Piecewise Testable Languages via Combinatorics on Words

Ondřej Klíma *

Department of Mathematics and Statistics, Masaryk University
Kotlářská 2, 611 37 Brno, Czech Republic

Abstract. A regular language L over an alphabet A is called *piecewise testable* if it is a finite boolean combination of languages of the form $A^*a_1A^*a_2A^*\dots A^*a_\ell A^*$, where $a_1,\dots,a_\ell \in A$, $\ell \geq 0$. An effective characterization of piecewise testable languages was given in 1972 by Simon who proved that a language L is piecewise testable if and only if its syntactic monoid is $\mathcal{J}$-trivial. Nowadays there exist several proofs of this result based on various methods from algebraic theory of regular languages. Our contribution adds a new purely combinatorial proof.
Keywords: piecewise testable languages, syntactic congruence
2000 Classification: 68Q45 Formal languages and automata

1 Introduction

Piecewise testable languages occur as languages which are recognized by a special model of automaton which can be called *Hydra automaton*. Such an automaton has a finite number (say k) of heads, which are ordered and each head can read a letter of the input word. So, together they can read a subword of the input word of length at most k and put it into the memory. (Note that in the whole paper the term subword means scattered subword.) Finally, the automaton accepts the input word if the collection of words in the memory is from a given list of possible sets of subwords.

More formally, a language L over an alphabet A is piecewise testable if and only if there exists a natural number k such that L is a union of classes of the equivalence relation $\sim_k$ defined in the following way: $u \sim_k v$ if and only if words $u, v \in A^*$ have the same subwords of length at most k.

The first basic observation (see Lemma 1) says that a language L over an alphabet A is piecewise testable if and only if it is a finite boolean combination of languages of the form

$$A^*a_1A^*a_2A^*\dots A^*a_\ell A^*, \text{ where } a_1,\dots,a_\ell \in A,\ \ell \geq 0\ .$$

The property is used as a formal definition of piecewise testable languages in some papers. Unfortunately, such a characterization is not effective. It is not

* The author was supported by the ESF program AutoMathA and by the Grant no. 201/09/1313 of the Grant Agency of the Czech Republic.

clear how one can recognize whether a language (e.g. given by an automaton) is piecewise testable or not.

An effective characterization of piecewise testable languages was given by Simon [7, 8] who proved that a language L is piecewise testable if and only if its syntactic monoid is $\mathcal{J}$-trivial. The direct implication in this statement is quite easy. The difficulty is contained in the converse implication. There exist several proofs of the converse implication which use different techniques: the original combinatorial proof by Simon [7, 8], see Pin [4] for a slightly improved version, the proof by Straubing and Thérien [9] using ideas concerning ordered monoids, the proof by Almeida [1] using sophisticated profinite topology and the proof by Higgins [3] working with transformation semigroups. Many other interesting papers on the topic were written — we refer to the survey paper by Pin [6] and the book by Almeida [2] for more information.

The content of our contribution is a new proof of Simon's result. In fact, we show a straightforward proof of the converse implication based just on combinatorics on words. The whole paper is self-contained, in particular we give not only the proof of the crucial statement (Lemma 3), but we also repeat the proofs of all used statements (Lemma 1 and Lemma 2), which can be find in any paper concerning piecewise testable languages. The complete proof is contained in Section 3, while technical notation is summarized in Section 2.

2 Notation

Let A^* denote the set of all words over an alphabet A including the empty one, denoted by λ. The length of a word $u \in A^*$ is denoted by $|u|$. For words $u, v \in A^*$ we write $u \triangleleft v$ if and only if u is a subword of v, i.e. there are letters $a_1, a_2 \dots, a_\ell \in A$ and words $v_0, v_1, \dots, v_\ell \in A^*$ such that $u = a_1 a_2 \dots a_\ell$ and $v = v_0 a_1 v_1 a_2 \dots a_\ell v_\ell$. For $v \in A^*$ we denote $\mathsf{Sub}_k(v) = \{u \in A^* \mid u \triangleleft v, |u| \leq k\}$. We define the relation $\sim_k$ on A^* by the rule

$$u \sim_k v \quad \text{ if and only if } \quad \mathsf{Sub}_k(u) = \mathsf{Sub}_k(v) \ .$$

A language L over an alphabet A is a set $L \subseteq A^*$ and the complement of L is denoted by L^c, i.e. $L^c = A^* \setminus L$. Further, for a given word $u \in A^*$ we denote by L_u the language of all words which contain the word u as a subword, i.e. $L_u = \{v \in A^* \mid u \triangleleft v\}$. If $u = a_1 a_2 \dots a_\ell$, where $a_1, a_2, \dots, a_\ell \in A$, then we can write

$$L_u = A^* a_1 A^* a_2 A^* \dots A^* a_\ell A^* \ .$$

For a regular language $L \subseteq A^*$ we define the relation $\sim_L$ on A^* as follows: for $u, v \in A^*$ we have

$$u \sim_L v \quad \text{ if and only if } \quad (\ \forall\, p, q \in A^*\ )\ (\ puq \in L \iff pvq \in L\ )\ .$$

It is easy to see that the relation $\sim_L$ is a congruence on A^*, i.e. $\sim_L$ is an equivalence relation on A^* which satisfies

$$u \sim_L v \quad \Longrightarrow \quad wu \sim_L wv,\ uw \sim_L vw$$

for every $u, v, w \in A^*$. The relation $\sim_L$ is called the *syntactic congruence* of L and the corresponding quotient monoid $A^*/\sim_L$ is called the *syntactic monoid* of L. A basic observation in algebraic theory of regular languages says that the monoid $A^*/\sim_L$ is isomorphic to the transformation monoid of the minimal automaton of the language L. In particular, the monoid $A^*/\sim_L$ is finite and consequently $\sim_L$ has a finite index. Further, it is easy to see that the language L is a union of some classes in the partition given by $\sim_L$. The reader can see the survey papers [5] or [6] for an introduction to syntactic methods, however it is not needed for understanding the paper.

The last definition which we will need is that of $\mathcal{J}$-trivial monoids. To make the presentation as simple as possible we rephrase this notion for congruences. We say that a congruence $\sim$ on A^* is $\mathcal{J}$-trivial if and only if

$$w_1 w_2 u w_3 w_4 \sim u \quad \Longrightarrow \quad w_2 u w_3 \sim u$$

for every words $u, w_1, w_2, w_3, w_4 \in A^*$.

3 A Proof of Simon's Theorem

The following lemma can be found in any paper concerning piecewise testable languages.

Lemma 1. *Let A be an alphabet and L be a regular language over A. Then the following two conditions are equivalent.*
(i) There exists a natural number k such that $\sim_k \subseteq \sim_L$.
(ii) The language L is a finite boolean combination of languages L_u, $u \in A^*$.

Proof. If $\sim_k \subseteq \sim_L$ then each class of the partition $A^*/\sim_L$ is a union of classes of the partition $A^*/\sim_k$. Since L is a union of classes of $A^*/\sim_L$, it is enough to show that each class of the partition $A^*/\sim_k$ can be written as a combination of languages of the form L_u, $u \in A^*$. If we take $v \in A^*$ then for the class $v{\sim_k} = \{w \in A^* \mid w \sim_k v\}$ the following expression is easy to see:

$$v \sim_k = \bigcap_{u \in \mathsf{Sub}_k(v)} L_u \qquad \cap \qquad \bigcap_{u \notin \mathsf{Sub}_k(v), |u| \leq k} L_u^c \ .$$

Now let L satisfy condition (ii), i.e. L be a finite union of finite intersections of languages of the form L_u and L_u^c where $u \in A^*$. Let k be a natural number such that $|u| \leq k$ for all words u used in this expression. We would like to prove $\sim_k \subseteq \sim_L$. So, let $v, w \in A^*$ be such that $v \sim_k w$, i.e. $\mathsf{Sub}_k(v) = \mathsf{Sub}_k(w)$. Let $p, q \in A^*$ be arbitrary words such that $pvq \in L$. Our goal is to prove $pwq \in L$. We can assume that $pvq \in K = L_{u_1} \cap \cdots \cap L_{u_m} \cap L_{v_1}^c \cap \cdots \cap L_{v_n}^c$ where K is one of the summands in the considered expression of L. All mentioned words $u_1, \ldots, u_m, v_1, \ldots, v_n$ have length at most k. For each $i = 1, \ldots, m$ we have $pvq \in L_{u_i}$ and for each $j = 1, \ldots, n$ we have $pvq \notin L_{v_j}$. Now for each $i = 1, \ldots, m$ we have $u_i \lhd pvq$ and one can deduce that $u_i \lhd pwq$ because $\mathsf{Sub}_k(v) = \mathsf{Sub}_k(w)$.

This means $pwq \in L_{u_i}$ for each $i = 1, \ldots, m$. Further $pwq \notin L_{v_j}$ for each $j = 1, \ldots, n$, because the fact $pwq \in L_{v_j}$ implies $pvq \in L_{v_j}$, which is not true. So, we have proved that $pwq \in K$ and finally $pwq \in L$. If we exchange v and w, we obtain also the proof of the converse implication $pwq \in L \implies pvq \in L$ and the proof is complete. □

Recall that both conditions in the previous lemma are used in literature to define that a language L is piecewise testable. The following lemma uses quite a standard technique from semigroup theory and the proof is not new.

Lemma 2. *Let A be a finite alphabet and L be a piecewise testable language over A. Then $\sim_L$ is a $\mathcal{J}$-trivial congruence on A^*.*

Proof. Let L be a piecewise testable language, i.e. we have $\sim_k \subseteq \sim_L$ for some k. Assume that for words $u, w_1, w_2, w_3, w_4 \in A^*$ we have $w_1w_2uw_3w_4 \sim_L u$. Since $\sim_L$ is a congruence, we have

$$(w_1w_2)^2u(w_3w_4)^2 \ \sim_L w_1w_2uw_3w_4 \sim_L u \ .$$

If we denote $u_n = (w_1w_2)^n u(w_3w_4)^n$ then it is easy to prove (by induction on n) that $u_n \sim_L u$ for every natural number n. It is clear that $u \lhd u_1 \lhd u_2 \lhd \ldots$, hence we have $\mathsf{Sub}_k(u) \subseteq \mathsf{Sub}_k(u_1) \subseteq \mathsf{Sub}_k(u_2) \subseteq \ldots$. Since there are only finitely many possible sets of the form $\mathsf{Sub}_k(v)$, $v \in A^*$, we see that $\mathsf{Sub}_k(u_n) = \mathsf{Sub}_k(u_{n'})$ for some $n < n'$. Then $\mathsf{Sub}_k(u_n) \subseteq \mathsf{Sub}_k(w_2u_nw_3) \subseteq \mathsf{Sub}_k(u_{n'})$, so the equality $\mathsf{Sub}_k(u_n) = \mathsf{Sub}_k(w_2u_nw_3)$ follows. This means $u_n \sim_k w_2u_nw_3$. Since $u_n \sim_L u$ and $\sim_k \subseteq \sim_L$, we can conclude with $u \sim_L w_2uw_3$. We have proved that $\sim_L$ is a $\mathcal{J}$-trivial congruence. □

The following lemma formulates the difficult part of Simon's result. The proof of Lemma 3 is an essence of our contribution.

Lemma 3. *Let A be a finite alphabet and L be a regular language over A such that $\sim_L$ is a $\mathcal{J}$-trivial congruence on A^*. Then L is piecewise testable.*

Proof. Assume that $L \subseteq A^*$ is such that $\sim_L$ is a $\mathcal{J}$-trivial congruence. Let m be the index of this congruence. We show that $\sim_k \subseteq \sim_L$ for $k = 2m - 2$.

Let $u = a_1a_2 \ldots a_p$ and $v = b_1b_2 \ldots b_q$, where $a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_q \in A$, $p, q \geq 0$, be such words that $u \sim_k v$. We consider all the prefixes of u, namely $u_i = a_1a_2 \ldots a_i$ for each $i = 0, \ldots, p$, where $u_0 = \lambda$. Since $\sim_L$ is $\mathcal{J}$-trivial, we know that the fact $u_i \sim_L u_j$ for some given $i < j$ implies $u_i \sim_L u_{i'}$ for each $i' \in \{i, i+1, \ldots, j\}$. We call an index $i \in \{1, \ldots, p\}$ blue if $u_{i-1} \not\sim_L u_{i-1}a_i$. Since the number of classes in the partition $A^*/\sim_L$ is m, there are at most $m - 1$ blue indices $i_1 < i_2 < \cdots < i_r$ in u, where $r \leq m - 1$. For an index i which is not blue we have $u_{i-1} \sim_L u_{i-1}a_i$. So, for a blue index i_t and an arbitrary index $i \in \{i_t + 1, \ldots, i_{t+1} - 1\}$ we have

$$u_{i_t} \sim_L u_{i_t}a_{i_t+1} \sim_L \ldots \sim_L u_{i_t}a_{i_t+1} \ldots a_{i-2} \sim_L u_{i-1} \sim_L u_i \ .$$

Since $\sim_L$ is a congruence and $u_{i_t} \sim_L u_{i-1}$ we get $u_{i_t}a_i \sim_L u_i \sim_L u_{i_t}$. Hence we can state the following observation.

Claim 1: Let u' be a subword of the word u which contains all occurrences of letters at the blue positions (and some others). Then $u' \sim_L a_{i_1}a_{i_2}\dots a_{i_r} \sim_L u$.

Moreover, the blue indices denote the leftmost occurrence of the word $u_{\mathsf{left}} = a_{i_1}a_{i_2}\dots a_{i_r}$ as a subword of the word u, i.e. for an arbitrary $r' \leq r$ the word $a_{i_1}\dots a_{i_{r'}}$ is not a subword of the word $u_{i_{r'}-1}$. Since $u \sim_k v$, we can consider also the leftmost occurrence of the word u_{left} in the word v and we denote the appropriate indices by $\overline{i}_1 < \overline{i}_2 < \dots < \overline{i}_r$ from the set $\{1,\dots,q\}$ blue indices in the word v.

Now we use the dual construction for the word v. We consider red indices j such that $b_j v_{j+1} \not\sim_L v_{j+1}$, where v_{j+1} is the suffix of v starting after the j-th letter. For red indices $j_1 < j_2 < \dots < j_s$, $s \leq m-1$, we have the dual property, i.e. they are indices which determine the rightmost occurrence of the word $v_{\mathsf{right}} = b_{j_1}b_{j_2}\dots b_{j_s}$ in v. We consider the rightmost occurrence of v_{right} in u too and we speak about red indices $\overline{j}_1 < \overline{j}_2 < \dots < \overline{j}_s$ in u.

Now we formulate the crucial claim which says that the leftmost occurrence of the word u_{left} and the rightmost occurrence of the word v_{right} are shuffled in the same way in both words u and v.

Claim 2: Let $\overline{u}$ be the subword of u consisting of the occurrences of letters at blue and red positions in u and similarly let $\overline{v}$ be the subword of v consisting of the occurrences of letters at blue and red positions in v. Then $\overline{u} = \overline{v}$.

Proof of Claim 2: Consider the occurrence of $a_{i_{r'}}$ at a blue position $i_{r'}$ and the occurrence of $b_{j_{s'}}$ at a red position $\overline{j}_{s'}$ in u. If $a_{i_{r'}}$ and $b_{j_{s'}} = a_{\overline{j}_{s'}}$ are different letters then $i_{r'} < \overline{j}_{s'}$ if and only if

$$w_{r's'} = a_{i_1}\dots a_{i_{r'}}b_{j_{s'}}\dots b_{j_s} = a_{i_1}\dots a_{i_{r'}}a_{\overline{j}_{s'}}\dots a_{\overline{j}_s}$$

is a subword of u. Indeed, the direct implication is trivial and for the converse implication assume that $w_{r's'} \lhd u$. So, let appropriate indices be $\ell_1 < \dots < \ell_{r'} < \ell'_{s'} < \dots < \ell'_s$, i.e. $w_{r's'} = a_{\ell_1}\dots a_{\ell_{r'}}a_{\ell'_{s'}}\dots a_{\ell'_s}$. Since blue indices $i_1,\dots,i_s$ denote the leftmost occurrence of the word u_{left} in u one can prove by an induction on $t = 1,\dots,r'$ that $i_t \leq \ell_t$. Dually, for $t = s',\dots,s$, we have $\ell'_t \leq \overline{j}_t$. Hence $i_{r'} \leq \ell_{r'} < \ell'_{s'} \leq \overline{j}_{s'}$.

Now assume that $a_{i_{r'}}$ and $b_{j_{s'}}$ are the same letter. We can state the following.
(i) If $w_{r's'}$ is a subword of u then $i_{r'} < \overline{j}_{s'}$.
(ii) If $w_{r's'}$ is not a subword of u, but $\overline{w}_{r's'} = a_{i_1}\dots a_{i_{r'}}b_{j_{s'+1}}\dots b_{j_s} \lhd u$, then $i_{r'} = \overline{j}_{s'}$, i.e. the considered index is both blue and red at the same time.
(iii) If $\overline{w}_{r's'}$ is not a subword of u then $\overline{j}'_s < i_{r'}$.
Altogether we see that the relative position of the considered blue and red indices is given by $\mathsf{Sub}_k(u)$, because all words $w_{r's'}$ and $\overline{w}_{r's'}$ are not longer than $|u_{\mathsf{left}}v_{\mathsf{right}}| \leq 2m-2 = k$. So, the statement of the claim follows from the assumption $u \sim_k v$.

Finally, we proved that $\overline{u} = \overline{v}$ and we have $\overline{u} \sim_L u$ by Claim 1. Similarly, $\overline{v} \sim_L v$, and we can conclude with $u \sim_L \overline{u} = \overline{v} \sim_L v$. □

Note that we slightly improve the original estimate of Simon who proved the inclusion $\sim_k \subseteq \sim_L$ for $k = 2m - 1$. In fact, the parameter m can be the length of the longest chain of ideals in the syntactic monoid (see [4]).

When we put the lemmas together, we obtain the result of Simon.

Theorem 1 (Simon [8]). *Let A be a finite alphabet. Then a regular language L is piecewise testable if and only if $\sim_L$ is a $\mathcal{J}$-trivial congruence on A^*.*

Acknowledgments

The author greatly benefited from an AutoMathA exchange grant to visit the University of Porto.

References

1. J. Almeida, Implicit operations on finite $\mathcal{J}$-trivial semigroups and a conjecture of I. Simon, *J. Pure Appl. Algebra* **69** (1990), 205–218
2. J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, 1994
3. P. Higgins, A proof of Simon's Theorem on piecewise testable languages, *Theoret. Comput. Sci.* **178** (1997), 257–264
4. J.-E. Pin, *Varieties of Formal Languages*, North Oxford Academic, Plenum, 1986
5. J.-E. Pin, Finite semigroups and recognizable languages: an introduction, in *NATO Advanced Study Institute Semigroups, Formal Languages and Groups*, J.Fountain ed., 1–32, Kluwer Academic Publisher, 1995
6. J.-E. Pin, Syntactic semigroups, Chapter 10 in *Handbook of Formal Languages*, G. Rozenberg and A. Salomaa eds, Springer, 1997
7. I. Simon, Hierarchies of events of dot-depth one, Ph.D. thesis, University of Waterloo, 1972
8. I. Simon, Piecewise testable events, in *Proc. ICALP 1975*, LNCS Vol. 33 (1975), 214–222
9. H. Straubing and D. Thérien, Partially ordered finite monoids and a theorem of I. Simon, *J. Algebra* **119** (1988), 393–399