(P) = \{w \in \Delta^+ \mid \text{there exist } p \in P, p', p'' \in \Delta^{++} \text{ such that } p = p' \ominus w \ominus p''\}$. The column language of P , $COL(P)$, is analogously defined.

Theorem 3 (Saturation of components). $ROW(DC_k) = D_k^{Row}, COL(DC_k) = D_k^{Col}$.

Proof. It is enough to prove that $D_k^{Row} \subseteq ROW(DC_k)$, since the other inclusion is obvious and the case for columns is symmetrical. Without loss of generality, we consider only the case $k = 1$. We prove by induction on $n \geq 2$, that for every word $w \in D_1^{Row}$ of length n there exists a picture $p \in DC_1$ of the form $w_1 \ominus w_2 \ominus w \ominus w_3$ for $w_1, w_2, w_3 \in D_1^{Row}$. There are two base cases, the words ab and cd . The word ab is (also) the third row in the DC_1 picture $ab \ominus cd \ominus ab \ominus cd$, while cd is (also) the third row in the DC_1 picture $ab \ominus ab \ominus cd \ominus cd$. The induction step has three cases: a word $w \in D_1^{Row}$ of length $n > 2$ has the form $w'w''$, or the form $aw'b$ or the form $cw'd$, for some $w', w'' \in D_1^{Row}$ of length less than n . Let p', p'' be the pictures verifying the induction hypothesis for w' and w'' , respectively. The case of concatenation $w'w''$ is obvious (just consider the picture $p' \oplus p''$). The case $aw'b$ can be solved by considering the picture $(a \ominus c \ominus a \ominus c) \oplus p' \oplus (b \ominus d \ominus b \ominus d)$, which is in DC_1 . Similarly, for the case $cw'd$ just consider the DC_1 picture $(a \ominus a \ominus c \ominus c) \oplus p' \oplus (b \ominus b \ominus d \ominus d)$. \square

3.1 Matching-Graph Circuits

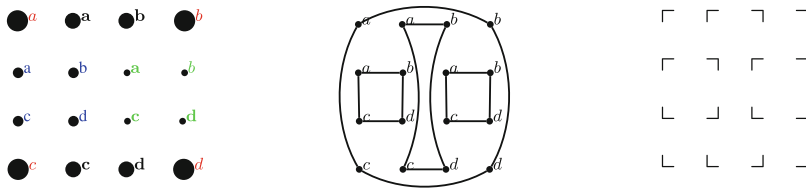


Fig. 1. (Left) A DC_1 picture whose cells are partitioned into 4 quadruples of matching symbols, identified by the same node size (color). (Middle) An alternative visualization by a graph using edges that connect matching symbols (see Definition 3). (Right) The use of corner symbols instead of letters highlights the row and column couplings of rectangle vertexes. (Color figure online)

Indeed, some interesting and surprising patterns may occur in DC_k pictures. The simplest patterns are found in pictures that are partitioned into rectangular circuits connecting four elements, see, e.g., Fig. 1, middle, where an edge connects two symbols on the same row (or column) which match in the row (column) Dyck word. Notice that the graph made by the edges contains four disjoint circuits of length four, called *rectangles* for brevity. Three of the circuits are nested inside the outermost one.

We formally define the graph, situated on the picture grid, made by such circuits.

Definition 3 (Matching graph). *The matching graph associated with a picture $p \in DC_k$, of size (m, n) , is a pair (V, E) where the set V of nodes is the set $\{1, \dots, n\} \times \{1 \dots m\}$ with the obvious labeling over D_k , and the set E of edges is partitioned in two sets of row and column edges defined as follows, for all $1 \leq i \leq n, 1 \leq j \leq m$:*

- for all pairs of matching letters $p_{i,j}, p_{i,j'}$ in Δ_k^{Row} , with $j < j' \leq m$, there is a row (horizontal) edge connecting (i, j) and (i, j') ,
- for all pairs of matching letters $p_{i,j}, p_{i',j}$ in Δ_k^{Col} , with $i < i' \leq n$, there is a column (vertical) edge connecting (i, j) and (i', j) .

Therefore, there is a horizontal edge connecting two matching letters a_i, b_i or c_i, d_i that occur in the same row; analogously, there is a vertical edge connecting two matching letters a_i, c_i or b_i, d_i , that occur in the same column.

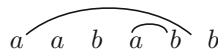
Theorem 4 (Matching circuits). *Let p be a picture in DC_k . Then:*

1. *its matching graph is partitioned into simple circuits, called matching circuits;*
2. *for all $1 \leq j \leq k$, the clockwise visit of a matching circuit, starting from any of its nodes with label a_j , yields a word in $(a_j b_j d_j c_j)^+$, called the circuit label.*

Proof. Part (1): By Definition 3, every node of G has degree 2, with one row edge and one column edge, since its corresponding row and column in picture p are Dyck words. Every node must be on a circuit, otherwise there would be a node of degree 1. Each circuit must be simple and the sets of nodes on two circuits are disjoint, else one of the nodes would have degree greater than 2. Part (2) is obvious, since from a node labeled a_j there is a row edge connecting with a node labeled b_j , for which there is a column edge connecting with a d_j , then a row edge connecting d_j with c_j , etc., finally closing the circuit with a column edge connecting a c_j with the original a_j . □

Notice that when a picture on Δ_1 is represented by its matching graph, the node labels are redundant since they are uniquely determined on each circuit.

Theorem 4 has a simple interpretation in the case of Dyck words: the associated matching graph of a Dyck word is the well-known, so-called rainbow representation, e.g.,



of the syntax tree of the word. A matching graph then corresponds to the binary relation induced by the rainbow arcs and a matching circuit just to an arc.

Remark 2. The following elementary property of Dyck words immediately generalizes to crosswords. Let $x a_i y b_i w$ be a Dyck word, where a_i, b_i match; then, for any coupled pair $a_j, b_j, 1 \leq j \leq k$, the string $x a_j y b_j w$ is a Dyck word. For crosswords, the statement is that, by replacing a matching circuit labeled $a_i b_i d_i c_i$ in a picture in DC_k with a matching circuit labeled $a_j b_j d_j c_j$, the result is still in DC_k .

A natural question is whether there are pictures with more complex matching circuits than rectangular ones. It is maybe unexpected that moving from 1D to 2D the

circuit length is not just 2×2 , but may increase without an upper bound. Two examples of pictures in DC_1 with matching circuits longer than four are in Fig. 2: (left), with a circuit of length 12 labeled by the word $(abdc)^3$, and (right) with a circuit of length 36.

The pictures of DC_k , like the ones in Figs. 1 and 4, that are devoid of circuits longer than four make a proper subset that we define for later convenience.

Definition 4 (Quaternate DC_k). A Dyck crossword picture such that all its matching circuits are of length 4 is called quaternate; the corresponding language, denoted by DQ_k , is the quaternate Dyck language.

Corollary 1. Quaternate Dyck languages DQ_k are strictly included in Dyck crosswords DC_k for all $k \geq 1$.

The structure of quaternate pictures having only rectangular circuits is made more evident by an alternative typography for the Dyck alphabet, using so-called *corner symbols* instead of Latin letters. Let Δ_1 be the alphabet $\{\lrcorner, \ulcorner, \llcorner, \lrcorner\}$ with the correspondence: $a = \lrcorner$, $b = \ulcorner$, $c = \llcorner$, $d = \lrcorner$. Thus, the picture $\begin{array}{cccc} \lrcorner & \ulcorner & \lrcorner & \ulcorner \\ \ulcorner & \lrcorner & \ulcorner & \lrcorner \\ \llcorner & \lrcorner & \llcorner & \lrcorner \\ \lrcorner & \ulcorner & \lrcorner & \ulcorner \end{array}$ is the same as $\begin{array}{cccc} a & b & a & b \\ c & d & c & d \end{array}$. Another example is in Fig. 1, right.

Section 4 studies the quaternate pictures and defines a sublanguage where the containment relation of rectangles defines a partial order.

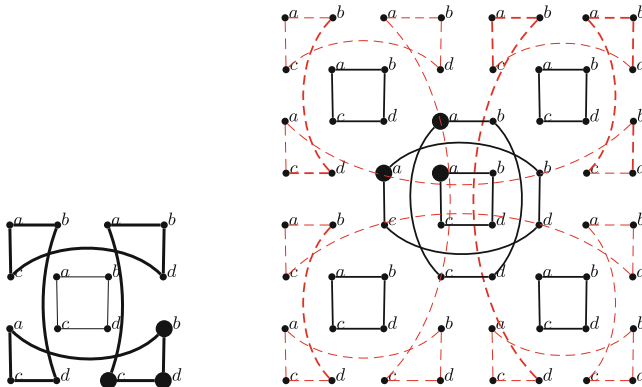


Fig. 2. Two pictures in DC_1 . (Left) The picture is partitioned into two circuits of length 12 and 4. (Right) The picture includes a circuit of length 36 and seven rectangular circuits. Its pattern embeds four partial copies (direct or rotated) of the left picture; in, say, the NW copy of the evidenced “triangle”, the letters b, d, c have been changed to a, a, a , also evidenced by larger dots. Such a transformation can be reiterated to grow a series of pictures.

We continue with the study of longer circuits.

Theorem 5 (Unbounded circuit length). For all $h \geq 0$ there exist a picture $p_{(h)}$ in DC_k that contains a matching circuit of length $4 + 8h$.

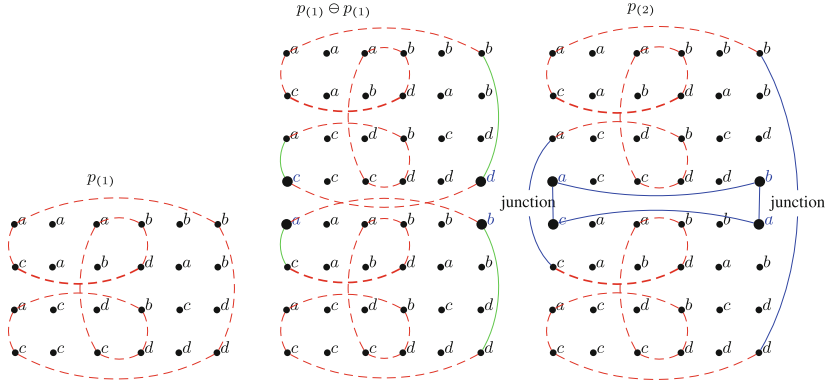


Fig. 3. (Left) Picture $p_{(1)}$ used as induction basis of Theorem 5. It is covered by a circuit of length $4 + 8 \cdot 1 = 12$ and by 3 rectangles (not shown). (Middle) Picture $p_{(1)} \ominus p_{(1)}$, the four arcs to be deleted are in green (solid lines), and the four nodes to be relabeled are in blue (also larger dots). (Right) Inductive step: picture $p_{(2)}$ is obtained from $p_{(1)} \ominus p_{(1)}$ by canceling the four green arcs, relabeling the four larger blue nodes as shown (the corresponding rectangle is in blue) and finally adding two solid arcs (blue) that join the double-noise circuits. A circuit of length $4 + 8 \cdot 2$ results. All length 4 circuits of $p_{(h-1)}$ and $p_{(1)}$ (not shown for clarity) are unchanged in $p_{(h)}$. (Color figure online)

Proof. We prove the statement for DC_1 , since $DC_1 \subseteq DC_k$. The case $h = 0$ is obvious. The case $h > 0$ is proved by induction on a sequence of pictures $p_{(1)}, \dots, p_{(h)}$ using as basis the DC_1 picture $p_{(1)}$ in Fig. 3 (left), that has size $(m_{(1)}, 6)$, where $m_{(1)} = 4$, and contains a circuit of length $12 = 4 + 8$, referred to as double-noose.

Induction step. It extends picture $p_{(h-1)}$, $h > 1$, by appending a copy of $p_{(1)}$ underneath and making a few changes essentially defined in Fig. 3 (right). It is easy to see that the result is a picture $p_{(h)}$ of size $(m_{(h-1)} + 4, 6)$ such that: $p_{(h)} \in DC_1$ and $p_{(h)}$ contains a circuit of length $4 + 8h$. \square

Another series of pictures that can be enlarged indefinitely is the one in Fig. 2, where the first two terms of the series are shown.

An examination of Fig. 3 in the next example shows that there are subsets of DC_k that are in REC, yet they contain quite complex matching circuits.

Example 1. The language L composed of all pictures $p_{(h)}$, for all $h \geq 1$, of Theorem 5 is in the REC family. We first extend the alphabet of L to $\{a, b, c, d, a_1, b_1, c_1, d_1\}$ so that the circuits longer than 4 are over the alphabet $\{a_1, b_1, c_1, d_1\}$ and the remaining circuits are over $\{a, b, c, d\}$. The resulting pictures $p'_{(h)}$, constituting a language L' , have only 6 distinct rows, here identified (from top to bottom) with the letters 1, ..., 6:

- 1: $a_1 a a_1 b_1 b b_1$, 2: $c_1 a b d_1 a b$, 3: $a_1 c d b_1 c d$, 4: $a c c_1 d_1 d b$, 5: $c a a_1 b_1 b d$, 6: $c_1 c c_1 d_1 d d_1$.

It is clear from the construction of the pictures p_h for $h > 1$ that L' can be defined as $1 \ominus (2 \ominus 3 \ominus 4 \ominus 5)^* \ominus 2 \ominus 3 \ominus 6$. Since each of rows 1, ..., 6 can be seen as a finite language (thus, in REC) and tiling systems are closed by vertical concatenation and closure, also

L' is in REC. By closure of REC under projection, also L is in REC (by projecting a_1 to a , b_1 to b , etc.)

From an elementary property of Dyck word languages it follows that the distance on the picture grid between two nodes connected by an edge is an odd number, to let room for an even number of letters. This suggests the following Lemma 1.

Given a picture p over an alphabet Γ , let $x=p_{i,j}$, for $x \in \Gamma$. We say that the occurrence of x in position (i, j) has *row parity 1* if i is even, row parity -1 otherwise; similarly, x in (i, j) has *column parity 1* if j is even, column parity -1 otherwise.

Lemma 1 (Circuit property). *Let γ be a matching circuit of a picture in DC_k , with label in $(a_i b_i d_i c_i)^+$.*

- i) *All occurrences of a_i and b_i have the same row parity, but they have opposite row parity to every occurrence of c_i and d_i ;*
- ii) *All occurrences of a_i and c_i have the same column parity, but they have opposite column parity to every occurrence of b_i and d_i .*

Proof. Without loss of generality, let $k = 1$. Let an occurrence of a in γ be in a position row (r, s) . The vertical matching symbol c of a (in the same column s) must occur in a row of the form $2n + 1 + r$, for some $n \geq 0$, since there must be an even number of positions in p between the occurrence of a and c . The same happens for the symbol d matching c and for the b matching the above occurrence of a . The circuit γ continue alternating between odd and even rows, and between odd and even columns, without modifying the row and column parity of each occurrence of the same letter. \square

An application of Lemma 1 follows.

Let p be a picture in DC_k and G its matching graph. A matching circuit that visits all the nodes of G is called *Hamiltonian*.

Theorem 6 (Hamiltonian circuits). *The only existing DC_k pictures with a Hamiltonian matching circuit are defined by the set $\left\{ \begin{matrix} a_i & b_i \\ c_i & d_i \end{matrix} \mid 1 \leq i \leq k \right\}$.*

Proof. Without loss of generality, let $k = 1$. By contradiction, assume that a picture $p \in DC_1$, of size greater than $(2, 2)$, has a Hamiltonian circuit. The first row of any picture is a Dyck word over $\{a, b\}$ and the leftmost column is a (vertical) word over $\{a, c\}$. By Lemma 1, the second row must be a word over $\{c, d\}$ and the second column from the left is over $\{b, d\}$. Therefore, the subpicture $(p(1, 1) \ominus p(2, 1))$ must be $a \ominus c$ (a cannot occur in the second row, therefore the row must begin with the open letter c for rows) and similarly the subpicture $(p(1, 2) \ominus p(2, 2))$ is $b \ominus d$.

Therefore p contains the subpicture $\begin{matrix} a & b \\ c & d \end{matrix}$ in the top, left corner, i.e., it has a matching circuit of length 4, a contradiction with the existence of a Hamiltonian circuit for a picture of size greater than $(2, 2)$. \square

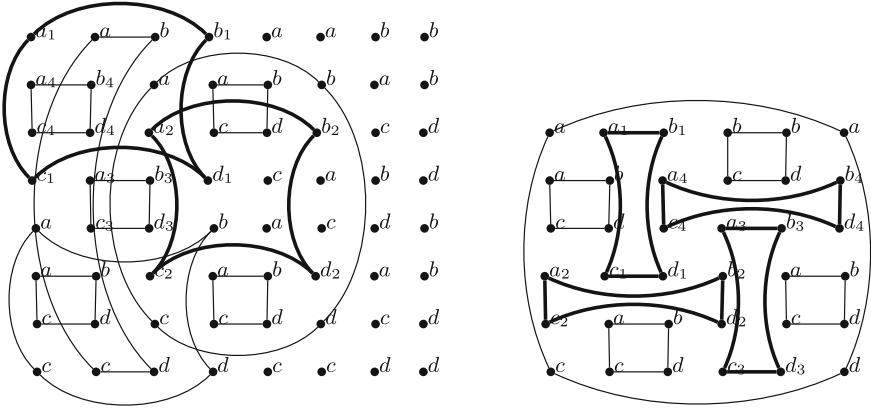


Fig. 4. Both pictures are quaternate but not partially ordered, hence not neutralizable by Theorem 7. (Left) To avoid clogging, the rectangles in the specular right half of the left picture are not drawn. (Right) The vertices relevant for illustrating partial containment are indexed.

4 A Sublanguage Preserving Characteristic Dyck Words Properties

This section only deals with quaternate pictures, whose circuits we call “rectangles”. We show that the standard definition of Dyck words by means of the cancellation rule³ can be extended to a sublanguage of quaternate pictures that is characterized by a geometrical relation of containment between the rectangles.

The absence of long and intricate circuits will permit to define a partial containment relation between the rectangles present in a picture, and then to define a partial order if such a relation is acyclic. The corresponding language is called *partially ordered quaternate*, DPO_k . We also define a subset of Dyck crosswords, named *neutralizable* (DN_k), by means of a cancellation rule suitably transformed into a neutralization operation. At last we prove that the partially ordered and the neutralizable languages are the same, and we list some of their properties.

Preliminarily, we transform the cancellation rule for words $a_i b_i \rightarrow \varepsilon$, which erases innermost matching letters, into a length preserving rule, since in 2D the erasure of an internal subpicture would create a “hole”, producing an object that no longer qualifies as a picture. The Dyck cancellation rule is rephrased as the *neutralization rule* $a_i b_i \rightarrow NN$, where N is a new “neutral” (i.e., not coupled) letter; in this way a Dyck word is mapped to a word in N^+ by a series of neutralization steps.

Geometrical Containment Relation. Consider two rectangles R_1 and R_2 with vertexes, resp., a_1, b_1, c_1, d_1 and a_2, b_2, c_2, d_2 (the letters are distinct to simplify reference).

We say that R_1 is *partially contained* in R_2 , writing $R_1 < R_2$, if some vertexes of R_1 are inside or on a side of R_2 . The partial containment relation of a picture is the set of all such relations. Notice that the number of vertexes contained in R_2 may be 1, 2 or 4, but not 3 which is geometrically impossible.

³ In [4] the property of well nesting of parentheses is also reformulated for quaternate pictures.

Figure 4, left, illustrates (among others) the following containment relations: $R_1 < R_2, R_2 < R_1, R_3 < R_1, R_3 < R_2, R_4 < R_1$.

Definition 5 (Partially ordered quaternate picture). A quaternate picture in DQ_k is called partially ordered if its partial containment relation “ $<$ ” is acyclic. The language of such pictures is denoted by DPQ_k .

We observe that the pictures in Fig. 4 are not partially ordered, because they respectively contain the cycles $R_1 < R_2 < R_1$ and $R_1 < R_2 < R_3 < R_4 < R_1$. On the other hand, the picture presented in Example 2 below is partially ordered since its partial containment relation (displayed in the example) is acyclic.

Neutralizable Dyck Languages

We introduce a neutralization rule mapping the letters in a quadruple, representing the corners of a subpicture, to a new neutral letter N . The *neutralizable Dyck* language DN_k is obtained by iterating neutralization, starting from 2-by-2 subpictures, until the picture is wholly neutralized. Given a picture p , all subpictures of the form $\begin{matrix} \lrcorner & \ulcorner \\ \llcorner & \lrcorner \end{matrix}$ are

neutralized, i.e., replaced in p by the subpicture $\begin{matrix} N & N \\ N & N \end{matrix}$. If p includes a subpicture with four matching corners and having its interior and sides completely neutralized, then also

$$\begin{matrix} \ulcorner & N \dots N & \urcorner \\ N & \dots & N \end{matrix}$$

the four corners are neutralized. This is schematized by the subpicture $\begin{matrix} \vdots & N \dots N & \vdots \\ N & \dots & N \\ \llcorner & N \dots N & \lrcorner \end{matrix}$ that

is replaced by a subpicture of the same size having only N as letters. The procedure successfully terminates when the resulting picture is in N^{++} .

Definition 6 (Neutralizable Dyck language). Let N be a new symbol not in Δ_k . The neutralization relation $\xrightarrow{\nu} \subseteq (\{N\} \cup \Delta_k)^{++} \times (\{N\} \cup \Delta_k)^{++}$, is the smallest relation such that for every pair of pictures p, p' in $(\{N\} \cup \Delta_k)^{++}$, $p \xrightarrow{\nu} p'$ if there are $m, n \geq 2$ and $1 \leq i \leq k$, such that p' is obtained from p by replacing a subpicture of p of the form:

$$(a_i \ominus N^{m-2,1} \ominus c_i) \oplus N^{m,n-2} \oplus (b_i \ominus N^{m-2,1} \ominus d_i). \tag{2}$$

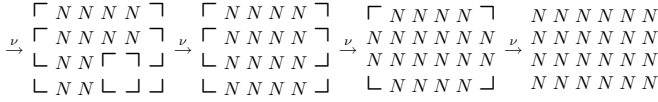
with the picture (of the same size) $N^{m,n}$.

The neutralizable Dyck language, denoted by $DN_k \subseteq \Delta_k^{++}$, is the set of pictures p such that there exists $p' \in N^{++}$ with $p \xrightarrow{\nu^+} p'$.

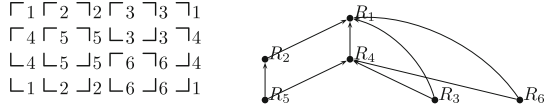
To sum up, a DN_k picture is transformed into a picture in N^{++} by a series of neutralizations, applied in any order. Clearly, every neutralizable picture is a quaternate.

Example 2 (Neutralizations). The following picture p on the alphabet Δ_1 is in DN_1 since it reduces to the neutral one by means of a sequence of six neutralization steps:

$$p = \begin{matrix} \ulcorner \ulcorner \ulcorner \ulcorner \ulcorner & \ulcorner \ulcorner \ulcorner \ulcorner \ulcorner & \ulcorner N N \ulcorner \ulcorner \ulcorner \\ \ulcorner \ulcorner \ulcorner \ulcorner \ulcorner & \ulcorner N N \ulcorner \ulcorner \ulcorner & \ulcorner N N \ulcorner \ulcorner \ulcorner \\ \llcorner \llcorner \llcorner \llcorner \llcorner & \llcorner N N \llcorner \llcorner \llcorner & \llcorner N N \llcorner \llcorner \llcorner \\ \llcorner \llcorner \llcorner \llcorner \llcorner & \llcorner \llcorner \llcorner \llcorner \llcorner & \llcorner N N \llcorner \llcorner \llcorner \end{matrix}$$



Neutralizations have been applied in a left to right order. We show the partial containment relation “<”, with the numbering below.



The relation represented by the graph is acyclic and defines a partial order on the set of rectangles, thus proving that this picture is in DPO_1 .

It is no coincidence that the picture of Example 2 is both neutralizable and partially ordered; the next theorem proves that the two definitions define the same set of pictures.

Theorem 7 (Partially ordered equals neutralizable). *A quaternate picture is neutralizable if, and only if, it is partially ordered, i.e., $DN_k = DPO_k$.*

Proof. Let relation < be acyclic. Then sort the rectangles in topological order and apply neutralization starting from a rectangle without predecessors. When a rectangle is checked, all of its predecessors have already been neutralized, and neutralization can proceed until all rectangles are neutralized. The converse is obvious: if relation < has a cycle, no rectangle in the cycle can be neutralized. □

This result supports the analogy between the neutralization rule for Dyck words and the rule of the same name for pictures: both rely on a partial order relation such that any topological sorting order can be applied to perform neutralization. For Dyck words, the order is a tree partial order, whereas for pictures it is a directed acyclic graph.

Properties of Neutralizable Picture Languages. The result on row/column language saturation (Theorem 3) remains valid, i.e., $ROW(DN_k) = D_k^{Row}$, $COL(DN_k) = D_k^{Col}$, since the languages used in the proof of that theorem are also in DN_k .

Similarly, by a proof almost identical to the one of Theorem 2, since the language $T^{\Theta+}$ can be obtained from DN_k by intersection with a recognizable language, we have:

Theorem 8 (Comparison with REC). *The languages DN_k and DQ_k are not in REC for every $k \geq 1$.*

From Theorem 7 and from the examples of Fig. 4, we have the inclusions:

Theorem 9 (Hierarchy). $DN_k \subsetneq DQ_k \subsetneq DC_k$.

5 Conclusion

In our opinion, the mathematical study of the properties of 2D Dyck languages is a promising research area, where much remains to be understood, for the general case of (simple) Dyck crosswords containing matching circuits of any length. Very diverse patterns may occur in such crosswords, that we have been able to start classifying just in the simpler case of rectangular circuits. In fact, the variety of patterns depends on quite a few circuit parameters such as the circuit length, the number of crossings in a circuit or between different circuits, and, more vaguely, the relative positions of circuits on the grid. We mention a few specific open problems.

First, by Theorem 4 the length of circuits in DC_1 pictures is unbounded, of the form $4 + 8h$ for all values $h \geq 0$. The question is whether, for each $n \geq 1$, there is a DC_1 picture containing a circuit of length $4n$.

Second, it seems that every picture in DC_k has at least one circuit of length 4.

Third, the number of circuits present in a picture is unbounded for the picture series used in the proof of Theorem 4. This raises the more general question whether, by bounding the number of circuits present in a picture, the number of such pictures is also bounded. (Theorem 6 bounds the number of pictures with only one circuit.)

Another question concerns the properties of those DC_k sublanguages that are in REC. For instance, Example 1, though visually complex, satisfies such a hypothesis.

At last, we mention a related future research direction on context-free crosswords, having as baseline the present work on Dyck crosswords and the variant of the Chomsky-Schützenberger Theorem [1] that characterizes the context-free word languages as the non-erasing homomorphism of the intersection of a Dyck language and a regular one.

References

1. Chomsky, N., Schützenberger, M.: The algebraic theory of context-free languages. In: Bradford, H. (ed.) *Computer Programming and Formal Systems*, pp. 118–161. North-Holland, Amsterdam (1963)
2. Crespi Reghizzi, S., Giammarresi, D., Lonati, V.: Two-dimensional models. In: Pin, J. (ed.) *Handbook of Automata Theory*, pp. 303–333. European Mathematical Society Publishing House (2021)
3. Crespi-Reghizzi, S., Pradella, M.: Tile rewriting grammars and picture languages. *Theor. Comput. Sci.* **340**(1), 257–272 (2005). <https://doi.org/10.1016/j.tcs.2005.03.041>
4. Crespi Reghizzi, S., Restivo, A., San Pietro, P.: Two-dimensional Dyck words. *CoRR* abs/2307.16522 (2023)
5. Drewes, F.: *Grammatical Picture Generation: A Tree-Based Approach*. Springer, Heidelberg (2006). <https://doi.org/10.1007/3-540-32507-7>
6. Fenner, S.A., Padé, D., Thierauf, T.: The complexity of regex crosswords. *Inf. Comput.* **286**, 104777 (2022). <https://doi.org/10.1016/j.ic.2021.104777>
7. Giammarresi, D., Restivo, A.: Two-dimensional languages. In: Rozenberg, G., Salomaa, A. (eds.) *Handbook of Formal Languages*, vol. 3, pp. 215–267. Springer, Heidelberg (1997). https://doi.org/10.1007/978-3-642-59126-6_4
8. Latteux, M., Simplot, D.: Recognizable picture languages and domino tiling. *Theor. Comput. Sci.* **178**(1–2), 275–283 (1997). [https://doi.org/10.1016/S0304-3975\(96\)00283-6](https://doi.org/10.1016/S0304-3975(96)00283-6)

9. Matz, O.: Regular expressions and context-free grammars for picture languages. In: Reischuk, R., Morvan, M. (eds.) STACS 1997. LNCS, vol. 1200, pp. 283–294. Springer, Heidelberg (1997). <https://doi.org/10.1007/BFb0023466>
10. Nivat, M., Saoudi, A., Subramanian, K.G., Siromoney, R., Dare, V.R.: Puzzle grammars and context-free array grammars. *Int. J. Pattern Recogn. Artif. Intell.* **5**, 663–676 (1991)
11. Průša, D.: Two-dimensional Languages. Ph.D. thesis, Charles University, Faculty of Mathematics and Physics, Czech Republic (2004)
12. Siromoney, R., Subramanian, K.G., Dare, V.R., Thomas, D.G.: Some results on picture languages. *Pattern Recogn.* **32**(2), 295–304 (1999)