Biautomata for k-Piecewise Testable Languages

Ondřej Klíma and Libor Polák*

Department of Mathematics and Statistics, Masaryk University
Kotlářská 2, 611 37 Brno, Czech Republic
{klima,polak}@math.muni.cz
http://www.math.muni.cz

Abstract. An effective characterization of piecewise testable languages was given by Simon in 1972. A difficult part of the proof is to show that if L has a $\mathcal{J}$-trivial syntactic monoid $\mathsf{M}(L)$ then L is k-piecewise testable for a suitable k. By Simon's original proof, an appropriate k could be taken as two times the maximal length of a chain of ideals in $\mathsf{M}(L)$. In this paper we improve this estimate of k using the concept of biautomaton: a kind of finite automaton which arbitrarily alternates between reading the input word from the left and from the right. We prove that an appropriate k could be taken as the length of the longest simple path in the canonical biautomaton of L. We also show that this bound is better than the known bounds which use the syntactic monoid of L.

Key words: biautomata, k-piecewise testable languages, $\mathcal{J}$-trivial monoids

1 Introduction

A language L over a non-empty finite alphabet A is called piecewise testable if it is a Boolean combination of languages of the form

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

Simon's celebrated theorem [12] states that a regular language L is piecewise testable if and only if the syntactic monoid $\mathsf{M}(L)$ of L is $\mathcal{J}$-trivial. Here we are interested in a finer question, namely to decide, for a given non-negative integer k, the k-piecewise testability, i.e. whether L can be written as a Boolean combination of languages of the form $(*)$ with $\ell \leq k$. Although there exist several proofs of Simon's result based on various methods from algebraic and combinatorial theory of regular languages (e.g. proofs due to Almeida [1], Straubing and Thérien [13], Higgins [4], Klíma [5]; see the survey paper by Pin [9] for more information), little attention has been paid to this problem.

* The authors were supported by the Institute for Theoretical Computer Science (GAP202/12/G061), Czech Science Foundation.

The least k such that a given piecewise testable language L is k-piecewise testable, can be found by brute-force algorithms. The first one uses the fact that for each fixed k and a fixed alphabet A, there are only finitely many k-piecewise testable languages over A. A more sophisticated algorithm can apply Eilenberg's correspondence; it tests whether the syntactic monoid of L belongs to the pseudovariety $\mathbb{J}_k$ of finite monoids corresponding to the variety of all k-piecewise testable languages. But both methods are unrealistic in practice.

A natural question, considering $\mathbb{J}_k$, is the existence of a finite basis of identities for this class of monoids; in the positive case one can test those identities in the syntactic monoids. Such a finite basis exists for $k = 1$ since $\mathbb{J}_1$ is formed by semilattices. Furthermore, Simon [11] and Blanchet-Sadri [2, 3] found finite sets of identities for $\mathbb{J}_2$ and $\mathbb{J}_3$. Unfortunately, it was proved in [2, 3] that a finite basis of identities for $\mathbb{J}_k$ does not exist for $k \geq 4$.

Our ambition, in this paper, is not to decide the k-piecewise testability in a reasonable computational time. Instead of that, for a given piecewise testable language L, we would like to find a good estimate, i.e. a (possibly small) number k, such that L is k-piecewise testable. Such a bound is implicitly contained in the original Simon's proof [12]. Namely, it is shown that k could be taken to be equal to $2n - 1$ where n is the maximal length of a $\mathcal{J}$-chain, i.e. the maximal length of a chain of ideals, in the syntactic monoid of L (see the proof of Corollary 1.7 in [10]). Note that a similar estimate was also established in the first author's combinatorial proof of Simon's result [5]: k could be taken as $\ell + r - 2$ where ℓ and r are the maximal lengths of chains for the orderings $\leq_{\mathcal{L}}$ and $\leq_{\mathcal{R}}$.

In this paper we consider a different proof of Simon's result using a new notion of biautomaton introduced recently by the authors in [7]. The biautomaton is, simply speaking, a finite automaton which arbitrarily alternates between reading the input word from the left and from the right. In the formal definition of a biautomaton there are some compatibility assumptions which ensure that the acceptance of an input does not depend on the way how the input is read. One application of biautomata in [7] gives a characterization of prefix-suffix testable languages. Other result in [7] was an alternative characterization of piecewise testable languages: L is piecewise testable if and only if its canonical biautomaton $\mathcal{C}(L)$ is acyclic. The core of the proof was to show that if $\mathcal{C}(L)$ has m states then L is $2m$-piecewise testable. Here we improve this result in two directions, namely, we eliminate the coefficient 2 and we replace the size of $\mathcal{C}(L)$ by the length of the longest simple path in this acyclic biautomaton, which is called the depth of the biautomaton. The main result of this paper can be phrased as follows.

Theorem 1. *Let L be a piecewise testable language with an (acyclic) canonical biautomaton of depth k. Then L is k-piecewise testable.*

A quite delicate and technical proof of this result is not fully presented here (see appendix in [8]), Section 3 contains only a sketch of this proof. Instead of presenting a complete proof we prefer to add some examples and some additional results. First of all, for each k, we present an easy example of a piecewise testable language which has the canonical biautomaton of depth k and which is not $(k -$

1)-piecewise testable. This shows that the estimate given by Theorem 1 cannot be improved in terms of the depth of the biautomaton. Furthermore, in Section 4 we compare our new estimate with those using the syntactic monoid. We show that the depth of the canonical biautomaton is never larger than the mentioned characteristics $2n-1$ and $\ell+r-2$ for the syntactic monoid of the language. Moreover, we show that there are languages for which these characteristics are arbitrarily larger than the depth of the canonical biautomaton. In the last section of the paper we also establish a lower bound on k using another characteristic of $\mathcal{C}(L)$, namely the length of the shortest simple path from the initial state to an absorbing state.

2 Preliminaries

2.1 Piecewise Testable Languages and Syntactic Monoids

Let A^* be the free monoid over a non-empty finite alphabet A with the neutral element λ; its elements are called *words*. For $u, v \in A^*$, we write $u \triangleleft v$ if u is a *subword* of v, i.e. $u = a_1 \dots a_\ell$, $a_1, \dots, a_\ell \in A$ and there are words $v_0, v_1, \dots, v_\ell \in A^*$ such that $v = v_0 a_1 v_1 \dots a_\ell v_\ell$. Furthermore, for a given word $u \in A^*$, we denote by L_u the language of all words which contain u as a subword, i.e. $L_u = \{\, v \in A^* \mid u \triangleleft v \,\}$. Alternatively, for $u = a_1 \dots a_n$, we can write $L_u = A^* a_1 A^* \dots A^* a_n A^*$. For such u we call n the *length* of the word u, in notation $|u|$, and $\{a_1, \dots, a_n\}$ the *content* of u, in notation $\mathsf{c}(u)$. The complement of a language $L \subseteq A^*$ is denoted by L^{c}.

Definition 1. *A regular language is* k-piecewise testable *if it is a Boolean combination of languages of the form L_u where all u's satisfy $|u| \leq k$. A regular language is piecewise testable if it is* k-piecewise testable *for some k.*

We will use further notation. For $v \in A^*$, we let $\mathsf{Sub}_k(v) = \{\, u \in A^+ \mid u \triangleleft v,\ |u| \leq k \,\}$. We define the equivalence relation $\sim_k$ on A^* by the rule: $u \sim_k v$ if and only if $\mathsf{Sub}_k(u) = \mathsf{Sub}_k(v)$. Note that $\mathsf{Sub}_1(u) = \mathsf{c}(u)$. An easy consequence of the definition of piecewise testable languages is the following lemma. A proof can be found e.g. in [11], [6]. Note that the usual formulation concerns the class of all piecewise testable languages.

Lemma 1. *A language L is k-piecewise testable if and only if L is a union of classes in the partition $A^*/\!\sim_k$.*

Example 1. Let $A = \{a, b\}$. Then $L_{aba} \cup L_{bab} = L_{ab} \cap L_{ba}$ is a 2-piecewise testable language. The language L_{aba} is not 2-piecewise testable because $\mathsf{Sub}_2(abab) = \mathsf{Sub}_2(baab) = A^2$, i.e. $abab \sim_2 baab$ but $abab \in L_{aba}$ and $baab \notin L_{aba}$.

Example 2. For each k, we can consider the word $u = a^k$ over an arbitrary alphabet containing the letter a. Then the language L_u is k-piecewise testable but it is not $(k-1)$-piecewise testable. Indeed, $u = a^k \sim_{k-1} a^{k-1}$, $u \in L_u$ and $a^{k-1} \notin L_u$. Among others, this easy example shows that the classes of k-piecewise testable languages are different for different k's.

In an arbitrary monoid M, we define Green's relations $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{J}$ as follows: for $a, b \in M$, we have $a\mathcal{R}b$ if and only if $aM = bM$, $a\mathcal{L}b$ if and only if $Ma = Mb$, $a\mathcal{J}b$ if and only if $MaM = MbM$. Furthermore, $a \leq_{\mathcal{R}} b$ if and only if $aM \subseteq bM$, $a <_{\mathcal{R}} b$ if and only if $aM \subset bM$. Similarly for $\mathcal{L}$ and $\mathcal{J}$. The monoid M is $\mathcal{J}$-trivial if, for each $a, b \in M$, $a\mathcal{J}b$ implies $a = b$. If, for $a \in M$, we have $MaM = \{a\}$, then a is called a *zero* and it is denoted by 0.

An $\mathcal{R}$-*chain* is a sequence $a_0 <_{\mathcal{R}} a_1 <_{\mathcal{R}} \cdots <_{\mathcal{R}} a_r$. Its *length* is the number $r+1$. The monoid M is of $\mathcal{R}$-*height* r if $r+1$ is the maximal length of an $\mathcal{R}$-chain in M; we write $\mathcal{R}\text{-}\mathsf{height}(M) = r$. Similarly for $\mathcal{L}$ and $\mathcal{J}$.

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

$$u \equiv_L v \quad \text{if and only if} \quad (\,\forall\, p, r \in A^*\,)\,(\,pur \in L \iff pvr \in L\,)\,.$$

The relation $\equiv_L$ is a congruence on A^*; it is called the *syntactic congruence* of L and the quotient structure $\mathsf{M}(L) = A^*/\equiv_L \; = \{\, [u]_{\equiv_L} \mid u \in A^* \,\}$ is the *syntactic monoid* of L. Moreover, the monoid $\mathsf{M}(L)$ is finite whenever L is a regular language. The natural mapping $\eta_L : A^* \to \mathsf{M}(L)$ given by $\eta_L(u) = [u]_{\equiv_L}$, for $u \in A^*$, is called the *syntactic* homomorphism. The language L is a union of certain classes of the partition $A^*/\equiv_L$. If we denote $F = \eta_L(L)$ the set of these classes, then $L = \{\, u \in A^* \mid \eta_L(u) \in F \,\}$. When L is fixed, we will write simply M, $[u]$ and η instead of $\mathsf{M}(L)$, $[u]_{\equiv_L}$ and η_L.

The result by Simon follows.

Theorem 2 (Simon [11, 12]). *A regular language L is piecewise testable if and only if its syntactic monoid $\mathsf{M}(L)$ is $\mathcal{J}$-trivial.*

We also mention two results which are proved in Corollary 1.7 in [10] and in the second author's paper [5] respectively.

Proposition 1 ([10],[5]). *Let L be a piecewise testable language with syntactic monoid $\mathsf{M}(L)$. Then L is k-piecewise testable for $k = 2 \cdot \mathcal{J}\text{-}\mathsf{height}(\mathsf{M}(L)) + 1$ and also for $k = \mathcal{R}\text{-}\mathsf{height}(\mathsf{M}(L)) + \mathcal{L}\text{-}\mathsf{height}(\mathsf{M}(L))$.*

Note that the relation $\mathcal{R}\text{-}\mathsf{height}(\mathsf{M}(L)) + \mathcal{L}\text{-}\mathsf{height}(\mathsf{M}(L)) \leq 2 \cdot \mathcal{J}\text{-}\mathsf{height}(\mathsf{M}(L))$ is obvious.

2.2 Biautomata for Piecewise Testable Languages

The authors' paper [7] initialized the study of biautomata. We recall now the basic notions and results which we will need here.

Definition 2. *A biautomaton over a non-empty finite alphabet A is a six-tuple $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ where*

- *Q is a non-empty set of states,*
- *$\cdot : Q \times A \to Q$, extended to $\cdot : Q \times A^* \to Q$ by $q \cdot \lambda = q$, $q \cdot (ua) = (q \cdot u) \cdot a$, where $q \in Q$, $u \in A^*$, $a \in A$,*

- $\circ : Q \times A \to Q$, *extended to* $\circ : Q \times A^* \to Q$ *by* $q \circ \lambda = q$, $q \circ (av) = (q \circ v) \circ a$, *where* $q \in Q$, $v \in A^*$, $a \in A$ *(such actions are marked by dotted lines in diagrams),*
- $i \in Q$ *is the initial state,*
- $T \subseteq Q$ *is the set of terminal states,*
- *for each* $q \in Q$, $a, b \in A$, *we have* $(q \cdot a) \circ b = (q \circ b) \cdot a$,
- *for each* $q \in Q$, $a \in A$, *we have* $q \cdot a \in T$ *if and only if* $q \circ a \in T$.

The language recognized *by* $\mathcal{B}$ *is the regular language* $L_{\mathcal{B}} = \{ u \in A^* \mid i \cdot u \in T \}$.

The following two properties, which generalize the last conditions in the definition, follow immediately (see [7], Lemma 2.2).

- For each $q \in Q$, $u, v \in A^*$, we have $(q \cdot u) \circ v = (q \circ v) \cdot u$.
- For each $q \in Q$, $u \in A^*$, we have $q \cdot u \in T$ if and only if $q \circ u \in T$.

A crucial property is the following lemma which says that to decide whether $u \in L_{\mathcal{B}}$ it is possible to consider an arbitrary reading of u in $\mathcal{B}$.

Lemma 2 ([7], Lemma 2.3). *Having a biautomaton* $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$, $p \in Q$ *and* $u \in A^+$, *dividing* $u = u_1 \dots u_k v_k \dots v_1$ *arbitrarily,* $u_1, \dots, u_k, v_k, \dots, v_1 \in A^*$, *when reading from* p, *the words* u_1 *first, then* v_1, *then* u_2, *and so on, i.e. we move from* p *to the state* $q = ((\dots((((p \cdot u_1) \circ v_1) \cdot u_2) \circ v_2) \dots) \cdot u_k) \circ v_k$, *then* $q \in T$ *if and only if* $p \cdot u \in T$.

For our propose, we recall the basic construction from [7].

Definition 3. *For a regular language* $L \subseteq A^*$ *and* $u, v \in A^*$, *we put* $u^{-1}Lv^{-1} = \{ w \in A^* \mid uwv \in L \}$ *and* $C = \{ u^{-1}Lv^{-1} \mid u, v \in A^* \}$. *We define* $\mathcal{C}(L) = (C, A, \cdot, \circ, L, T)$, *where* $q \cdot a = a^{-1}q$, $q \circ a = qa^{-1}$ *and* $T = \{ u^{-1}Lv^{-1} \mid \lambda \in u^{-1}Lv^{-1} \}$.

The structure $\mathcal{C}(L)$ is a biautomaton recognizing L and it is called the *canonical biautomaton* of the language L. A useful property of $\mathcal{C}(L)$ is that all states are reachable (from the initial state). More formally, we say that a state $q \in Q$ of a biautomaton $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ is *reachable* if there is a pair of words $u, v \in A^*$ such that $(i \cdot u) \circ v = q$. For an arbitrary state $p \in Q$, we denote $Q_p = \{ (p \cdot u) \circ v \mid u, v \in A^* \}$ and we put $\mathcal{B}_p = (Q_p, A, \cdot, \circ, p, T)$. This definition is correct because, for $u, v \in A^*$ and $a \in A$, we have $((p \cdot u) \circ v) \circ a = (p \cdot u) \circ av \in Q_p$ and $((p \cdot u) \circ v) \cdot a = ((p \cdot u) \cdot a) \circ v = (p \cdot ua) \circ v \in Q_p$. Hence $\mathcal{B}_p$ is a biautomaton with all states reachable.

Example 3. [Continuation of Example 1] The canonical biautomaton of $L_{aba} \cup L_{bab}$ is depicted in Figure 1 and the canonical biautomaton of L_{aba} is depicted in Figure 2. We are using the construction described in Definition 3. Note that both biautomata are very similar; in fact, the only difference is how the letters act on the initial state. But as we saw in Example 1, the first is 2-piecewise testable and the second one is not.

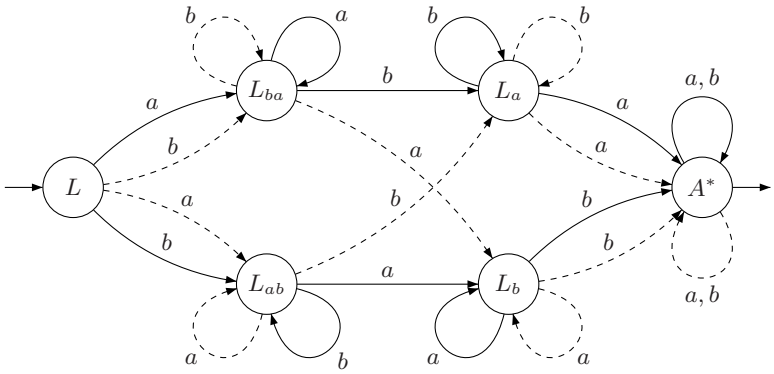

Fig. 1. The canonical biautomaton of the language $L = L_{aba} \cup L_{bab}$.

Let $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ be a biautomaton. A sequence $(q_0, q_1, \ldots, q_n)$ of states is called a *path* in $\mathcal{B}$ if for each $j \in \{1, \ldots, n\}$ there is $a_j \in A$ such that $q_j = q_{j-1} *_j a_j$ where $*_j$ is $\cdot$ or $\circ$. A path $(q_0, q_1, \ldots, q_n)$ is *simple* if the states $q_0, \ldots, q_n$ are pairwise different and it is a *cycle* if $n \geq 2$ and $q_n = q_0 \neq q_1$. The biautomaton $\mathcal{B}$ is called *acyclic* if there is no cycle in $\mathcal{B}$. Note that "loops" are not cycles and for the acyclic biautomaton $\mathcal{B}$, each biautomaton $\mathcal{B}_p$ is also acyclic.

The first major application of biautomata was the following statement.

Theorem 3 ([7]). *Let $L \subseteq A^*$ be a regular language. Then L is piecewise testable if and only if the canonical biautomaton of L is acyclic.*

We say that the state q of a biautomaton $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ is *absorbing* if, for every $a \in A$, we have $q \cdot a = q \circ a = q$. It is clear that in each acyclic biautomaton there is some absorbing state and every simple path can be prolonged to such a state. Furthermore, each simple path in a biautomaton with all states reachable can be prolonged in such a way that it starts in i. We define the following two characteristics of an acyclic biautomaton $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ with all states reachable. The *depth* of $\mathcal{B}$, $\mathsf{depth}(\mathcal{B})$ in notation, is the maximal number n such that there is a simple path $(i, q_1, \ldots, q_n)$ in $\mathcal{B}$ where q_n is an absorbing state. Similarly, $\mathsf{diam}(\mathcal{B})$ is the minimal number n for which such simple path exists. We call this characteristic the *diameter* of $\mathcal{B}$.

Example 4 (Continuation of Examples 1 and 3). For both biautomata in Figure 1 and 2 we have $\mathsf{depth}(\mathcal{B}) = \mathsf{diam}(\mathcal{B}) = 3$.

Example 5 (Continuation of Example 2). If $u = a^k$, then it is not hard to see that states in $\mathcal{C}(L_u)$ are exactly L_{a^ℓ} where $\ell \leq k$. In particular, there is the unique terminal state $L_{a^0} = L_\lambda = A^*$ which is also the unique absorbing state.

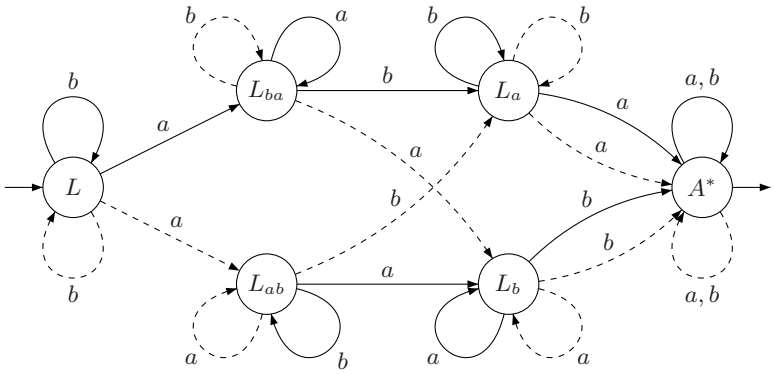

Fig. 2. The canonical biautomaton of the language $L = L_{aba}$.

For each $0 < \ell \leq k$, we have $L_{a^\ell} \cdot a = L_{a^\ell} \circ a = L_{a^{\ell-1}}$ and $L_{a^\ell} \cdot b = L_{a^\ell} \circ b = L_{a^\ell}$ for each letter $b \neq a$. Hence $\mathsf{depth}(\mathcal{C}(L_u)) = k$.

The previous example shows that the estimate given by Theorem 1 is in some sense optimal (at least in terms of the depth of the biautomaton).

3 Proof of the theorem

Due to the space limitation, a complete proof is situated in Appendix. Here we just try to explain the main idea and techniques of the proof, which are, in fact, the same as in the original proof of Theorem 3. But the proof is more delicate here since it is built from weaker assumptions. Basically, the statement of Theorem 1 is a consequence of the following proposition and Lemma 1.

Proposition 2. *Let $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ be an acyclic biautomaton with all states reachable and with $\mathsf{depth}(\mathcal{B}) = \ell$. Then, for every $u, v \in A^*$, such that $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$, we have*

$$u \in L_{\mathcal{B}} \quad \textit{if and only if} \quad v \in L_{\mathcal{B}}\,.$$

Proof (Sketch.). We prove the statement by induction with respect to ℓ in such a way that the induction assumption will be applied on subbiautomata $\mathcal{B}_p$'s of the biautomaton $\mathcal{B}$ which have smaller depth whenever $p \neq i$. One can find a complete discussion of cases $\ell = 0$ and $\ell = 1$ in Appendix.

Assume that $\ell \geq 2$ and that the statement holds for all $\ell' < \ell$, and assume that it is not true for ℓ. Then there is a pair of words $u, v \in A^*$ such that

$$\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v) \text{ and } i \cdot u \in T \text{ and } i \cdot v \notin T\ . \tag{1}$$

We will show that these assumptions lead to a contradiction.

Our complete proof consists of numerous steps. At each stage we have certain set of assumptions and we are adding a new one to them. After a detailed analysis we show that this new additional assumption leads to a contradiction. This means we could add the negation of the last assumption to our actual family of assumptions and consider this new family in the next stage. At the end of the process we will have enough strong assumptions which will lead to a final contradiction. This process is demonstrated here by the beginning of the detailed proof together with one (quite significant and typical) step.

In the state i, we read from the left both words u and v, and we are interested in the positions in the words u and v where we leave the initial state i. First assume that $i \cdot u = i$, i.e. we do not leave the state i. Recall that the assumption $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$ implies $\mathsf{c}(u) = \mathsf{c}(v)$. Thus we have $i \cdot v = i \in T$ – a contradiction. From this moment we may assume that

$$i \cdot u \neq i \text{ and also } i \cdot v \neq i, \text{ and moreover dually, } i \circ u \neq i \text{ and } i \circ v \neq i \ . \quad (2)$$

So we really leave the state i and there are $u', u'' \in A^*$, $a \in A$ such that

$$u = u'au'', \text{ for each } x \in \mathsf{c}(u') \text{ we have } i \cdot x = i, \text{ and } i \cdot a \neq i, \ a \notin \mathsf{c}(u') \ . \quad (3)$$

Similarly, let $v', v'' \in A^*$, $b \in A$ be such that

$$v = v'bv'', \text{ for each } x \in \mathsf{c}(v') \text{ we have } i \cdot x = i, \text{ and } i \cdot b \neq i, \ b \notin \mathsf{c}(v') \ . \quad (4)$$

The assumption $a = b$ leads to a contradiction (see the full version for the argument) and therefore we may assume that

$$a \neq b \ . \quad (5)$$

Let us assume, for a moment, that $i \cdot a = i \cdot b = p$. We will consider the first occurrence of b in u. Since in the biautomaton $\mathcal{B}$, when we read u from the left we move from the initial state by a, it is clear that the first occurrence of b in u is behind the first occurrence of a in u. More formally, $u = u'au_0''bu_1''$ where $a \notin \mathsf{c}(u')$ and $b \notin \mathsf{c}(u'au_0'')$. Similarly, $v = v'bv_0''av_1''$ where $b \notin \mathsf{c}(v')$ and $a \notin \mathsf{c}(v'bv_0'')$.

Now from the assumption (1), i.e. $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$, and since mentioned occurrences of a and b are the first occurrences of these letters in u and v we get $\mathsf{Sub}_{\ell-1}(u_0''bu_1'') = \mathsf{Sub}_{\ell-1}(v_1'')$. Indeed, if $w \in \mathsf{Sub}_{\ell-1}(u_0''bu_1'')$ then $aw \in \mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$ from which we obtain $w \in \mathsf{Sub}_{\ell-1}(v_1'')$. One proves the opposite inclusion similarly. Thus we can deduce that $\mathsf{Sub}_{\ell-1}(u_0''bu_1'') = \mathsf{Sub}_{\ell-1}(v_1'') \subseteq \mathsf{Sub}_{\ell-1}(v_0''av_1'') = \mathsf{Sub}_{\ell-1}(u_1'') \subseteq \mathsf{Sub}_{\ell-1}(u_0''bu_1'')$. Therefore $\mathsf{Sub}_{\ell-1}(u_0''bu_1'') = \mathsf{Sub}_{\ell-1}(v_0''av_1'')$. We have $i \cdot u = p \cdot u_0''bu_1'' \in T$ and $i \cdot v = p \cdot v_0''av_1'' \notin T$. This is a contradiction to the induction assumption applying to the biautomaton $\mathcal{B}_p$ and the pair of words $u_0''bu_1'$ and $v_0''av_1''$. Altogether we have that $i \cdot a \neq i \cdot b$ and we can add this formula to the actual set of assumptions.

Then we continue in the way described above. Note that in other steps we need to discuss more complicated situations. For example, we consider also the

positions in the words, where we leave the initial state i when we read both words u and v from the right. This leads (in one case) to factorizations of words u and v of the form $u = u_1au_2bu_3du_4cu_5$ and $v = v_1bv_2av_3cv_4dv_5$, where mentioned occurrences of a and b are the first occurrences of these letters and the mentioned occurrences of c and d are the last occurrences of these letters. Then one uses the real power of the notion of biautomata because we read u in such a way that we read u_1a from the left first and then cu_5 from the right. This means we move to a certain state p for which $\mathcal{B}_p$ has depth at most $\ell - 2$ (which is ensured by certain additional assumptions added during the proof). Then the induction assumption is applied on this p (and, in fact, to certain other states which must be considered in this case). □

4 Estimates using $\mathcal{J}$-trivial monoids

We compare the estimates form Theorem 1 with those from Proposition 1.

Proposition 3. *Let L be a piecewise testable language and let $\mathsf{M}(L)$ be its ($\mathcal{J}$-trivial) syntactic monoid and $\mathcal{C}(L)$ be its (acyclic) canonical biautomaton. Then*

$$\mathsf{depth}(\mathcal{C}(L)) \leq \mathcal{R}\text{-}\mathsf{height}(\mathsf{M}(L)) + \mathcal{L}\text{-}\mathsf{height}(\mathsf{M}(L)) \leq 2 \cdot \mathcal{J}\text{-}\mathsf{height}(\mathsf{M}(L)) \ .$$

Proof. The second inequality is trivially satisfied in every monoid. We use the construction of a biautomaton from a monoid described in Remark 2.10 in [7]. Let $M = \mathsf{M}(L)$ be a syntactic ($\mathcal{J}$-trivial) monoid of a piecewise testable language L and $\eta : A^* \to M,\ u \mapsto [u]$ be the syntactic homomorphism and let $F = \eta(L)$. Then the biautomaton $\mathcal{B}_\eta = (B_\eta, A, \cdot, \circ, i, T)$, where

- $B_\eta = M \times M$,
- for every $a \in A$ and $p, r \in M$, we set $(p, r) \cdot a = (p[a], r)$,
- for every $a \in A$ and $p, r \in M$, we set $(p, r) \circ a = (p, [a]r)$,
- $i = ([\lambda], [\lambda]) = (1, 1)$,
- $T = \{\, (p, r) \mid pr \in F \,\}$,

recognizes L. Since M is $\mathcal{J}$-trivial the biautomaton $\mathcal{B}_\eta$ is acyclic. Moreover, $\mathcal{C}(L)$ has minimum depth among all biautomata recognizing L (see [7], Section 2.4), so it is enough to prove that $\mathsf{depth}(\mathcal{B}_\eta) \leq \mathcal{R}\text{-}\mathsf{height}(M) + \mathcal{L}\text{-}\mathsf{height}(M)$. Now, assume that $\mathsf{depth}(\mathcal{B}_\eta) = k$. Thus there is a simple path $(q_0 = i, q_1, q_2, \ldots, q_k)$ in $\mathcal{B}_\eta$. In particular, for each $j \in \{1, \ldots, n\}$ we have $q_{j-1} \neq q_j$ and there is $a_j \in A$ such that $q_{j-1} *_j a_j = q_j$, where $*_j$ is $\cdot$ or $\circ$. Let $q_j = (m_j, n_j)$ for $j = 0, \ldots, k$.

Now we have $1 \geq_{\mathcal{R}} m_1 \geq_{\mathcal{R}} m_2 \geq_{\mathcal{R}} \cdots \geq_{\mathcal{R}} m_k$ and $1 \geq_{\mathcal{L}} n_1 \geq_{\mathcal{L}} n_2 \geq_{\mathcal{L}} \cdots \geq_{\mathcal{L}} n_k$. For each j, there are two possibilities: 1) $m_{j-1} = m_j$ and $n_{j-1} \neq n_j$, i.e. $n_{j-1} >_{\mathcal{L}} n_i$ or 2) $n_{j-1} = n_j$ and $m_{j-1} \neq m_j$. So, if we omit repeated occurrences of elements of M in the sequences $(1, m_1, \ldots, m_k)$ (and $(1, n_1, \ldots, n_k)$ respectively) we obtain the chains of $>_{\mathcal{R}}$ (and $>_{\mathcal{L}}$ respectively) related elements. Thus $k \leq \mathcal{R}\text{-}\mathsf{height}(M) + \mathcal{L}\text{-}\mathsf{height}(M)$ and the statement follows. □

In the following example we demonstrate that the described inequalities are strict for some languages. Namely, for each integer n, we find a language L (over the alphabet having $3n$ letters) such that its canonical biautomaton has depth 4 but $\mathcal{J}$-$\mathsf{height}(\mathsf{M}(L))$ is at least n.

Example 6. For an arbitrary n, we denote $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_n\}$ and $C = \{c_1, \ldots, c_n\}$ (altogether $3n$ pairwise different letters). Let K be the language of all words which does not contain neither two letters from the subalphabet A nor two letters from the subalphabet C. More formally, K is a 2-piecewise testable language given by the following expression

$$K = \bigcap_{i,j=1}^{n} L^{\mathsf{c}}_{a_i a_j} \cap \bigcap_{i,j=1}^{n} L^{\mathsf{c}}_{c_i c_j}\,.$$

For each $i = 1, \ldots, n$, we put $L_i = L_{a_i b_i c_i} \cap K$ and we define $L = \bigcup_{i=1}^{n} L_i$.

Deciding, for a given $u \in A^*$, whether $u \in L$ using biautomaton $\mathcal{C}(L)$, we can ignore b's from the left and from the right (i.e. no non-trivial moves in $\mathcal{C}(L)$, in fact we are staying in the initial state L) until we read some a_i form the left or some c_i from the right. Then the index i is fixed and $u \in L$ if and only if $u \in L_i$. The last condition is checked in $\mathcal{C}(L_i)$. To illustrate further computation, we assume that $i = 1$ and we describe the biautomaton $\mathcal{C}(L_1)$. Its part is depicted in Figure 3. First, all letters from B act identically on all states, with the exception of the states p_a, p_c and p_{ac} where only letters from B different from b_1 act identically, and b_1 acts as depicted. Secondly, all actions by letters from $A \cup C$ which are not shown in the figure move from a state to the unique non-terminal absorbing state $\emptyset$ which is not on the image. We get a decision whether $u \in L_1$ using at most three non-trivial moves. The canonical biautomaton of the language L can be seen as the union of n copies of the biautomata for L_i's where the initial states are merged to the initial state L and all non-terminal absorbing states are also merged. We see that $\mathsf{depth}(\mathcal{C}(L)) = \mathsf{depth}(\mathcal{C}(L_i)) = 4$.

On the other hand, in the syntactic monoid of L, there is a $\mathcal{J}$-chain:

$$1 >_{\mathcal{R}} [b_1] >_{\mathcal{R}} \cdots >_{\mathcal{R}} [b_1 \ldots b_n] >_{\mathcal{R}} [b_1 \ldots b_n c_1] >_{\mathcal{L}} [a_1 b_1 \ldots b_n c_1] >_{\mathcal{R}} 0.$$

Indeed, let $v_i = b_1 \ldots b_i$ for $i = 0, \ldots, n$. Then $a_{i+1} \cdot v_i \cdot c_{i+1} \notin L$ and $a_{i+1} \cdot v_{i+1} \cdot c_{i+1} \in L$. Therefore $[v_i] \neq [v_{i+1}]$ and $[v_i] >_{\mathcal{R}} [v_{i+1}]$ follows from the $\mathcal{J}$-triviality of $\mathsf{M}(L)$. The last three relations are obtained similarly.

Hence $\mathcal{J}$-$\mathsf{height}(\mathsf{M}(L)) \geq n+3$. In fact, one can show that $\mathcal{J}$-$\mathsf{height}(\mathsf{M}(L)) = n + 3$.

5 Concluding Remarks

The goal of this paper was to give, for a piecewise testable language L, a good estimate of the minimum number k such that L is k-piecewise testable. The estimate from Theorem 1 is a tight upper bound in terms of the depth of the

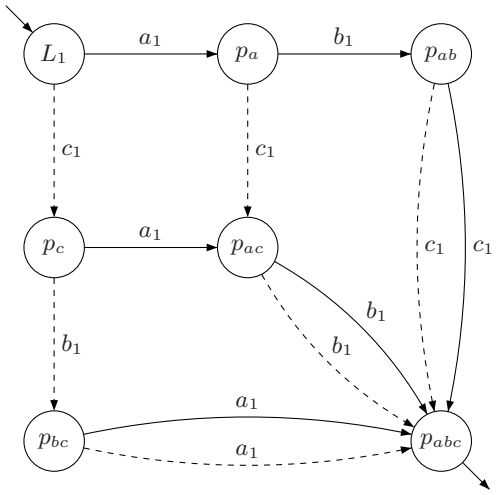

Fig. 3. A part of $\mathcal{C}(L_1)$ where $L_1 = L_{a_1b_1c_1} \cap \bigcap_{i,j} L^{\mathsf{c}}_{a_ia_j} \cap \bigcap_{i,j} L^{\mathsf{c}}_{c_ic_j}$.

canonical biautomaton of the language L as we saw in Examples 2 and 5. We also saw in Section 4 that the estimate from Theorem 1 is better than those from Proposition 1. But we should say that we are still far from the optimal value of k because there are languages for which depth of the canonical biautomaton is larger then the optimal k as we demonstrate in the following example.

Example 7. Let $A = \{a_1, a_2, \ldots, a_n\}$ and let ℓ be an integer. We consider $L = \{a_1^\ell a_2^\ell \ldots a_n^\ell\}$ consisting of a single word. One can easily check that this language is given by the expression

$$L = \bigcap_{i=1}^{n} L_{a_i^\ell} \cap \bigcap_{i=1}^{n} L^{\mathsf{c}}_{a_i^{\ell+1}} \cap \bigcap_{i<j} L^{\mathsf{c}}_{a_ja_i}\,.$$

In particular, L is a $(\ell+1)$-piecewise testable language. It is clear that L is not ℓ-piecewise testable, because $u = a_1^\ell a_2^\ell \ldots a_n^\ell \in L$, $v = a_1^{\ell+1} a_2^\ell \ldots a_n^\ell \notin L$ and $u \sim_\ell v$.

If we consider the canonical biautomaton of L, then each state, as a language, is $u^{-1}Lv^{-1}$, where $u, v \in A^*$, and it consists of at most one word. Thus one can see that $\mathsf{depth}(\mathcal{B}(L)) = \ell \cdot n + 1$.

The existence of languages like in the previous example requests the need of some lower bounds for k-piecewise testability. The first attempt is the content of the following result.

Proposition 4. *Let L be a piecewise testable language over the n-element alphabet. If $kn < \mathsf{diam}(\mathcal{C}(L))$, then L is not k-piecewise testable.*

Proof. Let $A = \{a_1, \ldots, a_n\}$. We prove the statement by induction with respect to $\mathsf{diam}(\mathcal{C}(L))$. For $\mathsf{diam}(\mathcal{C}(L)) = 0$ there is nothing to prove and therefore assume that $1 \leq \mathsf{diam}(\mathcal{C}(L)) \leq n$ and $kn < \mathsf{diam}(\mathcal{C}(L))$. Then $k = 0$ and it is clear that 0-piecewise testable languages over the alphabet A are just A^* and $\emptyset$ which both have trivial canonical biautomata, i.e. $\mathsf{diam}(\mathcal{C}(L)) = 0$ – a contradiction. Thus we have proved the statement for each L such that $\mathsf{diam}(\mathcal{C}(L)) \leq n$.

Let $s = \mathsf{diam}(\mathcal{C}(L)) > n$ and let k be an arbitrary number such that $kn < s$. We can look at $q = i{\cdot}a_1a_2 \ldots a_n$. Let $\mathcal{B}_q$ be the subbiautomaton of $\mathcal{C}(L)$ consisting from all states reachable from the state q. Then $\mathsf{diam}(\mathcal{B}_q) \geq s - n$ and by the induction assumption the language L' recognized by the biautomaton $\mathcal{B}_q$ is not $(k-1)$-piecewise testable language. This means that there is a pair of words u', v' such that $u' \in L'$, $v' \notin L'$ and $u' \sim_{k-1} v'$. Now we consider $u = a_1a_2 \ldots a_nu'$ and $v = a_1a_2 \ldots a_nv'$ for which we claim that $u \sim_k v$. Indeed, since the prefix $a_1 \ldots a_n$ of u contains all letters, we see that each word w, satisfying $w \in \mathsf{Sub}_k(u)$, can be factorized in two parts $w = w_0w_1$ in such a way that $w_0 \in \mathsf{Sub}_k(a_1 \ldots a_n)$ and $w_1 \in \mathsf{Sub}_{k-1}(u')$. Thus $w_1 \in \mathsf{Sub}_{k-1}(v')$ and we can conclude $w \in \mathsf{Sub}_k(v)$. Hence we have a pair of words u, v such that $u \sim_k v$, $u \in L$ and $v \notin L$, which implies that L is not k-piecewise testable by Lemma 1. □

References

1. Almeida, J.: Implicit operations on finite $\mathcal{J}$-trivial semigroups and a conjecture of I. Simon. J. Pure Appl. Algebra 69 (1990), 205–218
2. Blanchet-Sadri, F.: Games, equations and the dot-depth hierarchy. Comput. Math. Appl. 18 (1989), 809–822
3. Blanchet-Sadri, F.: Equations and monoids varieties of dot-depth one and two. Theoret. Comput. Sci. 123 (1994), 239–258
4. Higgins, P.: A proof of Simon's Theorem on piecewise testable languages. Theoret. Comput. Sci. 178 (1997), 257–264
5. Klíma, O.: Piecewise testable languages via combinatorics on words. Discrete Mathematics 311 (2011), 2124–2127
6. Klíma, O., Polák, L.: Hierarchies of piecewise testable languages. International Journal of Foundations of Computer Science 21 (2010), 517–533
7. Klíma, O., Polák, L.: On biautomata. To appear in RAIRO, available at `http://math.muni.cz/~klima/Math/publications.html` (previous version: Non-Classical Models for Automata and Applications, NCMA 2011, 153–164)
8. Klíma, O., Polák, L.: present paper with appendix, available at `http://math.muni.cz/~klima/Math/publications.html`
9. Pin, J.-E.: Syntactic semigroups, Chapter 10 in Handbook of Formal Languages, G. Rozenberg and A. Salomaa eds, Springer, 1997
10. Pin, J.-E.: Varieties of Formal Languages, North Oxford Academic, Plenum, 1986
11. Simon, I.: Hierarchies of events of dot-depth one. Ph.D. thesis. U. Waterloo, (1972)
12. Simon, I.: Piecewise testable events. In Proc. ICALP 1975 LNCS 33 (1975), 214–222
13. Straubing H., Thérien, D.: Partially ordered finite monoids and a theorem of I. Simon, J. Algebra 119 (1988), 393–399

Appendix

Here we present a complete and detailed proof of Proposition 2. We prove a few technical lemmas first.

Lemma 3. *Let $\ell \geq 1$ and let $u, v \in A^*$ be such that $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$. Let a be a letter from $\mathsf{c}(u)$. Then there are uniquely determined words u', u'' and v', v'' such that $u = u'au''$ and $v = v'av''$ and $a \notin \mathsf{c}(u')$, $a \notin \mathsf{c}(v')$. Moreover, for u'' and v'' we have $\mathsf{Sub}_{\ell-1}(u'') = \mathsf{Sub}_{\ell-1}(v'')$.*

Proof. Note that the equality $\mathsf{c}(u) = \mathsf{c}(v)$ follows from the assumption $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$. We consider the first occurrence of the letter a in u and in v respectively and hence u' and v' respectively must be a factor before this first occurrence of a. Thus we have $u = u'au''$ and $v = v'av''$ satisfying $a \notin \mathsf{c}(u')$ and $a \notin \mathsf{c}(v')$. Now if $w \in \mathsf{Sub}_{\ell-1}(u'')$, then $aw \in \mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$ from which $w \in \mathsf{Sub}_{\ell-1}(v'')$ follows. This means that $\mathsf{Sub}_{\ell-1}(u'') \subseteq \mathsf{Sub}_{\ell-1}(v'')$ and the opposite inclusion can be proved in the same way. □

We will also use the dual version of the previous lemma where instead of the first occurrences of a letter in words u and v we consider the last occurrences. Furthermore, in the same way, one can prove the following two-sided version of Lemma 3.

Lemma 4. *Let $\ell \geq 2$ and $u, v \in A^*$ be such that $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$. Let $a, d \in A$ be letters such that $ad \in \mathsf{Sub}_\ell(u)$. Then there are words $u_1, u_2, u_3, v_1, v_2, v_3 \in A^*$ satisfying $u = u_1au_2du_3$ and $v = v_1av_2dv_3$ where $a \notin \mathsf{c}(u_1)$, $a \notin \mathsf{c}(v_1)$, $d \notin \mathsf{c}(u_3)$ and $d \notin \mathsf{c}(v_3)$. Moreover, we have $\mathsf{Sub}_{\ell-2}(u_2) = \mathsf{Sub}_{\ell-2}(v_2)$.* □

Recall the formulation of Proposition 2.

Proposition 2. *Let $\mathcal{B} = (Q, A, \cdot, \circ, i, T)$ be an acyclic biautomaton with all states reachable and with $\mathsf{depth}(\mathcal{B}) = \ell$. Then, for every $u, v \in A^*$, such that $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$, we have*

$$u \in L_{\mathcal{B}} \iff v \in L_{\mathcal{B}}\,.$$

Proof. At the beginning of the proof we recall that the condition $u \in L_{\mathcal{B}}$ means that $i \cdot u \in T$, but we can also consider other possible reading of u in $\mathcal{B}$ and we must again finish in a terminal state. This property is used often in what follows without special references.

We prove the statement by induction with respect to the depth of the biautomaton.

For $\ell = 0$, there is nothing to prove, because the assumption $\mathsf{depth}(\mathcal{B}) = 0$ means that $\mathcal{B}$ is a trivial biautomaton, i.e. $Q = \{i\}$, and hence $L_{\mathcal{B}} = A^*$ or $L_{\mathcal{B}} = \emptyset$, depending on the fact whether $i \in T$ or not.

Now let $\mathsf{depth}(\mathcal{B}) = \ell = 1$. Since each state is reachable from i by a simple path of length at most 1 we have that, for each $q \in Q \setminus \{i\}$, that $q = i \cdot a$ or $q = i \circ a$ for some $a \in A$. Moreover, every state $q \in Q \setminus \{i\}$ is absorbing. If, for $a, b \in A$,

$p = i \cdot a \neq i$ and $q = i \circ b \neq i$ then $p = p \circ b = (i \cdot a) \circ b = (i \circ b) \cdot a = q \cdot a = q$. Thus there are only two possible cases: there is only one state in $Q \setminus \{i\}$ or there are more than two states in $Q \setminus \{i\}$, but then all actions leading to different states contained in $Q \setminus \{i\}$ are all given by $\cdot$ or by $\circ$. The second possibility means that either, for all $a \in A$, we have $i \circ a = i$ or, for all $a \in A$, we have $i \cdot a = i$. In this case we have $L_{\mathcal{B}} = A^*$ or $L_{\mathcal{B}} = \emptyset$, depending on the fact whether $i \in T$ or not. Let us consider the first case when $Q = \{i, q\}$, with q being an absorbing state. If $T = Q$ or $T = \emptyset$ then again $L_{\mathcal{B}} = A^*$ or $L_{\mathcal{B}} = \emptyset$. Thus we can assume that $i \notin T$ and $q \in T$, because the other possibility describes the complementary language. We denote $C = \{\, a \in A \mid i \cdot a = q \,\}$ and we see that $L_{\mathcal{B}} = \bigcup_{a \in C} L_a$. Since this language is 1-piecewise testable, Lemma 1 completes the proof.

Assume for the rest of the proof that $\ell \geq 2$ and that the statement holds for all $\ell' < \ell$. And furthermore, assume that the statement is not true for ℓ. We will reach a contradiction by strengthening our assumptions. Let there be a pair of words $u, v \in A^*$ such that

$$\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v) \text{ and } i \cdot u \in T \text{ and } i \cdot v \notin T \ . \tag{1}$$

In the state i, we read from the left both words u and v, and we are interested in the positions in the words, where we leave the initial state i. First assume that $i \cdot u = i \in T$, i.e. we do not leave the state i. Recall that the assumption $\mathsf{Sub}_\ell(u) = \mathsf{Sub}_\ell(v)$ implies $\mathsf{c}(u) = \mathsf{c}(v)$. Thus we have $i \cdot v = i \in T$ – a contradiction. From this moment we may assume that

$$i \cdot u \neq i \text{ and also } i \cdot v \neq i, \text{ and moreover dually, } i \circ u \neq i \text{ and } i \circ v \neq i \ . \tag{2}$$

So we really leave the state i and there are $u', u'' \in A^*$, $a \in A$ such that

$$u = u'au'', \text{ for each } x \in \mathsf{c}(u') \text{ we have } i \cdot x = i, \text{ and } i \cdot a \neq i, \ a \notin \mathsf{c}(u') \ . \tag{3}$$

Similarly, let $v', v'' \in A^*$, $b \in A$ be such that

$$v = v'bv'', \text{ for each } x \in \mathsf{c}(v'), \text{ we have } i \cdot x = i \text{ and } i \cdot b \neq i, \ b \notin \mathsf{c}(v') \ . \tag{4}$$

Assume for a moment that $a = b$. We denote $p = i \cdot a = i \cdot u'a = i \cdot v'a$ and we consider the biautomaton $\mathcal{B}_p$. It is clear that the depth of $\mathcal{B}_p$ is at most $\ell - 1$. By our assumptions $i \cdot u = p \cdot u'' \in T$ and $i \cdot v = p \cdot v'' \notin T$. By Lemma 3 we have $\mathsf{Sub}_{\ell-1}(u'') = \mathsf{Sub}_{\ell-1}(v'')$. Now we obtain a contradiction to the induction assumption applied on the biautomaton $\mathcal{B}_p$ and the pair of words u'' and v''. Therefore we may assume that

$$a \neq b \ . \tag{5}$$

We will consider the first occurrence of b in u. When we read u (in the biautomaton $\mathcal{B}$) from the left, we move from the initial state only when we reach the letter a for the first time. Therefore the first occurrence of b in u is after the first occurrence of a in u. More formally,

$$u = u'au_0''bu_1'' \text{ where } a \notin \mathsf{c}(u') \text{ and } b \notin \mathsf{c}(u'au_0'') \ . \tag{6}$$

Similarly,

$$v = v'bv_0''av_1'' \text{ where } b \notin \mathsf{c}(v') \text{ and } a \notin \mathsf{c}(v'bv_0'')\,. \tag{7}$$

Now, by Lemma 3, we have $\mathsf{Sub}_{\ell-1}(u_0''bu_1'') = \mathsf{Sub}_{\ell-1}(v_1'') \subseteq \mathsf{Sub}_{\ell-1}(v_0''av_1'') = \mathsf{Sub}_{\ell-1}(u_1'') \subseteq \mathsf{Sub}_{\ell-1}(u_0''bu_1'')$, hence $\mathsf{Sub}_{\ell-1}(u_0''bu_1'') = \mathsf{Sub}_{\ell-1}(v_0''av_1'')$.

Assume, for a moment, that $i \cdot a = i \cdot b = p$. We have $i \cdot u = p \cdot u_0''bu_1'' \in T$ and $i \cdot v = p \cdot v_0''av_1'' \notin T$. Again this is a contradiction to the induction assumption applying to the biautomaton $\mathcal{B}_p$ and the pair of words $u_0''bu_1'$ and $v_0''av_1''$. Altogether we have that

$$i \cdot a \neq i \cdot b\,. \tag{8}$$

Now we will change our strategy a bit: in the state i we read both u and v from the right and we are interested in the position in the words, where we leave the state i.

Recall our assumption (2). Let $w', w'' \in A^*$, $c \in A$ be such that

$$u = w'cw'', \text{ for every } x \in \mathsf{c}(w'') \text{ we have } i \circ x = i, \text{ and } i \circ c \neq i\,. \tag{9}$$

Similarly, let $z', z'' \in A^*$, $d \in A$ be such that

$$v = z'dz'', \text{ for every } x \in \mathsf{c}(z''), \text{ we have } i \circ x = i, \text{ and } i \circ d \neq i\,. \tag{10}$$

With respect to the previous paragraph, we can further assume that

$$i \circ c \neq i \circ d \tag{11}$$

and we know that the last occurrence of d in u is before the last occurrence of c in u and vice verse for v.

Now we are interested in the relative positions of the considered occurrences of letters a and d in words u and v . Since the mentioned occurrence of a is the first occurrence of this letter in u and the occurrence d is the last occurrence of the letter d in u, we see that the occurrence of a is before the occurrence of d if and only if $ad \in \mathsf{Sub}_2(u)$.

First assume that $ad \notin \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$. We distinguish two possibilities: $a = d$ and $a \neq d$.

Case $a = d$: From the assumption $ad \notin \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$ we know that a has exactly one occurrence in u and exactly one occurrence in v. By Lemma 3 and its dual, we have $\mathsf{Sub}_{\ell-1}(u'') = \mathsf{Sub}_{\ell-1}(z'')$ and $\mathsf{Sub}_{\ell-1}(u') = \mathsf{Sub}_{\ell-1}(z')$, in particular $\mathsf{c}(u'') = \mathsf{c}(z'')$ and $\mathsf{c}(u') = \mathsf{c}(z')$. Thus from our assumptions (10) and (3), we have $i \circ u'' = i \circ z'' = i$ and $i \cdot z' = i \cdot u' = i$. Since $i \cdot u \in T$ we have also $(i \cdot u') \circ au'' \in T$, i.e. $(i \cdot u') \circ au'' = i \circ au'' = (i \circ u'') \circ a = i \circ a \in T$. Similarly, for v, we get $(i \circ z'') \cdot z'a \notin T$ and thus $(i \circ z'') \cdot z'a = i \cdot z'a = (i \cdot z') \cdot a = i \cdot a \notin T$. We obtained a contradiction since $i \circ a \in T$ and $i \cdot a \notin T$ cannot be true at the same moment in a biautomaton.

Case $a \neq d$: Now the first occurrence of a in u is after the last occurrence of d in u and the same is true for v. We use Lemma 3 again. The first occurrence of a in v is somewhere (see (10)) in the suffix z'' of v and hence $z'' = z_0''az_1''$ where

$a \notin \mathsf{c}(z'dz_0'')$. Then $\mathsf{Sub}_{\ell-1}(z_1'') = \mathsf{Sub}_{\ell-1}(u'')$ and it follows that $\mathsf{Sub}_{\ell-1}(u'') \subseteq \mathsf{Sub}_{\ell-1}(z'')$, in particular $\mathsf{c}(u'') \subseteq \mathsf{c}(z'')$. Thus for each $x \in \mathsf{c}(u'')$ we have $i \circ x = i$ and since a occurs in z'' we have also $i \circ a = i$. In the same way we obtain that $\mathsf{c}(z') \subseteq \mathsf{c}(u')$ and for each $x \in \mathsf{c}(z')$ we have $i \cdot x = i$ and also $i \cdot d = i$. This means that both u and v can be read without leaving the state i. Indeed, $(i \cdot u') \circ au'' = i \circ au'' = i$ and $(i \cdot z'd) \circ z'' = i \circ z'' = i$. This is a contradiction to the assumption that $u \in L_{\mathcal{B}}$ and $v \notin L_{\mathcal{B}}$, because the state i cannot be terminal and non-terminal at the same moment. We have finished the case $ad \notin \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$. Note that if we interchange u with v the previous part of the proof can be used for the situation when $bc \notin \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$.

For the rest of the proof we can assume

$$ad \in \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v) \text{ and } bc \in \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)\,. \tag{12}$$

Now we will discuss the case when $i \cdot a \neq i \circ d$. Denote $p_1 = i \cdot a \neq i$ and $p_2 = i \circ d \neq i$ for which we have $p_1 \neq p_2$ and consider $p_3 = p_1 \circ d = (i \cdot a) \circ d = (i \circ d) \cdot a = p_2 \cdot a$. The state p_3 need not be different from p_1 and p_2, but is it important that there is a simple path of length two from i to p_3. Hence both biautomata $\mathcal{B}_{p_1}$ and $\mathcal{B}_{p_2}$ has depth at most $\ell - 1$ and the biautomaton $\mathcal{B}_{p_3}$ has depth at most $\ell - 2$.

Since $ad \in \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$ we can consider the following factorizations of u and v: $u = u_1au_2du_3$, $v = v_1av_2dv_3$, where $u_1, u_2, u_3, v_1, v_2, v_3 \in A^*$ are such that $a \notin \mathsf{c}(u_1)$, $a \notin \mathsf{c}(v_1)$, $d \notin \mathsf{c}(u_3)$, $d \notin \mathsf{c}(v_3)$. Recall that here $u_1 = u'$ satisfying conditions from assumption (3), and $v_3 = z''$ from (10). By Lemmas 3 and 4 we have $\mathsf{Sub}_{\ell-1}(u_2du_3) = \mathsf{Sub}_{\ell-1}(v_2dv_3)$, $\mathsf{Sub}_{\ell-1}(u_1au_2) = \mathsf{Sub}_{\ell-1}(v_1av_2)$ and $\mathsf{Sub}_{\ell-2}(u_2) = \mathsf{Sub}_{\ell-2}(v_2)$.

The following part of the proof is illustrated in Figure 4.

By the induction assumption applied on $\mathcal{B}_{p_1}$ and the pair of words u_2du_3 and v_2dv_3 we have $p_1 \cdot u_2du_3 \in T$ if and only if $p_1 \cdot v_2dv_3 \in T$. The first condition is satisfied because $p_1 \cdot u_2du_3 = i \cdot u_1au_2bu_3 = i \cdot u$. Hence $p_1 \cdot v_2dv_3 \in T$ and consequently $(p_1 \circ dv_3) \cdot v_2 \in T$. Now $p_1 \circ v_3 = (i \cdot a) \circ v_3 = (i \circ v_3) \cdot a = i \cdot a = p_1$ which explains why in Figure 4 there is a loop labeled by v_3 in the state p_1. Thus for the terminal state $(p_1 \circ dv_3) \cdot v_2$ we get $(p_1 \circ dv_3) \cdot v_2 = ((p_1 \circ v_3) \circ d) \cdot v_2 = (p_1 \circ d) \cdot v_2 = p_3 \cdot v_2$. Therefore $p_3 \cdot v_2 \in T$.

Analogously, we have $p_2 \circ v_1av_2 = i \circ v \notin T$ and hence $p_2 \circ u_1au_2 \notin T$ follows from the induction assumption applied on the biautomaton $\mathcal{B}_{p_2}$ and the pair of words v_1av_2 and u_1au_2. We deduce that $p_2 \cdot u_1au_2 = (p_2 \cdot u_1a) \cdot u_2 \notin T$. Now we can see that $p_2 \cdot u_1a = ((i \circ d) \cdot u_1) \cdot a = ((i \cdot u_1) \circ d) \cdot a = (i \circ d) \cdot a = p_2 \cdot a = p_3$ from which $p_2 \cdot u_1au_2 = p_3 \cdot u_2 \notin T$ follows. We have observed $p_3 \cdot v_2 \in T$ in the previous paragraph and $p_3 \cdot u_2 \notin T$ here. This is a contradiction to the induction assumptions that $\mathcal{B}_{p_3}$ is a acyclic biautomaton of depth at most $\ell - 2$ and the equality $\mathsf{Sub}_{\ell-2}(u_2) = \mathsf{Sub}_{\ell-2}(v_2)$. We have finished the case when $i \cdot a \neq i \circ d$.

Finally assume that $i \cdot a = i \circ d$. Since the previous argument can be done dually, we assume that

$$i \cdot a = i \circ d, \quad \text{and} \quad i \cdot b = i \circ c\,. \tag{13}$$

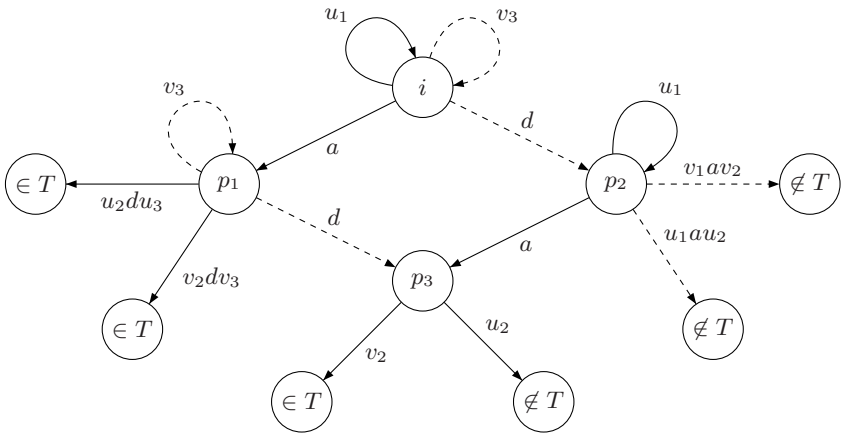

Fig. 4. States in the case $ad \in \mathsf{Sub}_2(u)$, $i \cdot a \neq i \circ d$.

During the proof we collected assumptions (1) – (13). We recall some of them now. By (6) and (7) the first occurrence of a in u is before the first occurrence of b and vise verse in v. Also from assumptions (9) and (10) we get that the last occurrence of d in u is before the last occurrence of c in u and vise verse in v. Furthermore, we assumed in (12) that $ad \in \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$ and $bc \in \mathsf{Sub}_2(u) = \mathsf{Sub}_2(v)$. Since the first occurrence of b in v is before the first occurrence of a in v and $ad \in \mathsf{Sub}_2(v)$, we have $bd \in \mathsf{Sub}_2(v)$. Hence $bd \in \mathsf{Sub}_2(u)$ and the first occurrence of b in u is before the last occurrence of d in u. Similarly we can get $ac \in \mathsf{Sub}_2(v)$. Therefore we deduce the following factorizations of words u and v:

$$u = u_1au_2bu_3du_4cu_5, \quad v = v_1bv_2av_3cv_4dv_5\,,$$

where mentioned occurrences of a and b are the first ones and the occurrences of c and d are the last ones in the words u and v. By Lemma 4 we get

$$\mathsf{Sub}_{\ell-2}(u_2bu_3du_4) = \mathsf{Sub}_{\ell-2}(v_3) \subseteq$$
$$\subseteq \mathsf{Sub}_{\ell-2}(v_2av_3cv_4) = \mathsf{Sub}_{\ell-2}(u_3) \subseteq \mathsf{Sub}_{\ell-2}(u_2bu_3du_4)\,.$$

Thus all inclusions hold as equalities and we can deduce

$$\mathsf{Sub}_{\ell-2}(u_3) = \mathsf{Sub}_{\ell-2}(bu_3) = \mathsf{Sub}_{\ell-2}(u_2bu_3du_4) =$$
$$= \mathsf{Sub}_{\ell-2}(v_3) = \mathsf{Sub}_{\ell-2}(v_3c) = \mathsf{Sub}_{\ell-2}(v_2av_3cv_4)\,.$$

Now we return our attention to the biautomaton $\mathcal{B}$. Recall that we have assumed in (13) and (8) that $i \circ d = i \cdot a \neq i \cdot b = i \circ c$. Denote $p_1 = i \circ d = i \cdot a$,

$p_2 = i \cdot b = i \circ c$ and further $q_1 = p_1 \circ c$, $q_2 = p_2 \circ d$ and $q_3 = q_1 \cdot b$. From the equalities among states we also get $q_1 = p_1 \circ c = (i \cdot a) \circ c = (i \circ c) \cdot a = p_2 \cdot a$, $q_2 = p_2 \circ d = (i \cdot b) \circ d = (i \circ d) \cdot b = p_1 \cdot b$ and $q_3 = q_1 \cdot b = (p_1 \circ c) \cdot b = (p_1 \cdot b) \circ c = q_2 \circ c$. The following part of the proof is illustrated in Figure 5. We know that p_1, p_2

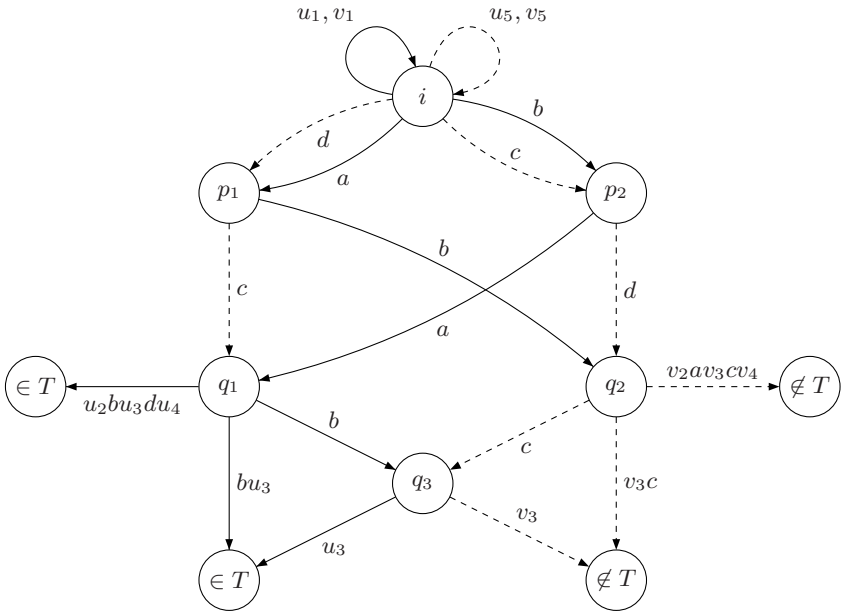

Fig. 5. States in the case $i \cdot a = i \circ d$.

and i are three distinct states. Hence there is a simple path of length at least 2 from i to each of q_1, q_2 and q_3. Thus the biautomata $\mathcal{B}_{q_1}$, $\mathcal{B}_{q_2}$ and $\mathcal{B}_{q_3}$ have all depth at most $\ell - 2$. We have $i \cdot u \in T$ and hence $((((i \cdot u_1) \circ u_5) \cdot a) \circ c) \cdot u_2bu_3du_4 = ((i \cdot a) \circ c) \cdot u_2bu_3du_4 = q_1 \cdot u_2bu_3du_4 \in T$. If we apply the induction assumption on the biautomaton $\mathcal{B}_{q_1}$ and the pair of words $u_2bu_3du_4$ and bu_3, we obtain $q_1 \cdot bu_3 \in T$. Thus we have $q_1 \cdot bu_3 = (q_1 \cdot b) \cdot u_3 = q_3 \cdot u_3 \in T$. In a similar way we get $i \circ v \notin T$ from which we conclude that $q_2 \circ v_2av_3cv_4 \notin T$. By induction assumption for $\mathcal{B}_{q_2}$ and the pair of words $v_2av_3cv_4$ and v_3c we obtain $q_2 \circ v_3c = q_3 \circ v_3 \notin T$. Now this is a contradiction to the induction assumption for $\mathcal{B}_{q_3}$ and the pair of words u_3 and v_3.

The proof of Proposition 2 is now complete. □