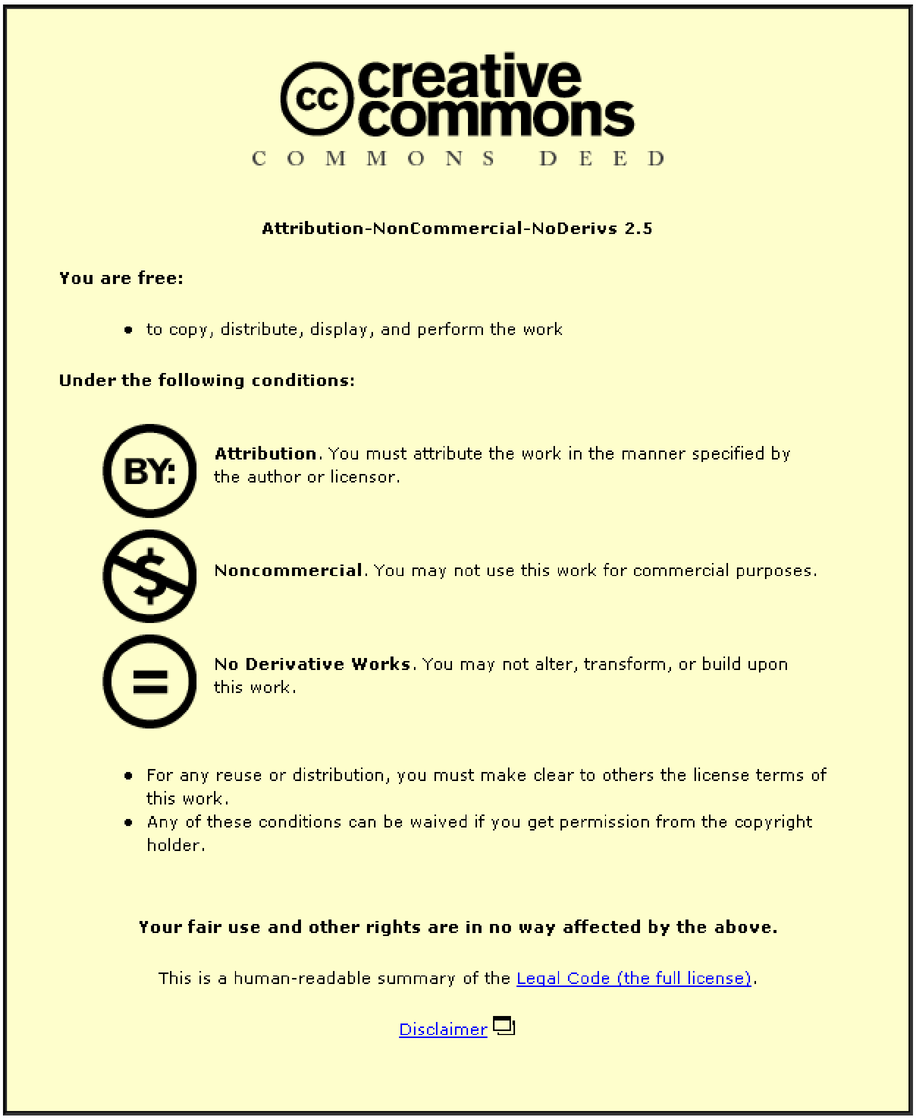

On the Equivalence Problem for E-pattern Languages over Small Alphabets

Daniel Reidenbach[*]

Fachbereich Informatik, Technische Universität Kaiserslautern,
Postfach 3049, 67653 Kaiserslautern, Germany
`reidenba@informatik.uni-kl.de`

Abstract. We contribute new facets to the discussion on the equivalence problem for E-pattern languages (also referred to as extended or erasing pattern languages). This fundamental open question asks for the existence of a computable function that, given any pair of patterns, decides whether or not they generate the same language. Our main result disproves Ohlebusch and Ukkonen's conjecture (*Theoretical Computer Science* 186, 1997) on the equivalence problem; the respective argumentation, that largely deals with the nondeterminism of pattern languages, is restricted to terminal alphabets with at most four distinct letters.

1 Introduction

Patterns—finite strings that consist of variables and terminal symbols—are compact and "natural" devices for the definition of numerous regular and nonregular formal languages. A pattern generates a word by a uniform substitution of the variables with arbitrary strings of terminal symbols, and, accordingly, its language is the set of all words that can be obtained by suchlike morphisms. For instance, the language generated by the pattern $\alpha = x_1\, x_1\, \mathtt{a}\, \mathtt{b}\, x_2$ (with variables x_1, x_2 and terminals $\mathtt{a}$, $\mathtt{b}$) includes all words where the prefix consists of two occurrences of the same string, followed by the string $\mathtt{ab}$ and concluded by an arbitrary suffix. Thus, the language of α contains, e.g., the words $w_1 = \mathtt{b\,a\,b\,a\,a\,b\,a}$ and $w_2 = \mathtt{a\,b\,b\,b}$, whereas $v_1 = \mathtt{b\,b\,b\,b\,a}$ and $v_2 = \mathtt{b\,a\,a\,b\,b}$ are not covered by α.

The investigation of patterns in strings—initiated by Thue [16, 17]—may be seen as a classical topic in the research on word monoids and combinatorics of words (cf., e.g., [2], a survey is given in [3]). Contrary to this, the definition of *pattern languages* as described above—introduced by Angluin [1]—originally has been motivated by considerations on algorithmic language learning within the scope of *inductive inference*. Since then, however, the properties of pattern languages have been intensively studied from a language theoretical point of view as well, e.g. by Jiang, Kinber, Salomaa, Salomaa, Yu [7, 8]; for a survey see [10]. These examinations reveal that the characteristics of languages generated by a definition which disallows the substitution of variables with the empty word—as given by Angluin—and of those produced by a definition allowing the empty

[*] Supported by the Deutsche Forschungsgemeinschaft (DFG), Grant Wi 1638/1-3

substitution (as applied when generating w_2 in our example) differ significantly. Languages of the latter type have been introduced by Shinohara [15]; they are called *extended*, *erasing*, or simply *E*-pattern languages.

In spite of the wide range of profound examinations, a number of fundamental properties of E-pattern languages is still unresolved; one of the best-known open problems among these is the decidability of the *equivalence*, i.e. the question on the existence of a total computable function that, given any pair of patterns, decides whether or not they generate the same language. This problem, that for Angluin's pattern languages has a trivial answer in the affirmative, has been tackled several times (cf. [5, 7, 8, 4, 11, 13]), contributing a number of positive results on subclasses, properties, conjectures, and conditions, but no comprehensive answer. Consequently, the anticipation of a positive outcome, as expressed in [8], so far could be neither verified nor refuted.

The current state of knowledge on E-pattern languages reveals that several of their properties strongly depend on the size of the terminal alphabet. For instance, the subclass generated by terminal-free patterns is learnable *if and only if* the alphabet is not binary (cf. [12]), whereas the full class is learnable for unary alphabets, but not for those with two, three or four letters (cf. [13]). Consequently, and particularly for small alphabets, E-pattern languages show a variety of (frequently fairly surprising) discontinuities. This phenomenon is brought about by the fact that especially those words over only a few distinct letters tend to be *ambiguous*, i.e. one and the same pattern can generate such a word by different substitutions. The influence of this nondeterminism of E-pattern languages—that is even more complex provided that the patterns do not consist of variables only—on several open problems is not completely understood yet, and therefore most corresponding partial results are restricted to those cases where ambiguity of words is somewhat easy to grasp (cf., e.g., [7, 8, 11]).

These observations establish the background of the present paper, that provides new insight into the consequences of nondeterminism of pattern languages. We apply our approach to the prevailing conjecture on the equivalence problem for E-pattern languages—given by Ohlebusch and Ukkonen [11]—according to which, for terminal alphabets with at least three distinct letters, two arbitrary patterns α and β generate the same language if and only if there exist terminal-preserving morphisms ϕ and ψ such that $\phi(\alpha) = \beta$ and $\psi(\beta) = \alpha$. This conjecture, that we recently have claimed to be incorrect for alphabets with exactly three letters (cf. [13]), in the present paper is disproven for alphabets of size 4.

2 Preliminaries

We now proceed formally. For notions and preliminary results not given in this paper we refer to [14] or, if appropriate, to the respective referenced literature.

$\mathbb{N}$ is the set of natural numbers, $\{0, 1, 2, \dots\}$. A *word* is a finite string of symbols. For an arbitrary set A of symbols, A^+ denotes the set of all non-empty words over A and A^* the set of all (empty and non-empty) words over A. Any set $L \subseteq A^*$ is a *language* over an alphabet A. We designate the *empty* word as

ε. For the word that results from the n-fold concatenation of a letter $\mathtt{a}$ or of a word w we write $\mathtt{a}^n$ or $(w)^n$, respectively. $|\cdot|$ denotes the size of a set or the length of a word, respectively, and $|w|_{\mathtt{a}}$ the frequency of a letter $\mathtt{a}$ in a word w.

The following notion allows to address certain parts of a word w over an alphabet A: If w contains n, $n \geq 1$, occurrences of a subword u then for every i, $1 \leq i \leq n$, $u\langle i\rangle$ is the ith occurrence (from the left) of u in w. For that case, the subword $[w/u\langle i\rangle]$ is the prefix of w up to (but not including) the leftmost letter of $u\langle i\rangle$ and the subword $[u\langle i\rangle\backslash w]$ is the suffix of w beginning with the first letter that is to the right of $u\langle i\rangle$. Moreover, for every word w that contains at least i occurrences of a subword u, j occurrences of subword v and that satisfies $w = w_1\, u\langle i\rangle\, w_2\, v\langle j\rangle\, w_3$ with $w_1, w_2, w_3 \in A^*$, we use $[u\langle i\rangle\backslash w/v\langle j\rangle]$ as an abbreviation for $[u\langle i\rangle\backslash[w/v\langle j\rangle]]$. Thus, for appropriate u, v, w, the specified subwords satisfy $w = [w/u\langle i\rangle]\ u\langle i\rangle\ [u\langle i\rangle\backslash w]$ or $w = [w/u\langle i\rangle]\ u\langle i\rangle\ [u\langle i\rangle\backslash w/v\langle j\rangle]\ v\langle j\rangle\ [v\langle j\rangle\backslash w]$, respectively; e.g., with $w = \mathtt{a\,b\,c\,a\,b\,b}$, $u = \mathtt{a}$ and $v = \mathtt{a\,b}$, the definition leads to $[w/u\langle 2\rangle] = \mathtt{a\,b\,c}$, $[u\langle 2\rangle\backslash w] = \mathtt{b\,b}$, and $[u\langle 1\rangle\backslash w/v\langle 2\rangle] = \mathtt{b\,c}$.

We proceed with the pattern specific terminology. Σ is a finite alphabet of *terminal* symbols and $X = \{x_1, x_2, x_3, \dots\}$ an infinite set of *variables*, $\Sigma \cap X = \emptyset$. Henceforth, we use lower case letters in typewriter font, e.g. $\mathtt{a}, \mathtt{b}, \mathtt{c}$, as terminal symbols exclusively, and terminal words are named as u, v, or w.

A *pattern* is a non-empty word over $\Sigma \cup X$, a *terminal-free pattern* is a non-empty word over X; naming patterns we use lower case letters from the beginning of the Greek alphabet such as α, β, γ. $\mathrm{var}(\alpha)$ denotes the set of all variables of a pattern α. We write Pat_Σ for the set of all patterns over the union of X and a specific alphabet Σ or, if there is no need to emphasise the terminal alphabet, Pat for short.

Following [4] we designate two patterns α, β as *similar* if and only if $\alpha = \alpha_0\, u_1 \alpha_1\, u_2\, \dots\, \alpha_{m-1}\, u_m \alpha_m$ and $\beta = \beta_0\, u_1 \beta_1\, u_2\, \dots\, \beta_{m-1}\, u_m \beta_m$ with $m \in \mathbb{N}$, $\alpha_i, \beta_i \in X^+$ for $1 \leq i < m$, $\alpha_0, \beta_0, \alpha_m, \beta_m \in X^*$ and $u_i \in \Sigma^+$ for $i \leq m$; in other words, we call patterns similar if their terminal substrings are identical and occur in the same order in the patterns.

A morphism $\phi : (\Sigma \cup X)^* \longrightarrow (\Sigma \cup X)^*$ is *terminal-preserving* if and only if, for every $\mathtt{a} \in \Sigma$, $\phi(\mathtt{a}) = \mathtt{a}$. If additionally, for a terminal-preserving morphism ϕ and for all $x_i \in X$, $\phi(x_i) \in X^*$ then we call ϕ *similarity-preserving.* We say that patterns α, β are *(morphically) coincident* if there exist similarity-preserving morphisms ϕ and ψ such that $\phi(\alpha) = \beta$ and $\psi(\beta) = \alpha$; we call them *(morphically) semi-coincident* if there is either such a ϕ or such a ψ, and, for the case that there is neither such a ϕ nor such a ψ, they are designated as *(morphically) incoincident.*

A terminal-preserving morphism σ is a *substitution* if and only if, for every $x_i \in X$, $\sigma(x_i) \in \Sigma^*$. The *E-pattern language* $L_\Sigma(\alpha)$ of a pattern α is defined as the set of all $w \in \Sigma^*$ such that $\sigma(\alpha) = w$ for some substitution σ. For any word $w = \sigma(\alpha)$ we say that σ *generates* w, and for any language $L = L_\Sigma(\alpha)$ we say that α generates L. If Σ is understood then we denote the E-pattern language of a pattern α simply as $L(\alpha)$. We use ePAT as an abbreviation for the full class of E-pattern languages (or ePAT_Σ if the corresponding alphabet is of interest).

We designate a pattern α as *succinct* if and only if $|\alpha| \leq |\beta|$ for all patterns β with $L(\beta) = L(\alpha)$. The pattern $\beta = x_1 x_2 x_1 x_2$, e.g., generates the same language as $\alpha = x_1 x_1$, and therefore β is not succinct; α is succinct as $L(x_1) \neq L(\alpha)$.

According to [9] we denote a word w as *ambiguous* (in respect of a pattern α) if and only if it can be generated by several substitutions of α, i.e. there exist substitutions σ and σ', $\sigma(x_i) \neq \sigma'(x_i)$ for some $x_i \in \text{var}(\alpha)$, such that $\sigma(\alpha) = w = \sigma'(\alpha)$. Correspondingly, we call a word w *unambiguous* (in respect of α) if and only if there is exactly one substitution σ with $\sigma(\alpha) = w$. The word $w_1 = \mathtt{aaba}$, for instance, is ambiguous in respect of $\alpha = x_1 \, \mathtt{a} \, x_2$ since it can be generated by, e.g., σ and σ' with $\sigma(x_1) = \mathtt{a}$, $\sigma(x_2) = \mathtt{ba}$ and $\sigma'(x_1) = \varepsilon$, $\sigma'(x_2) = \mathtt{aba}$. The example word $w_2 = \mathtt{ba}$ is unambiguous in respect of α.

We now proceed with some decidability problems on E-pattern languages: Let ePAT* be any set of E-pattern languages. We say that the *inclusion problem* for ePAT* is *decidable* if and only if there exists a computable function which, given two arbitrary patterns α, β with $L(\alpha), L(\beta) \in \text{ePAT}^\star$, decides whether or not $L(\alpha) \subseteq L(\beta)$. Accordingly, the *equivalence problem* is decidable if and only if there exists another computable function which for every pair of patterns α, β with $L(\alpha), L(\beta) \in \text{ePAT}^\star$ decides whether or not $L(\alpha) = L(\beta)$. Obviously, the decidability of the inclusion implies the decidability of the equivalence. As mentioned in Section 1, the decidability of the equivalence problem for ePAT has not been resolved yet, but there is a number of positive results on subclasses given in [11]. The inclusion problem is known to be undecidable (cf. [8]). Under certain circumstances, however, the inclusion problem is decidable; this results from the following fact:

Fact 1 ([11]). *Let Σ be an alphabet and α, β two arbitrary similar patterns such that Σ contains two distinct letters not occurring in α and β. Then $L_\Sigma(\beta) \subseteq L_\Sigma(\alpha)$ iff there exists a similarity-preserving morphism ϕ such that $\phi(\alpha) = \beta$.*

In particular, Fact 1 implies the decidability of the inclusion problem for the class of terminal-free E-pattern languages if $|\Sigma| \geq 2$ (proven in [5] and [8]).

The following theorem shows that any consideration on the equivalence problem can be restricted to similar patterns. Therefore, Fact 1 implies the decidability of the equivalence for all pairs of patterns if one of the two patterns does not contain at least two distinct letters of the alphabet.

Fact 2 ([5] and [7]). *Let Σ be an alphabet, $|\Sigma| \geq 3$, and let $\alpha, \beta \in \text{Pat}_\Sigma$. If $L_\Sigma(\alpha) = L_\Sigma(\beta)$ then α and β are similar.*

Moreover, Fact 2 suggests a possible approach to the equivalence problem, that has been addressed by [4] and [11]: Obviously, the equivalence of E-pattern languages is decidable provided that Fact 1 holds for all similar patterns (and not only for those satisfying the additional condition).

We conclude this section with a definition that originates in [11] and that is motivated by the facts stated above: Let Σ be an alphabet and define $\Sigma' := \Sigma \cup \{\mathtt{a}\}$ for an arbitrary $\mathtt{a} \notin \Sigma$. We say that the *equivalence for* ePAT$_\Sigma$ *is preserved under alphabet extension* if and only if, for every pair $\alpha, \beta \in \text{Pat}_\Sigma$, $L_\Sigma(\alpha) = L_\Sigma(\beta)$ implies $L_{\Sigma'}(\alpha) = L_{\Sigma'}(\beta)$ and vice versa.

3 On Ohlebusch and Ukkonen's conjecture

The equivalence problem for E-pattern languages has first been examined in [5] and [7] and later in [8], [4], and [11]. The latter authors give a procedure that for every pattern computes a shortest *normal form.* They conjecture that, for alphabets with at least three letters, two patterns generate the same language if and only if their normal forms are the same, and the authors paraphrase their conjecture as follows:

Conjecture 1 ([11]*).* For an alphabet Σ, $|\Sigma| \geq 3$, and patterns $\alpha_1, \alpha_2 \in \mathrm{Pat}_\Sigma$, $L_\Sigma(\alpha_1) = L_\Sigma(\alpha_2)$ if and only if α_1 and α_2 are morphically coincident.

Furthermore, as a consequence of Fact 1 and Fact 2, the authors annotate that the equivalence problem is decidable for $|\Sigma| \geq 3$ if the equivalence for ePAT_Σ is preserved under alphabet extension (cf. [11], Open Question 2).

The choice of alphabet size 3 as a lower bound in Conjecture 1 might be caused by the following observations: The patterns $\alpha_1 = x_1\,\mathtt{a}\,x_2\,\mathtt{b}\,x_3$ and $\alpha_2 = x_1\,\mathtt{a}\,\mathtt{b}\,x_2$, for instance, generate the same language if $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ (although they are semi-coincident) since for every word in $\{\sigma_i(\alpha_1) \mid \sigma_i(x_2) \neq \varepsilon\}$ a second substitution σ'_i can be given with $\sigma'_i(\alpha_1) = \sigma_i(\alpha_1)$ and $\sigma'(x_2) = \varepsilon$. Thus, this specific ambiguity of words, that is caused by the small alphabet and by the composition of variables and terminal symbols in α_1, brings about the equivalence of $L_\Sigma(\alpha_1)$ and $L_\Sigma(\alpha_2)$. Contrary to this, for $|\Sigma| \geq 3$, the existence of analogue examples seems to be rather implausible since for every variable in a pattern at least one occurrence can be chosen for assigning a substitution that contains a letter which differs from the terminal symbols to the left and to the right of the variable (cf., e.g., the proof of Fact 2 as given in [7]). Consequently, Conjecture 1 suggests that such patterns do not exist for alphabets containing at least three letters.

As a by-product of learning theoretical studies, [13] anticipates that—at least for alphabets with *exactly* three letters—this conjecture is incorrect. More precisely, the paper claims that, for $\Sigma := \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, $\Sigma' := \Sigma \cup \{\mathtt{d}\}$ and

$$\begin{aligned}\tilde{\alpha}_{\mathtt{abc},1} &:= x_1 \ \mathtt{a} \ x_2\, x_3^2\, x_4^2\, x_5^2\, x_6^2\, x_7^2\, x_8 \ \mathtt{b} \ x_9 \ \mathtt{a} \ x_2\, x_{10}^2\, x_4^2\, x_5^2\, x_6^2\, x_{11}^2\, x_8 \ \mathtt{b} \ x_{12},\\ \tilde{\alpha}_{\mathtt{abc},2} &:= x_1 \ \mathtt{a} \ x_2\, x_3^2\, x_4^2\, x_7^2\, x_8 \ \mathtt{b} \ x_9 \ \mathtt{a} \ x_2\, x_{10}^2\, x_4^2\, x_{11}^2\, x_8 \ \mathtt{b} \ x_{12},\end{aligned}$$

$L_\Sigma(\tilde{\alpha}_{\mathtt{abc},1}) = L_\Sigma(\tilde{\alpha}_{\mathtt{abc},2})$, but $L_{\Sigma'}(\tilde{\alpha}_{\mathtt{abc},1}) \supset L_{\Sigma'}(\tilde{\alpha}_{\mathtt{abc},2})$, and $\tilde{\alpha}_{\mathtt{abc},1}$ and $\tilde{\alpha}_{\mathtt{abc},2}$ are semi-coincident.

The present paper actually dis*proves* Conjecture 1; to this end, however, we regard different alphabets, namely those of size 4. Thus, we establish an additional result to that in [13]. Besides, the chosen alphabet size is by far more challenging, and therefore it requires a significantly more elaborate and instructive reasoning. Hence, our presumably unexpected main result reads as follows:

Theorem 1. *Let Σ be an alphabet, $|\Sigma| = 4$. Then the equivalence for* ePAT_Σ *is not preserved under alphabet extension.*

Theorem 2. *Let Σ be an alphabet, $|\Sigma| = 4$. Then there exist morphically incoincident patterns $\alpha_1, \alpha_2 \in \mathrm{Pat}_\Sigma$ such that $L_\Sigma(\alpha_1) = L_\Sigma(\alpha_2)$.*

Referring to these statements we can conclude that for "small" alphabets (i.e. for those with at most four distinct letters) the equivalence of E-pattern languages has some common properties, which nicely contrast with the expectations (potentially) involved in Fact 2, Conjecture 1, and Theorem 7.2 of [8]:

Corollary 1. *Let Σ be an alphabet, $|\Sigma| \leq 4$. Then the equivalence for* ePAT_{Σ} *is not preserved under alphabet extension.*

Corollary 2. *Let Σ be an alphabet, $|\Sigma| \leq 4$. Then there exist morphically incoincident or semi-coincident patterns $\alpha_1, \alpha_2 \in \mathrm{Pat}_{\Sigma}$ such that $L_{\Sigma}(\alpha_1) = L_{\Sigma}(\alpha_2)$.*

The proof of Theorem 1 and Theorem 2, that for the given patterns $\tilde{\alpha}_{\mathtt{abc},1}$ and $\tilde{\alpha}_{\mathtt{abc},2}$ can be adapted to the case $|\Sigma| = 3$ with little effort, is accomplished in Section 3.1. Its underlying principle follows the course indicated above: We compose two sophisticated incoincident example patterns—each of them consisting of 82 variables and terminals—and identify "decisive" words in their languages. Then we precisely examine the ambiguity of these words and reveal that all of them can be generated by substitutions assigning the empty word to at least one among two specific variables; thereby we can conclude that both patterns generate the same language. In other words, we analyse the nondeterminism of E-pattern languages, that has been the subject, e.g., of [9]. However, the prevailing point of view in literature does not exactly meet our requirements as it investigates the ambiguity of pattern *languages*, i.e. the maximum ambiguity among all words in the language, whereas we ask for the existence of particular alternative substitutions for selected words. Thus, our method is rather related to the research on *equality sets* (cf. [6]).

In spite of the extensive argumentation required even for a single alphabet, we expect our method to be useful for future examinations of Conjecture 1 with regard to different alphabet sizes as well. Moreover, we suggest that the (supposably necessary) complexity of our example patterns explains the lack of comprehensive results on the equivalence problem so far, and we consider the subsequent section to provide an insight into the extraordinary combinatorial depth of E-pattern languages.

Obviously, with the present state of knowledge on the subject, the given results do not imply the non-decidability of the equivalence problem for ePAT_{Σ} with $|\Sigma| = 4$. They show, first, that the expected lower bound in terms of alphabet size for a uniform behaviour of E-pattern languages concerning the decidability of the equivalence—as expressed in Conjecture 1—needs to be redetermined (provided such a bound exists at all). Second, they suggest that any decision procedure for $|\Sigma| = 4$ (if any) presumably needs to be more elaborate than that given in [11]—which, by the way, still might be applicable to $|\Sigma| \geq 5$. Additional remarks on suchlike aspects are given in Section 3.2.

3.1 Proof of the Main Results

The present section contains four lemmata. Lemma 1 and Lemma 4 prove Theorem 1; the argumentation on Theorem 2 is accomplished by Lemma 1 again and, additionally, Lemma 3—which, in turn, utilises Lemma 2.

We begin with the example patterns that constitute the core of our reasoning:

Definition 1 (first version). *The patterns $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ are given by*

$$
\begin{aligned}
\tilde{\alpha}_{\mathtt{abcd},1} := {} & x_1 \; \mathtt{a} \; x_2 \, x_3^2 \, x_4^2 \, x_5^2 \, x_6^2 \, x_7 \, \mathtt{b} \; x_8 \; \mathtt{a} \; x_2 \, x_9^2 \, x_4^2 \, x_5^2 \, x_{10}^2 \, x_7 \; \mathtt{b} \; x_{11} \\
& \mathtt{c} \; x_{12} \, x_{13}^2 \, x_{14}^2 \, x_{15}^2 \, x_{16}^2 \, x_{17} \, \mathtt{d} \; x_{18} \; \mathtt{c} \; x_{12} \, x_{19}^2 \, x_{14}^2 \, x_{15}^2 \, x_{20}^2 \, x_{17} \; \mathtt{d} \; x_{21} \\
& x_{14}^2 \, x_{15}^2 \, x_{14}^2 \, x_{15}^2 \, x_{14}^2 \, x_{15}^2 \, x_{22} \, x_4^2 \, x_5^2 \, x_4^2 \, x_5^2 \, x_4^2 \, x_5^2 \, x_{23} \, x_4 \, x_{14} \, x_{24} \\
\tilde{\alpha}_{\mathtt{abcd},2} := {} & x_1 \; \mathtt{a} \; x_2 \, x_3^2 \, x_4^2 \, x_5^2 \, x_6^2 \, x_7 \, \mathtt{b} \; x_8 \; \mathtt{a} \; x_2 \, x_9^2 \, x_4^2 \, x_5^2 \, x_{10}^2 \, x_7 \; \mathtt{b} \; x_{11} \\
& \mathtt{c} \; x_{12} \, x_{13}^2 \, x_{14}^2 \, x_{15}^2 \, x_{16}^2 \, x_{17} \, \mathtt{d} \; x_{18} \; \mathtt{c} \; x_{12} \, x_{19}^2 \, x_{14}^2 \, x_{15}^2 \, x_{20}^2 \, x_{17} \; \mathtt{d} \; x_{21} \\
& x_{14}^2 \, x_{15}^2 \, x_{14}^2 \, x_{15}^2 \, x_{14}^2 \, x_{15}^2 \, x_{22} \, x_4^2 \, x_5^2 \, x_4^2 \, x_5^2 \, x_4^2 \, x_5^2 \, x_{23} \, x_{14} \, x_4 \, x_{24}.
\end{aligned}
$$

Since $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ might be regarded as fairly intricate we give a second version of Definition 1 revealing the structure of the patterns:

Definition 1 (second version). *Consider the patterns*

$$
\begin{aligned}
\gamma_1 &:= x_4^2 \, x_5^2, \\
\gamma_2 &:= x_{14}^2 \, x_{15}^2, \\
\beta_1 &:= x_2 \, x_3^2 \, \gamma_1 \, x_6^2 \, x_7, \\
\beta_1' &:= x_2 \, x_9^2 \, \gamma_1 \, x_{10}^2 \, x_7, \\
\beta_2 &:= x_{12} \, x_{13}^2 \, \gamma_2 \, x_{16}^2 \, x_{17}, \\
\beta_2' &:= x_{12} \, x_{19}^2 \, \gamma_2 \, x_{20}^2 \, x_{17} \\
\hat{\alpha}_1 &:= x_1 \; \mathtt{a} \; \beta_1 \; \mathtt{b} \; x_8 \; \mathtt{a} \; \beta_1' \; \mathtt{b} \; x_{11} \; \mathtt{c} \; \beta_2 \; \mathtt{d} \; x_{18} \; \mathtt{c} \; \beta_2' \; \mathtt{d} \; x_{21}, \\
\hat{\alpha}_2 &:= (\gamma_2)^3 \, x_{22} \, (\gamma_1)^3.
\end{aligned}
$$

Then $\tilde{\alpha}_{\mathtt{abcd},1} := \hat{\alpha}_1 \, \hat{\alpha}_2 \, x_{23} \, x_4 \, x_{14} \, x_{24}$ and $\tilde{\alpha}_{\mathtt{abcd},2} := \hat{\alpha}_1 \, \hat{\alpha}_2 \, x_{23} \, x_{14} \, x_4 \, x_{24}$.

In order to facilitate the understanding of our reasoning we give some brief informal explanatory remarks before proceeding with the actual proof of Theorems 1 and 2: Evidently, concerning the question whether or not $L(\tilde{\alpha}_{\mathtt{abcd},1})$ and $L(\tilde{\alpha}_{\mathtt{abcd},2})$ are different, only those words are of interest that are generated by a substitution which is not empty for both x_4 and x_{14}, as the order of the last occurrences of these variables is the only difference between the patterns. Therefore, the components of $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ are tailor-made for ensuring the ambiguity of all words generated by a substitution σ that satisfies $\sigma(x_4) \neq \varepsilon \neq \sigma(x_{14})$.

With regard to the subpatterns of $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$, we first examine the kernels, i.e. γ_1 and γ_2. Obviously, for any substitution σ, if $\sigma(\gamma_1)$ or $\sigma(\gamma_2)$ do not contain at least two different letters then $\sigma(\gamma_1)$ or $\sigma(\gamma_2)$, respectively, are ambiguous and can be generated simply by x_5 or x_{15}, respectively; thus, x_4 or x_{14} can be substituted empty. In our formal argumentation, we utilise this fact only for $\sigma(\gamma_1) \in \{\mathtt{c}\}^* \cup \{\mathtt{d}\}^*$ or $\sigma(\gamma_2) \in \{\mathtt{a}\}^* \cup \{\mathtt{b}\}^*$. The other cases are covered by the ambiguity of $\sigma(x_1 \, \mathtt{a} \, \beta_1 \, \mathtt{b} \, x_8)$ (resp. $\sigma(x_{11} \, \mathtt{c} \, \beta_2 \, \mathtt{d} \, x_{18})$) whenever $\sigma(\gamma_1)$ (resp. $\sigma(\gamma_2)$) contains—possibly among others—the letters $\mathtt{a}$ or $\mathtt{b}$ (resp. $\mathtt{c}$ or $\mathtt{d}$), leading again to an optional empty substitution for x_4 (resp. x_{14}). Thus, σ only can generate a decisive word if $\sigma(\gamma_1)$ consists of $\mathtt{c}$ and $\mathtt{d}$ and $\sigma(\gamma_2)$ of $\mathtt{a}$ and $\mathtt{b}$. Such a choice of a substitution utilising letters that are distinguishable from the terminals to the left and to the right of the corresponding variable subword in the pattern probably is the most natural option and is used frequently (see, e.g., proof on Theorem 7.2 in [8]). However, for that case, $\sigma(\hat{\alpha}_2) = w_0\, \mathtt{a\,b}\, w_1\, \mathtt{a\,b}\, w_2\, \mathtt{c\,d}\, w_3\, \mathtt{c\,d}\, w_4$ for some words w_i, $i \leq 4$. Consequently, $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$ and $\sigma(\tilde{\alpha}_{\mathtt{abcd},2})$ can be generated by the (sub-)pattern $x_1\ \mathtt{a\,b}\ x_8\ \mathtt{a\,b}\ x_{11}\ \mathtt{c\,d}\ x_{18}\ \mathtt{c\,d}\ x_{21}$, and therefore x_4 and x_{14} can be substituted empty again. The variables with single occurrences, such as x_8 and x_{23}, are used to compensate the side effects of the empty substitution of x_4 or x_{14}; the modified repetitions of β_1 (as β_1') and β_2 (as β_2') and, particularly, those variables that distinguish β_1 from β_1' (e.g. x_3) and β_2 from β_2' (e.g. x_{13}) guarantee that $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ are incoincident. The latter point, one of the statements of Theorem 2, is discussed in Lemma 3.

As the ambiguity of decisive words affects $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ in the same way, the stated phenomenon allows us to prove the following, crucial lemma:

Lemma 1. *Let* $\Sigma_1 = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$*. Then* $L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},1}) = L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},2})$.

Proof. We first prove $L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},1}) \subseteq L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},2})$. Hence, let σ be an arbitrary substitution that is applicable to $\tilde{\alpha}_{\mathtt{abcd},1}$. We show that there exists a substitution σ' such that $\sigma'(\tilde{\alpha}_{\mathtt{abcd},2}) = \sigma(\tilde{\alpha}_{\mathtt{abcd},1})$. To this end, we refer to $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ as declared in the second version of Definition 1 and regard the following cases—that evidently can be restricted to a consideration of $\sigma(\gamma_1)$ and $\sigma(\gamma_2)$:

Case 1 $\sigma(\gamma_1) \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}^+ \setminus \{\mathtt{b}, \mathtt{c}, \mathtt{d}\}^+$:

Define
$$\begin{aligned}
\sigma'(x_1) &:= \sigma(x_1\ \mathtt{a}\ x_2\ x_3^2)\ [\sigma(\gamma_1)/\,\mathtt{a}\langle 1\rangle],\\
\sigma'(x_2) &:= [\mathtt{a}\langle 1\rangle \setminus \sigma(\gamma_1)],\\
\sigma'(x_8) &:= \sigma(x_8\ \mathtt{a}\ x_2\ x_9^2)\ [\sigma(\gamma_1)/\,\mathtt{a}\langle 1\rangle],\\
\sigma'(x_{22}) &:= \sigma(x_{22}\ (\gamma_1)^3),\\
\sigma'(x_{23}) &:= \sigma(x_{23}\ x_4),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \mathrm{var}(\beta_2\ \beta_2') \cup \{x_6, x_7, x_{10}, x_{11}, x_{18}, x_{21}, x_{24}\},\\
\sigma'(x_j) &:= \varepsilon,\ x_j \in \mathrm{var}(\gamma_1) \cup \{x_3, x_9\}.
\end{aligned}$$

Case 2 $\sigma(\gamma_1) \in \{\mathtt{b}, \mathtt{c}, \mathtt{d}\}^+ \setminus \{\mathtt{c}, \mathtt{d}\}^+$:

Define
$$\begin{aligned}
\sigma'(x_7) &:= [\sigma(\gamma_1)/\,\mathtt{b}\langle 1\rangle],\\
\sigma'(x_8) &:= [\mathtt{b}\langle 1\rangle \setminus \sigma(\gamma_1)]\ \sigma(x_6^2\ x_7\ \mathtt{b}\ x_8),\\
\sigma'(x_{11}) &:= [\mathtt{b}\langle 1\rangle \setminus \sigma(\gamma_1)]\ \sigma(x_{10}^2\ x_7\ \mathtt{b}\ x_{11}),\\
\sigma'(x_{22}) &:= \sigma(x_{22}\ (\gamma_1)^3),\\
\sigma'(x_{23}) &:= \sigma(x_{23}\ x_4),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \mathrm{var}(\beta_2\ \beta_2') \cup \{x_1, x_2, x_3, x_9, x_{18}, x_{21}, x_{24}\},\\
\sigma'(x_j) &:= \varepsilon,\ x_j \in \mathrm{var}(\gamma_1) \cup \{x_6, x_{10}\}.
\end{aligned}$$

Case 3 $\sigma(\gamma_1) \in \{\mathtt{c}\}^* \cup \{\mathtt{d}\}^*$:

Define
$$\begin{aligned}
\sigma'(x_4) &:= \varepsilon,\\
\sigma'(x_5) &:= \sigma(x_4\ x_5),\\
\sigma'(x_{23}) &:= \sigma(x_{23}\ x_4),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \mathrm{var}(\tilde{\alpha}_{\mathtt{abcd},2}) \setminus (\mathrm{var}(\gamma_1) \cup \{x_{23}\}).
\end{aligned}$$

Case 4 $\sigma(\gamma_1) \in \{\mathtt{c}, \mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$ and $\sigma(\gamma_2) \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}^+ \setminus \{\mathtt{a}, \mathtt{b}, \mathtt{d}\}^+$:

$$\begin{aligned}
\text{Define}\ \sigma'(x_{11}) &:= \sigma(x_{11}\ \mathtt{c}\ x_{12}\ x_{13}^2)\ [\sigma(\gamma_2) / \ \mathtt{c}\langle 1\rangle],\\
\sigma'(x_{12}) &:= [\mathtt{c}\langle 1\rangle \setminus \sigma(\gamma_2)],\\
\sigma'(x_{18}) &:= \sigma(x_{18}\ \mathtt{c}\ x_{12}\ x_{19}^2)\ [\sigma(\gamma_2) / \ \mathtt{c}\langle 1\rangle],\\
\sigma'(x_{22}) &:= \sigma((\gamma_2)^3\ x_{22}),\\
\sigma'(x_{24}) &:= \sigma(x_{14}\ x_{24}),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \operatorname{var}(\beta_1\ \beta_1') \cup \{x_1, x_8, x_{16}, x_{17}, x_{20}, x_{21}, x_{23}\},\\
\sigma'(x_j) &:= \varepsilon,\ x_j \in \operatorname{var}(\gamma_2) \cup \{x_{13}, x_{19}\}.
\end{aligned}$$

Case 5 $\sigma(\gamma_1) \in \{\mathtt{c}, \mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$ and $\sigma(\gamma_2) \in \{\mathtt{a}, \mathtt{b}, \mathtt{d}\}^+ \setminus \{\mathtt{a}, \mathtt{b}\}^+$:

$$\begin{aligned}
\text{Define}\ \sigma'(x_{17}) &:= [\sigma(\gamma_2) / \ \mathtt{d}\langle 1\rangle],\\
\sigma'(x_{18}) &:= [\mathtt{d}\langle 1\rangle \setminus \sigma(\gamma_2)]\ \sigma(x_{16}^2\ x_{17}\ \mathtt{d}\ x_{18}),\\
\sigma'(x_{21}) &:= [\mathtt{d}\langle 1\rangle \setminus \sigma(\gamma_2)]\ \sigma(x_{20}^2\ x_{17}\ \mathtt{d}\ x_{21}),\\
\sigma'(x_{22}) &:= \sigma((\gamma_2)^3\ x_{22}),\\
\sigma'(x_{24}) &:= \sigma(x_{14}\ x_{24}),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \operatorname{var}(\beta_1\ \beta_1') \cup \{x_1, x_8, x_{11}, x_{12}, x_{13}, x_{19}, x_{23}\},\\
\sigma'(x_j) &:= \varepsilon,\ x_j \in \operatorname{var}(\gamma_2) \cup \{x_{16}, x_{20}\}.
\end{aligned}$$

Case 6 $\sigma(\gamma_1) \in \{\mathtt{c}, \mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$ and $\sigma(\gamma_2) \in \{\mathtt{a}\}^* \cup \{\mathtt{b}\}^*$:

$$\begin{aligned}
\text{Define}\ \sigma'(x_{14}) &:= \varepsilon,\\
\sigma'(x_{15}) &:= \sigma(x_{14}\ x_{15}),\\
\sigma'(x_{24}) &:= \sigma(x_{14}\ x_{24}),\\
\sigma'(x_j) &:= \sigma(x_j),\ x_j \in \operatorname{var}(\tilde{\alpha}_{\mathtt{abcd},2}) \setminus (\operatorname{var}(\gamma_2) \cup \{x_{24}\}).
\end{aligned}$$

Case 7 $\sigma(\gamma_1) \in \{\mathtt{c}, \mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$ and $\sigma(\gamma_2) \in \{\mathtt{a}, \mathtt{b}\}^+ \setminus (\{\mathtt{a}\}^+ \cup \{\mathtt{b}\}^+)$:

Consequently, $\sigma((\gamma_1)^3)$ contains at least two occurrences of the subword $\mathtt{c}\,\mathtt{d}$ and $\sigma((\gamma_2)^3)$ contains at least two occurrences of the subword $\mathtt{a}\,\mathtt{b}$. Furthermore, due to the shape of these subwords, their occurrences must be non-overlapping. Therefore σ' can be given as follows:

$$\begin{aligned}
\text{Define}\ \sigma'(x_1) &:= \sigma(\hat{\alpha}_1)\ [\sigma(\hat{\alpha}_2) / \ \mathtt{a}\,\mathtt{b}\langle 1\rangle],\\
\sigma'(x_8) &:= [\mathtt{a}\,\mathtt{b}\langle 1\rangle \setminus \sigma(\hat{\alpha}_2) / \ \mathtt{a}\,\mathtt{b}\langle 2\rangle],\\
\sigma'(x_{11}) &:= [\mathtt{a}\,\mathtt{b}\langle 2\rangle \setminus \sigma(\hat{\alpha}_2) / \ \mathtt{c}\,\mathtt{d}\langle 1\rangle],\\
\sigma'(x_{18}) &:= [\mathtt{c}\,\mathtt{d}\langle 1\rangle \setminus \sigma(\hat{\alpha}_2) / \ \mathtt{c}\,\mathtt{d}\langle 2\rangle],\\
\sigma'(x_{21}) &:= [\mathtt{c}\,\mathtt{d}\langle 2\rangle \setminus \sigma(\hat{\alpha}_2)]\ \sigma(x_{23}\ x_4\ x_{14}\ x_{24}),\\
\sigma'(x_j) &:= \varepsilon,\ x_j \in \operatorname{var}(\tilde{\alpha}_{\mathtt{abcd},2}) \setminus \{x_1, x_8, x_{11}, x_{18}, x_{21}\}.
\end{aligned}$$

With the annotations on the shape of $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ in mind, it is obvious that, in every of the seven cases, $\sigma'(\tilde{\alpha}_{\mathtt{abcd},2}) = \sigma(\tilde{\alpha}_{\mathtt{abcd},1})$. Thus, since σ has been chosen arbitrarily and as the cases are exhaustive, $L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},1}) \subseteq L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},2})$.

The proof for $L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},2}) \subseteq L_{\Sigma_1}(\tilde{\alpha}_{\mathtt{abcd},1})$ is similar: In the argumentation given above, it is sufficient to replace $\tilde{\alpha}_{\mathtt{abcd},1}$ by $\tilde{\alpha}_{\mathtt{abcd},2}$ and vice versa and, additionally, to adapt $\sigma'(x_{23})$ and $\sigma'(x_{24})$ in Cases 1-7 in an adequate manner such that it matches the shape of $\tilde{\alpha}_{\mathtt{abcd},1}$. The rest is verbatim the same. □

With Lemma 1, the crucial element of Theorem 1 and Theorem 2 is proven. In the next step we complete the proof of Theorem 2. As a prerequisite thereof, we proceed with an evident lemma that is of great use for the upcoming proof of Lemma 3 and that is a direct consequence of Lemma 1 in [13]:

Lemma 2. *Let α be a terminal-free pattern and let $\phi : X^* \longrightarrow X^*$ be a morphism with $\phi(\alpha) = \alpha$. Then either $\phi(x_j) = x_j$ for every $x_j \in \mathrm{var}(\alpha)$ or there is an $x_{j'} \in \mathrm{var}(\alpha)$ such that $|\phi(x_{j'})| \geq 2$ and $x_{j'} \in \mathrm{var}(\phi(x_{j'}))$.*

We call any $x_{j'}$ satisfying these two conditions an *anchor variable* (in respect of the morphism ϕ).

Now we can prove that there are no similarity-preserving morphisms mapping $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ onto each other:

Lemma 3. *$\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ are morphically incoincident.*

Proof. Assume to the contrary there is a similarity-preserving morphism ϕ with $\phi(\tilde{\alpha}_{\mathtt{abcd},1}) = \tilde{\alpha}_{\mathtt{abcd},2}$ or with $\phi(\tilde{\alpha}_{\mathtt{abcd},2}) = \tilde{\alpha}_{\mathtt{abcd},1}$. Then, obviously, $\phi(x_4) \neq x_4$ or $\phi(x_{14}) \neq x_{14}$. Consequently—since, e.g., β_1 and β_2 occur in $\tilde{\alpha}_{\mathtt{abcd},1}$ as well as in $\tilde{\alpha}_{\mathtt{abcd},2}$ and since necessarily $\phi(\beta_1) = \beta_1$ and $\phi(\beta_2) = \beta_2$—there must be an anchor variable $x_{j'}$ in β_1 or β_2 (cf. Lemma 2).

We start with β_1. First, for $j' \in \{3, 4, 5, 6\}$, $x_{j'}$ being an anchor variable implies that $\phi(x_{j'}^2) = x_k x_{k'} \delta x_k x_{k'} \delta$ with variables $x_k, x_{k'}$ and $\delta \in X^*$, but there is no substring in β_1 that equals the given shape of $\phi(x_{j'})$. Second, because of the necessity of $\phi(\beta_1') = \beta_1'$, x_2 cannot be an anchor variable since $\phi(x_2)$ had to equal both $x_2 x_3 \delta$ and $x_2 x_9 \delta$ for a $\delta \in X^*$. Finally, due to an analogous reason, $j' \neq 7$. Thus, there is no anchor variable in $\mathrm{var}(\beta_1)$. This contradicts $\phi(x_4) \neq x_4$.

With regard to β_2, the argumentation is equivalent, and, consequently, there is no anchor variable in $\mathrm{var}(\beta_2)$. Therefore, the assumption is incorrect. □

With Lemma 1 and Lemma 3, the proof of Theorem 2 is accomplished. Consequently, and referring to Fact 1, it is obvious that, for a terminal alphabet Σ_3 with at least six distinct letters, $L_{\Sigma_3}(\tilde{\alpha}_{\mathtt{abcd},1}) \neq L_{\Sigma_3}(\tilde{\alpha}_{\mathtt{abcd},2})$. Hence, for $|\Sigma| = 4$ or $|\Sigma| = 5$, the equivalence for ePAT$_\Sigma$ is not preserved under alphabet extension. In order to conclude the proof of Theorem 1, we therefore have to show explicitly that the given example patterns generate different languages for alphabets with exactly five letters:

Lemma 4. *Let $\Sigma_2 \supseteq \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$. Then $L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},1}) \neq L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},2})$.*

Proof. We show that there is a word in $L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},1}) \setminus L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},2})$. To this end, we refer to $\tilde{\alpha}_{\mathtt{abcd},1}$ and $\tilde{\alpha}_{\mathtt{abcd},2}$ as declared in the second version of Definition 1 and consider the substitution σ given by

$$\sigma(x_j) := \begin{cases} \mathtt{c}\,\mathtt{e}^{3j-2}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{3j-1}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{3j}\,\mathtt{c} & , \quad x_j \in \mathrm{var}(\beta_1\, \beta_1'), \\ \mathtt{a}\,\mathtt{e}^{3j-2}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{3j-1}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{3j}\,\mathtt{a} & , \quad x_j \in \mathrm{var}(\beta_2\, \beta_2'), \\ \varepsilon & , \quad \text{else.} \end{cases}$$

Then $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$ has the following suffix generated by $\sigma(\mathtt{d}\ x_{21}\ \hat{\alpha}_2\ x_{23}\ x_4\ x_{14}\ x_{24})$:

$$\begin{aligned}&\mathtt{d}\,((\mathtt{a}\,\mathtt{e}^{40}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{41}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{42}\,\mathtt{a})^2(\mathtt{a}\,\mathtt{e}^{43}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{44}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{45}\,\mathtt{a})^2)^3\\ &((\mathtt{c}\,\mathtt{e}^{10}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{11}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{12}\,\mathtt{c})^2(\mathtt{c}\,\mathtt{e}^{13}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{14}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{15}\,\mathtt{c})^2)^3\\ &\mathtt{c}\,\mathtt{e}^{10}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{11}\,\mathtt{c}\ \mathtt{c}\,\mathtt{e}^{12}\,\mathtt{c}\ \mathtt{a}\,\mathtt{e}^{40}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{41}\,\mathtt{a}\ \mathtt{a}\,\mathtt{e}^{42}\,\mathtt{a}\end{aligned}$$

and this is the only occurrence of that subword in $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$.

Now assume to the contrary there is a substitution σ' with $\sigma'(\tilde{\alpha}_{\mathtt{abcd},2}) = \sigma(\tilde{\alpha}_{\mathtt{abcd},1})$. As, due to $\sigma(x_j) \in \{\mathtt{a},\mathtt{c},\mathtt{e}\}^*$ for all $x_j \in \mathrm{var}(\tilde{\alpha}_{\mathtt{abcd},1})$, the letters $\mathtt{b}$ and $\mathtt{d}$ each occur exactly twice in $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$ we may conclude that $\sigma'(\beta) = \sigma(\beta)$ for $\beta \in \{\beta_1, \beta_1', \beta_2, \beta_2'\}$. Therefore—and since, according to Theorem 3 of [12], the patterns $\beta_1\,\beta_1'$ and $\beta_2\,\beta_2'$ are succinct—Lemma 1 of [12] is applicable, which shows that in the given case necessarily $\sigma'(x_j) = v_0\ \mathtt{c}\ \mathtt{c}\,\mathtt{e}^{3j-1}\ \mathtt{c}\ \mathtt{c}\ v_1$, $v_0, v_1 \in \Sigma^*$, for all $x_j \in \mathrm{var}(\beta_1\,\beta_1')$ and $\sigma'(x_j) = v_2\ \mathtt{a}\ \mathtt{a}\,\mathtt{e}^{3j-1}\ \mathtt{a}\ \mathtt{a}\ v_3$, $v_2, v_3 \in \Sigma^*$, for all $x_j \in \mathrm{var}(\beta_2\,\beta_2')$. Consequently, $\sigma'(x_{23}\,x_{14}\,x_4\,x_{24}) = v_4\ \mathtt{a}\,\mathtt{a}\,\mathtt{e}^{41}\ \mathtt{a}\,\mathtt{a}\,w\ \mathtt{c}\ \mathtt{c}\,\mathtt{e}^{11}\ \mathtt{c}\ \mathtt{c}\,v_5$, $v_4, v_5 \in \Sigma^*$, for some $w \in \{\mathtt{a},\mathtt{c},\mathtt{e}\}^*$. However, for every occurrence of this subword in $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$—or, more precisely, in $\sigma(\hat{\alpha}_2)$—we have $w = v_6\ \mathtt{a}\,\mathtt{e}^{44}\,\mathtt{a}\ v_7$, $v_6, v_7 \in \Sigma^*$ (see suffix of $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$ as depicted above). Thus, we may conclude $\sigma'(x_{15}) \neq v_8\ \mathtt{a}\,\mathtt{e}^{44}\,\mathtt{a}\ v_9$, $v_8, v_9 \in \Sigma^*$, since the frequency of the subword $\mathtt{a}\,\mathtt{e}^{44}\,\mathtt{a}$ in $\sigma(\tilde{\alpha}_{\mathtt{abcd},1})$ equals $|\,\tilde{\alpha}_{\mathtt{abcd},2}\,|_{x_{15}}$ and since at least one occurrence of $\mathtt{a}\,\mathtt{e}^{44}\,\mathtt{a}$—in fact, it is even all six occurrences in $\sigma(\hat{\alpha}_2)$—is contained in $\sigma'(x_{14}\,x_4)$. This contradicts the claim $\sigma'(x_j) = v_2\ \mathtt{a}\ \mathtt{a}\,\mathtt{e}^{3j-1}\ \mathtt{a}\ \mathtt{a}\ v_3$ for all $x_j \in \mathrm{var}(\beta_2\,\beta_2')$.

Consequently, there is no substitution σ' with $\sigma'(\tilde{\alpha}_{\mathtt{abcd},2}) = \sigma(\tilde{\alpha}_{\mathtt{abcd},1})$. □

Thus, with Lemma 1 and Lemma 4, Theorem 1 is proven. Moreover, the proof of Lemma 4 shows that our way of composing example patterns cannot directly be used for the transition between $|\Sigma| = 5$ and $|\Sigma| = 6$. The argumentation on Lemma 1 is based on the fact that every substitution either matches the "easier" Cases 1 - 6 or exactly reconstructs the terminal substring of the pattern (see Case 7). We are uncertain whether these substitutions can be avoided for all patterns—and not only for our examples—in case of $|\Sigma| \geq 5$.

3.2 Some Notes

The proof of Lemma 4 can be extended canonically such that in addition to $L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},2}) \not\supseteq L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},1})$ the opposite direction $L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},1}) \not\supseteq L_{\Sigma_2}(\tilde{\alpha}_{\mathtt{abcd},2})$ is shown. Consequently, both languages are incomparable, and it seems as if, for $|\Sigma| = 4$ and $|\Sigma'| > 4$, there is no pair of patterns $\alpha, \beta \in \mathrm{Pat}_\Sigma$ such that $L_\Sigma(\alpha) = L_\Sigma(\beta)$ and $L_{\Sigma'}(\alpha) \subset L_{\Sigma'}(\beta)$. In contrast to this, for smaller alphabets there are patterns that possess such a feature, for instance

- $\alpha = x_1^2$ and $\beta = x_1^2 x_2^2$ for the transition $|\Sigma| = 1$ vs. $|\Sigma| = 2$,
- $\alpha = x_1\,\mathtt{a}\,\mathtt{b}\,x_2$ and $\beta = x_1\,\mathtt{a}\,x_2\,\mathtt{b}\,x_3$ for the transition $|\Sigma| = 2$ vs. $|\Sigma| = 3$, and
- $\alpha = \tilde{\alpha}_{\mathtt{abc},2}$ and $\beta = \tilde{\alpha}_{\mathtt{abc},1}$ for the transition $|\Sigma| = 3$ vs. $|\Sigma| = 4$.

In this context, we conjecture that, for an alphabet Σ with four letters and morphically semi-coincident patterns $\alpha, \beta \in \mathrm{Pat}_\Sigma$, necessarily $L_\Sigma(\alpha) \neq L_\Sigma(\beta)$. Particularly with regard to Theorem 2, we consider this fairly counter-intuitive.

We conclude this paper with a hint on a potential problem concerning any common normal form for $\tilde{\alpha}_{\mathsf{abcd},1}$ and $\tilde{\alpha}_{\mathsf{abcd},2}$: We conjecture that both patterns are succinct for all alphabets with at least four letters. If this is correct then, for $|\Sigma| = 4$, not only the concrete algorithm in [11] has to fail (as shown in Theorem 2), but any suchlike approach as there are E-pattern languages that presumably do not have a "natural" unique shortest normal form.

Acknowledgements. The author wishes to thank Rolf Wiehagen and a referee for their helpful comments on a draft of this paper.

References

1. D. Angluin. Finding patterns common to a set of strings. *J. Comput. Syst. Sci.*, 21:46–62, 1980.
2. D.R. Bean, A. Ehrenfeucht, and G.F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. Math.*, 85:261–294, 1979.
3. C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, chapter 6, pages 329–438. Springer, 1997.
4. G. Dány and Z. Fülöp. A note on the equivalence problem of E-patterns. *Inf. Process. Lett.*, 57:125–128, 1996.
5. G. Filè. The relation of two patterns with comparable language. In *Proc. STACS 1988*, volume 294 of *LNCS*, pages 184–192, 1988.
6. T. Harju and J. Karhumäki. Morphisms. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, chapter 7, pages 439–510. Springer, 1997.
7. T. Jiang, E. Kinber, A. Salomaa, K. Salomaa, and S. Yu. Pattern languages with and without erasing. *Int. J. Comput. Math.*, 50:147–163, 1994.
8. T. Jiang, A. Salomaa, K. Salomaa, and S. Yu. Decision problems for patterns. *J. Comput. Syst. Sci.*, 50:53–63, 1995.
9. A. Mateescu and A. Salomaa. Finite degrees of ambiguity in pattern languages. *RAIRO Inform. théor.*, 28(3–4):233–253, 1994.
10. A. Mateescu and A. Salomaa. Patterns. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, chapter 4.6, pages 230–242. Springer, 1997.
11. E. Ohlebusch and E. Ukkonen. On the equivalence problem for E-pattern languages. *Theor. Comp. Sci.*, 186:231–248, 1997.
12. D. Reidenbach. A discontinuity in pattern inference. In *Proc. STACS 2004*, volume 2996 of *LNCS*, pages 129–140, 2004.
13. D. Reidenbach. On the learnability of E-pattern languages over small alphabets. In *Proc. COLT 2004*, volume 3120 of *LNAI*, pages 140–154, 2004.
14. G. Rozenberg and A. Salomaa. *Handbook of Formal Languages*, volume 1. Springer, Berlin, 1997.
15. T. Shinohara. Polynomial time inference of extended regular pattern languages. In *Proc. RIMS Symp.*, volume 147 of *LNCS*, pages 115–127, 1982.
16. A. Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.*, 7, 1906.
17. A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.*, 1, 1912.