# Regular and Context-Free Pattern Languages Over Small Alphabets[☆]

Daniel Reidenbach[a], Markus L. Schmid[b,*]

[a] *Department of Computer Science, Loughborough University,*
*Loughborough, Leicestershire, LE11 3TU, United Kingdom*
[b] *Fachbereich 4 – Abteilung Informatik, Universität Trier, D-54296 Trier, Germany*

**Abstract**

Pattern languages are generalisations of the copy language, which is a standard textbook example of a context-sensitive and non-context-free language. In this work, we investigate a counter-intuitive phenomenon: with respect to alphabets of size 2 and 3, pattern languages can be regular or context-free in an unexpected way. For this regularity and context-freeness of pattern languages, we give several sufficient and necessary conditions and improve known results.

*Keywords:* Pattern Languages, Regular Languages, Context-Free Languages

## 1. Introduction

Within the scope of this paper, a *pattern* is a finite sequence of terminal symbols and variables, taken from two disjoint alphabets $\Sigma$ and $X$. We say that such a pattern $\alpha$ generates a word $w$ if $w$ can be obtained from $\alpha$ by substituting arbitrary words of terminal symbols for all variables in $\alpha$, where, for any variable, the substitution word must be identical for all of its occurrences in $\alpha$. More formally, a substitution is therefore a *terminal-preserving* morphism, i. e., a morphism $\sigma : (\Sigma \cup X)^* \to \Sigma^*$ that satisfies $\sigma(a) = a$ for every $a \in \Sigma$. The *pattern language* $L(\alpha)$ is then simply the set of all words that can be obtained from $\alpha$ by arbitrary substitutions. For example, the language generated by $\alpha_1 := x_1 x_1 \mathtt{aba} x_2$ (where $\Sigma := \{\mathtt{a}, \mathtt{b}\}$ and $X \supset \{x_1, x_2\}$) is the set of all words over $\{\mathtt{a}, \mathtt{b}\}$ that have any square as a prefix, an arbitrary suffix and the factor $\mathtt{aba}$ in between. Hence, e. g., $w_1 := \mathtt{abbabbabaaa}$ and $w_2 := \mathtt{bbaba}$ are included in $L(\alpha_1)$, whereas $w_3 := \mathtt{abbababb}$ and $w_4 := \mathtt{bbbabaaa}$ are not.

Pattern languages were introduced by Angluin [1] in 1980 in order to formalise the process of computing commonalities of words in some given set. Her

[☆] A preliminary version [18] of this paper was presented at the conference DLT 2012.

[*] Corresponding author

*Email addresses:* D.Reidenbach@lboro.ac.uk (Daniel Reidenbach), MSchmid@uni-trier.de (Markus L. Schmid) original definition disallows the substitution of the empty word for the variables, and therefore these languages are also referred to as *nonerasing* pattern languages (or *NE*-pattern languages for short). This notion of pattern languages was soon afterwards extended by Shinohara [20], who included the empty word as an admissible substitution word, leading to the definition of *extended* or *erasing* pattern languages (or *E*-pattern languages for short). Thus, in the above example, $w_2$ is contained in the E-pattern language, but not in the NE-pattern language of $\alpha_1$. As revealed by numerous studies, the small difference between the definitions of NE- and E-pattern languages entails substantial differences between some of the properties of the resulting (classes of) formal languages (see, e. g., Mateescu and Salomaa [14] for a survey).

Pattern languages have not only been intensively studied within the scope of inductive inference (see, e. g., Lange and Wiehagen [12], Rossmanith and Zeugmann [19], Reidenbach [17] and, for a survey, Ng and Shinohara [15]), but their properties are closely connected to a variety of fundamental problems in computer science and discrete mathematics, such as for (un-)avoidable patterns (cf. Jiang et al. [10]), word equations (cf. Mateescu and Salomaa [13]), the ambiguity of morphisms (cf. Freydenberger et al. [7]), equality sets (cf. Harju and Karhumäki [8]) and extended regular expressions (cf. Câmpeanu et al. [4]). Therefore, quite a number of basic questions for pattern languages are still open or have been resolved just recently (see, e. g., Freydenberger and Reidenbach [6], Bremer and Freydenberger [3]).

If a pattern contains each of its variables once, then this pattern can be interpreted as a regular expression, and therefore its language is regular. In contrast to this, if a pattern has at least one variable with multiple occurrences, then its languages is a variant of the well known *copy language* $\{xx \mid x \in \Sigma^*\}$, which for $|\Sigma| \geq 2$ is a standard textbook example of a context-sensitive and non-context-free language. Nevertheless, there are some well-known example patterns of the latter type that generate regular languages. For instance, the NE-pattern language of $\alpha_2 := x_1x_2x_2x_3$ is regular for $|\Sigma| = 2$, since squares are unavoidable for binary alphabets, which means that the language is co-finite. Surprisingly, for terminal alphabets of size 2 and 3, there are even certain E- and NE-pattern languages that are context-free but not regular. This recent insight is due to Jain et al. [9] and solves a longstanding open problem.

It is the purpose of our paper to further investigate this counter-intuitive existence of languages that appear to be variants of the copy language, but are nevertheless regular or context-free. Thus, we wish to establish criteria where the seemingly high complexity of a pattern does not translate into a high complexity of its language. Since, as demonstrated by Jain et al., this phenomenon does not occur for E-pattern languages if the pattern does not contain any terminal symbols or if the size of the terminal alphabet is at least 4, our investigations focus on patterns with terminal symbols and on small alphabets of sizes 2 or 3.

## 2. Definitions and Known Results

Let $\mathbb{N} := \{1, 2, 3, \ldots\}$ and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For an arbitrary alphabet $A$, a *string* (*over* $A$) is a finite sequence of symbols from $A$, and $\varepsilon$ stands for the *empty string*. The notation $A^+$ denotes the set of all nonempty strings over $A$, and $A^* := A^+ \cup \{\varepsilon\}$. For the *concatenation* of two strings $w_1, w_2$ we write $w_1 \cdot w_2$ or simply $w_1 w_2$. We say that a string $v \in A^*$ is a *factor* of a string $w \in A^*$ if there are $u_1, u_2 \in A^*$ such that $w = u_1 \cdot v \cdot u_2$. If $u_1$ or $u_2$ is the empty string, then $v$ is a *prefix* (or a *suffix*, respectively) of $w$. The notation $|K|$ stands for the size of a set $K$ or the length of a string $K$.

If we wish to refer to the symbol at a certain position $j$, $1 \leq j \leq n$, in a string $w = \mathtt{a}_1 \cdot \mathtt{a}_2 \cdot \cdots \cdot \mathtt{a}_n$, $\mathtt{a}_i \in A$, $1 \leq i \leq n$, then we use $w[j] := \mathtt{a}_j$ and if the length of a string is unknown, then we denote its last symbol by $w[-] := w[|w|]$. Furthermore, for each $j, j'$, $1 \leq j < j' \leq |w|$, let $w[j, j'] := \mathtt{a}_j \cdot \mathtt{a}_{j+1} \cdot \cdots \cdot \mathtt{a}_{j'}$ and $w[j, -] := w[j, |w|]$.

For any alphabets $A, B$, a *morphism* is a function $h : A^* \to B^*$ that satisfies $h(vw) = h(v)h(w)$ for all $v, w \in A^*$; $h$ is said to be *nonerasing* if and only if, for every $a \in A$, $h(a) \neq \varepsilon$. Let $\Sigma$ be a finite alphabet of so-called *terminal symbols* and $X$ a countably infinite set of *variables* with $\Sigma \cap X = \emptyset$. We normally assume $X := \{x_1, x_2, x_3, \ldots\}$. A *pattern* is a nonempty string over $\Sigma \cup X$, a *terminal-free pattern* is a nonempty string over $X$ and a *word* is a string over $\Sigma$. For any pattern $\alpha$, we refer to the set of variables in $\alpha$ as $\mathrm{var}(\alpha)$ and for any $x \in \mathrm{var}(\alpha)$, $|\alpha|_x$ denotes the number of occurrences of $x$ in $\alpha$. A morphism $h : (\Sigma \cup X)^* \to \Sigma^*$ is called a *substitution* if $h(a) = a$ for every $a \in \Sigma$.

**Definition 1.** Let $\alpha \in (\Sigma \cup X)^*$ be a pattern. The *E-pattern language* of $\alpha$ is defined by $L_{\mathrm{E},\Sigma}(\alpha) := \{h(\alpha) \mid h : (\Sigma \cup X)^* \to \Sigma^* \text{ is a substitution}\}$. The *NE-pattern language* of $\alpha$ is defined by $L_{\mathrm{NE},\Sigma}(\alpha) := \{h(\alpha) \mid h : (\Sigma \cup X)^* \to \Sigma^* \text{ is a nonerasing substitution}\}$.

We denote the class of *regular* languages, *context-free* languages, E-*pattern* languages over $\Sigma$ and NE-*pattern* languages over $\Sigma$ by REG, CF, $\text{E-PAT}_\Sigma$ and $\text{NE-PAT}_\Sigma$, respectively. We use regular expressions as they are commonly defined (see, e. g., Yu [22]) and for any regular expression $r$, $L(r)$ denotes the language described by $r$.

We recapitulate regular and block-regular patterns as defined by Shinohara [21] and Jain et al. [9]. A pattern $\alpha$ is a *regular* pattern if, for every $x \in \mathrm{var}(\alpha)$, $|\alpha|_x = 1$. Every factor of variables of $\alpha$ that is delimited by terminal symbols is called a variable block. More precisely, for every $i, j$, $1 \leq i \leq j \leq |\alpha|$, $\alpha[i, j]$ is a *variable block* if and only if $\alpha[k] \in X$, $i \leq k \leq j$, $\alpha[i-1] \in \Sigma$ or $i = 1$ and $\alpha[j+1] \in \Sigma$ or $j = |\alpha|$. A pattern $\alpha$ is *block-regular* if in every variable block of $\alpha$ there occurs at least one variable $x$ with $|\alpha|_x = 1$. Let $\mathrm{Z} \in \{\mathrm{E}, \mathrm{NE}\}$. The class of Z-pattern languages defined by regular patterns and block-regular patterns are denoted by $\text{Z-PAT}_{\Sigma,\text{reg}}$ and $\text{Z-PAT}_{\Sigma,\text{b-reg}}$, respectively. To avoid any confusion, we explicitly mention that the term regular pattern always refers to a pattern with the syntactical property of being a regular pattern and a regular E- or NE-pattern language is a pattern language that is regular, but that is not necessarily given by a regular pattern.

In order to prove some of the technical claims in this paper, the following two versions of the pumping lemma for regular languages as stated by Yu [22] shall be used.

**Pumping Lemma 1.** *Let $L \subseteq \Sigma^*$ be a regular language. Then there is a constant $n$, depending on $L$, such that for every $w \in L$ with $|w| \geq n$ there exist $x, y, z \in \Sigma^*$ such that $w = xyz$ and*

1. $|xy| \leq n$,
2. $|y| \geq 1$,
3. $xy^k z \in L$ *for every* $k \in \mathbb{N}_0$.

**Pumping Lemma 2.** *Let $L \subseteq \Sigma^*$ be a regular language. Then there is a constant $n$, depending on $L$, such that for all $u, v, w \in \Sigma^*$, if $|w| \geq n$, then there exist $x, y, z \in \Sigma^*$, $y \neq \varepsilon$, such that $w = xyz$ and, for every $k \in \mathbb{N}_0$, $uxy^k zv \in L$ if and only if $uwv \in L$.*

We also need the following generalisation of Ogden's Lemma:

**Lemma 3 (Bader and Moura [2]).** *Let $L \subseteq \Sigma^*$ be a context-free language. Then there is a constant $n$, such that for every $z \in L$, if $d$ positions in $z$ are "distinguished" and $e$ positions are "excluded", with $d > n^{(e+1)}$, then there exist $u, v, w, x, y \in \Sigma^*$ such that $z = uvwxy$ and*

1. *$vx$ contains at least one distinguished position and no excluded positions,*
2. *if $r$ is the number of distinguished positions in $vwx$ and $s$ is the number of excluded positions in $vwx$, then $r \leq n^{(s+1)}$,*
3. $uv^i wx^i y \in L$ *for every* $i \in \mathbb{N}_0$.

Next, we give a summary of subclasses of patterns for which characterisations of the corresponding regular and context-free pattern languages are known.

It can be easily shown that every E- or NE-pattern language over a unary alphabet is a regular language (cf. Reidenbach [16] for further details). Hence, the classes of regular and context-free pattern languages over a unary alphabet are trivially characterised. Jain et al. [9] show that for any alphabet of cardinality at least 4, the regular and context-free E-pattern languages are characterised by the class of regular patterns.

**Theorem 4 (Jain et al. [9]).** *Let $\Sigma$ be an arbitrary alphabet. If $|\Sigma| \geq 4$, then* $(\text{E-PAT}_\Sigma \cap \text{REG}) = (\text{E-PAT}_\Sigma \cap \text{CF}) = \text{E-PAT}_{\Sigma,\text{reg}}$.

Unfortunately, the above mentioned cases are the only complete characterisations of regular or context-free pattern languages that are known to date. In particular, characterisations of the regular and context-free E-pattern languages with respect to alphabets with cardinality 2 and 3, and characterisations of the regular and context-free NE-pattern languages with respect to alphabets with cardinality at least 2 are still missing. In the following, we shall briefly summarise the known results in this regard and the reader is referred to Jain et al. [9] and Reidenbach [16] for further details.

Jain et al. [9] show that there exist regular E-pattern languages with respect to alphabet sizes 2 and 3 that cannot be described by regular patterns. Moreover, there exist non-regular context-free E-pattern languages with respect to alphabet sizes 2 and 3. Regarding NE-pattern languages, it is shown that, for every alphabet $\Sigma$ with cardinality at least 2, the class $(\text{NE-PAT}_\Sigma \cap \text{REG})$ is not characterised by regular patterns and with respect to alphabet sizes 2 and 3 it is not characterised by block-regular patterns either. Furthermore, for alphabet sizes 2 and 3, there exist non-regular context-free NE-pattern languages and for alphabets with cardinality of at least 4 this question is still open.

## 3. Regularity and Context-Freeness of Pattern Languages: Sufficient Conditions and Necessary Conditions

Since their introduction by Shinohara [21], it has been known that, for both the E and NE case and for any terminal alphabet, regular patterns can only describe regular languages. This is an immediate consequence of the fact that regular patterns do not use the essential mechanism of patterns, i. e., repeating variables in order to define sets of words that contain repeated occurrences of variable factors. Jain et al. [9] extend the concept of regular patterns to block-regular patterns, defined in Section 2. By definition, every regular pattern is a block-regular pattern. Furthermore, in the E case, every block-regular pattern $\alpha$ is equivalent to the regular pattern obtained from $\alpha$ by substituting every variable block by a single occurrence of a variable.

**Proposition 5.** *Let $\Sigma$ be some terminal alphabet and let $\alpha \in (\Sigma \cup X)^*$ be a pattern. If $\alpha$ is regular, then $L_{\text{NE},\Sigma}(\alpha) \in \text{REG}$. If $\alpha$ is block-regular, then $L_{\text{E},\Sigma}(\alpha) \in \text{REG}$.*

As mentioned in Section 2, for alphabets of size at least 4, both the class of regular patterns and the class of block-regular patterns characterise the set of regular and context-free E-pattern languages. However, in the NE case as well as in the E case with respect to alphabets of size 2 or 3, Jain et al. [9] demonstrate that block-regular patterns do not characterise the set of regular or context-free pattern languages.

Obviously, the regularity of languages given by regular patterns or block-regular patterns follows from the fact that there are variables that occur only once in the pattern. Hence, it is the next logical step to ask whether or not the existence of variables with only one occurrence is also necessary for the regularity or the context-freeness of a pattern language. With respect to terminal-free patterns, a positive answer to this question can be easily derived from existing results by Jain et al. [9].

**Theorem 6 (Jain et al. [9]).** *Let $\Sigma$ be a terminal alphabet with $|\Sigma| \geq 2$ and let $\alpha$ be a terminal-free pattern with $|\alpha|_x \geq 2$, for every $x \in \mathrm{var}(\alpha)$. Then $L_{\mathrm{E},\Sigma}(\alpha) \notin \mathrm{CF}$ and $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$.*

Proof. Let $\Sigma'$ be an alphabet with $|\Sigma'| = 2$ and $\Sigma' \subseteq \Sigma$. By Lemma 11 of [9], it follows that $L_{\mathrm{E},\Sigma'}(\alpha) \notin \mathrm{CF}$. Since $L_{\mathrm{E},\Sigma}(\alpha) \cap \Sigma'^*$ equals $L_{\mathrm{E},\Sigma'}(\alpha)$ and since the class of context-free languages is closed under intersection with regular sets, we can conclude that $L_{\mathrm{E},\Sigma}(\alpha) \notin \mathrm{CF}$.

In order to show $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$, we can apply the proof of Theorem 6.*a* of [9], which states that for any terminal alphabet $\Sigma'$ with $|\Sigma'| \geq 4$ and for any pattern $\beta$ that is not block-regular, $L_{\mathrm{NE},\Sigma'}(\beta)$ is not a regular language. However, for terminal-free patterns in which every variable occurs at least twice this proof also works for an alphabet of size 2 and 3, since we do not need the two terminal symbols to both sides of the variable block (see [9] for details). □

We can note that Proposition 5 and Theorem 6 characterise the regular and context-free E-pattern languages given by terminal-free patterns with respect to alphabets of size at least 2. More precisely, for every alphabet $\Sigma$ with $|\Sigma| \geq 2$ and for every terminal-free pattern $\alpha$, if $\alpha$ is block-regular, then $L_{\mathrm{E},\Sigma}(\alpha)$ is regular (and, thus, also context-free) and if $\alpha$ is not block-regular, then every variable of $\alpha$ occurs at least twice, which implies that $L_{\mathrm{E},\Sigma}(\alpha)$ is neither regular nor context-free.

However, for the NE case, we cannot hope for such a simple characterisation. This is due to the close relationship between the regularity of NE-pattern languages and the combinatorial phenomenon of unavoidable patterns, as already mentioned in Section 1.

In the following, we concentrate on E-pattern languages over alphabets of size 2 and 3 (since for all other alphabet sizes complete characterisations are known) that are given by patterns that are *not* terminal-free (since, as described above, the characterisation of regular and context-free E-pattern languages given by terminal-free patterns has been settled). Nevertheless, some of our results also hold for the NE case and we shall always explicitly mention if this is the case.

The next two results present a sufficient condition for the non-regularity and a sufficient condition for the non-context-freeness of pattern languages over small alphabets. More precisely, we generalise Theorem 6 to patterns that are not necessarily terminal-free. The first result states that for a pattern $\alpha$ (that may contain terminal symbols), if every variable in $\alpha$ occurs at least twice, then both the E- and NE-pattern language of $\alpha$, with respect to alphabets of size at least 2, is not regular.

**Theorem 7.** *Let $\Sigma$ be a terminal alphabet with $|\Sigma| \geq 2$, let $\alpha \in (\Sigma \cup X)^*$, and let $\mathrm{Z} \in \{\mathrm{E}, \mathrm{NE}\}$. If, for every $x \in \mathrm{var}(\alpha)$, $|\alpha|_x \geq 2$, then $L_{\mathrm{Z},\Sigma}(\alpha) \notin \mathrm{REG}$.*

Proof. We only prove that $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$ since $L_{\mathrm{E},\Sigma}(\alpha) \notin \mathrm{REG}$ can be shown in exactly the same way. To this end, we assume to the contrary that $L_{\mathrm{NE},\Sigma}(\alpha) \in \mathrm{REG}$ and we let $n$ be the constant from Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma}(\alpha)$. Furthermore, we assume that $\alpha := u_0 \cdot y_1 \cdot u_1 \cdot y_2 \cdot u_2 \cdot \cdots \cdot u_{k-1} \cdot y_k \cdot u_k$, where $y_i \in X$, $1 \leq i \leq k$, and $u_i \in \Sigma^*$, $0 \leq i \leq k$. Now, we let $w$ be the word obtained from $\alpha$ by substituting every variable by the word $\mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a}$, i. e.,

$$w = u_0 \cdot \mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a} \cdot u_1 \cdot \mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a} \cdot u_2 \cdot \cdots \cdot u_{k-1} \cdot \mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a} \cdot u_k \,.$$

By first applying Pumping Lemma 2 to the factor $\mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a}$ that results from $y_1$, then to the factor $\mathtt{b}\mathtt{a}^n\mathtt{b}^n\mathtt{a}$ that results from $y_2$ and so on, we can obtain the word

$$w' := u_0 \cdot \mathtt{b}\mathtt{a}^{n_1}\mathtt{b}^{n_2}\mathtt{a} \cdot u_1 \cdot \mathtt{b}\mathtt{a}^{n_3}\mathtt{b}^{n_4}\mathtt{a} \cdot u_2 \cdot \cdots \cdot u_{k-1} \cdot \mathtt{b}\mathtt{a}^{n_{2k-1}}\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k \,,$$

where $n \times |\alpha| < n_1$, and, for every $i$, $1 \leq i \leq 2k-1$, $n_i \times |\alpha| < n_{i+1}$. We shall now show that $w' \notin L_{\mathrm{NE},\Sigma}(\alpha)$. To this end, we assume to the contrary that there exists a substitution $h$ with $h(\alpha) = w'$. Let $p$, $1 \leq p \leq |\alpha|$, be such that $\alpha[p,-]$ is the shortest suffix of $\alpha$ such that $\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$ is a suffix of $h(\alpha[p,-])$. If $h(\alpha[p,-]) = v \cdot \mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$, $v \neq \varepsilon$, then $\alpha[p]$ must be a variable, since otherwise $\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$ is also a suffix of $h(\alpha[p+1,-])$ which implies that $\alpha[p,-]$ is not the shortest suffix of $\alpha$ such that $\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$ is a suffix of $h(\alpha[p,-])$. Moreover, for similar reasons, we can conclude that $h(\alpha[p]) = v \cdot v'$, where $v'$ is a non-empty prefix of $\mathtt{b}^{n_{2k}}$. If $h(\alpha[p])$ contains the whole factor $\mathtt{a}^{n_{2k-1}}$, then, since $\alpha[p]$ is a repeated variable in $\alpha$, there are two non-overlapping occurrences of factor $\mathtt{a}^{n_{2k-1}}$ in $h(\alpha)$, which is a contradiction, since there are no two non-overlapping occurrences of factor $\mathtt{a}^{n_{2k-1}}$ in $w'$. So we can conclude that either $h(\alpha[p,-]) = \mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$ or $h(\alpha[p,-]) = \mathtt{a}^m \cdot \mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$ and $\alpha[p]$ is a variable with $h(\alpha[p]) = \mathtt{a}^m \cdot \mathtt{b}^l$, $1 \leq m < n_{2k-1}$, $l \neq 0$.

There must exist at least one variable $x \in \mathrm{var}(\alpha)$ with $|h(x)| > n_{2k-1}$, since otherwise $|h(\alpha)| \leq |\alpha| \times n_{2k-1} < n_{2k} < |w'|$, which is a contradiction. Now let $z \in \mathrm{var}(\alpha)$ be such a variable, i. e., $|h(z)| > n_{2k-1}$. We recollect that $h(\alpha[1,p-1]) := u_0 \cdot \mathtt{b}\mathtt{a}^{n_1}\mathtt{b}^{n_2}\mathtt{a} \cdot u_1 \cdot \cdots \cdot u_{k-1} \cdot \mathtt{b}\mathtt{a}^{n_{2k-1}-m}$. If $z \in \mathrm{var}(\alpha[1,p-1])$, then there are two cases to consider. If, for some $i$, $1 \leq i \leq k-1$, $h(z)$ contains a factor $\mathtt{a}\mathtt{b}^{n_{2i}}\mathtt{a}$ or a factor $\mathtt{b}\mathtt{a}^{n_{2i-1}}\mathtt{b}$, then we obtain a contradiction, since in $w'$ there is exactly one occurrence of such a factor, but there are at least two occurrences of variable $z$ in $\alpha$. If, on the other hand, $h(z)$ contains no such factor, then $h(z)$ is a factor of the suffix $\mathtt{b}^{n_{2k-2}}\mathtt{a} \cdot u_{k-1} \cdot \mathtt{b}\mathtt{a}^{n_{2k-1}-m}$ of $h(\alpha[1,p-1])$. Since $|h(z)| > n_{2k-1}$, this implies that $h(z)$ must have a suffix $\mathtt{a}^q$, where $q > n_{2k-1} - (n_{2k-2} + |u_{k-1}| + 2)$. We observe that

$$\begin{aligned} &n_{2k-1} - (n_{2k-2} + |u_{k-1}| + 2) > n_{2k-1} - (3 \times n_{2k-2}) > \\ &|\alpha| \times n_{2k-2} - (3 \times n_{2k-2}) = (|\alpha| - 3) \times n_{2k-2} \,. \end{aligned}$$

We can therefore conclude that, since $(|\alpha|-3) \times n_{2k-2} > (|\alpha|-3) \times |\alpha| \times n_{2k-3} > n_{2k-3}$, $q > n_{2k-3}$. This directly implies that in $h(\alpha[1,p-1])$ there does not exist another occurrence of factor $\mathtt{a}^q$ and, thus, there is exactly one occurrence of variable $z$ in $\alpha[1,p-1]$, which implies that there must be another occurrence of variable $z$ in $\alpha[p,-]$. This particularly means that there is an occurrence of

$h(z)$ in $h(\alpha[p,-]) = \mathtt{a}^m \cdot \mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$. We recall that $h(z)$ contains $\mathtt{a}^q$ as a suffix, which implies that in $h(\alpha[p,-])$, $h(z)$ cannot end in $\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$, since this means that the whole suffix $\mathtt{a}^q$ is contained in $\mathtt{b}^{n_{2k}}\mathtt{a} \cdot u_k$. So $h(z)$ must entirely be contained in $\mathtt{a}^m$, which is a contradiction, since $|h(z)| > n_{2k-1}$ and $m < n_{2k-1}$.

This proves that the word $w'$ is not in $L_{\mathrm{NE},\Sigma}(\alpha)$, which, by Pumping Lemma 2, implies $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$. □

For alphabets of size at least 3 we can strengthen Theorem 7, i. e., if every variable in a pattern $\alpha$ occurs at least twice, then the E- and NE-pattern language of $\alpha$ is not context-free.

**Theorem 8.** *Let $\Sigma$ be a terminal alphabet with $|\Sigma| \geq 3$, let $\alpha \in (\Sigma \cup X)^+$, and let $Z \in \{\mathrm{E}, \mathrm{NE}\}$. If, for every $x \in \mathrm{var}(\alpha)$, $|\alpha|_x \geq 2$, then $L_{Z,\Sigma}(\alpha) \notin \mathrm{CF}$.*

Proof. Let $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \subseteq \Sigma$. We prove the theorem solely for the case that there exists a unique factorisation of $\alpha$ that reads as follows:

$$\alpha = u\,\beta\,f\,\alpha',$$

with $u \in \Sigma^*$, $\beta \in X^+$, $f \in \Sigma$ and $\alpha' \in (\Sigma \cup X)^*$. If this factorisation does not exist, then, for some $u' \in \Sigma^*$ and $\alpha'' \in X^*$, $\alpha = u'\,\alpha''$. For such a structure of $\alpha$, our reasoning can be adapted easily (and the statement of the theorem also follows from Lemma 11 in [9]). Furthermore, we assume that $f = \mathtt{c}$; if $f = \mathtt{a}$ or $f = \mathtt{b}$, then it is again straightforward to adapt the proof below.

We wish to prove the theorem by applying the contraposition of the generalisation of Ogden's Lemma stated in Lemma 3. Hence, we shall consider a word $z \in L_{Z,\Sigma}(\alpha)$ and label some of its letters as *distinguished* or *excluded* such that the conditions of Lemma 3 are satisfied. We then consider all possible factoristions $z = v_1 v_2 v_3 v_4 v_5$ that fit with the requirements of that lemma, and we shall demonstrate for each of them that $z' := v_1 v_2^0 v_3 v_4^0 v_5 \notin L_{Z,\Sigma}(\alpha)$. This then directly implies $L_{Z,\Sigma}(\alpha) \notin \mathrm{CF}$.

In order to produce an appropriate word $z$, let $\tau : (\Sigma \cup X)^* \to \Sigma^*$ be any substitution satisfying the following conditions:

1. For every $y \notin \mathrm{var}(\beta)$, $\tau(y) := \mathtt{a}$. Note that, in order to avoid a more involved proof, $f$ must not be contained in $\tau(y)$. Hence, if $f \neq \mathtt{c}$, then the definition of $\tau$ for the variables $y \notin \mathrm{var}(\beta)$ must be modified accordingly.
2. For every $y \in \mathrm{var}(\beta)$, $\tau(y) := \mathtt{a}\mathtt{b}^{1_y}\mathtt{a}\,\mathtt{a}\mathtt{b}^{2_y}\mathtt{a}\,\mathtt{a}\mathtt{b}^{3_y}\mathtt{a} \cdots \mathtt{a}\mathtt{b}^{g_y}\mathtt{a}$, $1_y, 2_y, \ldots, g_y \in \mathbb{N}$, where
   (a) $g := |\beta| + 2$ or, if $|\beta| \in \{1, 2\}$, $g := 5$ (note that this separate consideration of $|\beta| \in \{1, 2\}$ is merely for the sake of a uniform reasoning below),
   (b) for all $y, y' \in \mathrm{var}(\beta)$ and for all $i, i' \in \{1, 2, \ldots, g\}$, $i_y = i'_{y'}$ iff $y = y'$ and $i = i'$,
   (c) for every $y \in \mathrm{var}(\beta)$ and for every $i \in \{1, 2, \ldots, g\}$, $i_y > |\,\mathrm{res}(\alpha)|$, where $\mathrm{res}(\alpha)$ is the word that results from $\alpha$ by deleting all variables from it.

Note that if $f \neq \mathtt{c}$, then we simply choose any two distinct letters that differ from $f$ in order to define $\tau$ for the variables that occur in $\beta$. These two letters must exist since we assume $|\Sigma| \geq 3$. Furthermore, from now onwards, we call any word $\mathtt{ab}^{i_y}\mathtt{a}$, $i \in \{1, 2, \ldots, g\}$, a *segment* of $\tau(y)$, $y \in \operatorname{var}(\alpha)$.

3. Let $x$ be the leftmost variable in $\beta$ that satisfies, for every $y \in \operatorname{var}(\beta)$, $|\alpha|_x \leq |\alpha|_y$. Let $m_1$ be the number of terminal symbols in $\alpha$ between the leftmost occurrence and the second occurrence of $x$ in $\alpha$ (this second occurrence must exist since we assume, for every $y \in \operatorname{var}(\alpha)$, $|\alpha|_y \geq 2$), and let $m_2$ be number of segments between the leftmost and the second occurrence of $\mathtt{ab}^{2_x}\mathtt{a}$ in $\tau(\alpha)$. Let then $3_x$ (i. e., the number of occurrences of the letter $\mathtt{b}$ in the third segment of $\tau(x)$) be any number that satisfies
$$3_x > n^{m_1+2m_2+1_x+g_x+1},$$
where $n$ is the constant provided by Lemma 3.
4. Let $2_x$ (i. e., the number of occurrences of the letter $\mathtt{b}$ in the second segment of $\tau(x)$) be any number that satisfies
$$2_x > n^{|\operatorname{res}(\alpha)|+2g\sum_{y\in\operatorname{var}(\beta)}|\alpha|_y+1_x+3_x+g_x+1},$$
where $n$ again is the constant provided by Lemma 3.

Informally, we can summarise the properties of $\tau$ as follows: Every variable in $\beta$ is mapped to a word that consists of $|\beta| + 2$ unique segments $\mathtt{ab}^+\mathtt{a}$, and all variables that have no occurrence in $\beta$ are mapped to the letter $\mathtt{c}$. There are specific constraints regarding the minimum length of the second and the third segment of $\tau(x)$, where $x$ is the leftmost variable in $\beta$ that, among all variables in $\beta$, has the smallest number of occurrences in $\alpha$. More precisely, the length of these segments depends on the constant $n$ from Lemma 3, and the second segment is substantially longer than the third one.

Before we can apply Lemma 3 in the way summarised above, we need to label a number of positions in $z := \tau(\alpha)$:

5. The following positions are *excluded*:
   (a) all positions that result from the substitution of terminal symbols in $\alpha$,
   (b) all occurrences of the letter $\mathtt{a}$ in $z$ that are the first or the last letter of any segment of any $\tau(y)$, $y \in \operatorname{var}(\alpha)$,
   (c) all occurrences of the letter $\mathtt{b}$ in the first occurrence of the factors $\mathtt{ab}^{1_x}\mathtt{a}$ and $\mathtt{ab}^{g_x}\mathtt{a}$ in $z$, and
   (d) all occurrences of the letter $\mathtt{b}$ in the second occurrence of the factor $\mathtt{ab}^{3_x}\mathtt{a}$ in $z$.
6. The following positions are *distinguished*:
   (a) all occurrences of the letter $\mathtt{b}$ in the first occurrence of the factor $\mathtt{ab}^{2_x}\mathtt{a}$ and in the first occurrence of the factor $\mathtt{ab}^{3_x}\mathtt{a}$ in $z$.
7. All other positions in $z$ stay unlabeled.

Since $z \in L_{Z,\Sigma}(\alpha)$ is obviously true, we merely need to test whether the above conditions on the number of *excluded* and *distinguished* positions in $z$ satisfy the conditions of Lemma 3 before we apply this lemma: The number $d$ of *distinguished* positions in $z$ equals $2_x + 3_x$, whereas the number $e$ of *excluded* positions equals

$$|\operatorname{res}(\alpha)| + 2g \sum_{y \in \operatorname{var}(\beta)} |\alpha|_y + 1_x + 3_x + g_x.$$

Hence,

$$\begin{aligned} d &> n^{|\operatorname{res}(\alpha)|+2g\sum_{y\in\operatorname{var}(\beta)}|\alpha|_y+1_x+3_x+g_x+1} + 3_x \\ &> n^{|\operatorname{res}(\alpha)|+2g\sum_{y\in\operatorname{var}(\beta)}|\alpha|_y+1_x+3_x+g_x+1} \\ &= n^{e+1}, \end{aligned}$$

which implies that Lemma 3 is applicable.

We now consider any factorisation $z = v_1v_2v_3v_4v_5$ that satisfies conditions 1 and 2 of Lemma 3. We define $z' := v_1v_2^0v_3v_4^0v_5$, and we demonstrate that $z' \notin L_{Z,\Sigma}(\alpha)$: Since all positions in $z$ that result from a substitution of a terminal symbol in $\alpha$ are *excluded* (see Point 5a) and since $v_2v_4$ must not contain any *excluded* positions, $z'$ contains at most two segments that are shorter than their counterparts in $z$. Furthermore, as $v_2v_4$ must include at least one *distinguished* position, there must be at least one such segment, namely the first occurrence of $\mathtt{ab}^{2_x}\mathtt{a}$ or the first occurrence of $\mathtt{ab}^{3_x}\mathtt{a}$ in $z$ (due to Point 6). The second original segment in $z$ that is affected by pumping (if any) is not the second occurrence of $\mathtt{ab}^{3_x}\mathtt{a}$, as its positions are *excluded* (see Point 5d). Furthermore, if the first occurrence of $\mathtt{ab}^{2_x}\mathtt{a}$ in $z$ is shortened by pumping, then the second segment that is shortened (if any) is not the second occurrence of this segment – otherwise, there would be at least

$$3_x + 1 > n^{m_1+2m_2+1_x+g_x+1} + 1$$

*distinguished* positions in $v_2v_3v_4$, whereas the number of *excluded* positions in $v_2v_3v_4$ would be $m_1 + 2m_2 + 1_x + g_x$ (see Points 3 and 5), which conflicts with condition 2 of Lemma 3, stating that, for the number $r$ of *distinguished* positions and the number $s$ of *excluded* positions in $v_2v_3v_4$, $r \leq n^{s+1}$ needs to be satisfied. Incorporating the basic properties of $\tau$ stated in Points 2a to 2c, we can summarise these observations as follows:

*Claim.* The number of occurrences of the factor $w_x$ in the word $z'$ is strictly smaller than $|\alpha|_x$, where

$$w_x := \mathtt{ab}^{2'_x}\mathtt{aab}^{3'_x}\mathtt{aab}^{4'_x}\mathtt{a} \ldots \mathtt{ab}^{g'_x-1}\mathtt{a},$$

with $2'_x < 2_x$ or $3'_x < 3_x$, and $i'_x < i_x$ for at most two of the exponents $i'_x$, $2 \leq i \leq g-1$.

We now assume to the contrary that there exists a substitution $\tau'$ satisfying $\tau'(\alpha) = z'$, and we shall demonstrate that this assumption contradicts the Claim.

Let $\hat{w}_{\tau(\beta)}$ be the factor of $z'$ that results from the application of the pumping operation to $\tau(\beta)$. Since $\mathtt{a} \neq f \neq \mathtt{b}$, since $\tau(\beta) \in \{\mathtt{a}, \mathtt{b}\}^*$ (see Point 2) and since $\tau(y) = \mathtt{a}$ for all $y \in \mathrm{var}(\alpha) \setminus \mathrm{var}(\beta)$ (see Point 1), $\tau'(\beta) = \hat{w}_{\tau(\beta)}$ needs to be satisfied. Therefore there exists a *shortest* factor $\gamma$ of $\beta$ such that

- the word $w_x$ is a factor of $\tau'(\gamma)$ and
- there exists an $x' \in \mathrm{var}(\gamma)$ such that $\tau'(x')$ contains a segment of $\tau(x')$.

Note that such a factor $\gamma$ must exist. Furthermore, $x = x'$ is possible, but not necessary. More precisely, $\gamma$ can (but does not need to) contain $x$ and at most two variables other than $x$ satisfying the property of $x'$ – one to the right, and one to the left of $x$. For the sake of a less involved presentation, we assume that $x \in \mathrm{var}(\gamma)$, $x \neq x'$ and $\gamma = x\gamma' x'$, $\gamma' \in X^*$, and we note that our reasoning can be easily adapted to the other cases.

We now make use of the number of segments every variable is mapped to by $\tau$, as specified in Point 2 (note that this part of our proof is related to the the reasoning on the inclusion problem for terminal-free E-pattern languages provided by Filè [5] and Jiang et al. [11]): Since every variable is mapped by $\tau$ to $|\mathrm{var}(\beta)| + 2$ unique segments, since at most two of these segments have been shortened by pumping, and since $\tau'(\beta) = \hat{w}_{\tau(\beta)}$, we can select, for every $y \in \beta$, one full segment $s_y$ of $\tau(y)$ that is contained in $\tau'(x_y)$ for an $x_y \in \mathrm{var}(\beta)$. Furthermore, we can postulate, that $s_{x'}$ is contained in $\tau'(x')$. Note that this is not possible for any variables other than $x$ and $x'$, since otherwise $\gamma$ would not be the shortest factor of $\beta$ satisfying the conditions of the definition. We now delete from $\tau'(\gamma)$ all segments but the said $s_y$ for all $y \in \gamma$, and we call the resulting word $w''$. Since all $s_y$ are unique, we can define a morphism from $w''$ to $\gamma$, and this implies that we can define a morphism $\phi : X^* \to X^*$ mapping $\gamma$ to $\gamma$ that, due to our choice of segments $s_y$, has the following properties:

- $|\phi(x')| \geq 2$ and $x' \in \mathrm{var}(\phi(x'))$,
- for every $y \in \mathrm{var}(\gamma) \setminus \{x, x'\}$, $y \notin \mathrm{var}(\phi(y))$ and
- $\phi(x) = x$ or $\phi(x) = \varepsilon$.

Since $\phi$ is defined using unique segments that occur in both $\tau(\alpha)$ and $\tau'(\alpha)$, this implies that, for *every* occurrence of $x'$ in $\alpha$, $x'$ must occur in the factor $\gamma' x'$. We cannot necessarily conclude, solely from the definition of $\phi$, that every occurrence of $x'$ must even occur in the factor $\gamma = x\gamma' x'$, as $\phi(x) = x$ is possible. However, since

- $\tau'(\gamma' x')$ must contain at least the complete rightmost segment of $\tau(x)$ (otherwise $\gamma'$ and $x'$ would not have been included in $\gamma$),
- $\tau'(\gamma' x')$ must contain all segments of the leftmost variable in $\tau'(\gamma')$ (again due to the definition of $\gamma$), and

- the rightmost segment of $\tau(x)$, which is immediately to the right of the factor $w_x$ in $\tau(\gamma)$, cannot be affected by the pumping operation as its positions are *excluded* (see Point 5c), which means that it is longer than any factor over $\Sigma$ in $\alpha$,

we can indeed conclude that every occurrence of $x'$ is in a factor $\gamma = x\gamma' x'$. Since $x$ has been chosen as a variable in $\beta$ with a minimum number of occurrences in $\alpha$ (see Point 3), we also know that the same holds for every occurrence of $x$. This implies $|\alpha|_x = |\alpha|_{x'}$ and, hence, $|\gamma|_x = 1$.

Thus, the number of occurrences of $\tau'(\gamma)$ in $z' = \tau'(\alpha)$ equals the number of occurrences of $x$ in $\alpha$. Due to the definition of $\gamma$, we therefore have exactly $|\alpha|_x$ occurrences of the factor $w_x$ in $z'$. This is a contradiction to the Claim.

Hence, there is no substitution $\tau'$ satisfying $\tau'(\alpha) = z'$. According to Lemma 3, this means that $L_{Z,\Sigma}(\alpha) \notin \mathrm{CF}$. □

At this point, we recall that patterns, provided that they contain repeated variables, describe languages that are generalisations of the copy language, which strongly suggests that these languages are context-sensitive, but not context-free or regular. However, as stated in Section 1, for small alphabets this is not necessarily the case and the above results provide a strong indication of where to find this phenomenon of regular and context-free copy languages. More precisely, by Theorems 7 and 8, the existence of variables with only one occurrence is crucial. Furthermore, since, in the terminal-free case, regular and context-free E-pattern languages are characterised in a compact and simple manner, we should also focus on patterns containing terminal symbols.

Consequently, we concentrate on the question of how the occurrences of terminal symbols in conjunction with non-repeated variables can cause E-pattern languages to become regular. To this end, we shall now consider some simply structured examples of such patterns for which we can formally prove whether or not they describe a regular language with respect to terminal alphabets $\Sigma_2 := \{\mathtt{a}, \mathtt{b}\}$ and $\Sigma_{\geq 3}$, where $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \subseteq \Sigma_{\geq 3}$. Most parts of the following propositions require individual proofs, some of which, in contrast to the simplicity of the example patterns, are surprisingly involved. If, for some pattern $\alpha$ and $\mathrm{Z} \in \{\mathrm{E}, \mathrm{NE}\}$, $L_{\mathrm{Z},\Sigma_2}(\alpha) \notin \mathrm{REG}$, then $L_{\mathrm{Z},\Sigma_{\geq 3}}(\alpha) \notin \mathrm{REG}$. This follows directly from the fact that regular languages are closed under intersection. Hence, in the following examples, we consider $L_{\mathrm{Z},\Sigma_{\geq 3}}(\alpha)$ only if $L_{\mathrm{Z},\Sigma_2}(\alpha)$ is regular.

Firstly, we consider the pattern $x_1 \cdot d \cdot x_2 x_2 \cdot d' \cdot x_3$, which, for all choices of $d, d' \in \{\mathtt{a}, \mathtt{b}\}$, describes a regular E-pattern language with respect to $\Sigma_2$, but a non-regular E-pattern language with respect to $\Sigma_{\geq 3}$.

**Proposition 9.**

$$\begin{aligned}
L_{\mathrm{E},\Sigma_2}(x_1 \; \mathtt{a} \; x_2 \; x_2 \; \mathtt{a} \; x_3) &\in \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_{\geq 3}}(x_1 \; \mathtt{a} \; x_2 \; x_2 \; \mathtt{a} \; x_3) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_2}(x_1 \; \mathtt{a} \; x_2 \; x_2 \; \mathtt{b} \; x_3) &\in \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_{\geq 3}}(x_1 \; \mathtt{a} \; x_2 \; x_2 \; \mathtt{b} \; x_3) &\notin \mathrm{REG}\;.
\end{aligned}$$

Proof. Let $\alpha_1 := x_1\mathtt{a}x_2x_2\mathtt{a}x_3$ and let $\alpha_2 := x_1\mathtt{a}x_2x_2\mathtt{b}x_3$. It follows from Lemmas 16 and 14, respectively, that $L_{\mathrm{E},\Sigma_2}(\alpha_1)$ and $L_{\mathrm{E},\Sigma_2}(\alpha_2)$ are regular languages. Hence, it only remains to prove that $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_1) \notin \mathrm{REG}$ and $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2) \notin \mathrm{REG}$.

We assume that $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_1) \in \mathrm{REG}$ and we shall show that this assumption leads to a contradiction. Let $w := \mathtt{a} \cdot \mathtt{c}^n\mathtt{b} \cdot \mathtt{c}^n\mathtt{b} \cdot \mathtt{a} \in L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_1)$, where $n$ is the constant of Pumping Lemma 2 with respect to $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_1)$. By Pumping Lemma 2, there exists a word $w' := \mathtt{a}\cdot\mathtt{c}^n\mathtt{b}\cdot\mathtt{c}^{n'}\mathtt{b}\cdot\mathtt{a}$, $n < n'$, with $w' \in L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_1)$, which is obviously not the case.

Similarly, we can show that the assumption $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2) \in \mathrm{REG}$ leads to a contradiction. Let $v := \mathtt{a} \cdot \mathtt{c}^n\mathtt{b} \cdot \mathtt{c}^n\mathtt{b} \cdot \mathtt{b} \in L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$, where $n$ is odd and $n$ is greater than the constant of Pumping Lemma 2 with respect to $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$. By Pumping Lemma 2, there exists a word $v' := \mathtt{a} \cdot \mathtt{c}^n\mathtt{b} \cdot \mathtt{c}^{n'}\mathtt{b} \cdot \mathtt{b}$, $n < n'$, with $v' \in L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$, which is not the case, since for every factor $\mathtt{a} \cdot u \cdot \mathtt{b}$ in $v'$, $u$ is not a square. □

Next, we insert another occurrence of a terminal symbol in between the two occurrences of $x_2$, i. e., we consider $\beta := x_1 \cdot d \cdot x_2 \cdot d' \cdot x_2 \cdot d'' \cdot x_3$, where $d, d', d'' \in \{\mathtt{a}, \mathtt{b}\}$. Here, we find that $L_{\mathrm{Z},\Sigma}(\beta) \in \mathrm{REG}$ if and only if $\mathrm{Z} = \mathrm{E}$, $\Sigma = \Sigma_2$ and $d = d''$, $d \neq d' \neq d''$.

**Proposition 10.** *For every $Z \in \{\mathrm{E}, \mathrm{NE}\}$,*

$$\begin{aligned}
L_{Z,\Sigma_2}(x_1 \;\mathtt{a}\; x_2 \;\mathtt{a}\; x_2 \;\mathtt{a}\; x_3) &\notin \mathrm{REG}\,,\\
L_{Z,\Sigma_2}(x_1 \;\mathtt{a}\; x_2 \;\mathtt{a}\; x_2 \;\mathtt{b}\; x_3) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_2}(x_1 \;\mathtt{a}\; x_2 \;\mathtt{b}\; x_2 \;\mathtt{a}\; x_3) &\in \mathrm{REG}\,,\\
L_{\mathrm{NE},\Sigma_2}(x_1 \;\mathtt{a}\; x_2 \;\mathtt{b}\; x_2 \;\mathtt{a}\; x_3) &\notin \mathrm{REG}\,,\\
L_{Z,\Sigma_{\geq 3}}(x_1 \;\mathtt{a}\; x_2 \;\mathtt{b}\; x_2 \;\mathtt{a}\; x_3) &\notin \mathrm{REG}\;.
\end{aligned}$$

Proof. Let $\alpha_1 := x_1\mathtt{a}x_2\mathtt{a}x_2\mathtt{a}x_3$, $\alpha_2 := x_1\mathtt{a}x_2\mathtt{a}x_2\mathtt{b}x_3$ and $\alpha_3 := x_1\mathtt{a}x_2\mathtt{b}x_2\mathtt{a}x_3$. It follows from Proposition 13 that $L_{Z,\Sigma_2}(\alpha_1) \notin \mathrm{REG}$, $L_{Z,\Sigma_2}(\alpha_2) \notin \mathrm{REG}$ and $L_{Z,\Sigma_{\geq 3}}(\alpha_3) \notin \mathrm{REG}$. It remains to prove that $L_{\mathrm{E},\Sigma_2}(\alpha_3) \in \mathrm{REG}$ and $L_{\mathrm{NE},\Sigma_2}(\alpha_3) \notin \mathrm{REG}$. We shall first prove $L_{\mathrm{E},\Sigma_2}(\alpha_3) \in \mathrm{REG}$. To this end, we claim that $L_{\mathrm{E},\Sigma_2}(\alpha_3) = L(r)$, where $r := \Sigma_2^* \cdot \mathtt{a} \cdot (\mathtt{bb})^*\mathtt{b} \cdot \mathtt{a} \cdot \Sigma_2^*$. It can be easily verified that $L(r) \subseteq L_{\mathrm{E},\Sigma_2}(\alpha_3)$. In order to prove the converse, we let $h$ be an arbitrary substitution for $\alpha_3$. If $h(x_2) \in L(\mathtt{b}^*)$, then $h(\alpha_3) \in L(r)$. Thus, we assume that $h(x_2) = \mathtt{b}^n \cdot \widehat{u} \cdot \mathtt{b}^{n'}$, where $n, n' \in \mathbb{N}_0$, $\widehat{u} \in \Sigma_2^*$ and $\widehat{u}$ starts and ends with an occurrence of $\mathtt{a}$ (note that this includes the case $\widehat{u} = \mathtt{a}$). We note that $h(\alpha_3) = u \cdot \mathtt{a} \cdot \mathtt{b}^n \cdot \widehat{u} \cdot \mathtt{b}^{n+n'+1} \cdot \widehat{u} \cdot \mathtt{b}^{n'} \cdot \mathtt{a} \cdot v$, where $u := h(x_1)$ and $v := h(x_3)$. In order to prove that $h(\alpha_3) \in L(r)$ it is sufficient to identify a factor of the form $\mathtt{a}\mathtt{b}^k\mathtt{a}$ in $h(\alpha_3)$, where $k$ is odd. If $n$ is odd, then $\mathtt{a} \cdot \mathtt{b}^n \cdot \widehat{u}[1]$ is such a factor and if $n'$ is odd, then $\widehat{u}[-] \cdot \mathtt{b}^{n'} \cdot \mathtt{a}$ is such a factor. If both $n$ and $n'$ are even, then $\widehat{u}[-] \cdot \mathtt{b}^{n+n'+1} \cdot \widehat{u}[1]$ is a factor of the form $\mathtt{a}\mathtt{b}^k\mathtt{a}$, $k$ odd, since $n + n' + 1$ is odd. Hence, $h(\alpha_3) \in L(r)$ and $L_{\mathrm{E},\Sigma_2}(\alpha_3) \subseteq L(r)$ is implied, which concludes the proof.

Next, in order to prove $L_{\mathrm{NE},\Sigma_2}(\alpha_3) \notin \mathrm{REG}$, we assume to the contrary that $L_{\mathrm{NE},\Sigma_2}(\alpha_3) \in \mathrm{REG}$ and we define $w := \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{a}\mathtt{b}^n\mathtt{a} \cdot \mathtt{b} \cdot \mathtt{a}\mathtt{b}^n\mathtt{a} \cdot \mathtt{a} \cdot \mathtt{b} \in L_{\mathrm{NE},\Sigma_2}(\alpha_3)$, where $n$ is greater than the constant of Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma_2}(\alpha_3)$ and $n$ is even. By applying Pumping Lemma 2, we can obtain the word $w' := \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{ab}^n\mathtt{a} \cdot \mathtt{b} \cdot \mathtt{ab}^{n'}\mathtt{a} \cdot \mathtt{a} \cdot \mathtt{b}$, where $n < n'$ and $n'$ is even. It can be verified that for every factor of the form $\mathtt{a} \cdot u \cdot \mathtt{b} \cdot v \cdot \mathtt{a}$, $u, v \in \Sigma_2^+$, in $\mathtt{a} \cdot \mathtt{ab}^n\mathtt{a} \cdot \mathtt{b} \cdot \mathtt{ab}^{n'}\mathtt{a} \cdot \mathtt{a}$, $u \neq v$, which implies that $w' \notin L_{\mathrm{NE},\Sigma_2}(\alpha_3)$. Consequently, with Pumping Lemma 2, we can conclude that $L_{\mathrm{NE},\Sigma_2}(\alpha_3) \notin \mathrm{REG}$. □

The next type of pattern that we investigate is similar to the first one, but it contains two factors of the form $xx$ instead of only one, i. e., $\beta' := x_1 \cdot d \cdot x_2 x_2 \cdot d' \cdot x_3 x_3 \cdot d'' \cdot x_4$, where $d, d', d'' \in \{\mathtt{a}, \mathtt{b}\}$. Surprisingly, $L_{\mathrm{E},\Sigma_2}(\beta')$ is not regular if $d = d' = d''$, but regular in all other cases. However, if we consider the NE case or alphabet $\Sigma_{\geq 3}$, then $\beta'$ describes a non-regular language with respect to all choices of $d, d', d'' \in \{\mathtt{a}, \mathtt{b}\}$.

**Proposition 11.** *For every $Z \in \{\mathrm{E}, \mathrm{NE}\}$,*

$$\begin{aligned}
L_{Z,\Sigma_2}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{a}\ x_3\ x_3\ \mathtt{a}\ x_4) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_2}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{b}\ x_3\ x_3\ \mathtt{a}\ x_4) &\in \mathrm{REG}\,,\\
L_{\mathrm{NE},\Sigma_2}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{b}\ x_3\ x_3\ \mathtt{a}\ x_4) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_{\geq 3}}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{b}\ x_3\ x_3\ \mathtt{a}\ x_4) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_2}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{a}\ x_3\ x_3\ \mathtt{b}\ x_4) &\in \mathrm{REG}\,,\\
L_{\mathrm{NE},\Sigma_2}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{a}\ x_3\ x_3\ \mathtt{b}\ x_4) &\notin \mathrm{REG}\,,\\
L_{\mathrm{E},\Sigma_{\geq 3}}(x_1\ \mathtt{a}\ x_2\ x_2\ \mathtt{a}\ x_3\ x_3\ \mathtt{b}\ x_4) &\notin \mathrm{REG}\,.
\end{aligned}$$

Proof. We define $\alpha_1 := x_1\mathtt{a}x_2x_2\mathtt{a}x_3x_3\mathtt{a}x_4$, $\alpha_2 := x_1\mathtt{a}x_2x_2\mathtt{b}x_3x_3\mathtt{a}x_4$ and $\alpha_3 := x_1\mathtt{a}x_2x_2\mathtt{a}x_3x_3\mathtt{b}x_4$. We shall now prove the lemma by proving each of the 7 statements as individual claims.

*Claim.* $L_{\mathrm{Z},\Sigma_2}(\alpha_1) \notin \mathrm{REG}$, $Z \in \{\mathrm{E}, \mathrm{NE}\}$.

*Proof (Claim).* We first prove that $L_{\mathrm{NE},\Sigma_2}(\alpha_1) \notin \mathrm{REG}$. To this end, we assume to the contrary that $L_{\mathrm{NE},\Sigma_2}(\alpha_1)$ is a regular language and let $k \in \mathbb{N}$ be the constant from Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma_2}(\alpha_1)$. Furthermore, let $h$ be the substitution defined by $h(x_1) = h(x_4) = \mathtt{b}$, $h(x_2) := \mathtt{b}^n\mathtt{ab}$ and $h(x_3) := \mathtt{b}^m\mathtt{ab}$, where, $k < n$, $6n < m < 12n$ and both $n$ and $m$ are odd. We note that $h(\alpha_1) = \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{b}^n\mathtt{ab} \cdot \mathtt{b}^n\mathtt{ab} \cdot \mathtt{a} \cdot \mathtt{b}^m\mathtt{ab} \cdot \mathtt{b}^m\mathtt{ab} \cdot \mathtt{a} \cdot \mathtt{b}$. By applying Pumping Lemma 2 first to the second occurrence of factor $\mathtt{b}^n$ and then to the second occurrence of factor $\mathtt{b}^m$, we can obtain the word

$$w := \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{b}^n\mathtt{ab} \cdot \mathtt{b}^{n'}\mathtt{ab} \cdot \mathtt{a} \cdot \mathtt{b}^m\mathtt{ab} \cdot \mathtt{b}^{m'}\mathtt{ab} \cdot \mathtt{a} \cdot \mathtt{b}\,,$$

such that $2n < n' < 4n$ and $12n < m'$. Since we assume that $L_{\mathrm{NE},\Sigma_2}(\alpha_1) \in \mathrm{REG}$, we can conclude from Pumping Lemma 2 that $w \in L_{\mathrm{NE},\Sigma_2}(\alpha_1)$. Let $p_1, p_2, \ldots, p_7$ be exactly the positions in $w$ where there is an occurrence of $\mathtt{a}$. We shall now show that, for all $r, s, t$, $1 \leq r < s < t \leq 7$, the factor $w[p_r + 1, p_s - 1]$ is not a non-empty square or the factor $w[p_s + 1, p_t - 1]$ is not a non-empty square. This directly implies that there does not exist a substitution $g$ with $g(\alpha_1) = w$ and, thus, $w \notin L_{\mathrm{NE},\Sigma_2}(\alpha_1)$, which is a contradiction.

We can note that, for all $r, s$, with $1 \leq r < s \leq 7$, if $s - r$ is even, then $w[p_r + 1, p_s - 1]$ has an odd number of $\mathtt{a}$'s and, thus, it is not a square. Furthermore, since $n$ and $m$ are odd numbers, $w[p_1 + 1, p_2 - 1]$ and $w[p_4 + 1, p_5 - 1]$ cannot be squares and since $w[p_3 + 1, p_4 - 1] = w[p_6 + 1, p_7 - 1] = \mathtt{b}$, these cannot be squares either. The factor $w[p_1 + 1, p_4 - 1] = \mathtt{b}^n\mathtt{ab} \cdot \mathtt{b}^{n'}\mathtt{ab}$ is not a square since $n \neq n'$ and, since $m \neq m'$, the same holds for $w[p_4 + 1, p_7 - 1]$. The factor $w[p_1 + 1, p_6 - 1] = \mathtt{b}^n\mathtt{ab} \cdot \mathtt{b}^{n'}\mathtt{ab} \cdot \mathtt{a} \cdot \mathtt{b}^m\mathtt{ab} \cdot \mathtt{b}^{m'}$ cannot be a square, since $2n < n' < 4n$, $6n < m < 12n$ and $12n < m'$ implies that $n + n' + 2 < m + m' + 1$, and, with similar argumentations, we can conclude that factors $w[p_2 + 1, p_7 - 1]$, $w[p_2 + 1, p_5 - 1]$ and $w[p_3 + 1, p_6 - 1]$ are no squares either. We conclude that the only factors that can possibly be squares are $w[p_2 + 1, p_3 - 1]$ and $w[p_5 + 1, p_6 - 1]$. However, for all $r, s, t$, $1 \leq r < s < t \leq 7$, it is impossible that $(r, s) = (2, 3)$ and $(s, t) = (5, 6)$. Hence, we obtain a contradiction as described above and, thus, we can conclude that $L_{\mathrm{NE},\Sigma_2}(\alpha_1) \notin \mathrm{REG}$. Moreover, in exactly the same way, we can also prove that $L_{\mathrm{E},\Sigma_2}(\alpha_1) \notin \mathrm{REG}$. This is due to the fact that in the word $w$ there are no two occurrences of symbol $\mathtt{a}$ without occurrences of symbol $\mathtt{b}$ between them, i. e., we do not need to consider empty squares. So by exactly the same argumentation, we can show that $w$ is not in $L_{\mathrm{E},\Sigma_2}(\alpha_1) \notin \mathrm{REG}$, which, since $h(\alpha_1)$ clearly is in $L_{\mathrm{E},\Sigma_2}(\alpha_1)$, leads to a contradiction in the same way.

□ *(Claim)*

*Claim.* $L_{\mathrm{E},\Sigma_2}(\alpha_2) \in \mathrm{REG}$.

*Proof (Claim).* We claim that $L_{\mathrm{E},\Sigma_2}(\alpha_2) = L(r)$, where $r := \Sigma_2^* \cdot \mathtt{a} \cdot (\mathtt{bb})^* \cdot \mathtt{b} \cdot \mathtt{a} \cdot \Sigma_2^*$. First, we can note that $L(r) \subseteq L_{\mathrm{E},\Sigma_2}(\alpha_2)$ trivially holds. Now let $h$ be an arbitrary substitution. In order to prove that $h(\alpha_2) \in L(r)$, it is sufficient to show that in $h(\alpha_2)$ there occurs a factor of the form $\mathtt{a} \cdot \mathtt{b}^{2n-1} \cdot \mathtt{a}$, $n \in \mathbb{N}$.

We first consider the case that $h(x_2) = b^n \cdot u \cdot b^{n'}$, $n, n' \in \mathbb{N}_0$, where $u$ starts and ends with the symbol $\mathtt{a}$. We note that if $n$ is odd, then in $h(\alpha_2)$ there occurs the factor $\mathtt{a} \cdot \mathtt{b}^n \cdot \mathtt{a}$. If, on the other hand, $n$ is even and $n'$ is odd, then $n + n'$ is odd and in $h(\alpha_2)$ there occurs the factor $\mathtt{a} \cdot \mathtt{b}^{n+n'} \cdot \mathtt{a}$. Furthermore, if $n'$ and $n$ are even, then we cannot directly conclude that there exists a factor $\mathtt{a} \cdot \mathtt{b}^{2n-1} \cdot \mathtt{a}$, $n \in \mathbb{N}$, and we have to take a closer look at $h(x_3)$. If $h(x_3) \in L(\mathtt{b}^*)$, then we have the factor $\mathtt{a} \cdot \mathtt{b}^{n'} \cdot \mathtt{b} \cdot h(x_3) \cdot h(x_3) \cdot \mathtt{a}$ that necessarily is of form an $\mathtt{a} \cdot \mathtt{b}^{2n-1} \cdot \mathtt{a}$, $n \in \mathbb{N}$. If, on the other hand, $h(x_3) = \mathtt{b}^m \cdot v \cdot \mathtt{b}^{m'}$, $m, m' \in \mathbb{N}_0$, where $v$ starts and ends with an $\mathtt{a}$, then we have to consider several cases depending on whether $m$ and $m'$ is odd or even. If $m$ is even, then the factor $\mathtt{a} \cdot \mathtt{b}^{n'} \cdot \mathtt{b} \cdot \mathtt{b}^m \cdot \mathtt{a}$ occurs in $h(\alpha_2)$, where $n' + m + 1$ is odd. If, on the other hand, $m$ is odd and $m'$ is even, then the factor $\mathtt{a} \cdot \mathtt{b}^{m'+m} \cdot \mathtt{a}$ occurs in $h(\alpha_2)$, where $m' + m$ is odd. Finally, if $m'$ and $m$ are odd, then the factor $\mathtt{a} \cdot \mathtt{b}^{m'} \cdot \mathtt{a}$ occurs in $\alpha_2$. So we can conclude that if $h(x_2) = b^n \cdot u \cdot b^{n'}$, then there necessarily occurs a factor of the form $\mathtt{a} \cdot \mathtt{b}^{2n-1} \cdot \mathtt{a}$, $n \in \mathbb{N}$ in $h(\alpha_2)$.

It remains to consider the case where $h(x_2) \in L(\mathtt{b}^*)$. We first note that if also $h(x_3) \in L(\mathtt{b}^*)$, then the factor $\mathtt{a} \cdot h(x_2) \cdot h(x_2) \cdot \mathtt{b} \cdot h(x_3) \cdot h(x_3) \cdot \mathtt{a}$ occurs in $h(\alpha_2)$, which is of the form $\mathtt{a}\cdot\mathtt{b}^{2n-1}\cdot\mathtt{a}$, $n\in\mathbb{N}$. So we need to consider the case that $h(x_3)=\mathtt{b}^{m}\cdot v\cdot\mathtt{b}^{m'}$, $m,m'\in\mathbb{N}_0$, where $v$ starts and ends with $\mathtt{a}$. If $m$ is even, then the factor $\mathtt{a}\cdot h(x_2)\cdot h(x_2)\cdot\mathtt{b}\cdot\mathtt{b}^{m}\cdot\mathtt{a}$ occurs in $h(\alpha_2)$, that is of the form $\mathtt{a}\cdot\mathtt{b}^{2n-1}\cdot\mathtt{a}$, $n\in\mathbb{N}$. If $m'$ is odd, then the factor $\mathtt{a}\cdot\mathtt{b}^{m'}\cdot\mathtt{a}$ occurs and, finally, if $m$ is odd and $m'$ is even, then factor $\mathtt{a}\cdot\mathtt{b}^{m}\cdot\mathtt{b}^{m'}\cdot\mathtt{a}$ occurs in $h(\alpha_2)$. Consequently, $h(\alpha_2)$ necessarily contains a factor of the form $\mathtt{a}\cdot\mathtt{b}^{2n-1}\cdot\mathtt{a}$, $n\in\mathbb{N}$. Thus, $h(\alpha_2)\in L(r)$, which shows that $L(r)\subseteq L_{\mathrm{E},\Sigma_2}(\alpha_2)$ holds. □ *(Claim)*

*Claim.* $L_{\mathrm{NE},\Sigma_2}(\alpha_2)\notin\mathrm{REG}$.

*Proof (Claim).* We assume that $L_{\mathrm{NE},\Sigma_2}(\alpha_2)$ is a regular language and we define $w:=\mathtt{b}\cdot\mathtt{a}\cdot\mathtt{ab}^{n}\mathtt{a}\cdot\mathtt{ab}^{n}\mathtt{a}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{a}\cdot\mathtt{b}\in L_{\mathrm{NE},\Sigma_2}(\alpha_2)$, where $n$ is greater than the constant of Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma_2}(\alpha_2)$ and $n$ is even. By pumping, we can produce a word $w':=\mathtt{b}\cdot\mathtt{a}\cdot\mathtt{ab}^{n}\mathtt{a}\cdot\mathtt{ab}^{n'}\mathtt{a}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{a}\cdot\mathtt{b}$, where $n<n'$ and $n'$ is even. Now we can note that in $w'$, for every factor of the form $\mathtt{a}\cdot u\cdot\mathtt{b}\cdot v\cdot\mathtt{a}$, $u,v\in\Sigma'^{+}$, in $\mathtt{a}\cdot\mathtt{ab}^{n}\mathtt{a}\cdot\mathtt{ab}^{n'}\mathtt{a}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{a}$, $u$ is not a square or $v$ is not a square. This implies that $w'\notin L_{\mathrm{NE},\Sigma_2}(\alpha_2)$, which is a contradiction to Pumping Lemma 2. □ *(Claim)*

*Claim.* $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)\notin\mathrm{REG}$.

*Proof (Claim).* We assume that $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)\in\mathrm{REG}$ and we define $w:=\mathtt{a}\cdot\mathtt{c}^{n}\mathtt{b}\cdot\mathtt{c}^{n}\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{a}\in L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$, where $n$ is greater than the constant of Pumping Lemma 2 with respect to $L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$ and $n$ is odd. By pumping, we can produce a word $w':=\mathtt{a}\cdot\mathtt{c}^{n}\mathtt{b}\cdot\mathtt{c}^{n'}\mathtt{b}\cdot\mathtt{b}\cdot\mathtt{a}$, where $n<n'$. Since in $w'$ there is no factor of the form $\mathtt{a}\cdot vv\cdot\mathtt{b}$, $v\in\Sigma_{\geq 3}^{*}$, we can conclude that $w'\notin L_{\mathrm{E},\Sigma_{\geq 3}}(\alpha_2)$, which contradicts Pumping Lemma 2. □ *(Claim)*

*Claim.* $L_{\mathrm{E},\Sigma_2}(\alpha_3)\in\mathrm{REG}$.

*Proof (Claim).* We claim that $L_{\mathrm{E},\Sigma_2}(\alpha_3)=L(r)$, where $r:=\Sigma_2^{*}\cdot\mathtt{a}\cdot(\mathtt{bb})^{*}\cdot\mathtt{a}\cdot\mathtt{b}\cdot\Sigma_2^{*}$. First, we can note that $L(r)\subseteq L_{\mathrm{E},\Sigma_2}(\alpha_3)$ trivially holds. Let $h$ be an arbitrary substitution. We shall show that $h(\alpha_3)\in L(r)$, which implies that $L_{\mathrm{E},\Sigma_2}(\alpha_3)\subseteq L(r)$. If $h(x_2)$ starts with the symbol $\mathtt{a}$, $h(x_2)$ ends with the symbol $\mathtt{a}$ or $h(x_3)$ starts with the symbol $\mathtt{a}$, then the factor $\mathtt{a}\cdot\mathtt{a}\cdot\mathtt{b}$ occurs in $h(\alpha_3)$, which implies that $h(\alpha_3)\in L(r)$. Hence, we only need to consider the following case: if $h(x_2)$ is non-empty, then it starts and ends with the symbol $\mathtt{b}$ and if $h(x_3)$ is non-empty, then it starts with the symbol $\mathtt{b}$. Next, we can note that if $h(x_2)$ is empty or $h(x_2)=\mathtt{b}^{n}$, $n\in\mathbb{N}$, then, since $h(x_3)$ is either empty or it starts with $\mathtt{b}$, the factor $\mathtt{a}\cdot\mathtt{a}\cdot\mathtt{b}$ occurs in $h(\alpha_3)$ or the factor $\mathtt{a}\cdot\mathtt{b}^{2n}\cdot\mathtt{a}\cdot\mathtt{b}$ occurs in $h(\alpha_3)$, respectively, which implies that $h(\alpha_3)\in L(r)$. Therefore, we need to take a closer look at the case that $h(x_2)=\mathtt{b}^{n}\cdot u\cdot\mathtt{b}^{n'}$, $n,n'\in\mathbb{N}$, where $u$ starts and ends with the symbol $\mathtt{a}$. If $u$ contains the factor $\mathtt{a}\cdot\mathtt{a}$, then the factor $\mathtt{a}\cdot\mathtt{a}\cdot\mathtt{b}$ is contained in $h(\alpha_3)$, thus, $h(\alpha_3)\in L(r)$. If, on the other hand, $u$ does not contain the factor $\mathtt{a}\cdot\mathtt{a}$, i. e., every $\mathtt{a}$ in $u$ is followed by a $\mathtt{b}$, then we need to use a different argumentation. We note that in $h(\alpha_3)$ the factors $\mathtt{a}\cdot\mathtt{b}^{n}\cdot\mathtt{a}\cdot\mathtt{b}$ and $\mathtt{a}\cdot\mathtt{b}^{n'}\cdot\mathtt{b}^{n}\cdot\mathtt{a}\cdot\mathtt{b}$ occur. Furthermore, since $h(x_3)$ is either empty or it starts with

$\mathtt{b}$, we can also conclude that the factor $\mathtt{a} \cdot \mathtt{b}^{n'} \cdot \mathtt{a} \cdot \mathtt{b}$ occurs in $h(\alpha_3)$. We can now observe that if $n$ is even or $n'$ is even, then $h(\alpha_3) \in L(r)$. Furthermore, if $n$ is odd and $n'$ is odd, then $n + n'$ is even and, thus, $h(\alpha_3) \in L(r)$. Consequently, for all possible cases, $h(\alpha_3) \in L(r)$, which implies that $L_{\text{E},\Sigma_2}(\alpha_3) \subseteq L(r)$. □*(Claim)*

*Claim.* $L_{\text{NE},\Sigma_2}(\alpha_3) \notin \text{REG}$.

*Proof (Claim).* We assume that $L_{\text{NE},\Sigma_2}(\alpha_3)$ is a regular language and we define $w := \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{b}^n \mathtt{a} \cdot \mathtt{b}^n \mathtt{a} \cdot \mathtt{a} \cdot \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{a} \in L_{\text{NE},\Sigma_2}(\alpha_3)$, where $n$ is greater than the constant of Pumping Lemma 2 with respect to $L_{\text{NE},\Sigma_2}(\alpha_3)$ and $n$ is odd. By pumping, we can produce a word $w' := \mathtt{b} \cdot \mathtt{a} \cdot \mathtt{b}^n \mathtt{a} \cdot \mathtt{b}^{n'} \mathtt{a} \cdot \mathtt{a} \cdot \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{b} \cdot \mathtt{a}$, where $n < n'$ and $n'$ is odd. Now we can note that in $w'$ there is no factor of the form $\mathtt{a} \cdot vv \cdot \mathtt{a}$, $v \in \Sigma_2^+$. Thus, $w' \notin L_{\text{NE},\Sigma_2}(\alpha_3)$, which contradicts Pumping Lemma 2. □*(Claim)*

*Claim.* $L_{\text{E},\Sigma_{\geq 3}}(\alpha_3) \notin \text{REG}$.

*Proof (Claim).* This claim can be proved analogously to the claim $L_{\text{E},\Sigma_{\geq 3}}(\alpha_2) \notin \text{REG}$. □*(Claim)*

This concludes the proof of the proposition. □

We call two patterns $\alpha, \beta \in (\Sigma_2 \cup X)^*$ *almost identical* if and only if $|\alpha| = |\beta|$ and, for every $i$, $1 \leq i \leq |\alpha|$, $\alpha[i] \neq \beta[i]$ implies $\alpha[i], \beta[i] \in \Sigma_2$. The above examples show that even for almost identical patterns $\alpha$ and $\beta$, we can have the situation that $\alpha$ describes a regular and $\beta$ a non-regular language. Even if $\alpha$ and $\beta$ are almost identical and further satisfy $|\alpha|_{\mathtt{a}} = |\beta|_{\mathtt{a}}$ and $|\alpha|_{\mathtt{b}} = |\beta|_{\mathtt{b}}$, then it is still possible that $\alpha$ describes a regular and $\beta$ a non-regular language (cf. Proposition 10 above). This implies that the regular E-pattern languages over an alphabet with size 2 require a characterisation that caters for the exact order of terminal symbols in the patterns.

The examples considered in Propositions 9 and 11 mainly consist of factors of the form $d \cdot xx \cdot d'$, $d, d' \in \Sigma_2$, where $x$ does not have any other occurrence in the pattern. Hence, it might be worthwhile to investigate the question of whether or not patterns can also describe regular languages if we allow them to contain factors of the form $d \cdot x^k \cdot d'$, where $k \geq 3$ and there is no other occurrence of $x$ in the pattern. In the next result, we state that if a pattern $\alpha$ contains a factor $d \cdot x^k \cdot d'$ with $d = d'$, $k \geq 3$ and $|\alpha|_x = k$, then, for every $\text{Z} \in \{\text{E}, \text{NE}\}$, its Z-pattern language with respect to any alphabet of size at least 2 is not regular and, furthermore, for alphabets of size at least 3, we can show that this also holds for $d \neq d'$.

**Theorem 12.** *Let $\Sigma$ and $\Sigma'$ be terminal alphabets with $\{\mathtt{a}, \mathtt{b}\} \subseteq \Sigma$ and $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \subseteq \Sigma'$. Let $\alpha := \alpha_1 \cdot \mathtt{a} \cdot z^l \cdot \mathtt{a} \cdot \alpha_2$, let $\beta := \beta_1 \cdot \mathtt{a} \cdot z^l \cdot \mathtt{c} \cdot \beta_2$, where $z \in X$, $\alpha_1, \alpha_2 \in ((\Sigma \cup X) \setminus \{z\})^*$, $\beta_1, \beta_2 \in ((\Sigma' \cup X) \setminus \{z\})^*$ and $l \geq 3$. Then, for every $\text{Z} \in \{\text{E}, \text{NE}\}$, $L_{\text{Z},\Sigma}(\alpha) \notin \text{REG}$ and $L_{\text{Z},\Sigma'}(\beta) \notin \text{REG}$.*

Proof. We first prove that $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$. Let $k$ be the constant of Pumping Lemma 1 with respect to $L_{\mathrm{NE},\Sigma}(\alpha)$ and let $h$ be the substitution defined by $h(z) := \mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b}$, where $k' \geq k$, $k' \mod l = 1$, and $h(x) := \mathtt{b}$, $x \in \mathrm{var}(\alpha) \setminus \{z\}$. We can note that $w := h(\alpha) = u \cdot \mathtt{a} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^l \cdot \mathtt{a} \cdot v$, where $u$ and $v$ equal $h(\alpha_1)$ and $h(\alpha_2)$, respectively. Obviously, $|w| \geq k$ and $w \in L_{\mathrm{NE},\Sigma}(\alpha)$. We shall now show that for every factorisation $w = v_1 \cdot v_2 \cdot v_3$ with $|v_1 v_2| \leq k$ and $v_2 \neq \varepsilon$, there exists a $t \in \mathbb{N}_0$ such that $v_1 \cdot v_2^t \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$, which, by Pumping Lemma 1, proves that $L_{\mathrm{NE},\Sigma}(\alpha)$ is not regular. We first note that $|v_1 v_2| \leq k$ and $v_2 \neq \varepsilon$ implies that

- $v_2 = u'$, where $u'$ is a factor of $u$ with $1 \leq |u'| \leq k$ or
- $v_2 = u' \cdot \mathtt{a} \cdot \mathtt{b}^i$, where $u'$ is a suffix of $u$ and $0 \leq i \leq k - (|u'| + 1)$ or
- $v_2 = \mathtt{b}^i$, where $1 \leq i \leq k - (|u| + 1)$.

We first consider the case that $v_2 = \mathtt{b}^i$, $1 \leq i \leq k - (|u| + 1)$, and, furthermore, we assume that $i$ is a multiple of $l$, which implies that $k' - i$ is not a multiple of $l$, since $k'$ is not a multiple of $l$. Next, we consider the word $v_1 \cdot v_2^0 \cdot v_3 = u \cdot \mathtt{a} \cdot \mathtt{b}^{k'-i} \cdot \mathtt{a} \cdot \mathtt{b} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^{l-1} \cdot \mathtt{a} \cdot v$. We want to show that $v_1 \cdot v_2^0 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$. To this end, we first note that if there exists a substitution $g$ with $g(\alpha) = v_1 \cdot v_2^0 \cdot v_3$, then, since $u$ and $v$ are obtained by substituting all variables of $\alpha_1$ and $\alpha_2$ by a word of length 1, $u$ must be a prefix of $g(\alpha_1)$ and $v$ must be a suffix of $g(\alpha_2)$. This implies that, in order to conclude $v_1 \cdot v_2^0 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$, it is sufficient to show that every factor of the form $\mathtt{a} \cdot w \cdot \mathtt{a}$, $w \in \Sigma^*$, of $\mathtt{a} \cdot \mathtt{b}^{k'-i} \cdot \mathtt{a} \cdot \mathtt{b} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^{l-1} \cdot \mathtt{a}$ is not of the form $\mathtt{a} \cdot (w')^l \cdot \mathtt{a}$, $w' \in \Sigma^*$. We first note that the factor $\mathtt{a} \cdot \mathtt{b}^{k'-i} \cdot \mathtt{a} \cdot \mathtt{b} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^{l-1} \cdot \mathtt{a}$ is obviously not of this form. For all other factors $\mathtt{a} \cdot w \cdot \mathtt{a}$ of $\mathtt{a} \cdot \mathtt{b}^{k'-i} \cdot \mathtt{a} \cdot \mathtt{b} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^{l-1} \cdot \mathtt{a}$, where $|w|_{\mathtt{a}} \geq 1$, we have $|w|_{\mathtt{a}} \leq l - 1$, thus, they cannot be of the form $\mathtt{a} \cdot (w')^l \cdot \mathtt{a}$, $w' \in \Sigma^*$, either. Consequently, it remains to take a closer look at the factors $\mathtt{a} \cdot w \cdot \mathtt{a}$, where $|w|_{\mathtt{a}} = 0$. We can observe that for these factors the length of $w$ is either $k' + 1$, $k' - i$ or 1, and, since $l \geq 3$, neither $k' + 1$, $k' - i$ nor 1 is a multiple of $l$. This implies that these factors are also not of the form $\mathtt{a} \cdot (w')^l \cdot \mathtt{a}$, $w' \in \Sigma^*$, which proves that $v_1 \cdot v_2^0 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$.

Next, we consider the case that $v_2 = \mathtt{b}^i$, where $i$ is not a multiple of $l$. Now if $k' - i$ is not a multiple of $l$, then we can show in exactly the same way as before that $v_1 \cdot v_2^0 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$. If, on the other hand, $k' - i$ is a multiple of $l$, then, since $k' \mod l = 1$, we can conclude that $i \mod l = 1$ and, thus, $k' + i \mod l = 2$. We now consider the word $v_1 \cdot v_2^2 \cdot v_3 = u \cdot \mathtt{a} \cdot \mathtt{b}^{k'+i} \cdot \mathtt{a} \cdot \mathtt{b} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^{l-1} \cdot \mathtt{a} \cdot v$. As demonstrated above, $k' + i$ is not a multiple of $l$ and, thus, we can apply the same argumentation as before in order to show that $v_1 \cdot v_2^2 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$.

In order to conclude the proof, we have to consider the case that $v_2 = u'$, where $u'$ is a factor of $u$ with $1 \leq |u'| \leq k$ and the case that $v_2 = u' \cdot \mathtt{a} \cdot \mathtt{b}^i$, where $u'$ is a suffix of $u$ and $0 \leq i \leq k - (|u'| + 1)$. We first assume that $v_2 = u'$ with $u = q_1 \cdot u' \cdot q_2$, $1 \leq |u'| \leq k$, and consider the word $v_1 \cdot v_2^0 \cdot v_3 := q_1 \cdot q_2 \cdot \mathtt{a} \cdot (\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^l \cdot \mathtt{a} \cdot v$. If there exists a substitution $g$ with $g(\alpha) = v_1 \cdot v_2^0 \cdot v_3$, then, since $|q_1 \cdot q_2| < |u|$, we can conclude that $q_1 \cdot q_2 \cdot \mathtt{a}$ is a prefix of $g(\alpha_1)$, which implies that, in order to conclude $v_1 \cdot v_2^0 \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$, it is sufficient to show that every factor $\mathtt{a} \cdot w \cdot \mathtt{a}$, $w \in \Sigma^*$ of $(\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^l \cdot \mathtt{a}$ is not of the form $\mathtt{a} \cdot (w')^l \cdot \mathtt{a}$, $v \in \Sigma^*$. This can be easily seen, since $|(\mathtt{b}^{k'} \cdot \mathtt{a} \cdot \mathtt{b})^l \cdot \mathtt{a}|_{\mathtt{a}} \leq l + 1$ and, for every factor of the form $\mathtt{a} \cdot w \cdot \mathtt{a}$, where $|w|_{\mathtt{a}} = 0$, we can observe that $|w|$ equals either $k' + 1$ or 1, and, since $l \geq 3$, neither of these is a multiple of $l$. If $v_2 = u' \cdot \mathtt{a} \cdot \mathtt{b}^i$, where $u'$ is a suffix of $u$ and $0 \leq i \leq k - (|u'| + 1)$, then we can argue analogously. This proves that for every factorisation $w = v_1 \cdot v_2 \cdot v_3$ with $|v_1 v_2| \leq k$ and $v_2 \neq \varepsilon$, there exists a $t \in \mathbb{N}_0$ such that $v_1 \cdot v_2^t \cdot v_3 \notin L_{\mathrm{NE},\Sigma}(\alpha)$, which, by Pumping Lemma 1, implies that $L_{\mathrm{NE},\Sigma}(\alpha)$ is not regular.

It can be shown analogously that $L_{\mathrm{E},\Sigma}(\alpha) \notin \mathrm{REG}$. The only difference of the prove is that the substitution $h$ erases all variables of $\alpha_1$ and $\alpha_2$ instead of substituting them by $\mathtt{b}$. This is necessary to be able to assume that for any other substitution $g$, $h(\alpha_1)$ must be a prefix of $g(\alpha_1)$ and $h(\alpha_2)$ must be a suffix of $g(\alpha_2)$.

It remains to show that $L_{\mathrm{NE},\Sigma'}(\beta) \notin \mathrm{REG}$ and $L_{\mathrm{E},\Sigma'}(\beta) \notin \mathrm{REG}$. We shall first show that $L_{\mathrm{NE},\Sigma'}(\beta) \notin \mathrm{REG}$. Let $k$ be the constant of Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma'}(\beta)$ and let $h$ be the substitution defined by $h(z) := \mathtt{b}^k \cdot \mathtt{a}$ and $h(x) := \mathtt{b}$, $x \in \mathrm{var}(\beta) \setminus \{z\}$. We can note that $w := h(\beta) = u \cdot \mathtt{a} \cdot (\mathtt{b}^k \cdot \mathtt{a})^l \cdot \mathtt{c} \cdot v$, where $u$ and $v$ equal $h(\beta_1)$ and $h(\beta_2)$, respectively. Obviously, $|w| \geq k$ and $w \in L_{\mathrm{NE},\Sigma'}(\beta)$. By applying Pumping Lemma 2, we can obtain a word $w' := u \cdot \mathtt{a} \cdot \mathtt{b}^{k'} \cdot \mathtt{a} \cdot (\mathtt{b}^k \cdot \mathtt{a})^{l-1} \cdot \mathtt{c} \cdot v$ with $k < k'$. We shall now show that $w' \notin L_{\mathrm{NE},\Sigma'}(\beta)$. To this end, we first note that if there exists a substitution $g$ with $g(\beta) = w'$, then, since $u$ and $v$ are obtained by substituting all variables of $\beta_1$ and $\beta_2$ by a word of length 1, $u$ must be a prefix of $g(\beta_1)$ and $v$ must be a suffix of $g(\beta_2)$. This implies that, in order to conclude $w' \notin L_{\mathrm{NE},\Sigma'}(\beta)$, it is sufficient to show that every factor of the form $\mathtt{a} \cdot w \cdot \mathtt{c}$, $w \in \Sigma'^+$, in $\mathtt{a} \cdot \mathtt{b}^{k'} \cdot \mathtt{a} \cdot (\mathtt{b}^k \cdot \mathtt{a})^{l-1} \cdot \mathtt{c}$ is not of the form $\mathtt{a} \cdot (w')^l \cdot \mathtt{c}$, $w' \in \Sigma'^+$. It is easy to see that $\mathtt{a} \cdot \mathtt{b}^{k'} \cdot \mathtt{a} \cdot (\mathtt{b}^k \cdot \mathtt{a})^{l-1} \cdot \mathtt{c}$ is not of this form and for all other factors of the form $\mathtt{a} \cdot w \cdot \mathtt{c}$, $w \in \Sigma'^+$, we have $|w|_{\mathtt{a}} \leq l - 1$, which implies that $w$ cannot be of the form $(w')^l$, $w' \in \Sigma'^+$. This implies that $w' \notin L_{\mathrm{NE},\Sigma'}(\beta)$ and, thus, $L_{\mathrm{NE},\Sigma'}(\beta) \notin \mathrm{REG}$.

It can be shown analogously that $L_{\mathrm{E},\Sigma'}(\beta) \notin \mathrm{REG}$. The only difference of the prove is that the substitution $h$ erases all variables of $\beta_1$ and $\beta_2$ instead of substituting them by $\mathtt{b}$. □

In the examples of Propositions 9, 10 and 11 as well as in the above theorem, we did not consider the situation that two occurrences of the same variable are separated by a terminal symbol. In the next result, we state that, in certain cases, this situation implies non-regularity of pattern languages.

**Proposition 13.** *Let $\Sigma$ and $\Sigma'$ be terminal alphabets with $|\Sigma| \geq 2$ and $|\Sigma'| \geq 3$ and let $\mathrm{Z} \in \{\mathrm{E}, \mathrm{NE}\}$. Furthermore, let $\alpha_1 \in (\Sigma \cup X)^*$ and $\alpha_2 \in (\Sigma' \cup X)^*$ be patterns.*

1. *If there exists a $\gamma \in (\Sigma \cup X)^*$ with $|\mathrm{var}(\gamma)| \geq 1$ such that, for some $d \in \Sigma$,*
   - *$\alpha_1 = \gamma \cdot d \cdot \delta$ and $\mathrm{var}(\gamma) \subseteq \mathrm{var}(\delta)$,*
   - *$\alpha_1 = \gamma \cdot d \cdot \delta$ and $\mathrm{var}(\delta) \subseteq \mathrm{var}(\gamma)$ or*

- $\alpha_1 = \beta \cdot d \cdot \gamma \cdot d \cdot \delta$ *and* $\mathrm{var}(\gamma) \subseteq (\mathrm{var}(\beta) \cup \mathrm{var}(\delta))$,

*then* $L_{\mathrm{Z},\Sigma}(\alpha_1) \notin \mathrm{REG}$.

2. *If in* $\alpha_2$ *there exists a non-empty variable block, all the variables of which also occur outside this block, then* $L_{\mathrm{Z},\Sigma'}(\alpha_2) \notin \mathrm{REG}$.

Proof. We first prove point 1 of the proposition. To this end, we assume that $L_{\mathrm{NE},\Sigma}(\alpha)$ is a regular language. Furthermore, we assume that for $\alpha$ one of the three cases described in point 1 is satisfied with $d = \mathtt{b}$. Let $w$ be the word obtained from $\alpha$ by substituting all variables in $\mathrm{var}(\gamma)$ by $\mathtt{a}^n$, where $n$ is the constant of Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma}(\alpha)$, and all other variables by $\mathtt{a}$. By applying Pumping Lemma 2, we can obtain a word $w'$ from $w$ by pumping the part that results from $\gamma$ without pumping the other parts of the word. Since every variable of $\gamma$ occurs in the other parts as well, and since we only substituted the variables that do not occur in $\gamma$ by $\mathtt{a}$, we can conclude that $w'$ is not in $L_{\mathrm{NE},\Sigma}(\alpha)$, which proves that $L_{\mathrm{NE},\Sigma}(\alpha) \notin \mathrm{REG}$. Furthermore, the above proof can be applied in exactly the same way in order to show that $L_{\mathrm{E},\Sigma}(\alpha_1) \notin \mathrm{REG}$.

Point 2 of the proposition can be proved analogously. If in $\alpha_2$ there exists a variable block, all the variables of which also occur outside this block, then we can substitute all variables in this block by $\mathtt{a}^n$, where $n$ is the constant of Pumping Lemma 2 with respect to $L_{\mathrm{NE},\Sigma'}(\alpha)$ and, since $|\Sigma'| \geq 3$, we can assume that the variable block is not delimited by $\mathtt{a}$ to either side. Furthermore, we substitute all variables that do not occur in the variable block by $\mathtt{a}$. Now we can show in exactly the same way as before that the thus obtained word is not in $L_{\mathrm{NE},\Sigma'}(\alpha)$, which proves $L_{\mathrm{NE},\Sigma'}(\alpha) \notin \mathrm{REG}$ and $L_{\mathrm{E},\Sigma'}(\alpha) \notin \mathrm{REG}$ can be shown in exactly the same way. □

We conclude this section by referring to the examples presented in Propositions 9, 10 and 11, which, as described above, suggest that complete characterisations of the regular E-pattern languages over small alphabets might be extremely complex. In the next section, we wish to find out about the fundamental mechanisms of the above example patterns that are responsible for the regularity of their pattern languages. Intuitively speaking, some of the above example patterns describe regular languages, because they contain a factor that is less complex than it seems to be, e. g., for the pattern $\beta := x_1 \cdot \mathtt{a} \cdot x_2 x_2 \cdot \mathtt{a} \cdot x_3 x_3 \cdot \mathtt{b} \cdot x_4$ it can be shown that the factor $\mathtt{a} \cdot x_2 x_2 \cdot \mathtt{a} \cdot x_3 x_3 \cdot \mathtt{b}$ could be replaced by $\mathtt{a} \cdot x_{(\mathtt{bb})^*} \cdot \mathtt{a} \cdot \mathtt{b}$ (where $x_{(\mathtt{bb})^*}$ is a special variable that can only be substituted by a unary string over $\mathtt{b}$ of even length) without changing its E-pattern language with respect to $\Sigma_2$. This directly implies that $L_{\mathrm{E},\Sigma_2}(\beta) = L(\Sigma_2^* \cdot \mathtt{a}(\mathtt{bb})^*\mathtt{ab} \cdot \Sigma_2^*)$, which shows that $L_{\mathrm{E},\Sigma_2}(\beta) \in \mathrm{REG}$. In the next section, we generalise this observation.

## 4. Regularity of E-Pattern Languages: A Sufficient Condition Taking Terminal Symbols into Account

In this section we investigate the phenomenon that a whole factor in a pattern can be substituted by a less complex one, without changing the corresponding pattern language. This technique can be used in order to show that a complicated pattern is equivalent to one that can be easily seen to describe a regular language.

For the sake of a better presentation of our results, we slightly redefine the concept of patterns. A *pattern with regular expressions* is a pattern that may contain regular expressions. Such a regular expressions is then interpreted as a variable with only one occurrence that can only be substituted by words described by the corresponding regular expression. For example $L_{\mathrm{E},\Sigma_2}(x_1\mathtt{b}^*x_1\mathtt{a}^*) = \{h(x_1x_2x_1x_3) \mid h \text{ is a substitution with } h(x_2) \in L(\mathtt{b}^*), h(x_3) \in L(\mathtt{a}^*)\}$. Obviously, patterns with regular expressions exceed the expressive power of classical patterns. However, we shall use this concept exclusively in the case where a classical pattern is equivalent to a pattern with regular expressions. For example, the pattern $x_1 \cdot \mathtt{a} \cdot x_2x_3x_3x_2 \cdot \mathtt{a} \cdot x_4$ is equivalent to the pattern $x_1 \cdot \mathtt{a}(\mathtt{bb})^*\mathtt{a} \cdot x_2$ (see Lemma 16).

Next, we present a lemma that states that in special cases whole factors of a pattern can be removed without changing the corresponding pattern language.

**Lemma 14.** *Let* $\alpha := \beta\cdot y\cdot\beta'\cdot\mathtt{a}\cdot\gamma\cdot\mathtt{b}\cdot\delta'\cdot z\cdot\delta$*, where* $\beta, \delta \in (\Sigma_2\cup X)^*$*,* $\beta', \gamma, \delta' \in X^*$*,* $y, z \in X$ *and* $|\alpha|_y = |\alpha|_z = 1$*. Then* $L_{\mathrm{E},\Sigma_2}(\alpha) \subseteq L_{\mathrm{E},\Sigma_2}(\beta \cdot y \cdot \mathtt{ab} \cdot z \cdot \delta)$*. If, furthermore,* $\mathrm{var}(\beta'\cdot\gamma\cdot\delta')\cap\mathrm{var}(\beta\cdot\delta) = \emptyset$*, then also* $L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{ab}\cdot z\cdot\delta) \subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$*.*

Proof. Let $h$ be an arbitrary substitution. We obtain a substitution $g$ from $h$ in the following way. For every $x \in \mathrm{var}(\beta\cdot\delta)\setminus\{y, z\}$, we define $g(x) := h(x)$. If the last symbol in $h(\gamma)$ is $\mathtt{a}$, then we define $g(y) := h(y\cdot\beta')\cdot\mathtt{a}\cdot h(\gamma)[1, |h(\gamma)|-1]$ and $g(z) := h(\delta'\cdot z)$. If the first symbol in $h(\gamma)$ is $\mathtt{b}$, then we define $g(y) := h(y\cdot\beta')$ and $g(z) := h(\gamma)[2, |h(\gamma)|]\cdot\mathtt{b}\cdot h(\delta'\cdot z)$. If the last symbol in $h(\gamma)$ is $\mathtt{b}$ and the first symbol in $h(\gamma)$ is $\mathtt{a}$, then $h(\gamma) = u\cdot\mathtt{a}\cdot\mathtt{b}\cdot v$, $u, v \in \Sigma_2^*$. In this case, we define $g(y) := h(y\cdot\beta')\cdot\mathtt{a}\cdot u$ and $g(z) := v\cdot\mathtt{b}\cdot h(\delta'\cdot z)$. We observe that in all these cases we have $g(\beta\cdot y\cdot\mathtt{a}\cdot\mathtt{b}\cdot z\cdot\delta) = h(\alpha)$ and, thus, $L_{\mathrm{E},\Sigma_2}(\alpha) \subseteq L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}\cdot\mathtt{b}\cdot z\cdot\delta)$.

Next, we assume further that $\mathrm{var}(\beta'\cdot\gamma\cdot\delta')\cap\mathrm{var}(\beta\cdot\delta) = \emptyset$. Let $g$ be a substitution. Obviously, $g(\beta\cdot y\cdot\mathtt{a}\cdot\mathtt{b}\cdot z\cdot\delta) = h(\alpha)$, where $h(x) := g(x)$ if $x \in (\mathrm{var}(\beta\cdot\delta)\cup\{y, z\})$ and $h(x) := \varepsilon$ otherwise. This implies $L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}\cdot\mathtt{b}\cdot z\cdot\delta) \subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$. □

The fact that $L_{\mathrm{E},\Sigma_2}(x_1\cdot\mathtt{a}\cdot x_2x_2\cdot\mathtt{b}\cdot x_3) \in \mathrm{REG}$, which has already been stated in Proposition 9, is a simple application of Lemma 14, which implies $L_{\mathrm{E},\Sigma_2}(x_1\cdot\mathtt{a}\cdot x_2x_2\cdot\mathtt{b}\cdot x_3) = L_{\mathrm{E},\Sigma_2}(x_1\cdot\mathtt{ab}\cdot x_3)$. It is straightforward to construct more complex applications of Lemma 14 and it is also possible to apply it in an iterative way. For example, by applying Lemma 14 twice, we can show that

$$\begin{aligned}&L_{\mathrm{E},\Sigma_2}(x_1x_2x_3\cdot\mathtt{a}\cdot x_2x_4\cdot\mathtt{b}\cdot x_3x_4x_5x_6\cdot\mathtt{b}\cdot x_6x_7\cdot\mathtt{a}\cdot x_7x_8\cdot\mathtt{b}\cdot x_9\cdot\mathtt{a}\cdot x_{10}) =\\ &L_{\mathrm{E},\Sigma_2}(x_1\cdot\mathtt{ab}\cdot x_5x_6\cdot\mathtt{b}\cdot x_6x_7\cdot\mathtt{a}\cdot x_7x_8\cdot\mathtt{b}\cdot x_9\cdot\mathtt{a}\cdot x_{10}) =\\ &L_{\mathrm{E},\Sigma_2}(x_1\cdot\mathtt{ab}\cdot x_5\cdot\mathtt{ba}\cdot x_8\cdot\mathtt{b}\cdot x_9\cdot\mathtt{a}\cdot x_{10}) \in \mathrm{REG}\ .\end{aligned}$$

In the previous lemma, it is required that the factor $\gamma$ is delimited by different terminal symbols and, in the following, we shall see that an extension of the statement of Lemma 14 for the case that $\gamma$ is delimited by the same terminal symbols, is much more difficult to prove.

Roughly speaking, Lemma 14 holds due to the following reasons. Let $\alpha := y \cdot \beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{b} \cdot \delta' \cdot z$ be a pattern that satisfies the conditions of Lemma 14, then, for any substitution $h$ (with respect to $\Sigma_2$), $h(\alpha)$ necessarily contains the factor $\mathtt{ab}$. Conversely, since $y$ and $z$ are variables with only one occurrence and there are no terminals in $\beta' \cdot \gamma \cdot \delta'$, $\alpha$ can be mapped to every word that contains the factor $\mathtt{ab}$. On the other hand, for $\alpha' := y \cdot \beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{a} \cdot \delta' \cdot z$, $h(\alpha')$ does not necessarily contain the factor $\mathtt{aa}$ and it is not obvious if the factor $\beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{a} \cdot \delta'$ collapses to some simpler structure, as it is the case for $\alpha$. In fact, Theorem 12 states that if $\beta' = \delta' = \varepsilon$ and $\gamma = x^3$, then $L_{\mathrm{E},\Sigma_2}(\alpha') \notin \mathrm{REG}$.

However, by imposing a further restriction with respect to the factor $\gamma$, we can extend Lemma 14 to the case where $\gamma$ is delimited by the same terminal symbol. In order to prove this result, the next lemma is crucial, which states that for any terminal-free pattern that is delimited by two occurrences of symbols $\mathtt{a}$ and that has an even number of occurrences for every variable, if we apply any substitution to this pattern, we will necessarily obtain a word that contains a unary factor over $\mathtt{b}$ of even length that is delimited by two occurrences of $\mathtt{a}$.

**Lemma 15.** *Let $\alpha \in X^*$ such that, for every $x \in \mathrm{var}(\alpha)$, $|\alpha|_x$ is even. Then every $w \in L_{\mathrm{E},\Sigma_2}(\mathtt{a} \cdot \alpha \cdot \mathtt{a})$ contains a factor $\mathtt{ab}^{2n}\mathtt{a}$, $n \in \mathbb{N}_0$.*

Proof. First, we introduce the following definition that is convenient for this proof. A factor of the form $\mathtt{ab}^n\mathtt{a}$, $n \in \mathbb{N}_0$, is called a $\mathtt{b}$-*segment*. If $n$ is even, then $\mathtt{ab}^n\mathtt{a}$ is an *even* $\mathtt{b}$-*segment* and if $n$ is odd, then $\mathtt{ab}^n\mathtt{a}$ is an *odd* $\mathtt{b}$-*segment*. In a word $w \in \{\mathtt{a}, \mathtt{b}\}^*$, $\mathtt{b}$-segments that share exactly one occurrence of symbol $\mathtt{a}$ are considered to be distinct $\mathtt{b}$-segments, e. g., in $\mathtt{aab}^2\mathtt{ab}^4\mathtt{abab}^7\mathtt{a}$, there are 5 $\mathtt{b}$-segments, 3 of which are even $\mathtt{b}$-segments.

Before we can prove the statement of the lemma, we first prove the following claim:

*Claim.* Let $w_1 \in (\mathtt{a} \cdot \Sigma_2^*)$, $w_3 \in (\Sigma_2^* \cdot \mathtt{a})$, $w_2, v \in \Sigma_2^*$ and $v$ does not contain any even $\mathtt{b}$-segment. If $w_1 \cdot w_2 \cdot w_3$ has an odd number of even $\mathtt{b}$-segments, then $w_1 \cdot v \cdot w_2 \cdot v \cdot w_3$ has an odd number of even $\mathtt{b}$-segments as well.

*Proof (Claim).* We assume that for $w_1, w_2, w_3$ and $v$ the conditions of the lemma are satisfied and, for the sake of convenience, we define $w := w_1 \cdot w_2 \cdot w_3$ and $w' := w_1 \cdot v \cdot w_2 \cdot v \cdot w_3$. Intuitively, the statement of the lemma can be rephrased as follows. No matter where the two occurrences of $v$ are inserted into $w$, the total number of even $\mathtt{b}$-segments increases or decreases only by an even number. Since $v$ does not contain any even $\mathtt{b}$-segment, only the (possibly empty) prefix or suffix over $\mathtt{b}$ of $v$ can turn odd $\mathtt{b}$-segments of $w$ in even ones or vice versa. We shall first consider the case that $|w_2|_{\mathtt{a}} \geq 1$, i. e., $w_2$ contains at least one occurrence of symbol $\mathtt{a}$, and we recall that, since $w_1 \in (\mathtt{a} \cdot \Sigma_2^*)$ and $w_3 \in (\Sigma_2^* \cdot \mathtt{a})$, $w_1$ has a suffix of the form $\mathtt{ab}^*$ and $w_3$ has a prefix of the form $\mathtt{b}^*\mathtt{a}$. Furthermore, since $|w_2|_{\mathtt{a}} \geq 1$, $w_2$ has a prefix of the form $\mathtt{b}^*\mathtt{a}$ and a suffix of the form $\mathtt{ab}^*$. In summary, this implies that we can write $w'$ as

$$w' = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^n \cdot v \cdot \mathtt{b}^{n'} \cdot w'_2 \cdot \mathtt{b}^m \cdot v \cdot \mathtt{b}^{m'} \cdot \mathtt{a} \cdot w'_3\,,$$

where $n, n', m, m' \in \mathbb{N}_0$, $w'_2[1] = w'_2[-] = \mathtt{a}$, $w_1 = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^n$, $w_2 = \mathtt{b}^{n'} \cdot w'_2 \cdot \mathtt{b}^m$ and $w_3 = \mathtt{b}^{m'} \cdot \mathtt{a} \cdot w'_3$, and, furthermore,

$$w = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^{n+n'} \cdot w'_2 \cdot \mathtt{b}^{m+m'} \cdot \mathtt{a} \cdot w'_3 \,.$$

Obviously, all the even $\mathtt{b}$-segments in the factors $w'_1 \cdot \mathtt{a}$, $w'_2$ and $\mathtt{a} \cdot w'_3$ also occur in $w'$. Therefore, it is sufficient to compare the number of even $\mathtt{b}$-segments in the factors $\mathtt{a} \cdot \mathtt{b}^{n+n'} \cdot \mathtt{a}$ and $\mathtt{a} \cdot \mathtt{b}^{m+m'} \cdot \mathtt{a}$ with the number of even $\mathtt{b}$-segments in the factors $\mathtt{a} \cdot \mathtt{b}^n \cdot v \cdot \mathtt{b}^{n'} \cdot \mathtt{a}$ and $\mathtt{a} \cdot \mathtt{b}^m \cdot v \cdot \mathtt{b}^{m'} \cdot \mathtt{a}$.

If $v = \mathtt{b}^k$, $k \in \mathbb{N}_0$, then the $\mathtt{b}$-segment $\mathtt{a} \cdot \mathtt{b}^{n+n'} \cdot \mathtt{a}$ is changed into the $\mathtt{b}$-segment $\mathtt{a} \cdot \mathtt{b}^{n+k+n'} \cdot \mathtt{a}$ and the $\mathtt{b}$-segment $\mathtt{a} \cdot \mathtt{b}^{m+m'} \cdot \mathtt{a}$ is changed into the $\mathtt{b}$-segment $\mathtt{a} \cdot \mathtt{b}^{m+k+m'} \cdot \mathtt{a}$. If $k$ is even, then in $w'$ we have the same number of even $\mathtt{b}$-segments as in $w$, since $n + k + n'$ is even if and only if $n + n'$ is even, and $m + k + m'$ is even if and only if $m + m'$ is even. If, on the other hand, $k$ is odd, then $n + k + n'$ is even if and only if $n + n'$ is odd, and $m + k + m'$ is even if and only if $m + m'$ is odd. Thus, if $n + n'$ and $m + m'$ are both even or both odd, then the number of even $\mathtt{b}$-segments in $w'$ has decreased (or increased, respectively) by 2 compared to the number of even $\mathtt{b}$-segments in $w$. If, on the other hand, $n+n'$ is even and $m+m'$ is odd or the other way around, then in $w'$ there are as many even $\mathtt{b}$-segments as in $w$. So we can conclude that if $v = \mathtt{b}^k$, $k \in \mathbb{N}_0$, then the number of even $\mathtt{b}$-segments in $w'$ is odd.

We shall now assume that there is at least one occurrence of $\mathtt{a}$ in $v$, i. e., $v = \mathtt{b}^k \cdot u \cdot \mathtt{b}^{k'}$, $k, k' \in \mathbb{N}_0$, where $u[1] = u[-] = \mathtt{a}$. This implies

$$w' = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^{n+k} \cdot u \cdot \mathtt{b}^{k'+n'} \cdot w'_2 \cdot \mathtt{b}^{m+k} \cdot u \cdot \mathtt{b}^{k'+m'} \cdot \mathtt{a} \cdot w'_3 \,.$$

In the following we shall show that, for all possible choices of $n, n', m, m', k, k' \in \mathbb{N}_0$, the number of even $\mathtt{b}$-segments among the $\mathtt{b}$-segments $\mathtt{a} \cdot \mathtt{b}^{n+k} \cdot \mathtt{a}$, $\mathtt{a} \cdot \mathtt{b}^{k'+n'} \cdot \mathtt{a}$, $\mathtt{a} \cdot \mathtt{b}^{m+k} \cdot \mathtt{a}$ and $\mathtt{a} \cdot \mathtt{b}^{k'+m'} \cdot \mathtt{a}$ is even if and only if the number of even $\mathtt{b}$-segments among the $\mathtt{b}$-segments $\mathtt{a} \cdot \mathtt{b}^{n+n'} \cdot \mathtt{a}$ and $\mathtt{a} \cdot \mathtt{b}^{m+m'} \cdot \mathtt{a}$ is even. To this end, it is sufficient to note that if $(n+n')$ and $(m+m')$ are both even or both odd, then, for all possible choices of $n, n', m, m', k, k' \in \mathbb{N}_0$, either exactly 0, 2 or all 4 of the numbers $(n+k)$, $(k'+n')$, $(m+k)$ and $(k'+m')$ are even. If, on the other hand, one number of $(n+n')$ and $(m+m')$ is even and the other one is odd, then, for all possible choices of $n, n', m, m', k, k' \in \mathbb{N}_0$, either exactly 1 or 3 of the numbers $(n+k)$, $(k'+n')$, $(m+k)$ and $(k'+m')$ are even. This directly implies that the number of even $\mathtt{b}$-segments in $w'$ is odd, since, by assumption, the number of even $\mathtt{b}$-segments in $w$ is odd.

It remains to consider the case that $w_2 = \mathtt{b}^l$, $l \in \mathbb{N}_0$. We note that this implies the following.

$$w' = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^n \cdot v \cdot \mathtt{b}^l \cdot v \cdot \mathtt{b}^m \cdot \mathtt{a} \cdot w'_3 \,,$$

where $n, l, m, \in \mathbb{N}_0$, $w_1 = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^n$, $w_2 = \mathtt{b}^l$ and $w_3 = \mathtt{b}^m \cdot \mathtt{a} \cdot w'_3$, and, furthermore,

$$w = w'_1 \cdot \mathtt{a} \cdot \mathtt{b}^{n+l+m} \cdot \mathtt{a} \cdot w'_3 \,.$$

If $v = \mathtt{b}^k$, $k \in \mathbb{N}_0$, then $w' = w_1' \cdot \mathtt{a} \cdot \mathtt{b}^{n+k+l+k+m} \cdot \mathtt{a} \cdot w_3'$ and, since $(n+k+l+k+m)$ is even if and only if $(n + l + m)$ is even, we can directly conclude that $w'$ has as many even $\mathtt{b}$-segments as $w$.

If, on the other hand, $v = \mathtt{b}^k \cdot u \cdot \mathtt{b}^{k'}$, $k, k' \in \mathbb{N}_0$, where $u[1] = u[-] = \mathtt{a}$, then

$$w' = w_1' \cdot \mathtt{a} \cdot \mathtt{b}^{n+k} \cdot u \cdot \mathtt{b}^{k'+l+k} \cdot u \cdot \mathtt{b}^{k'+m} \cdot \mathtt{a} \cdot w_3' .$$

Similarly as before, we can show that, for all possible choices of $n, l, m, k, k' \in \mathbb{N}_0$, the number of even $\mathtt{b}$-segments among the $\mathtt{b}$-segments $\mathtt{a} \cdot \mathtt{b}^{n+k} \cdot \mathtt{a}$, $\mathtt{a} \cdot \mathtt{b}^{k'+l+k} \cdot \mathtt{a}$ and $\mathtt{a} \cdot \mathtt{b}^{k'+m} \cdot \mathtt{a}$ is even if and only if $\mathtt{a} \cdot \mathtt{b}^{n+l+m} \cdot \mathtt{a}$ is an odd $\mathtt{b}$-segment. To this end, it is sufficient to note that if $(n + l + m)$ is even, then, for all possible choices of $n, l, m, k, k' \in \mathbb{N}_0$, either exactly 1 or all 3 of the numbers $(n + k)$, $(k' + l + k)$ and $(k' + m)$ are even. If, on the other hand, $(n + l + m)$ is odd, then, for all possible choices of $n, l, m, k, k' \in \mathbb{N}_0$, either exactly 0 or 2 of the numbers $(n + k)$, $(k' + l + k)$ and $(k' + m)$ are even. This directly implies that the number of even $\mathtt{b}$-segments in $w'$ is odd, since, by assumption, the number of even $\mathtt{b}$-segments in $w$ is odd.

Hence, for all possible choices of $w_1, w_2, w_3$ and $v$, $w'$ has an odd number of even $\mathtt{b}$-segments, which concludes the proof. □ *(Claim)*

We are now ready to prove the statement of the lemma, i. e., for every $w \in L_{\mathrm{E},\Sigma_2}(\mathtt{a} \cdot \alpha \cdot \mathtt{a})$, $w$ contains an even $\mathtt{b}$-segment. Let $h$ be a substitution with $h(\mathtt{a} \cdot \alpha \cdot \mathtt{a}) = w$. Obviously, if, for some $x \in \mathrm{var}(\alpha)$, $h(x)$ contains an even $\mathtt{b}$-segment, then $h(\mathtt{a} \cdot \alpha \cdot \mathtt{a})$ contains an even $\mathtt{b}$-segment. Consequently, we only have to consider the case that, for every $x \in \mathrm{var}(\alpha)$, $h(x)$ does not contain an even $\mathtt{b}$-segment.

We can note that there are words $u_1, u_2, \ldots, u_k$, such that $u_1 = \mathtt{a} \cdot \mathtt{a}$, $u_k = h(\mathtt{a} \cdot \alpha \cdot \mathtt{a})$ and, for every $i$, $2 \leq i \leq k$, the word $u_i$ can be obtained by inserting two occurrences of a word $v$ into the word $u_{i-1}$. More precisely, we start with $u_1 = \mathtt{a} \cdot \mathtt{a}$ and insert two occurrences of $h(x_1)$ into $u_1$ in order to obtain $u_2$, then we repeat this step in order to construct $u_3$ and after $\frac{|\beta|_{x_1}}{2}$ such steps we stop. Next, we do the same for $\frac{|\beta|_{x_2}}{2}$ steps with respect to $h(x_2)$ and so on. Clearly, since, for every $x \in \mathrm{var}(\alpha)$, $|\alpha|_x$ is even, this can be done in such a way that $u_k = h(\mathtt{a} \cdot \alpha \cdot \mathtt{a})$ is satisfied. Furthermore, since $u_1$ has an odd number of even $\mathtt{b}$-segments, we can conclude with the above claim that, for every $i$, $1 \leq i \leq k$, the word $u_i$ has an odd number of even $\mathtt{b}$-segments, which implies that $u_k = h(\mathtt{a} \cdot \alpha \cdot \mathtt{a}) = w$ has at least one even $\mathtt{b}$-segments. This concludes the proof. □

By applying Lemma 15, we can show that if a pattern $\alpha := \beta \cdot y \cdot \beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{a} \cdot \delta' \cdot z \cdot \delta$ satisfies the conditions of Lemma 14, all variables in $\gamma$ have an even number of occurrences and there is at least one variable in $\gamma$ that occurs only twice, then the factor $y \cdot \beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{a} \cdot \delta' \cdot z$ can be substituted by a regular expression.

**Lemma 16.** *Let* $\alpha := \beta \cdot y \cdot \beta' \cdot \mathtt{a} \cdot \gamma \cdot \mathtt{a} \cdot \delta' \cdot z \cdot \delta$*, where* $\beta, \delta \in (\Sigma_2 \cup X)^*$*,* $\beta', \gamma, \delta' \in X^*$*,* $y, z \in X$*,* $|\alpha|_y = |\alpha|_z = 1$ *and, for every* $x \in \mathrm{var}(\gamma)$*,* $|\gamma|_x$ *is even. Then*

$L_{\mathrm{E},\Sigma_2}(\alpha) \subseteq L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}(\mathtt{bb})^*\mathtt{a}\cdot z\cdot\delta)$. *If, furthermore,* $\mathrm{var}(\beta'\cdot\gamma\cdot\delta')\cap\mathrm{var}(\beta\cdot\delta)=\emptyset$ *and there exists a* $z'\in\mathrm{var}(\gamma)$ *with* $|\alpha|_{z'}=2$*, then also* $L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}(\mathtt{bb})^*\mathtt{a}\cdot z\cdot\delta)\subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$*.*

Proof. Let $h$ be an arbitrary substitution. We first note that we can prove $h(\alpha)\in L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}\cdot(\mathtt{bb})^*\cdot\mathtt{a}\cdot z\cdot\delta)$ by showing that $h(y\cdot\beta'\cdot\mathtt{a}\cdot\gamma\cdot\mathtt{a}\cdot\delta'\cdot z)$ contains a factor of the form $\mathtt{a}\cdot\mathtt{b}^n\cdot\mathtt{a}$, where $n$ is even. We note that Lemma 15 directly implies that $h(\mathtt{a}\cdot\gamma\cdot\mathtt{a})$ contains such a factor. Thus, $L_{\mathrm{E},\Sigma_2}(\alpha)\subseteq L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}\cdot(\mathtt{bb})^*\cdot\mathtt{a}\cdot z\cdot\delta)$ follows.

In order to prove the second statement of the lemma, we assume that $\mathrm{var}(\beta'\cdot\gamma\cdot\delta')\cap\mathrm{var}(\beta\cdot\delta)=\emptyset$ and there exists a $z'\in(\mathrm{var}(\gamma)\setminus\mathrm{var}(\beta\cdot\beta'\cdot\delta'\cdot\delta))$ with $|\gamma|_{z'}=2$. Now let $h$ be an arbitrary substitution and let $h(\beta\cdot y\cdot\mathtt{a}\cdot(\mathtt{bb})^*\cdot\mathtt{a}\cdot z\cdot\delta)=h(\beta\cdot y)\cdot\mathtt{a}\cdot\mathtt{b}^{2n}\cdot\mathtt{a}\cdot h(z\cdot\delta)$, $n\in\mathbb{N}_0$. Obviously, $h(\beta\cdot y\cdot\mathtt{a}\cdot(\mathtt{bb})^*\cdot\mathtt{a}\cdot z\cdot\delta)=g(\alpha)$, where, for every $x\in(\mathrm{var}(\beta\cdot\delta)\cup\{y,z\})$, $g(x):=h(x)$, for every $x\in\mathrm{var}(\beta'\cdot\gamma\cdot\delta')\setminus\{z'\}$, $g(x):=\varepsilon$ and $g(z'):=\mathtt{b}^n$. This implies $L_{\mathrm{E},\Sigma_2}(\beta\cdot y\cdot\mathtt{a}\cdot(\mathtt{bb})^*\cdot\mathtt{a}\cdot z\cdot\delta)\subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$, which concludes the proof. □

Obviously, Lemmas 14 and 16 can also be applied in any order in the iterative way pointed out above with respect to Lemma 14. We shall illustrate this now in a more general way. Let $\alpha$ be an arbitrary pattern such that

$$\alpha:=\beta\cdot y_1\cdot\beta'_1\cdot\mathtt{a}\cdot\gamma_1\cdot\mathtt{a}\cdot\delta'_1\cdot z_1\cdot\pi\cdot y_2\cdot\beta'_2\cdot\mathtt{b}\cdot\gamma_2\cdot\mathtt{a}\cdot\delta'_2\cdot z_2\cdot\delta\,,$$

with $\beta,\pi,\delta\in(\Sigma_2\cup X)^*$, $\beta'_1,\beta'_2,\gamma_1,\gamma_2,\delta'_1,\delta'_2\in X^*$ and $y_1,y_2,z_1,z_2\in X$. If the factors $y_1\cdot\beta'_1\cdot\mathtt{a}\cdot\gamma_1\cdot\mathtt{a}\cdot\delta'_1\cdot z_1$ and $y_2\cdot\beta'_2\cdot\mathtt{b}\cdot\gamma_2\cdot\mathtt{a}\cdot\delta'_2\cdot z_2$ satisfy the conditions of Lemma 16 and Lemma 14, respectively, then we can conclude that $\alpha$ is equivalent to $\alpha':=\beta\cdot y_1\cdot\mathtt{a}(\mathtt{bb})^*\mathtt{a}\cdot z_1\cdot\pi\cdot y_2\cdot\mathtt{ba}\cdot z_2\cdot\delta$. This particularly means that the rather strong conditions

1. $\mathrm{var}(\beta'_1\cdot\gamma_1\cdot\delta'_1)\cap\mathrm{var}(\beta\cdot\pi\cdot\beta'_2\cdot\gamma_2\cdot\delta'_2\cdot\delta)=\emptyset$,
2. $\mathrm{var}(\beta'_2\cdot\gamma_2\cdot\delta'_2)\cap\mathrm{var}(\beta\cdot\beta'_1\cdot\gamma_1\cdot\delta'_1\cdot\pi\cdot\delta)=\emptyset$

must be satisfied. However, we can state that $L_{\mathrm{E},\Sigma_2}(\alpha)=L_{\mathrm{E},\Sigma_2}(\alpha')$ still holds if instead of conditions 1 and 2 from above the weaker condition $\mathrm{var}(\beta'_1\cdot\gamma_1\cdot\delta'_1\cdot\beta'_2\cdot\gamma_2\cdot\delta'_2)\cap\mathrm{var}(\beta\cdot\pi\cdot\delta)=\emptyset$ is satisfied. This claim can be easily proved by applying the same argumentations as in the proofs of Lemmas 14 and 16, and we can extend this result to arbitrarily many factors of the form $y_i\cdot\beta'_i\cdot c_1\cdot\gamma_i\cdot c_2\cdot\delta'_i\cdot z_i$, $c_1,c_2\in\Sigma_2$. Next, by the following definition, we formalise this observation in terms of a relation on patterns with regular expressions.

**Definition 2.** A pattern with regular expressions $\alpha$ is $\Sigma_2$*-reducible* to a pattern with regular expressions $\alpha'$ (denoted by $\alpha\rhd\alpha'$) if and only if the following conditions are satisfied.

- $\alpha$ contains factors $\alpha_i\in(\Sigma_2\cup X)^*$, $1\le i\le k$, where, for every $i$, $1\le i\le k$, $\alpha_i:=y_i\cdot\beta'_i\cdot d_i\cdot\gamma_i\cdot d'_i\cdot\delta'_i\cdot z_i$, with $\beta'_i,\gamma_i,\delta'_i\in X^+$, $y_i,z_i\in X$, $|\alpha|_{y_i}=|\alpha|_{z_i}=1$, $d_i,d'_i\in\Sigma_2$ and, if $d_i=d'_i$, then, for every $x\in\mathrm{var}(\gamma_i)$, $|\gamma_i|_x$ is even and there exists an $x'\in\mathrm{var}(\gamma_i)$ with $|\alpha|_{x'}=2$. Furthermore, the factors $\alpha_1,\alpha_2,\dots,\alpha_k$ can overlap by at most one symbol and the variables in the factors $\alpha_1,\alpha_2,\dots,\alpha_k$ occur exclusively in these factors.

- $\alpha'$ is obtained from $\alpha$ by substituting every $\alpha_i$, $1 \leq i \leq k$, by $y_i \cdot d_i d'_i \cdot z_i$, if $d_i \neq d'_i$ and by $y_i \cdot d_i(d''_i d''_i)^* d'_i \cdot z_i$, $d''_i \in \Sigma_2$, $d''_i \neq d_i$, if $d_i = d'_i$.

By generalising Lemmas 14 and 16, we can prove that if $\alpha$ is $\Sigma_2$-reducible to $\alpha'$, then $\alpha$ and $\alpha'$ describe the same E-pattern language with respect to $\Sigma_2$.

**Theorem 17.** *Let $\alpha$ and $\alpha'$ be patterns with regular expressions. If $\alpha \rhd \alpha'$, then $L_{\mathrm{E},\Sigma_2}(\alpha) = L_{\mathrm{E},\Sigma_2}(\alpha')$.*

Proof. We assume that $\alpha$ is $\Sigma_2$-reducible to $\alpha'$, which implies that $\alpha$ contains factors $\alpha_i \in (\Sigma_2 \cup X)^*$, $1 \leq i \leq k$, where, for every $i$, $1 \leq i \leq k$, $\alpha_i := y_i \cdot \beta'_i \cdot d_i \cdot \gamma_i \cdot d'_i \cdot \delta'_i \cdot z_i$, with $\beta'_i, \gamma_i, \delta'_i \in X^+$, $y_i, z_i \in X$, $|\alpha|_{y_i} = |\alpha|_{z_i} = 1$, $d_i, d'_i \in \Sigma_2$ and, if $d_i = d'_i$, then, for every $x \in \mathrm{var}(\gamma_i)$, $|\gamma_i|_x$ is even and there exists an $x' \in \mathrm{var}(\gamma_i)$ with $|\alpha|_{x'} = 2$. Furthermore, the factors $\alpha_1, \alpha_2, \ldots, \alpha_k$ can overlap by at most one symbol and the variables in the factors $\alpha_1, \alpha_2, \ldots, \alpha_k$ occur exclusively in these factors. Moreover, $\alpha'$ is obtained from $\alpha$ by substituting every $\alpha_i$, $1 \leq i \leq k$, by $\alpha'_i := y_i \cdot d_i \cdot d'_i \cdot z_i$, if $d_i \neq d'_i$ and by $\alpha'_i := y_i \cdot d_i \cdot (d''_i d''_i)^* \cdot d'_i \cdot z_i$, $d''_i \neq d_i$, if $d_i = d'_i$.

By Lemmas 14 and 16, we can conclude that $L_{\mathrm{E},\Sigma_2}(\alpha) \subseteq L_{\mathrm{E},\Sigma_2}(\pi_1)$, where $\pi_1$ is obtained from $\alpha$ by substituting $\alpha_1$ by $\alpha'_1$. In the same way, we can also conclude that $L_{\mathrm{E},\Sigma_2}(\pi_1) \subseteq L_{\mathrm{E},\Sigma_2}(\pi_2)$, where $\pi_2$ obtained from $\pi_1$ by substituting $\alpha_2$ by $\alpha'_2$. By repeating this argumentation, $L_{\mathrm{E},\Sigma_2}(\alpha) \subseteq L_{\mathrm{E},\Sigma_2}(\alpha')$ follows.

It remains to prove that $L_{\mathrm{E},\Sigma_2}(\alpha') \subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$. To this end, let $h$ be an arbitrary substitution. We shall show that $h(\alpha') \in L_{\mathrm{E},\Sigma_2}(\alpha)$ by defining a substitution $g$ that satisfies $h(\alpha') = g(\alpha)$. First, let $A \subseteq \{1, 2, \ldots, k\}$ be such that, for every $i$, $1 \leq i \leq k$, $d_i = d'_i$ if and only if $i \in A$. Moreover, for every $i \in A$, let $x_i$ be a variable that satisfies $x_i \in \mathrm{var}(\gamma_i)$ with $|\alpha|_{x_i} = 2$. Now, for every $x \in \mathrm{var}(\alpha) \setminus (\bigcup_{i=1}^{k} \mathrm{var}(\beta'_i \cdot \gamma_i \cdot \delta'_i))$, we define $g(x) := h(x)$. For every $x \in (\bigcup \mathrm{var}(\beta'_i \cdot \gamma_i \cdot \delta'_i) \setminus \{x_i \mid i \in A\})$, we define $g(x) := \varepsilon$. So it only remains to define $g(x_i)$, for every $x_i \in A$. To this end, we first note that, for every $i \in A$, $\alpha'_i = y_i \cdot d_i \cdot (d''_i d''_i)^* \cdot d'_i \cdot z_i$. Now, for every $i \in A$, let $n_i \in \mathbb{N}_0$ be such that $h$ maps $(d''_i d''_i)^*$ to $(d''_i)^{n_i}$. Finally, for every $i \in A$, we define $g(x_i) := (d''_i)^{n_i}$. It can be easily verified that $g(\alpha) = h(\alpha')$. Thus, $L_{\mathrm{E},\Sigma_2}(\alpha') \subseteq L_{\mathrm{E},\Sigma_2}(\alpha)$, which concludes the proof. □

We conclude this section by discussing a more complex example that illustrates how Definition 2 and Theorem 17 constitute a sufficient condition for the regularity of the E-pattern language of a pattern with respect to $\Sigma_2$. Let $\alpha$ be the following pattern.

$$\underbrace{x_1 \mathtt{a} x_2 x_3^2 \mathtt{b} x_4 x_3 x_5 x_6}_{\alpha_1 := y_1 \cdot \beta'_1 \cdot \mathtt{a} \cdot \gamma_1 \cdot \mathtt{b} \cdot \delta'_1 \cdot z_1} x_7^2 \underbrace{x_8 x_9 x_5 x_3 \mathtt{a} x_4 x_5 x_4 x_9 x_{10} \mathtt{b} x_{11}}_{\alpha_2 := y_2 \cdot \beta'_2 \cdot \mathtt{a} \cdot \gamma_2 \cdot \mathtt{b} \cdot \delta'_2 \cdot z_2} \mathtt{a} x_{12} \mathtt{b} x_{13} \mathtt{a} \underbrace{x_{14} x_{15} \mathtt{b} x_{15}^2 x_{16}^2 \mathtt{b} x_{17}}_{\alpha_3 := y_3 \cdot \beta'_3 \cdot \mathtt{a} \cdot \gamma_3 \cdot \mathtt{b} \cdot \delta'_3 \cdot z_3}.$$

By Definition 2, $\alpha \rhd \beta$ holds, where $\beta$ is obtained from $\alpha$ by substituting the above defined factors $\alpha_1$, $\alpha_2$ and $\alpha_3$ by factors $x_1 \cdot \mathtt{ab} \cdot x_6$, $x_8 \cdot \mathtt{ab} \cdot x_{11}$ and $x_{14} \cdot \mathtt{b}(\mathtt{aa})^* \mathtt{b} \cdot x_{17}$, respectively, i. e.,

$$\beta := x_1 \mathtt{ab} x_6 x_7 x_7 x_8 \mathtt{ab} x_{11} \mathtt{a} x_{12} \mathtt{b} x_{13} \mathtt{a} x_{14} \mathtt{b}(\mathtt{aa})^* \mathtt{b} x_{17}\,.$$

Furthermore, by Theorem 17, we can conclude that $L_{\mathrm{E},\Sigma_2}(\alpha) = L_{\mathrm{E},\Sigma_2}(\beta)$. However, we can also apply the same argumentation to different factors of $\alpha$, as pointed out below:

$$x_1 \mathtt{a} \underbrace{x_2 x_3^2 \mathtt{b} x_4 x_3 x_5 x_6 x_7^2 x_8 x_9 x_5 x_3 \mathtt{a} x_4 x_5 x_4 x_9 x_{10}}_{\alpha_1 := y_1 \cdot \beta_1' \cdot \mathtt{a} \cdot \gamma_1 \cdot \mathtt{b} \cdot \delta_1' \cdot z_1} \mathtt{b} x_{11} \mathtt{a} x_{12} \mathtt{b} x_{13} \mathtt{a} \underbrace{x_{14} x_{15} \mathtt{b} x_{15}^2 x_{16}^2 \mathtt{b} x_{17}}_{\alpha_2 := y_2 \cdot \beta_2' \cdot \mathtt{a} \cdot \gamma_2 \cdot \mathtt{b} \cdot \delta_2' \cdot z_2} .$$

Now, again by Definition 2, $\alpha \rhd \beta'$ is satisfied, where

$$\beta' := x_1 \mathtt{a} x_2 \mathtt{ba} x_{10} \mathtt{b} x_{11} \mathtt{a} x_{12} \mathtt{b} x_{13} \mathtt{a} x_{14} \mathtt{b} (\mathtt{aa})^* \mathtt{b} x_{17} .$$

Since every variable of $\beta'$ has only one occurrence, it can be easily seen that $L_{\mathrm{E},\Sigma_2}(\beta') \in \mathrm{REG}$ and, by Theorem 17, $L_{\mathrm{E},\Sigma_2}(\alpha) \in \mathrm{REG}$ follows.

## 5. Conclusions and Further Research Directions

In this paper, we have investigated the phenomenon of regular and context-free pattern languages over alphabets of size 2 or 3. In Section 3, we have pointed out by Theorem 7 that having at least one variable with only one occurrence is a necessary condition for a pattern to define a regular language. Regarding context-free languages, Theorem 8 states the same result, but only with respect to alphabets of size at least 3. In conjunction with the insights produced by Jain et al. [9], the difficult cases of the phenomenon of regular E-pattern languages have therefore been narrowed down to patterns with at least one terminal, with at least one variable with only one occurrence and with respect to alphabets of size 2 or 3. In the remainder of Section 3, numerous examples of such patterns have been provided, and we have shown how the interaction between the single occurrence variables and the terminal symbols can cause complicated structures in the pattern to collapse to simple structures in the words of the corresponding pattern language. Then, in Section 4, the thus gained insights have been generalised to a sufficient condition for a pattern to describe a regular E-pattern language over a two letter alphabet and it has been applied in order to demonstrate how rich the class of these patterns seems to be.

Unfortunately, we are still not able to characterise the class of regular E-pattern languages over an alphabet of size 2 or any other class of regular or context-free pattern languages for which characterisations are not already known.

The largest unsettled questions (which are also stated by Jain et al. [9]) arise with respect to the nonerasing case. In contrast to the E-case, for the NE-case we do not only lack characterisations of the regular and context-free pattern languages over small alphabets, but with respect to any fixed alphabet of size at least 2. In addition to this, the existence of context-free NE-pattern languages that are not regular is settled only for alphabets of size at most 3 and is open for larger alphabets.

## References

[1] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.

[2] C. Bader and A. Moura. A generalization of Ogden's Lemma. *Journal of the Association for Computing Machinery*, 29:404–407, 1982.

[3] J. Bremer and D. D. Freydenberger. Inclusion problems for patterns with a bounded number of variables. In *Proc. 14th International Conference on Developments in Language Theory, DLT 2010*, volume 6224 of *Lecture Notes in Computer Science*, pages 100–111, 2010.

[4] C. Câmpeanu, K. Salomaa, and S. Yu. A formal study of practical regular expressions. *International Journal of Foundations of Computer Science*, 14:1007–1018, 2003.

[5] G. Filè. The relation of two patterns with comparable language. In *Proc. 5th Annual Symposium on Theoretical Aspects of Computer Science, STACS 1988*, volume 294 of *Lecture Notes in Computer Science*, pages 184–192, 1988.

[6] D.D. Freydenberger and D. Reidenbach. Bad news on decision problems for patterns. *Information and Computation*, 208:83–96, 2010.

[7] D.D. Freydenberger, D. Reidenbach, and J.C. Schneider. Unambiguous morphic images of strings. *International Journal of Foundations of Computer Science*, 17:601–628, 2006.

[8] 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.

[9] S. Jain, Y. S. Ong, and F. Stephan. Regular patterns, regular languages and context-free languages. *Information Processing Letters*, 110:1114–1119, 2010.

[10] T. Jiang, E. Kinber, A. Salomaa, K. Salomaa, and S. Yu. Pattern languages with and without erasing. *International Journal of Computer Mathematics*, 50:147–163, 1994.

[11] T. Jiang, A. Salomaa, K. Salomaa, and S. Yu. Decision problems for patterns. *Journal of Computer and System Sciences*, 50:53–63, 1995.

[12] S. Lange and R. Wiehagen. Polynomial-time inference of arbitrary pattern languages. *New Generation Computing*, 8:361–370, 1991.

[13] A. Mateescu and A. Salomaa. Finite degrees of ambiguity in pattern languages. *RAIRO Informatique théoretique et Applications*, 28:233–253, 1994.

[14] 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.

[15] Y.K. Ng and T. Shinohara. Developments from enquiries into the learnability of the pattern languages from positive data. *Theoretical Computer Science*, 397:150–165, 2008.

[16] D. Reidenbach. *The Ambiguity of Morphisms in Free Monoids and its Impact on Algorithmic Properties of Pattern Languages*. PhD thesis, Fachbereich Informatik, Technische Universität Kaiserslautern, 2006. Logos Verlag, Berlin.

[17] D. Reidenbach. Discontinuities in pattern inference. *Theoretical Computer Science*, 397:166–193, 2008.

[18] D. Reidenbach and M. L. Schmid. Regular and context-free pattern languages over small alphabets. In *Proc. 16th International Conference on Developments in Language Theory, DLT 2012*, volume 7410 of *Lecture Notes in Computer Science*, pages 130–141, 2012.

[19] P. Rossmanith and T. Zeugmann. Stochastic finite learning of the pattern languages. *Machine Learning*, 44:67–91, 2001.

[20] T. Shinohara. Polynomial time inference of extended regular pattern languages. In *Proc. RIMS Symposia, Kyoto*, volume 147 of *LNCS*, pages 115–127, 1982.

[21] T. Shinohara. Polynomial time inference of pattern languages and its application. In *Proc. 7th IBM MFCS*, pages 191–209, 1982.

[22] S. Yu. Regular languages. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, chapter 2, pages 41–110. Springer, 1997.