Discrete Mathematics and Theoretical Computer Science DMTCS vol. **15:1**, 2013, 29–36

Automaticity of Primitive Words and Irreducible Polynomials

Anne Lacroix[†] and Narad Rampersad[‡]

Department of Mathematics, University of Liège, Belgium

received 4th July 2011, revised 20th December 2012, accepted 22nd January 2013.

If L is a language, the automaticity function $A_L(n)$ (resp. $N_L(n)$) of L counts the number of states of a smallest deterministic (resp. non-deterministic) finite automaton that accepts a language that agrees with L on all inputs of length at most n. We provide bounds for the automaticity of the language of primitive words and the language of unbordered words over a k-letter alphabet. We also give a bound for the automaticity of the language of base-b representations of the irreducible polynomials over a finite field. This latter result is analogous to a result of Shallit concerning the base-k representations of the set of prime numbers.

Keywords: automaticity, primitive word, unbordered word, irreducible polynomial

1 Introduction

Automaticity is a measure of how close a non-regular language is to being regular. We can approximate a non-regular language L by considering a regular language L' such that the words of length at most n in L are exactly the words of length at most n in L'. The automaticity of L is the number of states of a smallest deterministic finite automaton accepting some approximation L'. Non-deterministic automaticity can be defined similarly. Automaticity was first introduced by Trakhtenbrot [18] and later by Karp [7]. Shallit and Breitbart [17] wrote a survey of the basic results concerning automaticity known at the time.

In the first part of this article we give bounds for the non-deterministic automaticity of the language of primitive words and the language of unbordered words. A word is primitive if it is not a power of a smaller word. A word is unbordered if it has no non-trivial period. The language of primitive words has been well-studied (see the survey by Lischke [9], for example). It is not difficult to show that the language of primitive words is not regular, but it is a long-standing open problem to show that this language is not context-free. It is also not difficult to show that the language of unbordered words is not regular. For a proof that this language is not context-free see [12].

In the second part we give a bound on the automaticity of the set of irreducible polynomials over a finite field. The set of base-k representations of the prime numbers is not a regular language for any base

[†]Email: `A.Lacroix@ulg.ac.be`
[‡]Email: `narad.rampersad@gmail.com`

k. Shallit [16] gave a lower bound on the automaticity of the set of prime numbers in any base. We consider the same problem in the setting of polynomials over a finite field. Given a fixed non-constant polynomial b, one can also define the base-b representation for such polynomials (see for example [14]). Rigo and Waxweiler [15] proved that the set of base-b representations of the irreducible polynomials is again a non-regular language for any base b. We obtain our bound for the automaticity using arguments similar to those of [16].

There is an interesting connection between primitive words and irreducible polynomials over a finite field. The number of primitive words of length n over an alphabet of size q is

$$\sum_{d|n} \mu(d) q^{n/d}, \tag{1}$$

where μ is the Möbius function (see [10, Section 1.3]). Similarly, the number of monic irreducible polynomials of degree n over the finite field with q elements is

$$\frac{1}{n} \sum_{d|n} \mu(d) q^{n/d}.$$

This is equal to the number of equivalence classes of primitive words of length n under the conjugacy relation $x \sim y$ if x is a cyclic shift of y. For an explicit bijection between the set of irreducible polynomials and the set of primitive necklaces, see [13, Section 7.6.2].

2 Definitions

Let $L \subseteq \Sigma^*$. A language L' is an *n-th order approximation* to L if

$$L' \cap \Sigma^{\leq n} = L \cap \Sigma^{\leq n}.$$

We define the *automaticity* $A_L(n)$ of a language L to be the number of states of a smallest DFA accepting some n-th order approximation to L. Similarly, the *nondeterministic automaticity* $N_L(n)$ of a language L is the number of states of a smallest NFA accepting some n-th order approximation to L.

Let $x, y \in \Sigma^*$. We say that x and y are *n-similar for L* if for all $z \in \Sigma^*$ with $|xz|, |yz| \leq n$, we have $xz \in L$ if and only if $yz \in L$. If x and y are not n-similar, then they are *n-dissimilar for L*.

Theorem 1 ([8]) *Let $L \subseteq \Sigma^*$. For all $n \geq 0$, $A_L(n)$ is the maximum possible cardinality of a set of pairwise n-dissimilar words for L.*

Example 2 *Let $L = \{0^n 1^n : n \geq 0\}$. Then $A_L(n) \geq m+1$ for $n = 2m, 2m+1$, since $\{\varepsilon, 0, 00, \ldots, 0^m\}$ is a set of pairwise $2m$-dissimilar words for L. To see this, consider 0^j and 0^k for $0 \leq j < k \leq n$. Then $0^j 1^j \in L$ and $0^k 1^j \notin L$.*

Let U be a finite set of words. We say that U is a set of *uniformly n-dissimilar words for L* if for each $x \in U$ there exists z such that

- $|xz| \leq n$ and $xz \in L$; and
- for each $y \in U$ such that $x \neq y$, we have $|yz| \leq n$ and $yz \notin L$.

Theorem 3 ([3]) *Let $L \subseteq \Sigma^*$ and let U be a set of uniformly n-dissimilar words for L. Then $N_L(n) \geq |U|$.*

3 Automaticity of primitive and unbordered words

Let $k \geq 2$ be an integer. A word y is a *k-power* if y can be written as $y = x^k$ for some non-empty word x. If y cannot be so written for any $k \geq 2$, then y is *primitive*.

Bordered words are generalizations of powers. We say a word x is *bordered* if there exist words $u, v, w \in \Sigma^+$ such that $x = uv = wu$. In this case, the word u is said to be a *border* for x. Otherwise, x is *unbordered*.

Let $w = w_0 \cdots w_{\ell-1}$ and let $p < \ell$. The word w has a *period* p if $w_i = w_{i+p}$ for all $0 \leq i \leq \ell - p - 1$. Note that a word is unbordered if it has no period.

We recall the notation $O(\cdot)$ and $\Omega(\cdot)$. Let f and g be functions from $\mathbb{N}$ to $\mathbb{R}$. The function f is $O(g)$ if there exist $C > 0$ and n_0 such that for all $n > n_0$ we have $f(n) \leq C \cdot g(n)$. The function f is $\Omega(g)$ if there exist $C > 0$ and n_0 such that for all $n > n_0$ we have $f(n) \geq C \cdot g(n)$.

Theorem 4 *Let $\varepsilon > 0$ be a real number. The nondeterministic automaticity of the set Q_k of primitive words over the alphabet $\Sigma_k = \{0, \ldots, k-1\}$ is*

$$N_{Q_k}(n) \geq \beta_k k^{\lfloor n/2 \rfloor} + O((k+\varepsilon)^{n/4}),$$

where β_k is a constant that depends only on k.

Proof: For $n \geq 0$, we will define a set of uniformly n-dissimilar words as follows. Let

$$D_n = \{w \in \Sigma_k^{\lfloor n/2 \rfloor} : w \text{ is unbordered } \}.$$

To show that the words in D_n are uniformly n-dissimilar, let $x, y \in D_n$, $x \neq y$. Observe that xx is not primitive, but yx is, for if yx were not primitive, then $yx = z^\ell$ for some word z and some $\ell \geq 3$. In this case $|yx|/\ell$ is a period of y, and hence y is bordered, which contradicts the definition of D_n.

Guibas and Odlyzko [4, Theorem 7.2] gave the following formula for the size of D_n (see also [11]): there exists a constant β_k such that

$$|D_n| = \beta_k k^{\lfloor n/2 \rfloor} + O((k+\varepsilon)^{n/4}).$$

By Theorem 3 we have $N_{Q_k}(n) \geq |D_n|$, which is the desired result. $\square$

Note that Guibas and Odlyzko gave an explicit formula for the β_k, which permits one to calculate β_k to any desired degree of accuracy. For example, if $k = 2$, we have $\beta_2 = 0.26771654 \cdots$.

Next we give an upper bound on the nondeterministic automaticity of Q_k.

Theorem 5 *For $n \geq 0$,*

$$N_{Q_k}(n) \leq \left(\frac{2k^{3/2}}{(k-1)(\sqrt{k}-1)} \right) k^{n/2} + n^3 k^{n/3}$$

Proof: For each $n \geq 0$ we construct a deterministic automaton that accepts all words of length at most n in the complement of Q_k. The automaton is constructed as follows. First consider the language of square words (2-powers) of length i. We can construct an automaton accepting this language by first constructing the complete k-ary tree with $k^{i/2}$ leaves so that each path from the root to a leaf is labeled by a different

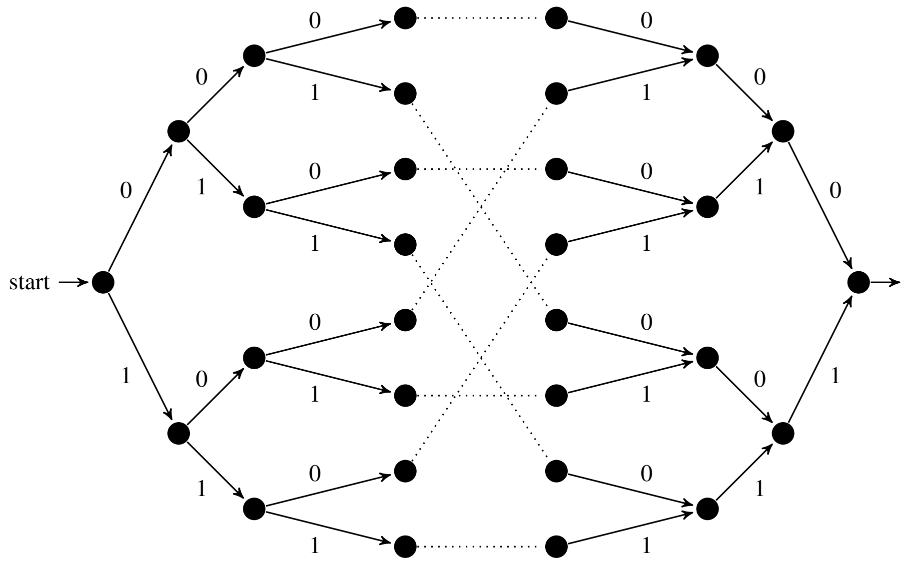

Fig. 1: An automaton accepting binary squares of length 6

word of length $i/2$. We then make a copy of this tree, but reflected, so that the arrows are directed away from the leaves towards the root of the tree. The leaves of the first tree are identified with the leaves of the second tree. This construction is illustrated in Figure 1, which shows the automaton accepting all binary squares of length 6. In the figure, dotted lines connect states to be identified, and transitions not shown go to a sink state.

The left tree has

$$\frac{k^{(i/2)+1}-1}{k-1}$$

states, so the automaton for the squares of length i has at most

$$2\left(\frac{k^{(i/2)+1}-1}{k-1}\right)$$

states.

For each $d > 2$, to accept d-powers of length i we simply construct a tree with $k^{i/d}$ leaves so that each path from the root to a leaf is labeled by a different d-power of length i. This tree has at most $ik^{i/d}$ states.

To create the automaton accepting all non-primitive words of length i, we can combine all of these automata, sharing edges and transitions whenever possible. The resulting automaton has at most

$$\begin{aligned}
& 2\left(\frac{k^{(i/2)+1}-1}{k-1}\right)+\sum_{\substack{d|i\\ d>2}} ik^{i/d} \\
\leq\ & \left(\frac{2k}{k-1}\right)k^{i/2}+\sum_{\substack{d|i\\ d>2}} ik^{i/d} \\
\leq\ & \left(\frac{2k}{k-1}\right)k^{i/2}+\sum_{\substack{d|i\\ d>2}} ik^{i/3} \\
\leq\ & \left(\frac{2k}{k-1}\right)k^{i/2}+i^2k^{i/3}
\end{aligned}$$

states.

We can therefore can construct an automaton accepting all non-primitive words of length at most n using at most

$$\begin{aligned}
& \sum_{i=1}^{n}\left(\left(\frac{2k}{k-1}\right)k^{i/2}+i^2k^{i/3}\right) \\
\leq\ & \frac{2k}{k-1}\sum_{i=1}^{n}k^{i/2}+\sum_{i=1}^{n}i^2k^{i/3} \\
\leq\ & \frac{2k}{k-1}\left(\frac{k^{(n+1)/2}-\sqrt{k}}{\sqrt{k}-1}\right)+\sum_{i=1}^{n}i^2k^{i/3} \\
\leq\ & \frac{2k}{k-1}\left(\frac{k^{(n+1)/2}-\sqrt{k}}{\sqrt{k}-1}\right)+n^3k^{n/3} \\
\leq\ & \left(\frac{2k^{3/2}}{(k-1)(\sqrt{k}-1)}\right)k^{n/2}+n^3k^{n/3}
\end{aligned}$$

states. Since this automaton is deterministic, the automaton accepting Q_k has at most this many states as well. □

Next we consider the language of unbordered words.

Theorem 6 *Let $\varepsilon > 0$. The nondeterministic automaticity of the set UB$_k$ of unbordered words over the alphabet Σ_k is*

$$N_{UB_k}(n) = \Omega((k-\varepsilon)^{n/2}).$$

Proof: For each $\varepsilon > 0$ there exists j such that the number of words of length m over a k-letter alphabet that avoid 1^j is $\Omega((k-\varepsilon)^m)$ (see for example the analysis given in the section "Longest runs" starting on p. 308 of [2]). Fix such a j. For $n \geq 2(j+2)$ we define

$$D_n = \{0w01^j : |w| = \lfloor n/2 \rfloor - (j+2) \text{ and } w \text{ does not contain } 1^j\}.$$

To show that the words in D_n are uniformly n-dissimilar, let $x, y \in D_n$, $x \neq y$. Since $x, y \in D_n$, there exist w_1 and w_2 such that

$$x = 0w_1 01^j \text{ and } y = 0w_2 01^j.$$

Clearly xx is bordered; however, xy is not bordered. Suppose to the contrary that xy has a border b. Since b is a non-empty prefix of xy, it must begin with 0; since it is also a suffix, it must end with 1^j. However xy contains only one occurrence of 1^j apart from the occurrence at the end. It follows that $b = x$, and since b is also a suffix of xy and $|x| = |y|$, we must also have $b = y$. Thus $x = y$, which is a contradiction. Since $|D_n| = \Omega((k-\varepsilon)^{n/2-(j+2)}) = \Omega((k-\varepsilon)^{n/2})$, we have the result by Theorem 3. □

4 Irreducible polynomials

In this section we consider the automaticity of the language of representations of irreducible polynomials over a finite field with respect to some base b.

Let $\mathbb{F}$ be a field with q elements. Let $\mathbb{F}[X]$ be the polynomial ring over $\mathbb{F}$. If $f \in \mathbb{F}[X]$ we denote its degree by $\deg f$. Let B be an integer and let $\mathbb{F}[X]_{<B}$ denote the set of polynomials over $\mathbb{F}$ of degree strictly less than B. If b is a fixed non-constant polynomial, then any polynomial f can be written uniquely as

$$f = \sum_{i=0}^{\ell} c_i b^{\ell-i}, \quad c_0 \neq 0,$$

where each c_i has degree less than $\deg b$.

We define a function $\Psi : \mathbb{F}[X]_{<B} \to \mathbb{F}^B$ by

$$\Psi(f) := (\underbrace{0, \ldots, 0}_{B-N-1}, F_0, \ldots, F_N)$$

if $f = F_0 X^N + \cdots + F_N$. The word $[f]_b := \Psi(c_\ell)\Psi(c_{\ell-1}) \cdots \Psi(c_0)$ over the alphabet $\mathbb{F}^B$ is the b-representation of f. By convention, the representation of the zero polynomial is ε. Given a b-representation $w \in (\mathbb{F}^B)^*$, we denote its value in $\mathbb{F}[X]$ by $\langle w \rangle_b$. Note that we have chosen to write f starting with the least significant "digit" and ending with the most significant "digit".

A set $\mathcal{T} \subseteq \mathbb{F}[X]$ is *b-recognizable* if the language

$$[\mathcal{T}]_b = \{[f]_b : f \in \mathcal{T}\} \subseteq (\mathbb{F}^B)^*$$

is regular. Rigo and Waxweiler [15] proved that for any base b, the set of irreducible polynomials over $\mathbb{F}$ is not b-recognizable.

Let $\mathcal{T} \subset \mathbb{F}[X]$ and let b be a non-constant polynomial. The *b-automaticity* of $\mathcal{T}$ is denoted by $A^b_{\mathcal{T}}(n)$ and is defined as the automaticity $A^b_L(n)$ of the language $L = \{[f]_b : f \in \mathcal{T}\}$.

Theorem 7 *There exists a constant B such that the set $\mathcal{S}$ of monic irreducible polynomials over $\mathbb{F}$ has b-automaticity $A^b_{\mathcal{S}}(n) \geq q^{Bn}/Bn + O(q^{Bn/2}/Bn)$.*

The main tool for the proof of this theorem is the following result of Hsu [6, Corollary 3.4].

Theorem 8 *Let a and m be polynomials over $\mathbb{F}$ such that $(a, m) = 1$. Let $\#S_N(a, m)$ denote the number of monic irreducible polynomials of degree N congruent to a modulo m and let $M = \deg m$. If $q^{N/2} > (3 + M)q^M$ then*

$$\#S_N(a, m) \geq 1.$$

The proof of the following lemma is similar to that of [16, Lemma 6], which is in turn based on an idea found in [5] and [1].

Lemma 9 *Let $d \in \mathbb{F}[X]$ such that $\deg d > 0$ and let $f, g \in \mathbb{F}[X]_{<\deg d}$ such that $f \neq g$ and $(d, f) = (d, g) = 1$. Then there exists a constant C_q and a polynomial h such that $hd + f$ is irreducible and $hd + g$ is not irreducible, where $\deg h \leq C_q \deg d$.*

Proof: By Theorem 8 there exist a constant C and a polynomial h_0 such that $r = h_0 d + f$ is irreducible, where $\deg h_0 \leq C \deg d$. If $s = h_0 d + g$ is reducible, we are done. Otherwise, since $(sd + r, sd) = 1$, by Theorem 8 again, there exists h_1 such that $h_1(sd) + (sd + r)$ is irreducible, where $\deg h_1 \leq C(\deg sd) = C(\deg s + \deg d)$. However, $h_1(sd) + (sd + s)$ is a multiple of s and hence is reducible. Therefore we set $h = s(h_1 + 1) + h_0$. Furthermore, we have

$$\begin{aligned}
\deg h &\leq \max\{\deg h_1 + \deg s, \deg h_0\} \\
&\leq \max\{C(2 \deg s + \deg d), C \deg d\} \\
&\leq C(2 \deg s + \deg d) \\
&\leq C(2(C \deg d) + 2(\deg d)) \\
&\leq 2C(C + 1)(\deg d).
\end{aligned}$$

We may thus take $C_q = 2C(C + 1)$ to complete the proof. □

Proof of Theorem 7: To prove Theorem 7 we will contruct a set D_n of n-dissimilar words for $[\mathcal{S}]_b$. Let C_q be as in Lemma 9. Let

$$D_n = \{[f]_b : f \in \mathcal{S}, (f, b) = 1, \deg f = (n \deg b)/(1 + C_q)\}.$$

Note that all words in D_n have the same length. Consider two elements $x, y \in D_n$. Let $f = \langle x \rangle_b$, $g = \langle y \rangle_b$. By Lemma 9, there exists h such that $hb^{|x|} + f$ is irreducible and $hb^{|x|} + g$ is not, where $\deg h = C_q \deg b^{|x|}$. Let $z = [h]_b$. Then $xz \in [\mathcal{S}]_b$ and $yz \notin [\mathcal{S}]_b$. Since $\deg h = C_q \deg b^{|x|}$, we have $|z| \deg b = C_q |x| \deg b$, and so $|z| = C_q |x|$. Since $\deg f = (n \deg b)/(1+C_q)$, we have $|x| = n/(1+C_q)$. Hence $|xz| = n/(1 + C_q) + C_q |x| = n/(1 + C_q) + C_q(n/(1 + C_q)) = n$.

We now estimate the size of D_n. Let $B = \deg b/(1+C_q)$. Note that there are $q^{Bn}/Bn + O(q^{Bn/2}/Bn)$ monic irreducible polynomials in $F[X]$ of degree Bn. Since $\deg b$ is a constant, there are at most a constant number of polynomials f that divide b. Hence $|D_n| = q^{Bn}/Bn + O(q^{Bn/2}/Bn)$. □

Acknowledgments

We thank Michel Rigo for suggesting to us the problem of the automaticity of the irreducible polynomials and for his helpful comments. We also thank Jeffrey Shallit for interesting discussions on the subject of automaticity.

References

[1] D. Allen Jr., On a characterization of the nonregular set of primes, J. Comput. System Sci. 2 (1968), 464–467.

[2] P. Flajolet, R. Sedgewick, Analytic Combinatorics, Cambridge Univ. Press 2009.

[3] I. Glaister, J. Shallit, Automaticity III: Polynomial automaticity and context-free languages, Comput. Complexity 7 (1998), 371–387.

[4] L. J. Guibas, A. M. Odlyzko, Periods in strings, J. Combin. Theory Ser. A 30 (1981), 19–42.

[5] J. Hartmanis, H. Shank, On the recognition of primes by automata, J. Assoc. Comput. Mach 15 (1968), 382–389.

[6] C.-N. Hsu, The distribution of irreducible polynomials in $\mathbb{F}_q[t]$, J. Number Theory 61 (1996), 85–96.

[7] R. M. Karp, Some bounds on the storage requirements of sequential machines and Turing machines, J. Assoc. Comput. Mach. 14 (1967), 478–489.

[8] J. Kaneps, R. Freivalds, Minimal nontrivial space complexity of probabilistic one-way Turing machines. In Proc. MFCS 1990, LNCS 452, pp. 355–361, Springer, 1990.

[9] G. Lischke, Primitive words and roots of words, Acta Univ. Sapientiae, Informatica, 3 (2011), 5–34.

[10] M. Lothaire, Combinatorics on Words, Cambridge Univ. Press 1983.

[11] P. Tolstrup Nielsen, A note on bifix-free sequences, IEEE Trans. Inform. Theory IT-19 (1973), 704–705.

[12] N. Rampersad, J. Shallit, M.-w. Wang, Inverse star, borders, and palstars, Inform. Process. Lett. 111 (2011), 420–422.

[13] C. Reutenauer, Free Lie Algebras, Oxford, 1993.

[14] M. Rigo, Syntactical and automatic properties of sets of polynomials over finite fields, Finite Fields Appl. 14 (2008), 258–276.

[15] M. Rigo, L. Waxweiler, Logical characterization of recognizable sets of polynomials over a finite field. To appear in Internat. J. Found. Comput. Sci.

[16] J. Shallit, Automaticity IV: Sequences, sets, and diversity, J. Théor. Nombres Bordeaux 8 (1996), 347–367.

[17] J. Shallit, Y. Breitbart, Automaticity I: Properties of a measure of descriptional complexity, J. Comput. System Sci. 53 (1996), 10-25.

[18] B. A. Trakhtenbrot, On an estimate for the weight of a finite tree, Sibirskii Matematicheskii Zhurnal 5 (1964), 186–191. In Russian.