International Journal of Foundations of Computer Science
Vol. 23, No. 8 (2012) 1711–1727

DOI: 10.1142/S0129054112400710

FROM REGULAR TO STRICTLY LOCALLY TESTABLE LANGUAGES*

STEFANO CRESPI REGHIZZI† and PIERLUIGI SAN PIETRO‡
Dipartimento di Elettronica e Informazione
Politecnico di Milano, P. Leonardo da Vinci, 32
Milano 20133, Italia
†*crespi@elet.polimi.it*
‡*sanpietro@elet.polimi.it*

Received 23 December 2011
Accepted 30 May 2012
Communicated by Štěpán Holub

A classical result (often credited to Y. Medvedev) states that every language recognized by a finite automaton is the homomorphic image of a local language, over a much larger so-called local alphabet, namely the alphabet of the edges of the transition graph. Local languages are characterized by the value $k = 2$ of the sliding window width in the McNaughton and Papert's infinite hierarchy of strictly locally testable languages (k-slt). We generalize Medvedev's result in a new direction, studying the relationship between the width and the alphabetic ratio telling how much larger the local alphabet is. We prove that every regular language is the image of a k-slt language on an alphabet of doubled size, where the width logarithmically depends on the automaton size, and we exhibit regular languages for which any smaller alphabetic ratio is insufficient. More generally, we express the trade-off between alphabetic ratio and width as a mathematical relation derived from a careful encoding of the states. At last we mention directions for theoretical development and application.

Keywords: Regular languages; strictly locally testable languages; homomorphic characterization; Medvedev Theorem; factor decodability.

1. Introduction

A classical result [15], often credited to Y. Medvedev [14], states that every regular language is the homomorphic image of a local language over a larger alphabet called *local*. In a local language the sentences are characterized by three sets: the initial letters, the final letters and the set of factors of length $k = 2$. Parameter k is the *width* of the simplest sliding window device introduced by McNaughton and Papert [13]. The result simply derives from the fact that the set of paths in an edge-labelled graph is a local language over the alphabet of the edges. Considering

*A preliminary version of this paper has appeared in WORDS 2011.

a finite automaton for the regular language, the local language of accepting paths can be naturally projected on the original language.

Our work originates from two observations. First, in the classic result the alphabet of the local language is larger than the source alphabet, by a multiplicative factor, called the *alphabetic ratio*, in the order of the square of the number of states. The simplicity of sliding window machines is very attractive, but the huge size of the local alphabet in Medvedev theorem makes their application impractical.

Then a natural question concerns the local alphabet in the classical result: how small can the alphabetic ratio be? A small alphabet may, for instance, allow to encode messages from a regular language into an slt language, to be transmitted over a communication channel, so that a more economical sliding window receiver can be used instead of a general finite state machine.

Second, the local languages are a member of McNaughton and Papert's [13] infinite hierarchy of *k-strictly locally testable*, for short k-slt, languages. Then, by considering k-slt, instead of just 2-slt (i.e., local languages), we raise a more general question: what is the minimum alphabetic ratio such that, for some finite parameter k, every regular language is the alphabetic homomorphism of a k-slt language? In that case, how big does the width parameter k need to be? More precisely, our main result, which generalizes Medvedev theorem, expresses the trade-off between two parameters: the alphabetic ratio and the width. We also spend a few lines to show the early but enduring interest for subfamilies of regular languages characterized by some form of local testability, without entering into details.

At the basis of formal language theory, the classical theorem of N. Chomsky and M.P. Schutzenberger characterizes context-free languages by a homomorphism applied to the intersection of a Dyck language and a 2-slt one. Several similar characterizations for other language families have later been proved. In mathematics, the slt languages have been applied in the theory of semigroups by A. de Luca and A. Restivo [1]. In linguistics, a persistent idea is that natural languages can be modeled, at various levels, by locally testable properties. For instance, the psychologist W. Wickelgren [16] made the observation that the set of English words are essentially a 3-slt (finite) language, and several brain scientists (in particular V. Braitenberg [4]) have suggested that sequences of finite length, such as the factors occurring in a locally testable language, can be easily stored and recognized by certain neural circuits (in particular the synfire chains of M. Abeles) that have been observed in the cortex. In computational linguistics locally testable definitions have proved to be useful at various levels of finite-state models. Many persons (e.g. [9]) working on language learning models have been attracted by the efficiency of learning algorithms for various types of locally testable languages. Contemporary comparative work on the aural pattern recognition cababilities of humans and animals [12] have called attention to the subregular hierarchies induced by local testability. In mathematical biology, in his seminal article on language theory and DNA [11], T. Head shows that certain splicing languages are precisely the slt languages.

The paper is organized as follows. After the basic definitions in Sect. 2, we introduce in Sect. 3 a new classification of regular languages based on their homomorphic characterization via a k-slt language over an alphabet of size m. In Sect. 3 we prove a lower bound on the alphabetic ratio. In Sect. 4 we prove the generalization of Medvedev theorem, including a mathematical analysis of the relationship between language complexity, alphabetic ratio, and width. The Conclusion presents an open problem and mentions conceivable developments and applications of the main result.

2. Preliminaries

The empty word is denoted by ε. Let A be a finite nonempty set, the *terminal alphabet*. For simplicity we deal only with languages in A^+, which do not contain the empty word. A nondeterministic finite automaton (NFA) M is a quintuple $M = (Q, A, E, q_0, F)$ where Q is a finite set of states, the transition relation (or graph) is $E \subseteq Q \times A \times Q$, $q_0 \in Q$ is the initial state; $F \subseteq Q$ is the set of final states, which does not contain q_0 (since only ε-free languages are considered). Therefore, in the following only NFA's with $|Q| > 1$ are considered. As usual, a NFA is called *deterministic* if for all $a \in A, p, p', q \in Q$ if both $(q, a, p) \in E$ and $(q, a, p') \in E$ then $p = p'$. Two transitions $(p, a, q), (p', a', q') \in E$ are *consecutive* if $q = p'$. A *path* $\eta = e_0 e_1 \dots e_{n-1}$ is a finite sequence of $n > 0$ consecutive transitions $e_0 = (p_0, a_0, p_1)$, $e_1 = (p_1, a_1, p_2), \dots, e_{n-1} = (p_{n-1}, a_{n-1}, p_n)$. The *origin* of η is $o(\eta) = p_0$, its *end* is $e(\eta) = p_n$, and its *label* is $l(\eta) = a_0 a_1 \dots a_{n-1}$. Paths $\eta_1, \eta_2, \dots, \eta_k$ of M, for $k \geq 2$, are called *consecutive* if $\eta_1 \eta_2 \dots \eta_k$ is also a path of M (i.e, $e(\eta_h) = o(\eta_{h+1})$, for all $1 \leq h \leq k-1$). A *successful* path is a path with origin q_0 and end in F. The language recognized by M, denoted $L(M)$, is the set of labels of all successful paths of M. We assume, without loss of generality, that the transition relation is *total*, i.e., for every $q \in Q, a \in A$, the set $\{p \in Q \mid (q, a, p) \in E\}$ is not empty (in case E is not total, just add a new sink state to Q). An automaton M is called *minimal* if no other automaton recognizing the same language has fewer states than M.

Given another finite alphabet B, an (*alphabetic*) *homomorphism* is a mapping $\pi : B \to A$. For a language $L' \subseteq B^+$, its (*homomorphic*) *image* under π is the language $L = \{\pi(x) \mid x \in L'\}$.

For every word $w \in A^+$, for every $k \geq 2$, let $i_k(w)$ and $t_k(w)$ denote the prefix and, respectively, the suffix of w of length k if $|w| \geq k$, or w itself if $|w| < k$. Let $f_k(w)$ denote the set of factors of w of length k. Extend i_k, t_k, f_k to languages as usual, i.e., $i_k(L) = \{i_k(w) \mid w \in L\}$, $t_k(L) = \{t_k(w) \mid w \in L\}$, and $f_k(L) = \bigcup_{w \in L} f_k(w)$. A factor of a word w starting at position h and ending at position h, with $1 \leq h, k \leq |w|$, is defined as follows:

$$\begin{cases} s_{h,k}(w) = \varepsilon & \text{if } \ k < h \\ s_{h,k}(w) = i_{k-h+1}\left(t_{|w|-h+1}(w)\right) & \text{otherwise}\,. \end{cases}$$

Hence, for $k \geq h$, $|s_{h,k}(w)| = k - h + 1$.

The next definition is equivalent to the one in [1, 3].

Definition 1. *A language L is* k-strictly locally testable, *shortly* k-slt, *if there exist finite sets $I_{k-1}, T_{k-1} \subseteq A^{k-1}$ and $F_k \subseteq A^k$ such that, for every $x \in A^k A^*$, the following condition holds:* $x \in L \iff i_{k-1}(x) \in I_{k-1} \wedge t_{k-1}(x) \in T_{k-1} \wedge f_k(x) \subseteq F_k$.

Integer k is called the *width* of L. The definition ignores words shorter than $k-1$, which can be checked directly against a finite set, if needed. The case $k = 2$ yields the well known family of *local languages* (see for instance [3] or [15]).

A language is *strictly locally testable* (slt) if it is k-slt for some k. The original name in [13] was "testable in the strict sense". This concept should not be confused with other language families based on local tests, see [5] for a recent account. The following example will be referred to later.

Example 2. *Language $L' = (a'a)^+ \cup (b'b)^+$ is 2-slt, i.e., local, since it can be defined by the sets $I_1 = \{a', b'\}$, $T_1 = \{a, b\}$, $F_2 = \{aa', a'a, bb', b'b\}$.*

It is known and straightforward to prove that the family of slt languages is strictly included in the family of regular languages, and it is an infinite strict hierarchy ordered by the width value. For instance, language $L_h = (ab^h)^+$ on $A = \{a, b\}$, with $h > 1$ a constant, is $(h+1)$-slt, but it is not h-slt. In fact, L_h is defined by the sets: $I_h = \{ab^{h-1}\}$, $T_h = \{b^h\}$, $F_{h+1} = \{b^i ab^{h-i} \mid 0 \le i \le h\}$. However, L_h is not h-slt: consider the words $ab^h \in L_h$ and $ab^{h+1} \notin L_h$: $i_{h-1}(ab^h) = i_{h-1}(ab^{h+1}) = ab^{h-2}$, $t_{h-1}(ab^h) = t_{h-1}(ab^{h+1}) = b^{h-1}$, $f_h(ab^h) = \{ab^{h-1}, b^h\} = f_h(ab^{h+1})$. Hence, the two words cannot be distinguished by using width h.

3. Lower Bounds

As said, every regular language, to be also named the *source*, is the image of a 2-slt language whose alphabet may be much larger than the one of the source. To talk precisely about the width of the slt language and of the ratio of the alphabet sizes, we introduce our central definition.

Definition 3. *For $k \ge 2, m \ge 1$, a language $L \subseteq A^+$ is* (m,k)-homomorphic *if there exist an alphabet B (called* local*) of cardinality m, a k-slt language $L' \subseteq B^+$, and a homomorphism $\pi : B \to A$ such that $L = \pi(L')$.*

Clearly, if $L \subseteq A^+$ is k-slt then it is trivially $(|A|, k)$-homomorphic. Otherwise, a local alphabet larger than A is needed. For instance, the language $L = (aa)^+ \cup (bb)^+$ is not slt but the language $L' = (a'a)^+ \cup (b'b)^+$ of Ex. 2 is 2-slt. By defining $\pi : \{a, a', b, b'\} \to \{a, b\}$ as $\pi(a) = \pi(a') = a$, $\pi(b) = \pi(b') = b$, then $L = \pi(L')$ and hence L is $(4, 2)$-homomorphic. The alphabetic ratio of L' and L is $4/2 = 2$.

The traditional construction (e.g. in [15]) of a 2-slt language L' starts from an NFA (Q, A, E, q_0, F) of size $n = |Q|$ for L, and uses set E as local alphabet, i.e., up to $n^2 \cdot |A|$ elements. Hence we can restate Medvedev's property saying that every regular language on A is $(n^2 \cdot |A|, 2)$-homomorphic (the alphabetic ratio is n^2). However, it is straightforward to show that the cardinality of the local alphabet can be reduced to $n \cdot |A|$.

Proposition 4. *Every regular language, accepted by an NFA with n states, is $(n \cdot |A|, 2)$-homomorphic.*

Proof. Let $M = (Q, A, E, q_0, F)$ be an NFA with n states, $n > 1$. Define two mappings $\pi : Q \times A \to A$ and $\rho : Q \times A \times Q \to Q \times A$ such that $\pi(\langle q, a\rangle) = a$, for every $a \in A$, $q \in Q$ and $\rho(p, a, q) = \langle p, a\rangle$ for every $p, q \in Q, a \in A$. The following sets define a 2-slt language $L' \subseteq (Q \times A)^+$:

$$I_1 = \{\langle q_0, a\rangle \mid a \in A\}; \qquad T_1 = \{\langle q, a\rangle \mid a \in A, \exists q' \in F : (q, a, q') \in E\};$$
$$F_2 = \{\langle q, a\rangle\langle q', b\rangle \mid a, b \in A, q, q' \in Q, (q, a, q') \in E\}.$$

We show first that $\pi(L') \subseteq L$. Let $w \in \pi(L')$. Hence, there exists $x \in L'$ such that $\pi(x) = w$. We claim that there exists a successful path η of M such that $x = \rho(\eta)$. Let $n = |w|$. Since $x \in L'$, there exist $q_1, q_2, \ldots q_{n-1} \in Q$, $a_0, a_1, \ldots, a_{n-1} \in A$ such that $x = \langle q_0, a_0\rangle\langle q_1, a_1\rangle \ldots \langle q_{n-1}, a_n\rangle$, and $w = a_0 a_1 \ldots a_{n-1}$. Since $\langle q_{n-1}, a_n\rangle \in T_1$, there exists $q \in F$ such that $(q_{n-1}, a_{n-1}, q) \in E$. Let η be (q_0, a_0, q_1) (q_1, a_1, q_2) $\ldots (q_{n-1}, a_{n-1}, q)$: η has label w, origin in q_0 and end in a final state; moreover, $\rho(\eta) = x$. By definition of F_2, every factor $\langle q_{i-1}, a_i\rangle\langle q_i, a_{i+1}\rangle$ of x, for $1 \leq i \leq n$, must be such that $(q_{i-1}, a_i, q_i) \in E$, hence all transitions of η are consecutive, i.e., η is a successful path of label w.

We show that $L \subseteq \pi(L')$. Let $w \in L$ be the label of a path η of M of the form

$$(q_0, a_0, q_1)(q_1, a_1, q_2) \ldots (q_{n-1}, a_{n-1}, q_n), \text{ with } q_n \in F \text{ and } a_0 \ldots a_{n-1} = w.$$

We claim that $\rho(\eta) \in L'$. In fact, $i_1(\rho(\eta)) = \langle q_0, a_0\rangle \in I_1$, $t_1(\rho(\eta)) = \langle q_{n-1}, a_{n-1}\rangle \in T_1$ and $f_2(\rho(\eta)) = \{\langle q_{i-1}, a_{i-1}\rangle\langle q_i, a_i\rangle \mid 1 \leq i \leq n\}$. Since each triple $(q_{i-1}, a_{i-1}, q_i) \in E$ is also a transition of path η, $f_2(\rho(\eta)) \subseteq F_2$. □

It is trivial to observe that the minimal DFA recognizing L' has a graph isomorphic to E, if edge labels are disregarded.

A natural question, to be later addressed, is whether, by allowing the width k to be larger than 2, it is possible to reduce the cardinality of the local alphabet to less than $n \cdot |A|$. Next we prove the simple, but perhaps unexpected result, that in general the local alphabet cannot be smaller than twice the size of the source one.

Theorem 5. *For every alphabet A, there exists a regular language $L \subseteq A^+$ such that for every $k \geq 2$ L is not $(2 \cdot |A| - 1, k)$-homomorphic.*

Proof. Let L be defined by the regular expression $\bigcup_{a \in A}(aa)^*$. By contradiction, assume that there exist $k \geq 2$ and a local alphabet B of cardinality $2|A| - 1$, a mapping $\pi : B \to A$ and a k-slt language $L' \subseteq B^+$ such that $\pi(L') = L$. Since $|B| = 2 \cdot |A| - 1$, there exists at least one symbol of A, say, a, such that there is only one symbol $b \in B$ with $\pi(b) = a$. Since the word $a^{2k} \in L$, there exists $x \in L'$ such that $\pi(x) = a^{2k}$. By definition of π and of B, $x = b^{2k}$. Consider the word $xb = b^{2k+1}$. Clearly, $\pi(xb) = a^{2k+1}$, which is not in L, since all words in L have even

length. Hence, $xb \notin L'$. But $i_{k-1}(x) = i_{k-1}(xb) = b^{k-1}$, $t_{k-1}(x) = t_{k-1}(xb) = b^{k-1}$, $f_k(x) = f_k(xb) = b^k$ and, by Def. 1, xb is in L', a contradiction. □

In other words, this negative property says that some regular languages cannot be generated as images of slt languages, if the alphabetic ratio is too small. The same result holds (with a very similar proof) if in the statement the class of *strictly* locally testable languages is replaced by the class of *locally testable* languages [13]. The question whether an alphabetic ratio of two is sufficient, is addressed in the next section, while in the remainder of this section we derive a lower bound, given a regular language R, on the width k of a slt language generating R.

Proposition 6. *Let $R \subseteq A^*$ be the homomorphic image of a k-slt language $L \subseteq B^*$, with $k, |B| > 1$. Then R is accepted by a NFA with at most $\frac{|B|^k-1}{|B|-1}$ states.*

Proof. Let M be a minimal DFA accepting L. Following [5], it is possible to define an equivalent DFA such that for all $w \in B^{k-1}$, the set of end states of paths labelled w contains at most one element. Hence, $|Q|$ is at most equal to the number of possible paths of M having origin q_0 and length up to $k-1$, which is given by the expression $1 + |B| + |B|^2 + \cdots + |B|^{k-1}$, a geometric series of sum $\frac{|B|^k-1}{|B|-1}$.

To prove the original statement, given an alphabet A, $|B| \geq |A|$, and a homomorphism $\pi : B \to A$, modify M into a NFA accepting $\pi(L)$ instead of L, by replacing, for every $b \in B$, each label b on every edge with $\pi(b)$. □

Theorem 7. *For every alphabet A, for every $h, k \geq 2$ and for every $n > h \cdot |A|$, if language $R \subseteq A^+$ is $(h \cdot |A|, k)$-homomorphic and is accepted by a minimal NFA with n states, then $k > \frac{\lg_2 n}{\lg_2 h + \lg_2 |A|}$.*

Proof. Let R be $(h \cdot |A|, k)$-homomorphic. Hence, there exist an alphabet B, with $|B| = h \cdot |A|$, and a k-slt language $L \subseteq B^+$ such that R is a homomorphic image of L. By Prop. 6, R is recognized by a NFA with a number s of states, where $s \leq \frac{|B|^k-1}{|B|-1}$. Since n is the number of states of a minimal NFA for R, it must be $s \geq n$. Therefore, since $\frac{|B|^k-1}{|B|-1} < |B|^k$, it follows that $|B|^k > n$, i.e., $k > \lg_{|B|} n = \frac{\lg_2 n}{\lg_2 |B|} = \frac{\lg_2 n}{\lg_2 h \cdot |A|} = \frac{\lg_2 n}{\lg_2 h + \lg_2 |A|}$. □

4. Main Result

The intuitive idea that by increasing the width one can use a smaller alphabet for the slt language, is studied in detail. Our approach consists of defining an slt language using a larger alphabet that encodes the states traversed by the original automaton into words of fixed length. Our main theorem states the relationship between the language complexity in terms of number of states, the alphabetic ratio, and the width of the slt language.

Theorem 8. *Given an alphabet A, if a language $R \subseteq A^+$ is accepted by a NFA with n states, then for every h, $2 \leq h < n$, R is $\left(h \cdot |A|, O(\frac{\lg n}{\lg h})\right)$-homomorphic.*

Combining the above result with Th. 7, one obtains:

Corollary 9. *Given an alphabet A, if a language $R \subseteq A^+$ is accepted by a* minimal *NFA with $n > 2$ states, then, for every $2 \leq h < n$, the minimal width k such that R is $(h \cdot |A|, k)$-homomorphic is $\Theta(\frac{\lg n}{\lg h})$,*

Therefore, the bound of Th. 8 is asymptotically optimal. The rest of the section is devoted to the proof of Th. 8. Special care is devoted to find an encoding of the original states into strings over the local alphabet, when the latter has the minimal alphabetic ratio. Since it may be important for applications, our encoding produces also a small, close to optimal, width of the slt language, at least when $h = 2$ and n is large enough. The proofs are organized so that the main lemmas hold, independently of the chosen encoding, which only affects the numerical results. This organization has the advantage that the proof would be unaffected by a change of encoding. The encoding is based on mapping alphabet A into a new alphabet $A \times D$, where set D has cardinality h. D can be assumed to be, e.g., the set $\{0, 1, 2, \ldots, h-1\}$. Only fixed-length encodings are considered.

Definition 10. *Let n, m, h be integers such that $2 \leq h < n$ and $m \geq \lceil lg_h(n) \rceil$. Let $M = (Q, A, E, q_0, F)$ be a NFA, where E is total and $n = |Q|$, and let D be a finite set with $|D| = h$. A* code *of Q into D of length m is a mapping $[\,] : Q \to D^m$ such that for every $p, q \in Q$, if $p \neq q$ then $[p] \neq [q]$*

Since the construction of the slt language and the proof of correctness are rather long and elaborate, we first present a small example (for the case $h = 2$) and its corresponding slt-language. To make the presentation easier, we illustrate the language by means of rather straightforward operations on automata.

Example 11. *The language $R = a(aa)^*b$ is accepted by the DFA $M = (Q = \{q_0, q_1, q_2\}, A = \{a, b\}, E, q_0, F = \{q_1\})$ shown in Fig.* 1*, left. For the sake of simplicity, the machine is not total. We choose to represent the states by the following* 5*-bit codes on alphabet $D = \{0, 1\}$: $[q_0] = 11100, [q_1] = 10100, [q_2] = 01100$. The discussion of the properties of good encodings, their length and their systematic construction is postponed to Lemma* 20*. Suffice to say that every chosen code ends in* 100*, and that in every concatenation of codes in $[q_0]\,([q_0] \cup [q_1] \cup [q_2])^*$ every factor of length* 9 *includes one, and only one, occurrence of factor* 100*. For language R, the proof of the main theorem defines a* 10*-slt-language L' over the alphabet $B = \{a, b\} \times \{0, 1\}$, such that the projection of L' on A is R, hence R is $(2, 10)$-slt. L' is there defined in terms of sets I_9, F_{10}, T_9, but, for the sake of illustration, we here use instead a simple machine M' that recognizes L'.*

M' is the product of machine M with another machine, the coder*, which recognizes the language $[q_0]\,([q_0] \cup [q_1] \cup [q_2])^+$. Each state of M' is obtained by pairing*

a state of M *and a* marked code*, such as* $11 \bullet 100$*; a code, such as* $[q_0]$*, can be marked placing the bullet in a position, ranging from* 1 *to* 5*; in* $11 \bullet 100$ *the position is* 3*, meaning that the next state transition will read the third bit of the code. The initial state is* $\frac{q_0}{\bullet 11100}$ *and the final states are all, and only, states with* q_2 *as first component; notice that all the final states are equivalent, but we keep them separate to make the systematic construction of* M' *more apparent.*

To define the edges of M'*, we distinguish two cases for a state, whether the bullet is in the last position or not. An example of the latter case is edge* $\frac{q_0}{\bullet 11100} \xrightarrow{a,1} \frac{q_1}{1 \bullet 1100}$*:* M *moves from* q_0 *to* q_1 *reading "a" and the first bit of* $[q_0]$ *is* 1*. An example of the former case is* $\frac{q_1}{1010 \bullet 0} \xrightarrow{a,0} \frac{q_0}{\bullet 11100}$*: every outgoing edge enters a state where a new code is started. Following our forthcoming results,* L' *is* 10*-slt, but in this case one can check that* L' *is actually a* 5*-slt language, by examining the only circuit present in* M'*, which, starting in* $\frac{q_0}{\bullet 11100}$*, is labeled* $l = 1110010100$ (*dropping the label component "a" which is clearly irrelevant*)*: all cyclic permutations of* l *are distinguishable by their length* 5 *prefixes.*

It is obvious that, dropping the second component, i.e., the code bits, of the edge labels, machine M' *becomes equivalent to* M*, the difference between the two being that the length* 2 *circuit* q_0, q_1 *of* M *has been unrolled into a length* 10 *circuit. Hence,* R *is* $(2, 5)$*-homomorphic.*

Actually, if we abandon our systematic code-based construction, there exists a simpler 2-slt generator L_2 for R, namely $L_2 = (a, 0)\,((a, 1)(a, 0))^* b$; thus R is

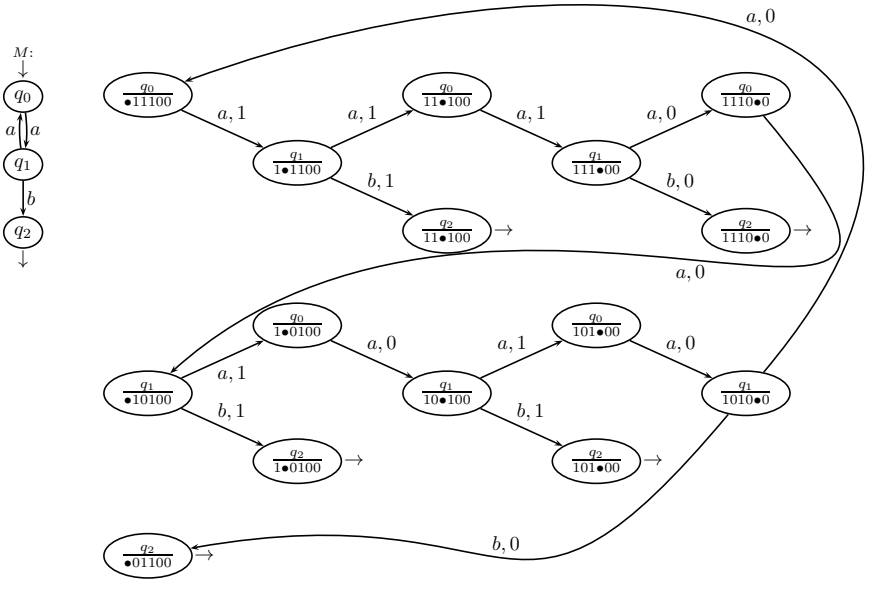

Fig. 1. Left: DFA M. Center: DFA M' recognizing the slt language; the states containing q_2 are all equivalent. The codes are $[q_0] = 11100$, $[q_1] = 10100$, $[q_2] = 01100$.

$(2,2)$-homomorphic. L_2 is essentially the 2-slt language obtained by the (revised) Medvedev's construction of Prop. 4. This does not contradict Th. 7, since n is smaller than $h \cdot |A|$.

Also when n is larger, the established lower bound is usually lower than the actual width. Consider for example generating the second power $R^2 = R \cdot R$ and the fifth power R^5 of the language of Ex. 11, using slt languages over a 4-letter alphabet (i.e., $h = 2$). The minimal DFAs of R^2 and R^5 have 5 and 11 states, respectively, giving a width $k \geq 2$ as lower bound in both cases. It would be a tedious exercise to show that in this case a slt-language for R^2 requires at least a width 4 (and a DFA with at least 7 states), while a slt-language for R^5 requires a width 8 (and a DFA with at least 23 states). Our construction in general would require 6-bit codes and width 12 (easily reduced to 9) for generating R^2, and 9 bits codes and width 18 (easily reduced to 15) for generating R^5.

The next definitions set the base for stating the properties a good encoding should have (see, e.g., [2]).

Consider a word x that is a factor of $[Q^+]$. We want to decode x to one state. This will be useful when defining a slt language whose homomorphic image is $L(M)$. If $|x| \geq 2m$, then x may include the concatenation of $[q]$ and $[p]$, for some $q, p \in Q$: it cannot be decoded to just one state symbol; moreover, if $|x| < 2m - 1$ then x may not contain any factor of the form $[q]$. However, if $|x|$ is exactly $2m - 1$, then the word must include one factor of the form $[q]$, for some $q \in Q$. If this factor is also unique, then x can be decoded to q. In general, however, uniqueness is not granted by the traditional notion of decodability (for every $x, y \in Q^+$, if $[x] = [y]$ then $x = y$): the latter assumes that the word to be decoded is a string in $[Q^+]$, while we need to consider a *factor* of $[Q^+]$ in any position of a word.

Definition 12. *A word $x \in D^{2m-1}$ is said to be* factor-decodable *if there exists one, and only one, position j, $1 \leq j \leq m-1$, such that there exists $q \in Q$: $s_{j,j+m}(x) = [q]$. A code $[\,] : Q \to D^m$ is* factor-decodable *if every word in $f_{2m-1}([Q^+])$ is factor-decodable.*

Lemma 13. *For all alphabets Q, D of cardinalities $n = |Q|$ and $h = |D|$, with $2 \leq h < n$, there exists a factor-decodable code of Q into D of length $m = \lceil g(h) + f(h) \lg_2 n \rceil \geq 3$, with:*

$$f(h) = \lg_2^{-1}\left(h - 1 + \sqrt{(h-1)(h+3)}\right) - 1$$

$$g(h) = 1 + \frac{f(h)}{2}\left(\lg_2(h-1) + \lg_2(h+3)\right).$$

Proof. Let $0 \in D$ be a symbol. The idea is to let code $[\,]$ be such that for every $q \in Q$, $[q]$ ends with the word 00, i.e., $s_{m-1,m}([q]) = 00$ and there is no other occurrence of 00 in $[q]$. Formally, for every i, $1 \leq i \leq m-1$, if $s_{i,i+1}([q]) = 00$ then $i = m - 1$. This condition is sufficient for factor-decodability. To find how large m must be as a function of h and n, first consider, for every $m \geq 2$, the set $S(m)$

of words in D^m such that $x \in S(m)$ if x has suffix 00 and in x there is no other occurrence of 00. If $|S(m)| \geq n$, then it is possible to assign a distinct word in $S(m)$ to every state of Q. The formal definition of $S(m)$ is by induction on $m \geq 2$. $S(2) = \{00\}$, i.e., the only word in $S(2)$ is 00. $S(3) = \{d00 \mid d \in D - \{0\}\}$. Given sets $S(m-1), S(m-2)$, let $S(m)$ be:

$$\{dy \mid d \in D - \{0\}, y \in S(m-1)\} \cup \{0dx \mid d \in D - \{0\}, y \in S(m-2)\}.$$

Hence, $|S(2)| = 1, |S(3)| = h-1$ and $|S(m)| = (h-1)\,|S(m-1)| + (h-1)\,|S(m-2)|$. This recurrence relation is strictly connected to the so-called Lucas sequence $U_m(M,N)$, where M, N are integers (see, e.g, p. 395 of [8]):

$$U_1(M,N) = 1, U_2(M,N) = M,$$
$$U_m(M,N) = M \cdot U_{m-1}(M,N) - N \cdot U_{m-2}(M,N), \text{for } m \geq 3.$$

For $M = 1, N = -1$ this is just a Fibonacci sequence. If $M^2 - 4N \geq 0$, a closed-form solution for every $m > 0$ is:

$$U_m(M,N) = \frac{a^m - b^m}{a-b}, a = \frac{M + \sqrt{M^2 - 4N}}{2}, b = \frac{M - \sqrt{M^2 - 4N}}{2} \tag{1}$$

It is immediate to see that, for every $m \geq 2$, $|S(m)| = U_{m-1}(h-1, 1-h)$. Therefore, by (1), $|S(m)| = \frac{a^{m-1} - b^{m-1}}{a-b}$, with:

$$a = \frac{h - 1 + \sqrt{(h-1)(h+3)}}{2},\ b = \frac{h - 1 - \sqrt{(h-1)(h+3)}}{2},\ a - b = \sqrt{(h-1)(h+3)}.$$

Notice that $a > 1$ and $-1 < b < 0$. In order for $|S(m)|$ not to be smaller than n, select $m \geq 3$ such that: $\frac{a^{m-1} - b^{m-1}}{a-b} \geq n$. Since $-1 < b < 0$, also $-1 < b^{m-1} < 1$: by choosing m such that $\frac{a^{m-1} - 1}{a-b} \geq n$, we have $lg_2(a^{m-1} - 1) \geq \lg_2(a-b) + \lg_2 n$. From $a > 1$, it follows that $\lg_2\left(a^{m-1} - 1\right) \leq \lg_2\left(a^{m-1}\right) - 1 = (m-1)\lg_2 a - 1$. Hence, $|S(m)| \geq n$ is satisfied if $m >= \lceil 1 + \frac{1 + \lg_2(a-b)}{\lg_2 a} + \frac{\lg_2 n}{\lg_2 a} \rceil$. With standard algebraic manipulations and by defining $f(h), g(h)$ as in the statement of the Lemma, one can derive that it is enough to choose m such that: $m = \lceil g(h) + f(h) \lg_2 n \rceil$. □

Remark 14. *Both $f(h)$ and $g(h)$ are monotonically decreasing with h, although very slowly for large h, with $\lim_{h\to\infty} f(h)\lg_2 h = 1$, $\lim_{h\to\infty} g(h) = 2$. In fact, $\lg_2 a$ is monotonically increasing with h, and $a - b = \sqrt{(h-1)(h+3)}$ grows more slowly with $h > 1$ than $a = \frac{h-1+\sqrt{(h-1)(h+3)}}{2}$. Hence both $f(h)$ and $g(h)$ are greatest for $h = 2$, corresponding to the values $a = \frac{1+\sqrt{5}}{2} \approx 1.44$, $a - b = \sqrt{5}$, with $\frac{\lg_2(a-b)}{\lg_2 a} \approx 1.67$. Hence, $0 < f(h) \lessapprox 1.44$, $2 \leq g(h) \lessapprox 4.11$. The expression for m is $O\left(\frac{\lg n}{\lg h}\right)$. By definition of a code, m cannot be smaller then $m_{min} = \lceil \frac{\lg_2 n}{\lg_2 h} \rceil$, i.e., m is $\Omega\left(\frac{\lg n}{\lg h}\right)$. In particular, the ratio m/m_{min}, where m is computed by the above formula, is dominated by term $f(h)\lg_2 h \lessapprox 1.44$, which is very close to 1 for $h \geq 3$. Hence, no encoding technique can significantly improve $f(h)$ (or $g(h)$), decreasing m/m_{min}. Examples of approximated values for $f(h)$, $g(h)$, and $f(h)\lg_2 h$ are shown in Table 1.*

Table 1. Some approximated values for $f(h)$, $g(h)$, and $f(h)\lg_2 h$.

h	2	3	4	10	100	1000
$f(h)$	1.44	0.68	0.52	0.29	0.15	0.10
$g(h)$	4.11	2.92	2.66	2.34	2.15	2.10
$f(h)\lg_2 h$	1.44	1.09	1.04	1.00	1.00	1.00

To prove Th. 8, a few more definitions are required.

Definition 15. *The alphabetic homomorphisms* $\alpha : A\times D \to A$ *and* $\delta : A\times D \to D$, *called respectively* projection on A *and* projection on D, *are defined as* $\alpha(a,d) = a$, $\delta(a,d) = d$ *for every* $a \in A, d \in D$.

In the following, a path of M of length $t \geq 0$ is called a *t-path*. Let $\div$ and mod denote, respectively, the integer division and integer remainder operations.

Definition 16. *Given a path* η *of* M *with* $|\eta| > m$, *let* $k = |\eta| \div m$, $j = (|\eta| \bmod m)$. *A sequence of* $k+1$ *consecutive paths of* M: $\eta_1, \eta_2, \ldots, \eta_k, \eta_{k+1}$ *is called the* canonical decomposition *of* η *if* $\eta = \eta_1\eta_2\ldots\eta_k\eta_{k+1}$, *each* η_h, $i \leq h \leq k$, *is a* m*-path and* η_{k+1} *is a* j*-path.*

The canonical decomposition $\eta_1, \eta_2, \ldots, \eta_k, \eta_{k+1}$ of a path η, verifying the above definition, is clearly unique. Moreover, $|\eta| = km + j$.

Definition 17. *With an abuse of notation, let* $[\ \ ] : (Q \times A \times Q)^* \to (A \times D)^*$ *be defined on paths as follows. Let* η *be a t-path. If* $t = 0$ *then* $[\eta] = \varepsilon$; *if* $1 \leq t \leq m$, *let* $[\eta]$ *be the unique word* z *in* $(A \times D)^m$ *such that* $\alpha(z) = l(\eta)$, $\delta(z) = i_t([o(\eta)])$ (*i.e.,* $\delta(z) = [o(\eta)]$ *if* η *is a* m*-path); finally, if* $t > m$ $[\eta] = [\eta_1][\eta_2]\ldots[\eta_k][\eta_{k+1}]$, *where* $\eta_1, \eta_2, \ldots, \eta_k, \eta_{k+1}$ *is the canonical decomposition of* η.

Example 18. *Consider DFA* M *of Ex.* 11, *where* $m = 5$ *and* $[q_0] = 11100, [q_1] = 10100$. *Define* η_0 *to be the* 5*-path:* $(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)$ *and* η_1 *to be the* 5*-path:* $(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)$. *Then,* $[\eta_0] = \langle a,1\rangle\langle a,1\rangle\langle a,1\rangle\langle a,0\rangle\langle a,0\rangle$, *i.e,* $\delta([\eta_0]) = [q_0]$, $[\eta_1] = \langle a,1\rangle\langle a,0\rangle\langle a,1\rangle\langle a,0\rangle\langle a,0\rangle$, *i.e.,* $\delta([\eta_1]) = [q_1]$. *Let* η_2 *be the* 2*-path* $(q_0,a,q_1)(q_1,b,q_2)$*: then* $[\eta_2] = \langle a,1\rangle\langle b,1\rangle$. *Any path longer than* 5 *admits a canonical decomposition; e.g., the* 12*-path* $\eta = (q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,a,q_0)(q_0,a,q_1)(q_1,b,q_2)$ *can be decomposed into* $\eta_0\eta_1\eta_2$. *Then,* $[\eta] = [\eta_0][\eta_1][\eta_2]$. *Notice that the labels of paths of machine* M' *of Ex.* 11 *are just encodings* (*using mapping* $[\ ]$ *of Def.* 17) *of paths of* M. *For instance,* $[\eta_0]$ *is the label of a path of* M' *of length* 5, *starting in state* $\frac{q_0}{\bullet 11100}$ *and ending in state* $\frac{q_1}{\bullet 10100}$.

Let L' be the $2m$-slt language defined by the following sets:

$$
\begin{aligned}
I_{2m-1} &= i_{2m-1}\left(\{[\eta'\eta''] \mid \eta', \eta'' \text{ are consecutive } m\text{-paths of } M \wedge \delta([\eta']) = [q_0]\}\right);\\
F_{2m} &= f_{2m}\left(\{[\eta'\eta''\eta'''] \mid \eta', \eta'', \eta''' \text{ are consecutive } m\text{-paths of } M\}\right);\\
T_{2m-1} &= t_{2m-1}(\{[\eta'\eta''\eta'''] \mid \eta', \eta'', \eta''' \text{ are consecutive paths of } M, |\eta'| = |\eta''| = m,\\
&\qquad 0 \le |\eta'''| < m, e(\eta''\eta''') \in F\}).
\end{aligned}
$$

L' is shown later to be a $2m$-slt language whose homomorphic image $\alpha(L')$ is $L(M)$, apart from a finite set. Notice that the definition of set F_{2m} crucially requires that the transition relation of M is total.

The proof of the following lemma follows from uniqueness of factor-decodability.

Lemma 19. *Let $[\ ] : Q \to D^m$ be a factor-decodable code. For all $z \in F_{2m}$, there exist a position j, $1 \le j \le m-1$, and two consecutive paths η_1, η_2 of M such that:*

(1) *η_1 is a m-path, and $t_{2m-j+1}(z) = [\eta_1][\eta_2]$;*
(2) *for any two consecutive paths of M, η_I, η_{II}, if η_I is a m-path and $[\eta_I][\eta_{II}]$ is a suffix of z then $[\eta_1] = [\eta_I]$ and $[\eta_2] = [\eta_{II}]$;*
(3) *if $\delta(i_m(z)) = [q]$ for some $q \in Q$, then $j = 1$, η_2 is a m-path and $o(\eta_1) = q$;*
(4) *if $t_{2m-1}(z) \in T_{2m-1}$, then $e(\eta_1\eta_2) \in F$.*

Proof. By definition of F_{2m}, there exist three consecutive m-paths η', η'', η''' such that $z \in f_{2m}([\eta'\eta''\eta'''])$. Let η_1, η_3 be defined as follows: a) $\eta_1 = \eta'$, $\eta_3 = \eta''$ if $z = i_{2m}([\eta'\eta''\eta'''])$; b) $\eta_1 = \eta''$, $\eta_3 = \eta'''$ otherwise. Since $|i_{2m-1}(z)| = 2m-1$, z must have $[\eta_1]$ as a factor. Let $w = \delta(i_{2m-1}(z))$. Since $|i_{2m-1}(z)| = 2m-1$, $i_{2m-1}(z)$ must have $[\eta_1]$ as a factor, hence w has a factor $\delta([\eta_1]) = [o(\eta_1)]$, beginning at a position j. Then, in case (a) it is $j = 1$ and in case (b) it is $2 \le j \le m-1$. Define η_2 to be the prefix of length $m - j + 1$ of η_3 (hence, in case (a) $\eta_2 = \eta_3$). Part (1) follows, by considering that $t_{2m-j+1}(z) = s_{j,j+m-1}(z)s_{j+m,2m}(z) = [\eta_1]i_{m-j+1}([\eta_3]) = [\eta_1][\eta_2]$. Part (2) is immediate by factor-decodability. Since $|w| = 2m-1$ and $w \in f_{2m-1}([\eta'\eta''\eta''']) \subseteq f_{2m-1}([Q^+])$, by uniqueness of factor-decodability every other position h, $1 \le h \le m-1$, and every state $q \in Q$ verifying $\delta(s_{h,h+m-1}(w)) = [q]$, are such that $j = h$, $[o(\eta_1)] = [q]$. Hence, $[\eta_I] = [\eta_1]$, and obviously also suffix $[\eta_{II}] = [\eta_2]$. Part (3) also follows by factor-decodability of w: it corresponds to case (a) of Part (1), hence $j = 1$ and $[\eta_1] = [q]$, i.e., $o(\eta_1) = q$. Part (4) also follows. Let $\eta_i, \eta_{ii}, \eta_{iii}$ be three consecutive paths verifying the definition of T_{2m-1} and such that $t_{2m-1}(z) = t_{2m-1}([\eta_i\eta_{ii}\eta_{iii}])$. Hence, $[\eta_{ii}\eta_{iii}]$ is a suffix of z: by Part (2), $[\eta_{ii}] = [\eta_1]$, and $[\eta_{iii}] = [\eta_2]$, i.e. $o(\eta_1) = o(\eta_{ii})$ since both are m-paths. Let η_{iv} be a path such that η_{iii}, η_{iv} are consecutive, $l(\eta_{iii}\eta_{iv}) = l(\eta''')$. Such a path must exist, since transition relation E is assumed to be total. Moreover, $[\eta_{iii}\eta_{iv}] = [o(\eta_{iii})]$. Consider case (b) of the proof of Part (1): z is a factor of $[\eta'\eta''\eta''']$, with $\eta_1 = \eta''$. Paths η', η_{ii} are consecutive: $e(\eta') = o(\eta'') = o(\eta_1)$ and, as noted above, $o(\eta_1) = o(\eta_{ii})$. Hence $\eta', \eta_{ii}, (\eta_{iii}\eta_{iv})$ are consecutive m paths: z is also in the set $f_{2m}([\eta'\eta_{ii}(\eta_{iii}\eta_{iv})]) \subseteq F_{2m}$. Consider case (a) of Part (1):

$z = i_{2m}([\eta'\eta''\eta'''']) = [\eta'\eta'']$, with $\eta_1 = \eta'$, $\eta_2 = \eta_3 = \eta''$. Since $[\eta_{ii}] = [\eta_1]$, and $[\eta_{iii}] = [\eta_2]$, then $z = [\eta'\eta''] = [\eta_{ii}\eta_{iii}]$. In both cases, it is enough to select in Part (1) $\eta_1 = \eta_{ii}$, $\eta_2 = \eta_{iii}$. □

Lemma 20. *There exists a finite language $L'' \subseteq A^+$ such that $\alpha(L') \cup L'' = L(M)$, where α is the projection defined in Def.* 15.

Proof. Let L'' be the set of words in $L(M)$ of length less than $3m$.

Part (I): $(L(M) - L'') \subseteq \alpha(L(M')$. Assume that $x \in L(M), |x| \geq 3m$. To show that there exists a successful path η of M such that $l(\eta) = x$, we first claim the following result for every path, whether successful or not:

$$(*) \text{ for all paths } \eta \text{ of } M, \text{ with } |\eta| \geq 3m, f_{2m}([\eta]) \subseteq F_{2m}.$$

The proof of (*) is by induction on the canonical decomposition $\eta_1, \eta_2, \ldots, \eta_k, \eta_{k+1}$ of η. We first prove (*) for the case $|\eta_{k+1}| = 0$, by induction on $k \geq 3$. The base $|\eta| = 3m$ is trivial, since in this case $\eta = \eta_1\eta_2\eta_3$, for some η_1, η_2, η_3 consecutive m-paths of M: $f_{2m}([\eta]) = f_{2m}([\eta_1\eta_2\eta_3]) \subseteq F_{2m}$. For $k > 3$, $f_{2m}([\eta]) = f_{2m}([\eta_1 \ldots \eta_k]) = f_{2m}([\eta_1] \ldots [\eta_k])$. Since each η_i is an m-path, $f_{2m}([\eta])$ is the union of set $f_{2m}([\eta_1 \ldots \eta_{k-1}])$, which is in F_{2m} by induction hypothesis, and of set $f_{2m}([\eta_{k-2}][\eta_{k-1}][\eta_k]) = f_{2m}([\eta_{k-2}\eta_{k-1}\eta_k]) \subseteq F_{2m}$. Therefore, $f_{2m}([\eta]) \subseteq F_{2m}$.

To prove (*) also for $j = |\eta_{k+1}| > 0$, consider that the transition relation of M is total: there exists a path η'' of M of length $m - j > 0$ such that $\eta_{k+1}\eta''$ of M is an m-path of M. Hence, also η, η'' are consecutive, and the length of $\eta' = \eta\eta''$ is a multiple of m. Then (*) holds for $[\eta']$: $f_{2m}([\eta']) \subseteq F_{2m}$. But clearly $f_{2m}([\eta]) \subseteq f_{2m}([\eta']) \subseteq F_{2m}$.

Part (I) can now be completed. For all $x \in L(M) - L''$, let η be a successful path of M with $l(\eta) = x$; moreover, let $\eta_1, \ldots, \eta_k, \eta_{k+1}$ be the canonical decomposition of η. By (*), $f_{2m}([\eta]) \subseteq F_{2m}$. But η is successful: $o(\eta) = o(\eta_1) = q_0$, hence $i_{2m-1}([\eta]) = i_{2m-1}([\eta_1][\eta_2]) \in I_{2m-1}$; $e(\eta) \in F$, hence $t_{2m-1}([\eta]) = t_{2m-1}([\eta_{k-1}\eta_k\eta_{k+1}]) \in T_{2m-1}$. Therefore, $[\eta] \in L'$.

Part (II): $\alpha(L') \subseteq L(M)$. The proof needs the assumption that code $[\,]$ is factor-decodable. The following property (+) is proved next by induction on $k \geq 2$, by applying Lemma 19:

$$(+) \text{ for all words } z \in (A \times D)^+, |z| \geq 2m, \text{ if } f_{2m}(z) \subseteq F_{2m} \text{ and } i_{2m-1}(z) \in I_{2m-1} \text{ then there exists a path } \eta \text{ of } M \text{ such that } z = [\eta] \text{ and } o(\eta) = q_0.$$

Let $z \in (A \times D)^+$, with $|z| = km + j$, for $k \geq 3$, $0 \leq j \leq m - 1$. Then z can be decomposed as $z = z_1 z_2 \ldots z_k z_{k+1}$, with $|z_1| = |z_2| = \cdots = |z_k|$ and $|z_{k+1}| = j$. We first prove (+) for the case $j = 0$, by induction on $k \geq 2$. The base case is $k = 2$. Since $z \in F_{2m}$, from Lemma 19, Part (1), it follows that $\delta(i_m(z)) = [q_0]$; by Part (2), there exist a unique position h, $1 \leq h \leq m - 1$, and two consecutive paths η_1, η_2 of M, with η_1 a m-path, such that $t_{2m-h+1}(z) = [\eta_1\eta_2]$. Since $i_{2m-1}(z) \in I_{2m-1}$, then $\delta(i_m(z)) = [q_0]$: by Part (3) of Lemma 19, it follows that $h = 1$ and

$o(\eta_1) = q_0$. Therefore, $z = [\eta_1\eta_2]$ and $o(\eta_1\eta_2) = q_0$. Let now assume $k > 2$ and let $z \in (A \times D)^{km}$. Let $z' = i_{(k-1)m}(z), z'' = t_m(z)$, hence $z = z'z''$. By induction hypothesis, there exists a path η' of M such that $z' = [\eta']$ and $o(\eta') = q_0$. Let $\eta_1, \eta_2, \ldots, \eta_{k-1}$ be the canonical decomposition of η'. By Def. 17 of $[\,]$ on m-paths, $\alpha([\eta_{k-1}]) = l(\eta_{k-1}), \delta([\eta_{k-1}]) = [o(\eta_{k-1})]$. Apply Lemma 19, Part (1), to $[\eta_{k-1}]z'' = t_{2m}(z'z'') = t_{2m}(z) \in f_{2m}(z) \subseteq F_{2m}$: there exists a position h, and two consecutive paths, say, η_I, η_{II} of M, with η_I a m-path, such that $s_{h,2m}([\eta_{k-1}]z'') = [\eta_I][\eta_{II}]$. However, $\delta(i_m(z)) = [o(\eta_{k-1})]$: by Part (3) of the same lemma, $h = 1$, $[\eta_{k-1}] = [\eta_I]$, $z'' = [\eta_{II}]$ and also η_{II} is a m-path. Since $o(\eta_I) = o(\eta_{k-1})$ and η_{k-2}, η_{k-1} are consecutive, then η_{k-2}, η_{II} are also consecutive. Therefore, $\eta_1, \ldots, \eta_{k-2}, \eta_I, \eta_{II}$ are consecutive m-paths, with $z = z'z'' = [\eta_1 \ldots \eta_{k-2}][\eta_{k-1}]z'' = [\eta_1 \ldots \eta_{k-2}][\eta_I\eta_{II}] = [\eta_1 \ldots \eta_{k-2}\eta_I\eta_{II}]$.

The case $j > 0$ now follows. Let z be such that $f_{2m}(z) \subseteq F_{2m}$ $i_{2m-1}(z) \in I_{2m-1}$, $|z| = km + j$, $k \geq 3, 0 \leq j \leq m-1$. Let z', z'' be such that $z = z'z''$, with $|z'| = km$ and $|z''| = j$. Hence, $f_{2m}(z') \subseteq f_{2m}(z) \subseteq F_{2m}$ and $i_{2m-1}(z') = i_{2m-1}(z) \subseteq I_{2m-1}$.

Apply Lemma 19, Part (1), to $t_{2m}(z'z'') = t_{2m}(z) \subseteq F_{2m}$. Then there exists one, and only one, position h such that there exist two consecutive paths η_I, η_{II} of M, with η_I a m-path, such that $t_{2m-h}(t_{2m}(z'z'')) = t_{2m-h}(z'z'')) = [\eta_I][\eta_{II}]$. Hence, $t_{2m}(z'z'') = t_{2m}([\eta_{k-1}][\eta_k]z'')$. By induction hypothesis, formula (+) holds for z'; hence, there exists a path η' of M such that $z' = [\eta']$ and $o(\eta') = q_0$. Let $\eta_1, \ldots, \eta_k$ be the canonical decomposition of η'. Since $t_{2m}(z'z'') = t_{2m}([\eta_{k-1}][\eta_k]z'')$, by Lemma 19, Part (2), $[\eta_k] = [\eta_I]$, $z'' = [\eta_{II}]$. Since $o(\eta_k) = o(\eta_I) = o(\eta_I\eta_{II})$ and η_{k-1}, η_k are consecutive, then also η_{k-1}, η_I are consecutive. Hence, also $\eta_1 \ldots \eta_{k-1}\eta_I\eta_{II}$ is a path of M, with $z = z'z'' = [\eta_1 \ldots \eta_{k-2}\eta_{k-1}\eta_k][\eta_{II}] = [\eta_1 \ldots \eta_{k-2}][\eta_{k-1}][\eta_k][\eta_{II}] = [\eta_1 \ldots \eta_{k-2}][\eta_{k-1}][\eta_I][\eta_{II}] = [\eta_1 \ldots \eta_{k-2}][\eta_{k-1}\eta_I\eta_{II}] = [\eta_1 \ldots \eta_{k-2}\eta_{k-1}\eta_I\eta_{II}]$, hence (+) holds for z.

The proof of Part (II) follows from (+). In fact, if $x \in \alpha(L')$, with $|x| \geq 3m$, then there exists $z \in L'$ such that $x = \alpha(z)$. Since in this case $i_{2m-1}(z) \in I_{2m-1}$, $f_{2m}(z) \subseteq F_{2m}$, $t_{2m-1}(z) \in T_{2m-1}$, by (+) there exists a path η of M with origin in q_0 and such that $z = [\eta]$. Let $\eta_1, \eta_2, \ldots, \eta_k, \eta_{k+1}$ be the canonical decomposition of η, with $|\eta| = km + j$, $k \geq 3$ and $0 \leq j \leq m-1$ (hence $|\eta_{k+1}| = j$). Let $w = t_{2m-1}([\eta_{k-1}][\eta_k][\eta_{k+1}])$ and consider $t_{2m-1}(z) = t_{2m-1}([\eta]) = t_{2m-1}([\eta_{k-1}][\eta_k][\eta_{k+1}]) = w$. Apply Lemma 19, Part (1), to $w \in F_{2m}, w \in T_{2m-1}$. Hence, there exists a position h and consecutive η', η'', with η' a m-path, such that $t_{2m-h}(w) = [\eta'][\eta'']$. Since $[\eta_k][\eta_{k+1}]$ (of length $m + j \leq 2m - 1$) is also a suffix of w, by Part (2) of Lemma 19, $[\eta_k] = [\eta'], [\eta_{k+1}] = [\eta'']$. Since $o(\eta_k) = o(\eta')$, also paths η_{k-1}, η' are consecutive. Hence, $z = [\eta] = [\eta_1 \ldots \eta_{k-1}\eta_k\eta_{k+1}] = [\eta_1 \ldots \eta_{k-1}][\eta_k\eta_{k+1}] = [\eta_1 \ldots \eta_{k-1}][\eta'\eta''] = [\eta_1 \ldots \eta_{k-1}\eta'\eta'']$. Therefore, path $\eta_1 \ldots \eta_{k-1}\eta'\eta''$ has label z, origin q_0, and end $e(\eta'\eta'')$ in F, i.e., it is successful: $x \in L$. □

By Lemmas 13 and 20, $m = \lceil g(h) + f(h) \lg_2 n \rceil$, and L' is $2m$-slt. Therefore, the following statement holds:

Table 2. Width for some values of number n of states and of alphabetic ratio h.

h \ n	10	10^3	10^6	10^9	10^{40}
2	18	38	66	94	392
3	12	20	34	48	190
4	10	16	28	38	144
10	8	12	18	24	86
100	6	8	12	14	46
1000	6	8	10	12	32

Theorem 21. *If a language $L \subseteq A^+$ is accepted by a NFA with n states, then for every $h \geq 2$, L is $(h \cdot |A|, 2\lceil g(h) + f(h) \lg_2 n \rceil)$-homomorphic.*

Th. 8 follows immediately, since $2\lceil g(h) + f(h) \lg_2 n \rceil$ is $O\left(\frac{\lg n}{\lg h}\right)$. Examples of width values are shown in Table 2.

Discussion. Our construction of a factor-decodable coding is focused on the case $h = 2$, since it requires a 00 separator. Clearly, it is pointless to consider an alphabetic ratio $h \geq n$, since for $h = n$ the simpler construction of Prop. 4 (encoding every state with one different symbol) produces a local language, recognized by a DFA with n states. If $h < n$ is very close to n, shorter encodings than the one of Lemma 13 may be defined, allowing to decrease the required width. For instance, when $h = n - 1$ (and n is large) the lower bound on the width is $k = 2$, while the above encoding requires $m = 4$, to define a 8-slt language: it is not difficult to define instead an encoding with $m = 2$ that allows the definition of a 3-slt language. Also, for many regular languages one can obtain, by direct constructions, a homomorphic definition that uses lower values of alphabetic ratio and/or width than those obtained by the main theorem. However, by Th. 7, the width has a lower bound which is logarithmic in the size of a minimal NFA for the language (see also the discussion after Ex. 11), hence a small width may require a large alphabet.

5. Conclusion

We have generalized Medvedev's homomorphic characterization of regular languages: instead of using as generator a local language over a large alphabet, which depends on the size of the minimal DFA, we can use a strictly locally testable language over a smaller alphabet that does not depend on size, but just on the source alphabet. We have proved that the smallest alphabet one can use in the generator is the double of the alphabet of the regular language; thus, for instan ce, four symbols suffice to homomorphically generate any regular binary language.

In the main proof we have offered a specific and fairly optimized construction of the strictly locally testable language, for which we have derived the relationship

between the width, the ratio of the alphabet sizes, and the state complexity of the source language. In our opinion, the construction should be of interest, as a new technique for simulating a NFA by means of a larger, yet strictly locally testable, machine. Our encoding is asymptotically optimal with respect to language complexity. But it is an open technical question whether a different construction would yield better values for the alphabetic ratio and the width.

Applications and developments of our result are conceivable in areas where a language characterization *à la* Medvedev has been found valuable, for instance, the next ones. *Picture languages*: a main family of 2-dimensional languages, the tiling systems [10], is defined by a 2-dimensional Medvedev characterization. Does our result extend to 2D languages? *Context-free languages*: combining our result with the Chomsky-Schutzenberger theorem it should be possible to obtain non-erasing homomorphic characterizations using a small alphabet. *Consensual languages* [7]: this generalization of finite-state machines motivated by modeling tightly connected concurrent computations uses homomorphism between words as its core mechanism. In [6], Th. 8 is used to prove that the family of regular languages coincides with the family of Consensual Languages generated by slt bases. *Information transmission* for reducing the receiver cost was already mentioned in the introduction.

Acknowledgments

Thanks to Aldo de Luca for suggesting relevant references. We also thank the anonymous reviewers for relevant technical improvements.

References

[1] A. de Luca and A. Restivo. A characterization of strictly locally testable languages and its applications to subsemigroups of a free semigroup. *Information and Control*, 44(3):300–319, March 1980.

[2] J. Berstel, D. Perrin, and C. Reutenauer. *Codes and Automata*, volume 129 of *Encyclopedia of Mathematics and its Applications*. CUP, November 2009. (619 pp.).

[3] J. Berstel and J.E. Pin. Local languages and the Berry-Sethi algorithm. *Th. Comput. Sc.*, 155(2):439–446, 1996.

[4] V. Braitenberg. *Das Bild der Welt im Kopf: Eine Naturgeschichte des Geistes*. LIT Verlag, Muenster, 2004.

[5] P. Caron. Families of locally testable languages. *Th. Comput. Sc.*, 242(1-2):361–376, 2000.

[6] S. Crespi Reghizzi and P. San Pietro. Strict local testability with consensus equals regularity (to appear). In *CIAA* 2012*, Porto, July* 17-20, 2012.

[7] S. Crespi Reghizzi and Pierluigi San Pietro. Consensual languages and matching finite-state computations. *RAIRO Theor. Informatics and Appl.*, 45:77–97, 2011.

[8] L.E. Dickson. *History of the Theory of Numbers*. Carnegie Institution of Washington, Online version at: http://www.archive.org/details/historyoftheoryo01dick, 1919.

[9] P. Garcia and E. Vidal. Inference of K-testable languages in the strict sense and application to syntactic pattern recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(9):920–925, September 1990.

[10] D. Giammarresi and A. Restivo. Two-dimensional languages. In G. Rozenberg and A. Salomaa, editors, *Handbook of formal languages, vol. 3: beyond words*, pages 215–267. Springer-Verlag New York, Inc., New York, NY, USA, 1997.

[11] T. Head. Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors. *Bulletin of Mathematical Biology*, 49:737–759, 1987.

[12] J. Rogers and G. Pullum. Aural pattern recognition experiments and the subregular hierarchy. *Journ. of Logic Language and Information*, to appear, 2011.

[13] R. McNaughton and S. Papert. *Counter-free Automata*. MIT Press, Cambridge, USA, 1971.

[14] Y. T. Medvedev. On the class of events representable in a finite automaton. In E. F. Moore, editor, *Sequential machines – Selected papers* (*translated from Russian*), pages 215–227. Addison-Wesley, New York, NY, USA, 1964.

[15] S. Eilenberg. *Automata, Languages, and Machines*. Academic Press, 1974.

[16] W. A. Wickelgren. Context-sensitive coding, associative memory and serial order in (speech) behavior. *Psychological Review*, 76(1), 1969.