Chomsky-Schützenberger-type representations of poly-slender context-free languages

Lucian Ilie

Turku Centre for Computer Science TUCS
FIN-20520 Turku, Finland

Grzegorz Rozenberg

Department of Computer Science, Leiden University
P.O.Box 9512, 2300 RA Leiden, The Netherlands
and
Department of Computer Science
University of Colorado at Boulder
Boulder, CO 80309, U.S.A.

Arto Salomaa

Turku Centre for Computer Science TUCS
FIN-20520 Turku, Finland

Turku Centre for Computer Science
TUCS Technical Report No 242
February 1999

ISBN 952-12-0388-9
ISSN 1239-1891

Abstract

For a non-negative integer k, we say that a language L is *k-poly-slender* if the number of words of length n in L is of order $\mathcal{O}(n^k)$. We give a characterization of the k-poly-slender context-free languages. The base for the structure of these languages is the structure of the Dyck language and so our characterization resembles the well-known Chomsky-Schützenberger characterization of context-free languages.

TUCS Research Group

Theory Group: Mathematical Structures in Computer Science

Introduction

A lot of attention has been devoted recently to considerations concerning the lengths of words in regular and context-free languages. An infinite sequence $(\#_L(n))_{n\geq 0}$ can be associated in a natural way to a language L: $\#_L(n)$ is the number of words of length n in L. The idea is by no means new; already in [3] (and also [11]) one considers the notion of the *population* function of a language L which associates, to every n, the number of words of length *at most* n in L.

When $\#_L(n)$ is bounded from above by a fixed constant, such languages are called *slender* in [1], where some of their applications to cryptography are presented. Independently of each other, [12] and [15] gave the same characterization of the slender regular languages, as finite unions of sets of the form uv^*w. A characterization of the slender context-free languages, as finite unions of sets of the form $\{uv^nwx^ny \mid n \geq 0\}$, was conjectured in [12] and shown to be true by [8]. Some generalizations have been considered in [6] and [7] in connection with formal power series and semilinear languages.

The case when $\#_L(n)$ is bounded by a polynomial has been considered for regular languages in [16] and for context-free languages in [13]. In this paper, we consider the situation when the degree of the polynomial is a fixed non-negative integer; the obtained languages are called *k-poly-slender*. Generalizing the result of [8], we give a characterization of the k-poly-slender context-free languages which resembles the well-known Chomsky-Schützenberger characterization of context-free languages since, as we show, the structure of the Dyck language is the base for the structure of these languages. The corresponding characterization given in [16] for regular languages follows immediately from ours.

Poly-slender languages

We first fix some notations. For a word w and a letter a, $|w|$ is the length of w, $|w|_a$ is the number of occurrences of a in w, and $\rho(w)$ denotes the primitive root of w. The conjugacy relation is denoted $\sim$; for two words u, v, $u \sim v$ iff $u = pq, v = qp$, for some p, q. For basic notions and results of combinatorics on words and formal languages we refer to [4, 10] and [14], respectively.

For a language L, denote $\#_L(n) = \mathrm{card}(\{w \in L \mid |w| = n\})$; this is referred to as the **complexity** (function) of L. For an integer $k \geq 0$, L is called **k-poly-slender** if $\#_L(n) = \mathcal{O}(n^k)$. L is **poly-slender** iff it is k-poly-slender, for some $k \geq 0$.

A language $L \subseteq \Sigma^*$ is called **bounded** if there are some words $w_1, w_2, \ldots, w_n \in \Sigma^*$ such that $L \subseteq w_1^* w_2^* \ldots w_n^*$.

It is clear that the class of poly-slender languages is the same with the class of languages with the population function polynomially limited. Therefore, the characterization theorem of [9] can be written as below (it appears in this form in [13]).

Theorem 1 ([9]). *A context-free language is poly-slender iff it is bounded.*

Dyck loops

Consider the **Dyck** language of order $k, k \geq 1$, $D_k \subseteq \{[_i,]_i \mid 1 \leq i \leq k\}^*$. D_k is generated by $S \longrightarrow SS \mid [_i S]_i \mid \varepsilon$. Consider also a word $z \in D_k$ with $|z|_{[_i} = |z|_{]_i} = 1, 1 \leq i \leq k$, and some words $u_i, v_i, w_i \in \Sigma^*$, where $\Sigma \cap \{[_i,]_i \mid 1 \leq i \leq k\} = \emptyset$. For any integers $n_i \geq 0, 1 \leq i \leq k$, define the morphism $h_{n_1,\ldots,n_k} : (\Sigma \cup \{[_i,]_i \mid 1 \leq i \leq k\})^* \longrightarrow \Sigma^*$, by $a \longrightarrow a$, for $a \in \Sigma$, $[_i \longrightarrow u_i^{n_i}$, $]_i \longrightarrow v_i^{n_i}$, $1 \leq i \leq k$. Put $z = z_1 z_2 \ldots z_{2k}, z_i \in \{[_i,]_i \mid 1 \leq i \leq k\}$.

Then $D \subseteq \Sigma^*$ is a $\boldsymbol{k}$**-Dyck loop** if, for some u_i, v_i, w_i, z as above,

$$D = \{h_{n_1,\ldots,n_k}(w_0 z_1 w_1 z_2 w_2 \ldots z_{2k} w_{2k}) \mid n_i \geq 0\}. \tag{1}$$

We shall call z an *underlying word* of D. (Clearly, z is not unique.) Also, h will stand for $h_{1,1,\ldots,1}$ and will be called an *underlying morphism* of D. D is a **Dyck loop** iff it is a k-Dyck loop, for some k. Notice that if $l < k$, then any l-Dyck loop is also a k-Dyck loop.

Example 2. For the underlying Dyck word $z = [_1]_1[_2[_3]_3[_4]_4]_2$, we can construct the Dyck loop

$$L_1 = \{a^{n_1} bb(abb)^{n_2} ab^{4n_3} a^{n_4} bba^{n_4} ba(aab)^{3n_2} \mid n_i \geq 0\}. \tag{2}$$

The underlying morphism h is clear if we just mention that the images of $]_1$ and $]_3$ are empty.

Example 3. The following more intricate example will be used also later. The underlying Dyck word is here $z = [_1]_1[_2[_3[_4]_4[_5]_5]_3[_6]_6[_7]_7]_2$, and we construct the Dyck loop

$$L_2 = \{a^{2n_1}(ba)^{n_1} b(ab)^{3n_2} b^{n_3} b^{5n_4} b^{n_5} a^{4n_5}(aba)^{n_3} b^{3n_6} a^{n_7} b^{2n_7} b^{n_2} \mid n_i \geq 0\}. \tag{3}$$

About the underlying morphism we mention only that the images of any of $]_4$ and $]_6$ are empty.

Bounded languages

Theorem 4 ([5]). *The family of bounded context-free languages is the smallest family which contains all finite languages and is closed under the following operations:*

(i) union,

(ii) catenation,

(iii) $(x, y) \star L = \bigcup_{n \geq 0} x^n L y^n$, *for* x, y *words.*

Remark 5. Clearly, Theorem 4 is still valid if, instead of finite languages, one starts from unary languages (that is, languages containing one word only). Moreover, if one starts from unary languages and uses only the operation (ii) and (iii) from Theorem 4, then what is obtained is always a Dyck loop. Conversely, any Dyck loop can be obtained in this way. Indeed, this is clear from the definition of the Dyck loops; the role of the production $S \longrightarrow SS$ is the same with the one of the catenation and the role of a production $S \longrightarrow [_i S]_i$ is the same with the one of the operation $\star$, in the sense that we use $(u_i, v_i) \star L$.

Therefore, we get from Theorems 1 and 4

Theorem 6. *For a context-free language* L*, the following assertions are equivalent:*

(i) L *is bounded,*

(ii) L *is poly-slender,*

(iii) L *is a finite union of Dyck loops.*

Proof. First, (i) and (ii) are equivalent by Theorem 1. Second, (iii) clearly implies any of (i) and (ii). Third, (i) implies (iii) follows from Remark 5 and the distributivity of the catenation and "$\star$" with respect to union. ∎

Characterization of k-poly-slenderness

By the well-known Chomsky-Schützenberger theorem (see, e.g., [2]) any context-free language is characterized as a homomorphic image of the intersection between the Dyck language and a regular one. We show below that the structure of the Dyck language is also the base for the structure of the poly-slender context-free languages.

In the above notations, the result of [8] is written as

Theorem 7 ([8]). *A context-free language is* 0*-poly-slender iff it is a finite union of* 1*-Dyck loops.*

In what follows, we shall generalize Theorem 7 to

Theorem 8. *For any $k \geq 0$, a context-free language is k-poly-slender iff it is a finite union of $(k+1)$-Dyck loops.*

Proof. One implication is obvious. For the other, the basic idea is to do the construction in Theorem 4 in a certain way, that is (roughly), anytime "$\star$" is applied, the highest power of n in the complexity function $\#_L(n)$ associated with the language L is increased by one and, anytime catenation is applied to two infinite languages, the highest exponent in the complexity of the resulting language is the sum of the former two.

Let us consider a context-free language L which is k-poly-slender. Then, by Theorem 6, L is a finite union of Dyck loops. If L contains no l-Dyck loop with $l > k+1$, then we are done. Assume then there is such a loop, say D, and that it has the form in (1) with l instead of k. We assume, with no loss of generality, that $u_i v_i \neq \varepsilon$, for any $1 \leq i \leq l$, as otherwise we have a Dyck loop of smaller order.

For three words u, w, v such that u and v are non-empty, we say that w **links** u with v, denoted $\mathsf{link}(u, w, v)$, iff $\rho(u)w = w\rho(v)$; it means that there are words p and q such that $\rho(u) = pq, \rho(v) = qp$, and $w \in (pq)^*p$. The idea is that w links u with v when in words of the form $u^n w v^m$ the period $|\rho(u)|$ is continued over w throughout v^m ($u^n w v^m$ is a prefix of $\rho(u)^\omega$), that is, one cannot really distinguish w there. Notice that, when w is empty, $\mathsf{link}(u, \varepsilon, v)$ means precisely that u and v are powers of the same word, in accordance with the intuitive meaning of link.

The following lemma will be a very useful tool in the proof.

Lemma 9. *Consider the words $x_i \in \Sigma^+, y_i \in \Sigma^*$, and some non-negative integers n_i, m_i. Denote $w = y_0 x_1^{n_1} y_1 x_2^{n_2} y_2 \ldots x_r^{n_r} y_r$, $w' = y_0 x_1^{m_1} y_1 x_2^{m_2} y_2 \ldots x_r^{m_r} y_r$ and assume that $\mathsf{link}(x_i, y_i, x_{i+1})$ holds for no i. Then there is a constant N_0, depending only on the lengths of the words x_i and y_i, such that, if all n_i, m_i are larger than N_0 and there is i with $n_i \neq m_i$, then $w \neq w'$.*

Proof. We prove that $N_0 = \frac{1}{\min |x_i|}(\max |y_i| + 4 \max |x_i|)$ is good. Assume $w = w'$. Without loss of generality, we may assume $n_1 > m_1$. If $(n_1 - m_1)|x_1| \geq |y_1| + |x_1 x_2|$, then, by Fine and Wilf's theorem and the choice of N_0, $\rho(x_1) \sim \rho(x_2)$ and, if $\rho(x_1) = pq, \rho(x_2) = qp$, then $y_1 \in (pq)^*p$. This means $\mathsf{link}(x_1, y_1, x_2)$, a contradiction. Otherwise, by the choice of N_0, we get, for some $s \geq 0$, $x_1^{n_1 - m_1} y_1 = y_1 \rho(x_2)^s$ thus again $\mathsf{link}(x_1, y_1, x_2)$, contradicting our assumption. The lemma is proved. ∎

We shall consider very much in the sequel the links made by w_i between the two adjacent images of h, $h(z_i)$ and $h(z_{i+1})$. But those images of h which are empty have no interest for the language (this is why link is undefined when the first or the third component is empty) and therefore, for any $1 \le i \le 2k-1$, we define $\mathsf{next}(i)$ as the smallest $j \ge i+1$ such that $h(z_j) \ne \varepsilon$. If there is no $\mathsf{link}(h(z_i), w_i w_{i+1} \dots w_{\mathsf{next}(i)-1}, h(z_{\mathsf{next}(i)}))$, then, by Lemma 9, any two different tuples of n_i's (n_i's are all assumed large enough – the complexity order is not affected) give different words of D. Therefore, $\#_D(n) \ne \mathcal{O}(n^{l-2})$, which contradicts the fact that L is k-poly-slender. Consequently, there are such links.

We then group together the linked powers and apply Lemma 9 to those. In order to make things clear, we define two relations on the set $\{i \mid 1 \le i \le 2k, h(z_i) \ne \varepsilon\}$. (Recall that the positions i with $h(z_i) = \varepsilon$ are ignored.) The first is $\mathsf{chain}(i, \mathsf{next}(i))$ iff $\mathsf{link}(h(z_i), w_i w_{i+1} \dots w_{\mathsf{next}(i)-1}, h(z_{\mathsf{next}(i)}))$. We say that i and j are in the same **chain** iff $\mathsf{chain}^\sharp(i,j)$, where $\mathsf{chain}^\sharp$ is the equivalence generated by chain. The second is $\mathsf{syst}(i,j)$ iff $\mathsf{chain}^\sharp(i,j)$ or $z_i = [_p, z_j =]_p$, for some $1 \le p \le l$. We say that i and j are in the same **system** iff $\mathsf{syst}^\sharp(i,j)$, where $\mathsf{syst}^\sharp$ is the equivalence generated by syst.

For the Dyck loop in (3), we have the following classes of $\mathsf{chain}^\sharp$: $\{1\}$, $\{2,3\}$, $\{4,5,7\}$, $\{8\}$, $\{9\}$, $\{10\}$, $\{12\}$, $\{13,14\}$ and the classes of $\mathsf{syst}^\sharp$ (unions of the former ones): $\{1,2,3,12,13,14\}$, $\{4,5,7,8,9\}$, $\{10\}$.

We next group together all powers in the same chain. For (3) this means that we write it as

$$L = \{a^{2n_1} b(ab)^{n_1+3n_2} b^{n_3+5n_4+n_5} a^{4n_5} (aba)^{n_3} b^{3n_6} a^{n_7} b^{2n_7+n_2} \mid n_i \ge 0\}.$$

Now, the application of Lemma 9 gives a different result in the sense that the powers in the lemma are no longer single powers of our Dyck loop but linear combinations of those in the same chain. This means for (3) that any new tuple $(2n_1, n_1+3n_2, n_3+5n_4+n_5, 4n_5, n_3, 3n_6, n_7, 2n_7+n_2)$ (recall that we assume the elements large enough) gives a new word. Incidentally, for (3), we get such a new tuple exactly when the tuple $(n_i)_{1 \le i \le 7}$ is new, due to the fact that the three systems corresponding to the classes of $\mathsf{syst}^\sharp$, namely

$$\begin{cases} 2n_1 = 0 \\ n_1 + 3n_2 = 0 \\ n_7 = 0 \\ 2n_7 + n_2 = 0 \end{cases} \qquad \begin{cases} n_3 + 5n_4 + n_5 = 0 \\ 4n_5 = 0 \\ n_3 = 0 \end{cases} \qquad \begin{cases} 3n_6 = 0 \end{cases}$$

have only trivial solutions. Hoewever, in general, this might not be the case.

Consider then, in the general case, the systems associated with the equivalence classes of $\mathsf{syst}^\sharp$, as seen in the above example. (Notice that each system has as many equations as the number of classes of $\mathsf{chain}^\sharp$ in the respective class of $\mathsf{syst}^\sharp$.) As we noticed, if all such systems have trivial solutions only, then we have the same situation as in the case of no links. Thus, the same contradiction with the k-poly-slenderness of L is obtained. Therefore, there are systems which have also non-trivial solutions.

Consider one class of $\mathsf{syst}^\sharp$, say $[i_0]_{\mathsf{syst}^\sharp}$ such that its associated system, say

$$\sum_{j=1}^{t} a_{ij} n_j = 0, \text{ for all } 1 \le i \le s, \tag{4}$$

has also non-trivial solutions. Here s is the number of chains (classes of $\mathsf{chain}^\sharp$) in $[i_0]_{\mathsf{syst}^\sharp}$ and t is the number of elements of $[i_0]_{\mathsf{syst}^\sharp}$. Therefore, if the matrix of the system is $A = (a_{ij})_{\substack{1 \le i \le s \\ 1 \le j \le t}}$, we have $\operatorname{rank}(A) \le s \le t$ but $\operatorname{rank}(A) < t$. (In fact, $\operatorname{rank}(A) \in \{s-1, s\}$, but we shall not need this.)

But now, we can write D as a finite union of $(l-1)$-Dyck loops. Indeed, denote the columns of A by $A_i, 1 \le i \le t$. As either $s < t$ or else $\operatorname{rank}(A) < s = t$, there exists a non-trivial linear combination of A_i's with integer coefficients. As we are interested in the values of the vector $\sum_{i=1}^t n_i A_i$, it is clear that these values can be obtained as combinations of fewer A_i. But, one A_i less means exactly one pair of parentheses less in the underlying Dyck word z. That means, the l-Dyck loop we started with can be written as a finite union of $(l-1)$-Dyck loops just because $l > k+1$.

The above reasoning can be iterated, if needed, and therefore concludes the proof. ■

We next give some examples of the construction in the proof of Theorem 8. Consider the 5-Dyck loop

$$L_1 = \{(ab)^{2n_1} a(ba)^{3n_2} aab^{n_2} b^{2n_3} b^{n_4} aba^{n_4} a^{3n_3} a^{n_5} a^{2n_1} \mid n_i \ge 0\},$$

the underlying Dyck word of which is $z = [_1[_2]_2[_3[_4]_4]_3[_5]_5]_1$ (we assume $h(]_5) = \varepsilon$). There are links and we have the classe of $\mathsf{chain}^\sharp$: $\{1,2\}$, $\{3,4,5\}$, $\{6,7,8,10\}$ and only one class of $\mathsf{syst}^\sharp$: $\{1,2,\ldots,8,10\}$. We write L_1 as

$$L_1 = \{a(ba)^{2n_1+3n_2} aab^{n_2+2n_3+n_4} aba^{2n_1+3n_3+n_4+n_5} \mid n_i \ge 0\}$$

and we have the matrix of the system (for the unique class of $\mathsf{syst}^\sharp$)

$$A = \begin{bmatrix} 2 & 3 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 2 & 0 & 3 & 1 & 1 \end{bmatrix}$$

with $\operatorname{rank}(A) = 3 = s < t = 5$. We have the relation between the columns $A_i, 1 \le i \le 5$, of A: $A_3 = 2A_4 + A_5$ which allows us to write L_1 as

$$L_1 = \{a(ba)^{2n_1+3n_2} aab^{n_2+n_3} aba^{4n_1+n_3+n_4} \mid n_i \ge 0\}.$$

The pair of parentheses $[_4]_4$ disappeared. But L_1 can be still reduced as the new matrix A' has still more columns than rows. After noticing the linear combination $3A'_1 + 2A'_3 = 2A''_2 + 8A'_4$, we get

$$\begin{aligned} L_1 = \; & \{a(ba)^{2n_1+3n_2}aab^{n_2+n_4}aba^{4n_1+n_4+n_5} \mid n_i \geq 0\} \cup \\ & \{a(ba)^{3n_2}aab^{n_2+2n_3+n_4}aba^{3n_3+n_4+n_5} \mid n_i \geq 0\} \end{aligned}$$

which cannot be reduced further as $\#_{L_1}(n) = \Theta(n^2)$.

Consider then the 4-Dyck loop

$$L_2 = \{a^{n_1}a^{n_2}ba^{n_2}a^{n_3}ba^{n_3}a^{n_4}ba^{n_4}a^{n_1} \mid n_i \geq 0\}.$$

As above, we reduce it to

$$\begin{aligned} L_2 = \; & \{a^{n_1}a^{n_2}ba^{n_2}ba^{n_4}ba^{n_4}a^{n_1} \mid n_i \geq 0\} \cup \\ & \{a^{n_2}ba^{n_2}a^{n_3}ba^{n_3}a^{n_4}ba^{n_4} \mid n_i \geq 0\}. \end{aligned}$$

We would like to notice that the 3-Dyck loop

$$L_3 = \{a^{n_1}a^{n_2}ba^{n_2}a^{n_3}ba^{n_3}a^{n_1} \mid n_i \geq 0\},$$

which is actually very similar with L_2 cannot be reduced.

Consequences

The first of the consequences of Theorem 8 is the characterization of k-poly-slender regular languages. We say that the k-Dyck loop D in (1) is **degenerate** if, for any $1 \leq i \leq k$, at most one of the words u_i and v_i is non-empty.

We get then immediately from Theorem 8 the following result which has been proved in [16].

Corollary 10. *For any $k \geq 0$, a regular language is k-poly-slender iff it is a finite union of degenerate $(k+1)$-Dyck loops.*

Consider next, for a language L, the following smoothening of the complexity function $\#_L(n)$:

$$\overline{\#}_L(n) = \max_{1 \leq i \leq n} \#_L(i).$$

We have then from Theorem 8

Corollary 11. *For any poly-slender context-free language L, there exists a $k \geq 0$ such that $\overline{\#}_L(n) = \Theta(n^k)$.*

This means there is nothing in between integer powers, e.g., there is no context-free language the complexity of which is of order $\Theta(n^{3.4})$, $\Theta(\log n)$, $\Theta(n \log\log n)$, etc.

References

[1] Andraşiu, M., J. Dassow, Gh. Păun, and A. Salomaa, Language-theoretic problems arising from Richelieu cryptosystems, *Theoret. Comput. Sci.* **116** (1993) 339 – 357.

[2] Autebert, J.-M., J. Bestel, and L. Boasson, Context-free languages and push-down automata, in G. Rozenberg, A. Salomaa, eds., *Handbook of Formal Languages* (Springer-Verlag, Berlin, Heidelberg, 1997) 111 – 174.

[3] Berstel, J., Sur la densité asymptotique de langages formels, in M. Nivat, ed., *Automata, Languages, and Programming* (North-Holland, 1972) 345 – 358.

[4] Choffrut, C., and J. Karhumäki, Combinatorics of Words, in G. Rozenberg, A. Salomaa, eds., *Handbook of Formal Languages* (Springer-Verlag, Berlin, Heidelberg, 1997) 329 – 438.

[5] Ginsburg, S., E. H. Spanier, Bounded ALGOL-like languages, *Trans. Amer. Math. Soc.* **113** (1964) 333 – 368.

[6] Honkala, J., On Parikh slender languages and power series, *J. Comput. System Sci.* **52** (1996) 185 – 190.

[7] Honkala, J., Decision problems concerning thiness and slenderness of formal languages, *Acta Inform.* **35** (1998) 625 – 636.

[8] Ilie, L., On a conjecture about slender context-free languages, *Theoret. Comput. Sci.* **132** (1994) 427 – 434.

[9] Latteux, M., and G. Thierrin, On bounded context-free languages, *Elektron. Informtionsverarb. Kybernet.* **20** (1984) 3 – 8.

[10] Lothaire, M., *Combinatorics on Words* (Addison-Wesley, Reading, MA., 1983).

[11] Maurer, H. A., and M. Nivat, Rational bijection of rational sets, *Acta Inform.* **13** (1980) 365 – 378.

[12] Păun, Gh., and A. Salomaa, Thin and slender languages, *Discrete Appl. Math.* **61** (1995) 257 – 270.

[13] Raz, D., Length considerations in context-free languages, *Theoret. Comput. Sci.* **183** (1997) 21 – 32.

[14] Salomaa, A., *Formal Languages* (Academic Press, New York, 1973).

[15] Shallit, J., Numeration systems, linear recurrences, and regular sets, *Inform. and Comput.* **113** (1994) 331 – 347.

[16] Szilard, A., S. Yu, K. Zhang, and J. Shallit, Characterizing regular languages with polynomial densities, *Proc. of the 17th MFCS* (Prague 1992) Lecture Notes in Comput. Sci., 629, Springer, Berlin-New York, 1992, 494 – 503.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**