Closure Properties of Pattern Languages[☆]

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

[a] *Department of Computer Science, Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK*
[b] *Fachbereich 4 – Abteilung Informatik, Universität Trier, D-54286 Trier, Germany*

Abstract

Pattern languages are a well-established class of languages, but very little is known about their closure properties. In the present paper we establish a large number of closure properties of the terminal-free pattern languages, and we characterise when the union of two terminal-free pattern languages is again a terminal-free pattern language. We demonstrate that the equivalent question for general pattern languages is characterised differently, and that it is linked to some of the most prominent open problems for pattern languages. We also provide fundamental insights into a well-known construction of E-pattern languages as unions of NE-pattern languages, and vice versa.

Keywords: Pattern languages, Closure properties

1. Introduction

Pattern languages were introduced by Dana Angluin [1] in order to model the algorithmic inferrability of patterns that are common to a set of words. In this context, a pattern is a sequence of variables and terminal symbols, and its language is the set of all words that can be generated from the pattern by a substitution that replaces all variables in the pattern by words of terminal symbols. Hence, more formally, a substitution is a terminal-preserving morphism, i. e., a morphism that maps every terminal symbol to itself. For example, the pattern language of the pattern $\alpha := x_1 x_1 \mathtt{a} x_2 \mathtt{b}$, where x_1, x_2 are variables and $\mathtt{a}, \mathtt{b}$ are terminal symbols, is the set of all words that have a square as a prefix, followed by an arbitrary suffix that begins with the letter `a` and ends with the letter `b`. Thus, e.g., `abbabbaab` is contained in the language of α, whereas `bbbaa` is not. It is a direct consequence of these definitions that a pattern language is either a singleton or infinite. Furthermore, it is worth noting that two basic types of pattern languages are considered in the literature, depending on whether the

[☆] A preliminary version [4] of this paper was presented at the conference DLT 2014.

[*] Corresponding author

Email addresses: `J.Day@lboro.ac.uk` (Joel D. Day), `D.Reidenbach@lboro.ac.uk` (Daniel Reidenbach), `MSchmid@uni-trier.de` (Markus L. Schmid)

variables must stand for nonempty words (referred to as non erasing or NE-pattern languages) or whether they may represent the empty word (so-called extended, erasing or simply E-pattern languages).

While the definition of pattern languages is simple, many of their properties are known to be related to complex phenomena in combinatorics on words, such as pattern avoidability (see Jiang et al. [9]) and ambiguity of morphisms (see Reidenbach [17]). Hence, the knowledge on pattern languages is still patchy, despite recent progress mainly regarding decision problems (see, e. g., Freydenberger, Reidenbach [7], Fernau, Schmid [5], Fernau et al. [6] and Reidenbach, Schmid [18]) and the relation to the Chomsky hierarchy (see Jain et al. [8] and Reidenbach, Schmid [19]).

Establishing the closure properties of a class of formal languages is one of the most classical and fundamental research tasks in formal language theory and any respective progress normally leads to insights and techniques that yield a better understanding of the class. In the case of pattern languages, it is known since Angluin's initial work that they are not closed under most of the usual operations, including union, intersection and complement. However, these non-closure properties can be shown by using very basic example patterns and exploiting peculiarities of the definition of pattern languages. For example, if a pattern does not contain a variable, then its language is a singleton; hence the union of any two distinct singleton pattern languages contains two elements, and therefore it cannot be a pattern language. Furthermore, the intersection of two pattern languages given by patterns that start with different terminal symbols is empty and the empty set, although a trivial language, is not a pattern language as well. Since, apart from a strong result by Shinohara [20] on the union of NE-pattern languages, hardly anything is known beyond such immediate facts, we can observe that in the case of pattern languages the existing closure properties fail to contribute to our understanding of their intrinsic properties.

It is the main purpose of this paper to investigate the closure properties of pattern languages more thoroughly. To this end, in Section 3, we consider the closure properties of two important subclasses of pattern languages, namely the classes of terminal-free NE- and E-pattern languages, i. e., pattern languages that are generated by patterns that do not contain any terminal symbols. This choice is motivated by the fact that terminal-free patterns have been a recent focus of interest in the research on pattern languages and, furthermore, most existing examples for non-closure of pattern languages (including the two examples for union and intersection given in the previous paragraph) do not translate to the terminal-free case. In Section 3.1, we completely characterise when the union of two terminal-free pattern languages is again a terminal-free pattern language and, in Section 3.2, we prove their non-closure under intersection, for which the situation is much more complicated compared to the operation of union.

We consider general pattern languages in Section 4, and we provide complex examples demonstrating that it is probably a very hard task to obtain full characterisations of those pairs of pattern languages whose unions or intersections are again a pattern language. In Section 4.3, we also study the question whether an E-pattern language can be expressed by the union of *nonerasing* pattern languages and, likewise, whether an NE-pattern language can be expressed by the union of *erasing* pattern languages. This question is slightly at odds with the classical investigation of closure properties, since we apply a language operation to members of one class and ask whether the resulting language is a member of another class. However, in the case of pattern languages, this makes sense, since every NE-pattern language is a finite union of E-pattern languages and every E-pattern language is a finite union of NE-pattern languages (see Jiang et al. [9]), a phenomenon that has been widely utilised in the context of inductive inference of pattern languages (see, e.g., Wright [22], Shinohara, Arimura [21]).

2. Definitions and Preliminary Results

The symbols $\cup$, $\cap$ and $\setminus$ denote the set operations of *union*, *intersection* and *set difference*, respectively. For sets U and B with $B \subseteq U$, $\overline{B} := U \setminus B$ is the *complement* of B.

Let $\mathbb{N} := \{1, 2, 3, \dots\}$ and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For an arbitrary alphabet A, a *word* (*over* A) is a finite sequence of symbols from A, and ε stands for the *empty word*. The notation A^+ denotes the set of all nonempty words over A, and $A^* := A^+ \cup \{\varepsilon\}$. For the *concatenation* of two words w_1, w_2 we write $w_1 \cdot w_2$ or simply $w_1 w_2$, and w^n stands for the n-fold concatenation of the word w. We say that a word $v \in A^*$ is a *factor* of a word $w \in A^*$ 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 word, then v is a *prefix* (or a *suffix*, respectively) of w. If $w = w_0 v_1 w_1 v_2 \cdots v_n w_n$ and $v = v_1 v_2 \cdots v_n$, for some $w_0, w_n \in A^*$, $w_i, v_j \in A^+$, $1 \leq i \leq n-1$, $1 \leq j \leq n$, then v is a *subsequence* of w. The notation $|K|$ stands for the size of a set K or the length of a word K. For a $w \in A^*$ and $a \in A$, $|w|_a$ denotes the number of occurrences of the symbol a in w. A word w is *primitive* if, for any u such that $w = u^k$, $k = 1$. The *primitive root* of a word w is the primitive word u such that $w = u^k$, $k \in \mathbb{N}$.

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, for every $a \in A$, $h(a) \neq \varepsilon$. A morphism h is *ambiguous (with respect to a word w)* if there exists a morphism g satisfying $g(w) = h(w)$ and, for a letter a in w, $g(a) \neq h(a)$. If such a morphism g does not exist, then h is called *unambiguous (with respect to w)*. A morphism $\sigma : A^* \to B^*$ is *periodic* if for some (primitive) word $w \in B^*$, $\sigma(x) \in \{w\}^*$ for every $x \in A$. The word w will be referred to as the *primitive root* of σ. If $|\sigma(x)| = 1$ for every $x \in A$, then σ is *1-uniform*.

Let Σ 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 $X := \{x_1, x_2, x_3, \dots\}$. A *pattern* is a nonempty word over $\Sigma \cup X$, a *terminal-free pattern* is a nonempty word over X; if a word contains symbols from Σ only, then we occasionally call it a *terminal word*. For any pattern α, we refer to the set of variables in α as $\mathrm{var}(\alpha)$. If the variables in a pattern α are labelled in the natural way, then it is said to be in *canonical form*, i. e., α is in canonical form if, for some $n \in \mathbb{N}$,

$\text{var}(\alpha) = \{x_1, x_2, \ldots, x_n\}$ and, for any $x_i, x_j \in \text{var}(\alpha)$ with $i < j$, there is a prefix β of α such that $x_i \in \text{var}(\beta)$ and $x_j \notin \text{var}(\beta)$. A pattern α is a *one-variable* pattern if $|\text{var}(\alpha)| = 1$. A pattern α is *periodicity forcing* if for any alphabet Σ and morphisms $g, h : \text{var}(\alpha)^* \to \Sigma^*$, $g(\alpha) = h(\alpha)$ implies g and h are periodic or $g = h$. A morphism $h : (\Sigma \cup X)^* \to (\Sigma \cup X)^*$ is *terminal-preserving* if $h(a) = a$ for every $a \in \Sigma$. The *residual* of a pattern α is the word $h_\varepsilon(\alpha)$, where $h_\varepsilon : (\Sigma \cup X)^* \to (\Sigma \cup X)^*$ is a terminal preserving morphism with $h_\varepsilon(x) := \varepsilon$ for every $x \in \text{var}(\alpha)$. A terminal-preserving morphism $h : (\Sigma \cup X)^* \to \Sigma^*$ is called a *substitution*.

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

Note that we call a pattern language terminal-free if there exists a terminal-free pattern that generates it.

Some parts of our reasoning in the subsequent sections is based on word equations, which are defined as follows. For a set of unknowns Y, a terminal alphabet Σ, and two words $\alpha, \beta \in (Y \cup \Sigma)^+$, the expression $\alpha = \beta$ is called a *word equation*. The solutions are terminal-preserving morphisms $\sigma : (Y \cup \Sigma)^* \to \Sigma^*$ such that $\sigma(\alpha) = \sigma(\beta)$. A solution is *periodic* if the morphism is periodic. The words $\sigma(\alpha)$ $(= \sigma(\beta))$ will be referred to as *solution-words*. It is often convenient to interpret variables from patterns as unknowns, and so word equations will often be formulated from two patterns.

This concludes the basic definitions of this paper. We now begin our investigation of the closure properties of the class of pattern languages. As a starting point, we refer to the corresponding result in the initial paper on pattern languages:

Theorem 1 (Angluin [1]). NE-*pattern languages are not closed under union, intersection, complement, Kleene plus, morphism and inverse morphism.* NE-*pattern languages are closed under concatenation and reversal.*

3. Terminal-free Patterns

As briefly explained in Section 1, the proof of Theorem 1 heavily relies on the fact that patterns can contain terminal symbols. In the present section, we therefore wish to study whether the situation changes if we consider the classes of terminal-free E-pattern languages and terminal-free NE-pattern languages.

3.1. Union

Simple examples show that neither the terminal-free NE-pattern languages nor the terminal-free E-pattern languages are closed under union:

Proposition 2. *Let Σ be an arbitrary alphabet. For every $Z, Z' \in \{\mathrm{E}, \mathrm{NE}\}$, there does not exist a pattern γ such that $L_{Z,\Sigma}(\gamma) = L_{Z',\Sigma}(x_1x_1) \cup L_{Z',\Sigma}(x_1x_1x_1)$.*

PROOF. The cases $Z \neq Z'$ are trivial since then either $\varepsilon \in L_{\mathrm{E},\Sigma}(x_1 x_1) \cup L_{\mathrm{E},\Sigma}(x_1 x_1 x_1) \setminus L_{\mathrm{NE},\Sigma}(\gamma)$ or $\varepsilon \in L_{\mathrm{E},\Sigma}(\gamma) \setminus L_{\mathrm{NE},\Sigma}(x_1 x_1) \cup L_{\mathrm{NE},\Sigma}(x_1 x_1 x_1)$.

If there exists a pattern γ with $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{E},\Sigma}(x_1 x_1) \cup L_{\mathrm{E},\Sigma}(x_1 x_1 x_1)$, then $\varepsilon \in L_{\mathrm{E},\Sigma}(\gamma)$; thus, γ is terminal-free. Furthermore, there is no $x \in \mathrm{var}(\gamma)$ with $|\gamma|_x = 1$, since otherwise $L_{\mathrm{E},\Sigma}(\gamma) = \Sigma^*$. In order to produce $\mathtt{aa}$ and $\mathtt{aaa}$, for some $\mathtt{a} \in \Sigma$, γ must contain a variable with exactly two occurrences and a variable with exactly three occurrences. This implies $\mathtt{a}^5 \in L_{\mathrm{E},\Sigma}(\gamma)$, which is a contradiction.

If there exists a pattern γ with $L_{\mathrm{NE},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(x_1 x_1) \cup L_{\mathrm{NE},\Sigma}(x_1 x_1 x_1)$, then $|\gamma| = 2$. If γ is not terminal-free, then γ is of the form bx_1, $x_1 b$ or bc for some (not necessarily different) $b, c \in \Sigma$, which obviously contradicts $L_{\mathrm{NE},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(x_1 x_1) \cup L_{\mathrm{NE},\Sigma}(x_1 x_1 x_1)$. Hence, $\gamma \in \{x_1 x_2, x_1 x_1\}$, but $x_1 x_2$ can generate the word $\mathtt{a}^5$, for some $\mathtt{a} \in \Sigma$, and $x_1 x_1$ cannot generate words of length 3. $\square$

It is worth noting that the above statement also provides a first minor insight into the topic of expressing E-pattern languages as unions of NE-pattern languages and vice versa. We shall study this subject in Section 4.3 for patterns with terminal symbols in much more detail. In the present section, we merely want to point out that the union of two terminal-free E-pattern languages is indeed never a terminal-free NE-pattern language, and the union of two terminal-free NE-pattern languages cannot be a terminal-free E-pattern language (in a similar way as in the proof of Proposition 2, this follows from the trivial fact that every terminal-free E-pattern language and no terminal-free NE-pattern language contains ε):

Proposition 3. *Let Σ be an arbitrary alphabet, and let α and β be terminal-free patterns. Then there does not exist a terminal-free pattern γ with $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$ or $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.*

In the remainder of this section we wish to prove a similarly strong result for the actual closure of the class of terminal-free E- or NE-pattern languages. Hence, we wish to characterise those pairs of terminal-free (NE-/E-)pattern languages where the union again is a terminal-free (NE-/E-)pattern language. Our results shall demonstrate that the union of two terminal-free E-pattern languages is only a terminal-free E-pattern language in the trivial case, namely if there is an inclusion relation between the two languages; the same holds for the NE-pattern languages.

Our reasoning on the E case is based on a construction given in [10] for a morphism τ_k such that, for two patterns α and β, the word $\tau_{|\beta|}(\alpha)$ is contained in $L_{\mathrm{E},\Sigma}(\beta)$ if and only if there exists a morphism φ from β to α.

Theorem 4. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let α and β be terminal-free patterns. There exists a terminal-free pattern γ with $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$ if and only if $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ or $L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\alpha)$.*

PROOF. The *if* direction is trivial. For every $k \in \mathbb{N}$, let the morphism $\tau_k : \{x_1, x_2, \ldots\}^* \to \{\mathtt{a}, \mathtt{b}\}^*$ be given by

$$\tau_k(x_i) := \mathtt{a}\mathtt{b}^{ki+1}\mathtt{a} \cdot \mathtt{a}\mathtt{b}^{ki+2}\mathtt{a} \cdots \mathtt{a}\mathtt{b}^{k(i+1)}\mathtt{a}\,,$$

for every $i \geq 1$.

We need the following claim, which is a simple extension of a reasoning from [10].

Claim 1[Jiang et al. [10]]: For every $k \in \mathbb{N}$ with $k \geq |\beta|$, $\tau_k(\alpha) \in L_{\mathrm{E},\Sigma}(\beta)$ if and only if there exists a morphism ϕ such that $\phi(\beta) = \alpha$.

We are now ready to prove the statement. Let $k := \max(|\alpha|, |\beta|)$. Note that, for the equivalence to hold, $\tau_k(\gamma) \in L_{\mathrm{E},\Sigma}(\alpha)$, or $\tau_k(\gamma) \in L_{\mathrm{E},\Sigma}(\beta)$. Without loss of generality let it be α. Then, by Claim 1, since $k \geq |\alpha|$, there exists a morphism ϕ such that $\phi(\alpha) = \gamma$. Thus $L_{\mathrm{E},\Sigma}(\gamma) \subseteq L_{\mathrm{E},\Sigma}(\alpha)$. Clearly, $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$, so the languages are equivalent and the statement holds. □

As illustrated by the following proposition, Theorem 4 does not hold for unary alphabets:

Proposition 5. *Let Σ be an alphabet, $|\Sigma| = 1$, and let $\alpha := x_1^2 x_2^5$, $\beta := x_1^3$ and $\gamma := x_1^2 x_2^3$. Then $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$, $L_{\mathrm{E},\Sigma}(\beta) \not\subseteq L_{\mathrm{E},\Sigma}(\alpha)$, and $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.*

Proof. Let $\Sigma := \{\mathtt{a}\}$. Obviously, $\mathtt{a}^2 \in L_{\mathrm{E},\Sigma}(\alpha) \setminus L_{\mathrm{E},\Sigma}(\beta)$ and $\mathtt{a}^3 \in L_{\mathrm{E},\Sigma}(\beta) \setminus L_{\mathrm{E},\Sigma}(\alpha)$, and therefore the two languages are incomparable. When comparing the languages of α and γ, we can observe that $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{E},\Sigma}(\alpha) \cup \{\mathtt{a}^3\}$. The word $\mathtt{a}^3$ is included in $L_{\mathrm{E},\Sigma}(\beta)$, and $L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$. Hence, $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$. □

It can be observed from simple examples that, in the nonerasing case, inclusion cannot be characterised by the existence of a morphism between the generating patterns. Thus, the argument used in the proof of Theorem 4 cannot be extended to the nonerasing case. However, a corresponding result can be obtained by looking at the shortest words in the nonerasing languages of α, β and γ. To this end, we define, for a pattern α, the set M_α to be $\{\sigma(\alpha) \mid \sigma : \mathrm{var}(\alpha)^* \to \Sigma^* \text{ is 1-uniform}\}$.

The set M_α has been used to positive effect in existing literature (see, e.g., Lange, Wiehagen [11]). It is particularly useful when considering nonerasing pattern languages, because it encodes exactly the original pattern α (up to a renaming of variables), as shown by the next lemma.

Lemma 6. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let $\alpha, \beta \in \mathbb{N}^+$ be patterns in canonical form. Then $M_\alpha = M_\beta$ if and only if $\alpha = \beta$.*

Proof. Clearly the *if* direction is trivial. Consider the case that $\alpha \neq \beta$. Since $|\mathrm{M}_\gamma| = |\Sigma|^{|\mathrm{var}(\gamma)|}$ for every pattern γ, $|\mathrm{var}(\alpha)| \neq |\mathrm{var}(\beta)|$ implies $|M_\alpha| \neq |M_\beta|$. Similarly, if $|\alpha| \neq |\beta|$, then the two sets contain words of different lengths, and therefore they cannot be equal. Thus it suffices to check that the statement holds for patterns of equal length, with equal numbers of variables. This is verified by Lemma 2.2 in [14]. □

Moreover, the set M_α has a number of convenient properties when considering the union of two NE-pattern languages. One such example is that if α is strictly shorter than β, then the set of shortest words in $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$ will be exactly M_α. Thus, if the union is itself the nonerasing language of some pattern γ, we have that $\gamma = \alpha$ up to a renaming of variables. The next lemma establishes a similar result for the case that $|\alpha| = |\beta|$ by considering $|M_\alpha \cup \mathrm{M}_\beta|$.

Lemma 7. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let α, β be terminal-free patterns in canonical form with $|\alpha| = |\beta|$. Suppose that γ is a terminal-free pattern (again in canonical form) with $\mathrm{M}_\alpha \cup \mathrm{M}_\beta = \mathrm{M}_\gamma$. Then $\gamma \in \{\alpha, \beta\}$.*

Proof. Note that if $|\operatorname{var}(\alpha)| > |\operatorname{var}(\gamma)|$ or $|\operatorname{var}(\beta)| > |\operatorname{var}(\gamma)|$, then either $|\mathrm{M}_\alpha| > |\mathrm{M}_\gamma|$ or $|\mathrm{M}_\beta| > |\mathrm{M}_\gamma|$ and the assumption that $\mathrm{M}_\alpha \cup \mathrm{M}_\beta = \mathrm{M}_\gamma$ cannot hold. Suppose that $|\operatorname{var}(\alpha)| < |\operatorname{var}(\gamma)|$ and $|\operatorname{var}(\beta)| < |\operatorname{var}(\gamma)|$. Obviously, $|\mathrm{M}_\alpha| + |\mathrm{M}_\beta| = |\Sigma|^{|\operatorname{var}(\alpha)|} + |\Sigma|^{|\operatorname{var}(\beta)|}$. If $|\operatorname{var}(\alpha)| \leq |\operatorname{var}(\beta)|$, then

$$\begin{aligned}
|\Sigma|^{|\operatorname{var}(\alpha)|} + |\Sigma|^{|\operatorname{var}(\beta)|} &\leq |\Sigma|^{|\operatorname{var}(\beta)|} + |\Sigma|^{|\operatorname{var}(\beta)|} \\
&= 2 \times |\Sigma|^{|\operatorname{var}(\beta)|} \\
&\leq |\Sigma|^{|\operatorname{var}(\beta)|+1} \\
&\leq |\Sigma|^{|\operatorname{var}(\gamma)|} \\
&= |\mathrm{M}_\gamma|.
\end{aligned}$$

The assumption $|\operatorname{var}(\beta)| \leq |\operatorname{var}(\alpha)|$ leads to $|\Sigma|^{|\operatorname{var}(\alpha)|} + |\Sigma|^{|\operatorname{var}(\beta)|} \leq |\mathrm{M}_\gamma|$ in an analogous way, and therefore $|\mathrm{M}_\alpha| + |\mathrm{M}_\beta| \leq |\mathrm{M}_\gamma|$. However, since $|\alpha| = |\beta|$ implies that $\mathtt{a}^{|\gamma|} \in \mathrm{M}_\alpha \cap \mathrm{M}_\beta$, where $\mathtt{a}$ is a letter in Σ, we can conclude that $|\mathrm{M}_\alpha \cup \mathrm{M}_\beta| < |\mathrm{M}_\alpha| + |\mathrm{M}_\beta| \leq |\mathrm{M}_\gamma|$. This contradicts the fact that $\mathrm{M}_\alpha \cup \mathrm{M}_\beta = \mathrm{M}_\gamma$. Thus $|\operatorname{var}(\alpha)| = |\operatorname{var}(\gamma)|$ or $|\operatorname{var}(\beta)| = |\operatorname{var}(\gamma)|$. Without loss of generality, let the pattern in question be α. Then $|\mathrm{M}_\alpha| = |\mathrm{M}_\gamma|$. Thus $\mathrm{M}_\alpha = \mathrm{M}_\gamma$, and by Lemma 6, $\alpha = \gamma$, which proves the statement. □

Consequently, with the help of Lemma 7, we can now verify the same statement for nonerasing languages as we have for erasing languages.

Theorem 8. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let α and β be terminal-free patterns. There exists a terminal-free pattern γ with $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$ if and only if $L_{\mathrm{NE},\Sigma}(\alpha) \subseteq L_{\mathrm{NE},\Sigma}(\beta)$ or $L_{\mathrm{NE},\Sigma}(\beta) \subseteq L_{\mathrm{NE},\Sigma}(\alpha)$.*

Proof. The *if* direction holds trivially. Suppose that $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$ for some terminal-free patterns α, β, $\gamma \in \mathbb{N}^+$ (in canonical form). Note that $|\alpha|$, $|\beta| \geq |\gamma|$ (otherwise the union will result in words shorter than those in the language of γ). Furthermore if $|\alpha| > |\gamma|$, then $\mathrm{M}_\beta = \mathrm{M}_\gamma$, and, by Lemma 6, $\beta = \gamma$. Therefore $|\alpha| = |\gamma|$ and, similarly, $|\beta| = |\gamma|$. Since we assume that $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$, this implies that $\mathrm{M}_\alpha \cup \mathrm{M}_\beta = \mathrm{M}_\gamma$. Consequently, the statement follows from Lemma 7. □

Note that Theorem 8 extends an equivalent result by Shinohara [20] that holds for alphabets with at least 3 letters.

For unary alphabets, Theorem 8 does not hold. Hence, the situation is equivalent to the erasing case, which is studied in Theorem 4 and Proposition 5 above. This can be verified with the following example of two incomparable, non-erasing pattern languages whose union is again a non-erasing pattern language.

Proposition 9. *Let Σ be an alphabet, $|\Sigma| = 1$, and let $\alpha := x_1 x_2 \cdots x_9$, $\beta := x_1^3 x_2^5$ and $\gamma := x_1 x_2 \cdots x_8$. Then $L_{\mathrm{NE},\Sigma}(\alpha) \not\subseteq L_{\mathrm{NE},\Sigma}(\beta)$, $L_{\mathrm{NE},\Sigma}(\beta) \not\subseteq L_{\mathrm{NE},\Sigma}(\alpha)$, and $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$.*

Proof. Let $\Sigma := \{\mathtt{a}\}$. Firstly, note that $L_{\mathrm{NE},\Sigma}(\alpha) = \{\mathtt{a}^k \mid k \geq 9,\, k \in \mathbb{N}\}$. Similarly, $L_{\mathrm{NE},\Sigma}(\beta) = \{\mathtt{a}^8, \mathtt{a}^{11}, \mathtt{a}^{13}, \mathtt{a}^{14}, \mathtt{a}^{16}, \mathtt{a}^{17}, \mathtt{a}^{18}, \mathtt{a}^{19}, \ldots\}$, which implies that the two languages are incomparable. Furthermore, $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = \{\mathtt{a}^k \mid k \geq 8,\, k \in \mathbb{N}\} = L_{\mathrm{NE},\Sigma}(\gamma)$. $\square$

Thus, summarising the two main results of this section given in Theorems 4 and 8, if the corresponding terminal alphabet contains at least two letters, then the languages of two terminal-free patterns only union together to produce a third in the trivial case.

Corollary 10. *Let $Z,\, Z' \in \{\mathrm{E}, \mathrm{NE}\}$. Let Σ be an alphabet, $|\Sigma| \geq 2$, and let $\alpha,\, \beta,\, \gamma$ be terminal-free patterns. Then $L_{Z,\Sigma}(\alpha) \cup L_{Z,\Sigma}(\beta) = L_{Z',\Sigma}(\gamma)$ if and only if $L_{Z,\Sigma}(\alpha) = L_{Z',\Sigma}(\gamma)$ and $L_{Z,\Sigma}(\beta) \subseteq L_{Z,\Sigma}(\alpha)$ or $L_{Z,\Sigma}(\beta) = L_{Z',\Sigma}(\gamma)$ and $L_{Z,\Sigma}(\alpha) \subseteq L_{Z,\Sigma}(\beta)$.*

It also is worth noting that, for terminal-free patterns, the inclusion problem – and therefore the question of closure under union – is decidable in the E case (see Jiang et al. [10], as explained above), but its decidability status is still open in the NE case.

3.2. Intersection

In the present section, we wish to investigate if the terminal-free NE- or E-pattern languages are closed under intersection. For the NE case, simple counterexamples can be used to prove that the corresponding class of pattern languages is not closed under intersection:

Proposition 11. *Let Σ be an arbitrary alphabet. Let $\alpha := x_1^2$ and let $\beta := x_2^3 x_3^5$. Then $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta)$ is not an NE-pattern language.*

Proof. It is known from Lyndon, Schützenberger [13] that the equation $x_1^2 = x_2^3 x_3^5$ has only periodic solutions. Thus, $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta)$ is the set of all words w^{2k} where $w \in \Sigma^+$, and $2k = 3n + 5m$ for $k, n, m \in \mathbb{N}$. Enumerating the first solutions to $2k = 3n + 5m$, we get $k = 4, 7, 8, \ldots$ Assume to the contrary that the intersection is an NE-pattern language with corresponding pattern γ. Then the shortest word in the $L_{\mathrm{NE},\Sigma}(\gamma)$ has length 8 (i.e., when $k = 4$ and $|w| = 1$). Thus $|\gamma| = 8$. Furthermore, since the second-shortest word has length 14, no variable can occur less than six times in γ. It follows that the only remaining option is that $\gamma = x_1^8$ (up to renaming). This is a contradiction, since no words of length 14 are in $L_{\mathrm{NE},\Sigma}(x_1^8)$. $\square$

We can obtain an equivalent result for the terminal-free E-pattern languages, but our reasoning is significantly more complex and requires the analysis of certain word equations. More precisely, for a restricted class of pairs of patterns, we are able to provide a characterisation of those pairs of pattern languages where the intersection is again a terminal-free E-pattern language, and we show that, for this class, the situation is non-trivial (i.e., there exist both positive and negative examples). We proceed by considering the link between word equations and intersections of pattern-languages.

If, for a word equation $\alpha = \beta$, the words α and β are over disjoint sets of unknowns, then the set of solutions $\sigma : (\mathrm{var}(\alpha) \cup \mathrm{var}(\beta))^* \to \Sigma^*$ corresponds exactly to the set of pairs of morphisms $\tau_1 : \mathrm{var}(\alpha)^* \to \Sigma^*$, $\tau_2 : \mathrm{var}(\beta)^* \to \Sigma^*$ such that $\tau_1(\alpha) = \tau_2(\beta)$. Thus, it also exactly describes the intersection $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$, as illustrated by the proof of Proposition 11 above. Furthermore, such an intersection is invariant under renamings of α and of β, so any intersection of E-pattern languages can be described in this way. We shall next characterise when the intersection of two terminal-free E-pattern languages is again a terminal-free E-pattern language in the restricted case that the corresponding word equation permits only periodic solutions (see Proposition 15). Note that, for α and β over disjoint alphabets, such solutions always exist. Before we can present this characterisation, we need to establish some preliminary results. The first such result is a consequence of the well documented defect effect (see Theorem 1.2.5 in Lothaire [12]).

Lemma 12 (Lothaire [12]). *Let x, y be unknowns and let α, $\beta \in \{x, y\}^+$. Provided $\alpha \neq \beta$, the word equation $\alpha = \beta$ permits only periodic solutions.*

Note that if α and β are identical, then any morphism is a solution. Non-trivial equations are therefore those for which $\alpha \neq \beta$. One immediate consequence of Lemma 12 is that no non-empty word can have two distinct primitive roots. Thus, the primitive root of a periodic morphism will always be the primitive root of its images.

Remark 13. *Let u be a primitive word, and suppose that u^n is a solution-word for some word equation which permits only periodic solutions. Then the corresponding solution σ has u as a primitive root. Furthermore, this means one can replace all occurrences of u in the definition of σ with a single terminal symbol $\mathtt{a}$, and thus $\mathtt{a}^n$ will also be a solution.*

Now, we are able to prove that if the erasing pattern language of a terminal-free pattern γ equals the intersection of two terminal-free erasing pattern languages (where the word equation constructed from the corresponding patterns only permits periodic solutions), then the erasing pattern language of γ is equal to the erasing pattern language of some pattern x_1^k. This result constitutes one half of the desired characterisation and is stated separately, since we shall use it again later.

Proposition 14. *Let Σ be an alphabet, $|\Sigma| \geq 2$. Let α, β be patterns over disjoint sets of variables, and suppose that the word equation $\alpha = \beta$ permits only periodic solutions. Furthermore, suppose that $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is a terminal-free E-pattern language $L_{\mathrm{E},\Sigma}(\gamma)$ for some $\gamma \in X^+$. Then $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{E},\Sigma}(x_1^k)$ for some $k \in \mathbb{N}$.*

Proof. Let δ be the primitive root of γ with $\gamma = \delta^k$. It follows that for some primitive word $u \in \Sigma^+$, the word u^k is a solution-word. By Remark 13, this implies that $\mathtt{a}^k$ is also a solution-word, and thus that $\mathtt{a}^k \in L_{\mathrm{E},\Sigma}(\gamma)$. Consequently, $|\delta|_x = 1$ for some $x \in \mathrm{var}(\gamma)$, and thus $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{E},\Sigma}(x_1^k)$. □

It is easy to see that the number k in Proposition 14 is the length of the shortest non-empty solution-word to the corresponding equation. This, in particular, means that if the word equation $\alpha = \beta$ permits only periodic solutions and $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(x_1^k)$, then it is necessary that every solution-word $u := \mathtt{a}^l$ to the equation satisfies that l is a multiple of k. The next proposition states that this necessary condition is also a sufficient one:

Proposition 15. *Let Σ be an alphabet, $|\Sigma| \geq 2$ and $\mathtt{a} \in \Sigma$. Let α, β be terminal-free patterns over disjoint sets of variables, and suppose that the word equation $\alpha = \beta$ permits only periodic solutions. Let w be the shortest non-empty solution-word. Then $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is a terminal-free E-pattern language if and only if, for every solution-word $u := \mathtt{a}^k$ to the equation, k is a multiple of $|w|$.*

Proof. The *only if* direction holds due to Proposition 14. Let α, $\beta \in X^+$ and suppose that the word equation $\alpha = \beta$ permits only periodic solutions. By Remark 13, there exists a $p \in \mathbb{N}$ such that $\mathtt{a}^p$ is a shortest non-empty solution-word. Clearly, since $\mathtt{a}^p$ is a solution-word, w^p is also a solution-word, for any word $w \in \Sigma^*$. Thus, if there does not exist a solution word $\mathtt{a}^k$, where $k \neq p \times q$ for some $q \in \mathbb{N}_0$, the set of solution words is exactly $\{w^{p \times q} \mid w \in \Sigma^*,\ q \in \mathbb{N}_0\} = L_{\mathrm{E},\Sigma}(x_1^p)$. This proves the *if* direction and the statement. □

We shall now utilise Proposition 15 in the following way. For example patterns α and β for which $\alpha = \beta$ permits only periodic solutions, we show that $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ does not satisfy the conditions of the characterisation of Proposition 15 and therefore $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ cannot be a terminal-free E-pattern language.

Proposition 16. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let $\alpha := x_1 x_2 x_1^2 x_2 x_1^3 x_2^2$ and $\beta := x_3 x_4^2 x_3^2 x_4^6 x_3^3$. Then $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is not a terminal-free E-pattern language.*

Proof. Note that since $\mathrm{var}(\alpha) \cap \mathrm{var}(\beta) = \emptyset$, the set $L := L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is equivalent to the set $\{\sigma(\alpha) \mid \sigma \text{ is a solution to the word equation } \alpha = \beta\}$. Thus, consider the equation

$$\overbrace{x_1 x_2 x_1 x_1 x_2}^{u} \overbrace{x_1 x_1 x_1 x_2 x_2}^{v} = \overbrace{x_3 x_4 x_4 x_3 x_3 x_4 x_4}^{w} \overbrace{x_4 x_4 x_4 x_4 x_3 x_3 x_3}^{x} \,. \tag{1}$$

Since u and v contain the same number of each variable, and likewise for w and x, it is possible to conclude that for any solution σ, $|\sigma(u)| = |\sigma(v)| = |\sigma(w)| = |\sigma(x)|$. Therefore $\sigma(u) = \sigma(w)$ and $\sigma(v) = \sigma(x)$; so the equation is equivalent to the following system of word equations:

$$x_1 x_2 x_1 x_1 x_2 = x_3 x_4 x_4 x_3 x_3 x_4 x_4$$

$$x_1 x_1 x_1 x_2 x_2 = x_4 x_4 x_4 x_4 x_3 x_3 x_3 \,,$$

which, by the substitution $x_5 := x_4 x_4$, is equivalent to the system:

$$x_1 x_2 x_1 x_1 x_2 = x_3 x_5 x_3 x_3 x_5 \tag{2}$$

$$x_1 x_1 x_1 x_2 x_2 = x_5 x_5 x_3 x_3 x_3 \tag{3}$$

$$x_4 x_4 = x_5 \,.$$

Note that by Lemma 12, this substitution does not alter the periodicity of solutions: any solution which is periodic over x_1, x_2, x_3, x_4 must also be periodic over all the variables x_1, x_2, x_3, x_4, and x_5. Similarly, any solution which is periodic over x_1, x_2, x_3 and x_5, will also be periodic over all the variables x_1, x_2, x_3, x_4 and x_5. From the fact that $x_1 x_2 x_1 x_1 x_2$ is a periodicity forcing word (see Culik II, Karhumäki [3]), Equation (2) has only solutions

1. which are periodic over x_1, x_2, x_3, and x_5 (and therefore, also x_4), or
2. such that $\sigma(x_1) = \sigma(x_3)$ and $\sigma(x_2) = \sigma(x_5) = \sigma(x_4 x_4)$.

Clearly, any solution which adheres to the first case corresponds to a periodic solution of Equation (1). Consider a solution which adheres to the second case. By substituting x_1 for x_3 and x_2 for x_5 in Equation (3), we obtain the equation $x_1 x_1 x_1 x_2 x_2 = x_2 x_2 x_1 x_1 x_1$, which is a non-trivial equation in two unknowns. Thus, by Lemma 12, any solution will be periodic over x_1 and x_2. Since $x_1 = x_3$ and $x_2 = x_5$ any solution will also be periodic over x_1, x_2, x_3, x_5 (and therefore also x_4). Consequently, all solutions to Equation (1) are periodic. The shortest solutions are clearly $\mathtt{a}^6$, for $\mathtt{a} \in \Sigma$. However, $\mathtt{a}^8$ is also a solution. Thus, by Proposition 15, the intersection is not a terminal-free E-pattern language. □

It is even possible to give a much stronger statement, showing the extent to which the ‘pattern-language mechanism’ is incapable of handling this seemingly uncomplicated set of solutions.

Corollary 17. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let $\alpha := x_1 x_2 x_1^2 x_2 x_1^3 x_2^2$ and $\beta := x_3 x_4^2 x_3^2 x_4^6 x_3^3$. Then $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is not a finite union of terminal-free E-pattern languages.*

Proof. By extending the proof of Proposition 14, any such union would, without loss of generality, be generated by patterns of the form x_1^k, $k \in \mathbb{N}$. Assume to the contrary that $\{x_1^{k_1}, x_1^{k_2}, \ldots, x_1^{k_n}\}$ is a finite set of patterns whose languages cover $L := L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$. Note that the case that $n = 1$ is covered by Proposition 16, so we may assume $n \geq 2$. Furthermore, for every even k

with $k > 26$, $\mathtt{a}^k \in L$. Also note that every k_i is 6 or larger. Thus, the word $\mathtt{a}^p$ is contained in L, where $p := 2 + (k_1 \times k_2 \times \cdots \times k_n)$. Clearly, p is not a multiple of any k_i, and therefore $\mathtt{a}^p$ is not in any language $L_{\mathrm{E},\Sigma}(x_1^{k_i})$. This is a contradiction and thus proves the statement. $\square$

It is worth noting that the approach above can be used to show that, for $\alpha' := x_1 x_2 x_1^2 x_2^2 x_1^3 x_2^3$ and $\beta' := x_3 x_4^2 x_3^2 x_4^7 x_3^3$, $L_{\mathrm{E},\Sigma}(\alpha') \cap L_{\mathrm{E},\Sigma}(\beta')$ equals $L_{\mathrm{E},\Sigma}(x_1^6)$. This demonstrates that the intersection of two E-pattern languages can in some cases be expressed as an E-pattern language, and therefore that the problem of whether the intersection of two E-pattern languages form an E-pattern language is nontrivial. However it is worth pointing out that a characterisation of this situation is probably very difficult to acquire due to the challenging nature of finding solution-sets of word equations.

3.3. Other closure properties

In this section, we show that regarding the closure under the operations of complementation, morphisms, inverse morphisms, Kleene plus and Kleene star, terminal-free pattern languages behave similarly to the full class of pattern languages.

Proposition 18. *Let Σ be an arbitrary alphabet. For every terminal-free pattern α, $\overline{L_{\mathrm{E},\Sigma}(\alpha)}$ is not a terminal-free* E*-pattern language and $\overline{L_{\mathrm{NE},\Sigma}(\alpha)}$ is not a terminal-free* NE*-pattern language.*

Proof. Every terminal-free E-pattern language contains ε and every terminal-free NE-pattern language does not contain ε. Since, for every language L, either L or $\overline{L}$ must contain ε, it is not possible that L and $\overline{L}$ are both terminal-free E-pattern languages or both terminal-free NE-pattern languages. $\square$

Proposition 18 does not only prove the non-closure of terminal-free E- and NE-pattern languages under complementation, but also characterises in a trivial way the terminal-free pattern languages whose complement is also a terminal-free pattern language.

We now investigate closure under morphisms and inverse morphisms. If we consider the class of terminal-free pattern languages with respect to a fixed alphabet Σ, then the non-closure under morphisms and inverse morphisms is obvious, since these operations may introduce new symbols that are not in Σ and therefore a morphic image or inverse morphic image of a terminal-free pattern language over Σ may not be a terminal-free pattern language over Σ anymore (simply because it is not a language over Σ). Therefore, we investigate the closure of terminal-free pattern languages with respect to a fixed alphabet Σ under morphisms and inverse morphisms that are defined over the same fixed alphabet Σ.

Proposition 19. *Let Σ be an alphabet, $|\Sigma| \geq 2$. Then the class of terminal-free* NE*-pattern languages over Σ and the class of terminal-free* E*-pattern languages over Σ are not closed under morphisms $h : \Sigma^* \to \Sigma^*$ or inverse morphisms $h^{-1} : \Sigma^* \to 2^{\Sigma^*}$.*

Proof. Let $\Sigma = \{a_1, a_2, \dots, a_k\}$ with $k \geq 2$ and let $h : \Sigma^* \to \Sigma^*$ be defined by $h(b) = a_1 a_2 \dots a_k$ for every $b \in \Sigma$. Obviously, $h(L_{\mathrm{NE},\Sigma}(x)) = \{(a_1 a_2 \dots a_k)^n \mid n \geq 1\}$ and $h(L_{\mathrm{E},\Sigma}(x)) = \{(a_1 a_2 \dots a_k)^n \mid n \geq 0\}$, which are both no pattern languages (in a non-singleton pattern language every possible factor occurs in at least one word, which is not the case for these languages). This proves the non-closure of the class of terminal-free NE- and the class of terminal-free E-pattern languages over Σ under morphisms $h : \Sigma^* \to \Sigma^*$.

Let $g : \Sigma^* \to \Sigma^*$ be defined by $g(b) = a_1$ for every $b \in \Sigma$. Then $L' := g^{-1}(L_{\mathrm{NE},\Sigma}(xx)) = \{w \mid w \in \Sigma^+, |w| \text{ is even}\}$, which is not a terminal-free NE-pattern language, since the shortest words in L' have length 2, so the hypothetical pattern that describes L' is either xx or xy, but xx cannot produce $a_1 a_2 \in L'$ and xy can produce $a_1 a_1 a_1 \notin L'$. Furthermore, $L'' := g^{-1}(L_{\mathrm{E},\Sigma}(xx)) = \{w \mid w \in \Sigma^*, |w| \text{ is even}\}$ is not a terminal-free E-pattern language, which can be seen as follows. If the hypothetical pattern α that describes L'' has a single-occurrence variable, then it can produce $a_1 \notin L''$ and if it has no single-occurrence variable, then it cannot produce $a_1 a_2 \in L''$. This proves the non-closure of the class of terminal-free NE- and the class of terminal-free E-pattern languages over Σ under inverse morphisms $h^{-1} : \Sigma^* \to 2^{\Sigma^*}$. □

We wish to point out that in the proof of Proposition 19, it is shown independently from the actual alphabet Σ that the morphic images and inverse morphic images of our pattern languages are no pattern languages over Σ and therefore they are no pattern languages over any alphabet. Consequently, Proposition 19 also proves the non-closure of the class of all pattern languages (over any alphabet) under morphisms and inverse morphisms.

With respect to the closure of terminal-free pattern languages under Kleene plus and Kleene star, we observe a dependency on the alphabet size, i. e., terminal-free pattern languages over alphabets that are at least binary are not closed under Kleene plus or star, whereas unary terminal-free pattern languages are. We first prove the negative closure property with respect to non-unary alphabets.

Proposition 20. *Let Σ be an alphabet, $|\Sigma| \geq 2$. The terminal-free* NE- *and* E-*pattern languages over Σ are not closed under Kleene plus and Kleene star.*

Proof. We assume that Σ contains the two distinct symbols $\mathtt{a}$ and $\mathtt{b}$. The language $L' := (L_{\mathrm{NE},\Sigma}(xx))^+$ is not a terminal-free NE-pattern language, which can be shown as follows. Since the shortest words in L' are of length 2, the hypothetical pattern that describes L' is either xx or xy, but xx cannot generate $\mathtt{aabb} \in L'$ and xy can generate $\mathtt{ab} \notin L'$. Moreover, no language L^* is an NE-pattern language, since $\varepsilon \in L^*$.

We claim that $L'' := (L_{\mathrm{E},\Sigma}(xx))^*$ is not a terminal-free E-pattern language over Σ. In order to prove this claim, we assume that α describes L'' and first note that every variable in α has at least two occurrences, since otherwise $L_{\mathrm{E},\Sigma}(\alpha) = \Sigma^* \neq L''$. Moreover, let $m := |\alpha|$ be even; the case that m is odd can be handled analogously. We now consider the word $w :=$

$\mathtt{a}^2\mathtt{b}^4\mathtt{a}^6\mathtt{b}^8\cdots\mathtt{a}^{2(m+1)}\mathtt{b}^{2(m+2)}\mathtt{a}^{2(m+3)}\mathtt{b}^{2(m+4)}$. Since $w \in L''$, there is a morphism h with $h(\alpha) = w$, and since w has $m+2$ factors of form $\mathtt{abb}\cdots\mathtt{ba}$ or $\mathtt{baa}\cdots\mathtt{ab}$ and α has only m occurrences of variables, there must be one variable $x \in \mathrm{var}(\alpha)$ such that $h(x)$ contains one of these factors. This is a contradiction since x is repeated in α, but every factor of this form has only one occurrence in w. Moreover, in the same way we can show that $(L_{\mathrm{E},\Sigma}(xx))^+$ is not a terminal-free E-pattern language. This proves that terminal-free NE- and E-pattern languages are not closed under Kleene plus and Kleene star. □

In contrast to Proposition 20, unary terminal-free E-pattern languages are closed under Kleene plus and Kleene star and unary terminal-free NE-pattern languages are closed under Kleene plus (note that unary terminal-free NE-pattern languages are obviously not closed under Kleene star, since the Kleene star closure of a language always contains ε). The reason for this is that in the unary case the Kleene plus or star does not change a terminal-free E-pattern language and the Kleene plus does not change a terminal-free NE-pattern language.

Proposition 21. *Let Σ be an alphabet, $|\Sigma| = 1$, and let α be a terminal-free pattern. Then $(L_{\mathrm{E},\Sigma}(\alpha))^+ = (L_{\mathrm{E},\Sigma}(\alpha))^* = L_{\mathrm{E},\Sigma}(\alpha)$ and $(L_{\mathrm{NE},\Sigma}(\alpha))^+ = L_{\mathrm{NE},\Sigma}(\alpha)$.*

Proof. Let $\Sigma := \{\mathtt{a}\}$. Since $(L_{\mathrm{E},\Sigma}(\alpha))^+ \cup \{\varepsilon\} = (L_{\mathrm{E},\Sigma}(\alpha))^*$ and $\varepsilon \in (L_{\mathrm{E},\Sigma}(\alpha))^+$, $(L_{\mathrm{E},\Sigma}(\alpha))^+ = (L_{\mathrm{E},\Sigma}(\alpha))^*$ follows. Obviously, $L_{\mathrm{E},\Sigma}(\alpha) \subseteq (L_{\mathrm{E},\Sigma}(\alpha))^*$, so we have to show the converse of this statement.

Without loss of generality, we can assume that $L_{\mathrm{E},\Sigma}(\alpha) = L_{\mathrm{E},\Sigma}(\beta)$ for some terminal-free pattern $\beta = x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$, $k \in \mathbb{N}$, $n_i \in \mathbb{N}$, $1 \leq i \leq k$. Now let $w, u \in L_{\mathrm{E},\Sigma}(\beta)$. This implies that $w = \mathtt{a}^{n_1 \times \ell_1} \mathtt{a}^{n_2 \times \ell_2} \cdots \mathtt{a}^{n_k \times \ell_k}$, for some $\ell_i \in \mathbb{N}_0$, $1 \leq i \leq k$, and $u = \mathtt{a}^{n_1 \times \ell'_1} \mathtt{a}^{n_2 \times \ell'_2} \cdots \mathtt{a}^{n_k \times \ell'_k}$, for some $\ell'_i \in \mathbb{N}_0$, $1 \leq i \leq k$. Hence, $w \cdot u = \mathtt{a}^{n_1 \times (\ell'_1 + \ell_1)} \mathtt{a}^{n_2 \times (\ell'_2 + \ell_2)} \cdots \mathtt{a}^{n_k \times (\ell'_k + \ell_k)} \in L_{\mathrm{E},\Sigma}(\beta)$. By induction, this implies $(L_{\mathrm{E},\Sigma}(\beta))^* \subseteq L_{\mathrm{E},\Sigma}(\beta)$.

The statement $(L_{\mathrm{NE},\Sigma}(\alpha))^+ = L_{\mathrm{NE},\Sigma}(\alpha)$ can be shown analogously. □

4. General Patterns

As explained in Section 1 and formally stated in Theorem 1, the closure properties of the full classes of NE-pattern languages and of E-pattern languages are understood. In the present section, we therefore wish to expand the more specific insights into the terminal-free pattern languages gained in Section 3 to the full classes. More precisely, with respect to the operations of complementation, intersection and union, we investigate those patterns that exhibit the property that their complement, intersection or union is again a pattern language and we try to characterise these patterns. Our strongest results are with respect to the operation of union.

4.1. Complement

With respect to the full class of E- and NE-pattern language, an analogue of Proposition 18 exists:

Proposition 22 (Bayer [2]). *Let Σ be an alphabet, $|\Sigma| \geq 2$. For every pattern α, $\overline{L_{\mathrm{E},\Sigma}(\alpha)}$ is not an* E*-pattern language and $\overline{L_{\mathrm{NE},\Sigma}(\alpha)}$ is not an* NE*-pattern language.*

With respect to Proposition 22, it is interesting to note that $\overline{L_{\mathrm{NE},\Sigma}(\alpha)}$ is not an NE-pattern language even for unary alphabets Σ (since every language or its complement contains ε), while this is not the case for erasing pattern languages:

Proposition 23. *Let $\Sigma = \{\mathtt{a}\}$. Then $\overline{L_{\mathrm{E},\Sigma}(x_1 x_1 \mathtt{a})} = L_{\mathrm{E},\Sigma}(x_1 x_1)$.*

In the same way as Proposition 18 does for terminal-free patterns, Proposition 22 yields a trivial characterisation of pattern languages with a complement that again is a pattern language.

4.2. Intersection

It is straightforward to construct patterns α and β such that $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is not an E-pattern language or $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta)$ is not an NE-pattern language. Furthermore, any two terminal-free patterns α and β are an example for the situation that $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is not an NE-pattern language and, as long as there are at least two symbols in Σ, also for the situation that $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta)$ is not an E-pattern language.

The patterns $\alpha = x_1 x_1$ and $\beta = x_1 x_1 x_1$ constitute examples of patterns for which the intersection of their pattern languages is again a pattern language, i.e., $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(x_1^6)$ and $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(x_1^6)$. Moreover, there are non-trivial examples of patterns α, β and γ, such that $L_{\mathrm{NE},\Sigma}(\alpha) \cap L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$:

- $L_{\mathrm{NE},\Sigma}(\mathtt{a}x) \cap L_{\mathrm{NE},\Sigma}(xx) = L_{\mathrm{E},\Sigma}(\mathtt{a}x\mathtt{a}x)$.
- $L_{\mathrm{NE},\Sigma}(x\mathtt{a}y) \cap L_{\mathrm{NE},\Sigma}(xxx) = L_{\mathrm{E},\Sigma}(x\mathtt{a}yx\mathtt{a}yx\mathtt{a}y)$.
- $L_{\mathrm{NE},\Sigma}(\mathtt{a}x\mathtt{a}) \cap L_{\mathrm{NE},\Sigma}(xx) = L_{\mathrm{E},\Sigma}(\mathtt{a}x\mathtt{a}\mathtt{a}x\mathtt{a})$.
- $L_{\mathrm{NE},\Sigma}(\mathtt{a}x\mathtt{a}x) \cap L_{\mathrm{NE},\Sigma}(x\mathtt{b}x\mathtt{b}) = L_{\mathrm{E},\Sigma}(\mathtt{a}x\mathtt{b}\mathtt{a}x\mathtt{b})$.
- $L_{\mathrm{NE},\Sigma}(\mathtt{a}xy) \cap L_{\mathrm{NE},\Sigma}(xxx) = L_{\mathrm{E},\Sigma}(\mathtt{a}x\mathtt{a}x\mathtt{a}x)$.

However, it is not known whether or not there are patterns α and β, such that $L_{\mathrm{E},\Sigma}(\alpha) \cap L_{\mathrm{E},\Sigma}(\beta)$ is an NE-pattern language. Moreover, we do not have any characterisations for the situation that the intersection of two pattern languages is again a pattern language.

4.3. Union

Examples of patterns α and β such that $L_{Z,\Sigma}(\alpha) \cup L_{Z,\Sigma}(\beta)$ is not a Z'-pattern language, for all $Z, Z' \in \{\mathrm{E}, \mathrm{NE}\}$, are provided by Proposition 2.

Corollary 10 is our strongest result in Section 3, as it shows that the union of terminal-free pattern languages can only be a terminal-free pattern language if one of the languages is contained in the other. At first glance it seems a reasonable hypothesis that a similar result might hold for the full class of pattern languages, but in the present section we show that this is not true.

For all but the union of pairs of E-pattern languages and the question of whether they can form an E-pattern language, suitable examples are not too hard to find:

Proposition 24.

- $L_{\mathrm{E},\{\mathtt{a},\mathtt{b}\}}(\mathtt{aa}x) \cup L_{\mathrm{E},\{\mathtt{a},\mathtt{b}\}}(\mathtt{ab}x) = L_{\mathrm{NE},\{\mathtt{a},\mathtt{b}\}}(\mathtt{a}x)$.
- $L_{\mathrm{NE},\{\mathtt{a},\mathtt{b},\mathtt{c}\}}(\mathtt{abc}) \cup L_{\mathrm{NE},\{\mathtt{a},\mathtt{b},\mathtt{c}\}}(\mathtt{a}x\mathtt{b}x\mathtt{c}x) = L_{\mathrm{E},\{\mathtt{a},\mathtt{b},\mathtt{c}\}}(\mathtt{a}x\mathtt{b}x\mathtt{c}x)$.
- $L_{\mathrm{NE},\{\mathtt{a},\mathtt{b}\}}(\mathtt{a}x_1) \cup L_{\mathrm{NE},\{\mathtt{a},\mathtt{b}\}}(\mathtt{b}x_1) = L_{\mathrm{NE},\{\mathtt{a},\mathtt{b}\}}(x_1x_2)$.

In the following, we study the question of whether the existence of a pattern γ satisfying $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$ implies that there is an inclusion relation between $L_{\mathrm{E},\Sigma}(\alpha)$ and $L_{\mathrm{E},\Sigma}(\beta)$. We present increasingly complex counterexamples for alphabet sizes 1, 2, 3, and 4. These examples are individually tailored for the size of the alphabet and they do not generalise to different alphabet sizes.

With respect to unary alphabets, we note that a suitable example is already provided by Proposition 23, i. e., for $\Sigma = \{\mathtt{a}\}$, $\alpha := x_1x_1\mathtt{a}$, $\beta := x_1x_1$ and $\gamma := x_1$, we have neither $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ nor $L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\alpha)$, but $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$. We now give such examples for alphabets of size 2, 3 and 4.

Proposition 25. *Let* $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, $\alpha := x_1\mathtt{a}x_2\mathtt{b}x_2\mathtt{a}x_3$, $\beta := x_1\mathtt{a}x_2\mathtt{bb}x_2\mathtt{a}x_3$ *and* $\gamma := x_1\mathtt{a}x_2\mathtt{b}x_3\mathtt{a}x_4$. *Then* $L_{\mathrm{E},\Sigma}(\alpha) \nsubseteq L_{\mathrm{E},\Sigma}(\beta)$, $L_{\mathrm{E},\Sigma}(\beta) \nsubseteq L_{\mathrm{E},\Sigma}(\alpha)$ *and* $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.

Proof. Obviously, $\mathtt{aba} \in L_{\mathrm{E},\Sigma}(\alpha) \setminus L_{\mathrm{E},\Sigma}(\beta)$ and $\mathtt{abba} \in L_{\mathrm{E},\Sigma}(\beta) \setminus L_{\mathrm{E},\Sigma}(\alpha)$; thus, $L_{\mathrm{E},\Sigma}(\alpha) \nsubseteq L_{\mathrm{E},\Sigma}(\beta)$ and $L_{\mathrm{E},\Sigma}(\beta) \nsubseteq L_{\mathrm{E},\Sigma}(\alpha)$. It remains to prove $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.

To this end, we first observe that there exist substitutions h and g for γ with $h(\gamma) = \alpha$ and $g(\gamma) = \beta$, i. e., $h(x_1) := x_1$, $h(x_2) := h(x_3) := x_2$, $h(x_4) := x_3$ and $g(x_1) := x_1$, $g(x_2) := x_2$, $g(x_3) := \mathtt{b}x_2$, $g(x_4) := x_3$. This directly implies that $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$ and $L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$. Hence, $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$.

Let h be some substitution for γ and let $w := h(\gamma)$. This means that, for some (possibly empty) words $p, q, r, s \in \Sigma^*$, $w = p\,\mathtt{a}\,q\,\mathtt{b}\,r\,\mathtt{a}\,s$. Obviously, $\mathtt{a}\,q\,\mathtt{b}\,r\,\mathtt{a}$ contains a factor $\mathtt{ab}^n\mathtt{a}$, with $n \geq 1$. If n is even, then $n = 2m$ for some $m \geq 1$

and $w = u\,\mathtt{a}\,\mathtt{b}^{m-1}\,\mathtt{bb}\,\mathtt{b}^{m-1}\,\mathtt{a}\,v$ for some $u, v \in \Sigma^*$. This means that w can be obtained from β by substituting x_1 by u, x_3 by v and x_2 by $\mathtt{b}^{m-1}$. If, on the other hand, n is odd, then $n = 2m + 1$ for some $m \geq 0$ and $w = u\,\mathtt{a}\,\mathtt{b}^m\,\mathtt{b}\,\mathtt{b}^m\,\mathtt{a}\,v$ for some $u, v \in \Sigma^*$. This means that w can be obtained from α by substituting x_1 by u, x_3 by v and x_2 by $\mathtt{b}^m$. Consequently, w is in $L_{\mathrm{E},\Sigma}(\alpha)$ or $L_{\mathrm{E},\Sigma}(\beta)$ and therefore $L_{\mathrm{E},\Sigma}(\gamma) \subseteq L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$, which concludes the proof. □

In the proof of the next two propositions, we make use of a notation that allows us to refer to specific factors in a word w over an alphabet A: If w contains $n \geq 1$ occurrences of a factor v then for every i, $1 \leq i \leq n$, $v\langle i\rangle$ is the ith occurrence (from the left) of v in w. Then the factor $[w/v\langle i\rangle]$ is the prefix of w up to (but not including) the leftmost letter of $v\langle i\rangle$ and the factor $[v\langle i\rangle\backslash w]$ is the suffix of w beginning with the first letter that is to the right of $v\langle i\rangle$. Moreover, for every word w that contains at least i occurrences of a factor u, j occurrences of factor v and that satisfies $w = w_1\,u\langle i\rangle\,w_2\,v\langle j\rangle\,w_3$ with $w_1, w_2, w_3 \in A^*$, we use $[u\langle i\rangle\backslash w/v\langle j\rangle]$ as an abbreviation for $[u\langle i\rangle\backslash[w/v\langle j\rangle]]$.

Proposition 26. *Let* $\Sigma := \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$,

$$\begin{aligned}
\alpha &:= x_1\mathtt{a}x_2x_3^6x_4^3x_5^6x_6\mathtt{b}x_7\mathtt{a}x_2x_8^{12}x_4^6x_9^{12}x_6\mathtt{b}x_{10},\\
\beta &:= x_1\mathtt{a}x_2x_3^6x_4^2x_5^5x_6^6x_7\mathtt{b}x_8\mathtt{a}x_2x_9^{12}x_4^4x_5^{10}x_{10}^{12}x_7\mathtt{b}x_{11} \text{ and}\\
\gamma &:= x_1\mathtt{a}x_2x_3^6x_4^2x_5^3x_6^6x_7\mathtt{b}x_8\mathtt{a}x_2x_9^{12}x_4^4x_5^6x_{10}^{12}x_7\mathtt{b}x_{11}.
\end{aligned}$$

Then $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$, $L_{\mathrm{E},\Sigma}(\beta) \not\subseteq L_{\mathrm{E},\Sigma}(\alpha)$ *and* $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.

Proof. The word $\mathtt{ac}^3\mathtt{bac}^6\mathtt{b}$ is obviously contained in $L_{\mathrm{E},\Sigma}(\alpha)$, as it is generated by a substitution that maps x_4 to the letter $\mathtt{c}$ and all other variables to the empty word. On the other hand, it is not included in $L_{\mathrm{E},\Sigma}(\beta)$, since the factor $\mathtt{c}^6$ would need to be generated by the variables x_2, x_4 or x_7 in β, and this would necessarily lead to a factor of either $\mathtt{c}^4$ or $\mathtt{c}^6$ between the first occurrence of $\mathtt{a}$ and $\mathtt{b}$ in the generated word. Hence, $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$.

A very similar reasoning shows that $\mathtt{ac}^2\mathtt{bac}^4\mathtt{b} \in L_{\mathrm{E},\Sigma}(\beta) \setminus L_{\mathrm{E},\Sigma}(\alpha)$, which directly implies $L_{\mathrm{E},\Sigma}(\beta) \not\subseteq L_{\mathrm{E},\Sigma}(\alpha)$.

We now explain why indeed $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$, i. e., we shall prove

1. $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$,
2. $L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$ and
3. $L_{\mathrm{E},\Sigma}(\gamma) \subseteq L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$.

Regarding point 1 it is sufficient to observe that there exist a terminal-preserving morphism ϕ satisfying $\phi(\gamma) = \phi(\alpha)$, which is known to be a sufficient condition for $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$ (see Jiang et al. [9]). This morphism can be defined by $\phi(x_i) := x_i$ for $1 \leq i \leq 3$, $\phi(x_4) := \varepsilon$, and $\phi(x_i) := x_{i-1}$ for $5 \leq i \leq 11$.

Point 2 is less straightforward to prove. Intuitively, it makes use of the fact that, whenever a substitution σ maps x_4 or x_5 in β to a word that contains $\mathtt{a}$ or

$\mathtt{b}$, then this substitution is ambiguous with respect to β; more precisely, there exists a substitution σ' that maps x_4 or x_5 to the empty word and nevertheless satisfies $\sigma'(\beta) = \sigma(\beta)$. In other words, the variables x_4 and x_5 are only required in β when they are mapped to the letter $\mathtt{c}$. This, in turn, allows, for every substitution σ, a substitution τ to be defined that satisfies $\tau(\gamma) = \sigma(\beta)$. Formally, τ is given as follows:

<u>Case 1</u> $\sigma(x_4x_5) \in \{\mathtt{c}\}^*$:

Define
$$\begin{aligned} \tau(x_4) &:= \mathtt{c}^{|\sigma(x_4)|+|\sigma(x_5)|}, \\ \tau(x_5) &:= \mathtt{c}^{|\sigma(x_5)|}, \\ \tau(x) &:= \sigma(x),\ x \in \mathrm{var}(\gamma) \setminus \{x_4, x_5\}. \end{aligned}$$

<u>Case 2</u> $\sigma(x_4)$ contains $k \geq 1$ occurrences of $\mathtt{a}$:

Define
$$\begin{aligned} \tau(x_1) &:= \sigma(x_1\mathtt{a}x_2x_3^6)[\sigma(x_4^2)/\mathtt{a}\langle 2k\rangle], \\ \tau(x_8) &:= \sigma(x_8\mathtt{a}x_2x_9^{12})[\sigma(x_4^4)/\mathtt{a}\langle 4k\rangle], \\ \tau(x_2) &:= [\mathtt{a}\langle k\rangle\backslash\sigma(x_4)], \\ \tau(x) &:= \sigma(x_5),\ x \in \{x_4, x_5\}, \\ \tau(x) &:= \varepsilon,\ x \in \{x_3, x_9\}, \\ \tau(x) &:= \sigma(x),\ x \in \{x_6, x_7, x_{10}, x_{11}\}. \end{aligned}$$

<u>Case 3</u> $\sigma(x_4)$ does not contain $\mathtt{a}$, and $\sigma(x_5)$ contains $k \geq 1$ occurrences of $\mathtt{a}$:

Define
$$\begin{aligned} \tau(x_1) &:= \sigma(x_1\mathtt{a}x_2x_3^6x_4^2)[\sigma(x_5^5)/\mathtt{a}\langle 5k\rangle], \\ \tau(x_8) &:= \sigma(x_8\mathtt{a}x_2x_9^{12}x_4^4)[\sigma(x_5^{10})/\mathtt{a}\langle 10k\rangle], \\ \tau(x_2) &:= [\mathtt{a}\langle k\rangle\backslash\sigma(x_5)], \\ \tau(x) &:= \varepsilon,\ x \in \{x_3, x_4, x_5, x_9\}, \\ \tau(x) &:= \sigma(x),\ x \in \{x_6, x_7, x_{10}, x_{11}\}. \end{aligned}$$

<u>Case 4</u> $\sigma(x_4x_5)$ does not contain $\mathtt{a}$, and $\sigma(x_4)$ contains $k \geq 1$ occurrences of $\mathtt{b}$:

Define
$$\begin{aligned} \tau(x_8) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^2x_5^5x_6^6x_7\mathtt{b}x_8)], \\ \tau(x_{11}) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^4x_5^{10}x_{10}^{12}x_7\mathtt{b}x_{11})], \\ \tau(x_7) &:= [\sigma(x_4)/\mathtt{b}\langle 1\rangle], \\ \tau(x) &:= \varepsilon,\ x \in \{x_4, x_5, x_6, x_{10}\}, \\ \tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_9\}. \end{aligned}$$

<u>Case 5</u> $\sigma(x_4x_5)$ does not contain $\mathtt{a}$, $\sigma(x_4)$ does not contain $\mathtt{b}$, and $\sigma(x_5)$ contains $k \geq 1$ occurrences of $\mathtt{b}$:

Define
$$\begin{aligned} \tau(x_8) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_5^5x_6^6x_7\mathtt{b}x_8)], \\ \tau(x_{11}) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_5^{10}x_{10}^{12}x_7\mathtt{b}x_{11})], \\ \tau(x_7) &:= [\sigma(x_5)/\mathtt{b}\langle 1\rangle], \\ \tau(x) &:= \varepsilon,\ x \in \{x_5, x_6, x_{10}\}, \\ \tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_9\}. \end{aligned}$$

In each of these cases, the actual verification of $\tau(\gamma) = \sigma(\beta)$ is straightforward.

In order to conclude our proof of Proposition 26, we now address point 3. Our reasoning is very similar to the one above, i. e., we construct, for any substitution σ, a substitution τ that satisfies $\tau(\alpha) = \sigma(\gamma)$ or $\tau(\beta) = \sigma(\gamma)$. This is again possible since most substitutions are ambiguous with respect to γ.

We consider the following cases:

Case 1 $\sigma(x_4) = \varepsilon$ and $\sigma(x_5) \in \{\mathtt{c}\}^*$:

The following substitution satisfies $\tau(\alpha) = \sigma(\gamma)$.

Define
$$\begin{aligned} \tau(x_i) &:= \sigma(x_i),\ i \in \{1, 2, 3\}, \\ \tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3\}. \end{aligned}$$

Case 2 $\sigma(x_4) \neq \varepsilon$ and $\sigma(x_4 x_5) \in \{\mathtt{c}\}^+$:

In this case, let $p, q \in \mathbb{N}_0$ be chosen such that $p2 + q5 = |\sigma(x_4^2 x_5^3)|$; it can be easily seen that such numbers always exist. Then, for the following substitution τ, $\tau(\beta) = \sigma(\gamma)$ holds true.

Define
$$\begin{aligned} \tau(x_4) &:= \mathtt{c}^p, \\ \tau(x_5) &:= \mathtt{c}^q, \\ \tau(x) &:= \sigma(x),\ x \in \mathrm{var}(\gamma) \setminus \{x_4, x_5\}. \end{aligned}$$

Case 3 $\sigma(x_4)$ contains $k \geq 1$ occurrences of $\mathtt{a}$:

Here, we can again employ α to generate $\sigma(\gamma)$. This is achieved by the following substitution.

Define
$$\begin{aligned} \tau(x_1) &:= \sigma(x_1 \mathtt{a} x_2 x_3^6)[\sigma(x_4^2)/\mathtt{a}\langle 2k\rangle], \\ \tau(x_7) &:= \sigma(x_8 \mathtt{a} x_2 x_9^{12})[\sigma(x_4^4)/\mathtt{a}\langle 4k\rangle], \\ \tau(x_2) &:= [\mathtt{a}\langle k\rangle\backslash\sigma(x_4)], \\ \tau(x) &:= \varepsilon,\ x \in \{x_3, x_8\}, \\ \tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3, x_7, x_8\}. \end{aligned}$$

Case 4 $\sigma(x_4)$ does not contain $\mathtt{a}$, and $\sigma(x_5)$ contains $k \geq 1$ occurrences of $\mathtt{a}$:

We define a substitution τ that satisfies $\tau(\beta) = \sigma(\gamma)$. Note that, alternatively, we could give a substitution for α, since $\sigma(\gamma)$ is also contained in $L_{\mathrm{E},\Sigma}(\alpha)$.

Define
$$\begin{aligned} \tau(x_1) &:= \sigma(x_1 \mathtt{a} x_2 x_3^6 x_4^2)[\sigma(x_5^3)/\mathtt{a}\langle 3k\rangle], \\ \tau(x_8) &:= \sigma(x_8 \mathtt{a} x_2 x_9^{12} x_4^4)[\sigma(x_5^6)/\mathtt{a}\langle 6k\rangle], \\ \tau(x_2) &:= [\mathtt{a}\langle k\rangle\backslash\sigma(x_5)], \\ \tau(x) &:= \varepsilon,\ x \in \{x_3, x_4, x_5, x_9\}, \\ \tau(x) &:= \sigma(x),\ x \in \{x_6, x_7, x_{10}, x_{11}\}. \end{aligned}$$

Case 5 $\sigma(x_4 x_5)$ does not contain $\mathtt{a}$, and $\sigma(x_4)$ contains $k \geq 1$ occurrences of $\mathtt{b}$:

This case is similar to Case 4, since $\sigma(\gamma)$ can be generated by both α and β. We again give a substitution τ that yields $\tau(\beta) = \sigma(\gamma)$.

Define $\quad \tau(x_8) := [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^2 x_5^3 x_6^6 x_7 \mathtt{b} x_8)]$,

$$\begin{aligned}
\tau(x_{11}) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^4 x_5^6 x_{10}^{12} x_7 \mathtt{b} x_{11})],\\
\tau(x_7) &:= [\sigma(x_4)/\mathtt{b}\langle 1\rangle],\\
\tau(x) &:= \varepsilon,\ x \in \{x_4, x_5, x_6, x_{10}\},\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_9\}.
\end{aligned}$$

Case 6 $\sigma(x_4 x_5)$ does not contain $\mathtt{a}$, $\sigma(x_4)$ does not contain $\mathtt{b}$, and $\sigma(x_5)$ contains $k \geq 1$ occurrences of $\mathtt{b}$:

In this final case, we must define a substitution for β.

Define $\quad \tau(x_8) := [\mathtt{b}\langle 1\rangle\backslash\sigma(x_5^3 x_6^6 x_7 \mathtt{b} x_8)]$,

$$\begin{aligned}
\tau(x_{11}) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_5^6 x_{10}^{12} x_7 \mathtt{b} x_{11})],\\
\tau(x_7) &:= [\sigma(x_5)/\mathtt{b}\langle 1\rangle],\\
\tau(x) &:= \varepsilon,\ x \in \{x_5, x_6, x_{10}\},\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_9\}.
\end{aligned}$$

Summarising these six cases, we can state that, for every word $w \in L_{\mathrm{E},\Sigma}(\gamma)$, there exists a substitution τ such that $\tau(\alpha) = w$ or $\tau(\beta) = w$. Hence, $L_{\mathrm{E},\Sigma}(\gamma) \subseteq L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$, and this concludes our proof. □

Proposition 27. *Let* $\Sigma := \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$,

$$\begin{aligned}
\alpha \;:=\;& x_1 \mathtt{a} x_2 x_3^2 x_4^2 x_5^2 x_6 \mathtt{b} x_7 \mathtt{a} x_2 x_8^2 x_4^2 x_9^2 x_6 \mathtt{b}\\
& x_{10} \mathtt{c} x_{11} x_{12}^2 x_{13}^2 x_{14}^2 x_{15}^2 x_{16} \mathtt{d} x_{17} \mathtt{c} x_{11} x_{18}^2 x_{13}^2 x_{14}^2 x_{19}^2 x_{16} \mathtt{d}\\
& x_{20} x_{13}^2 x_{14}^2 x_{13}^2 x_{14}^2 x_{13}^2 x_{14}^2 x_{21} x_4^6,\\
\beta \;:=\;& x_1 \mathtt{a} x_2 x_3^2 x_4^2 x_5^2 x_6^2 x_7 \mathtt{b} x_8 \mathtt{a} x_2 x_9^2 x_4^2 x_5^2 x_{10}^2 x_7 \mathtt{b}\\
& x_{11} \mathtt{c} x_{12} x_{13}^2 x_{14}^2 x_{15}^2 x_{16} \mathtt{d} x_{17} \mathtt{c} x_{12} x_{18}^2 x_{14}^2 x_{19}^2 x_{16} \mathtt{d}\\
& x_{20} x_{14}^6 x_{21} x_4^2 x_5^2 x_4^2 x_5^2 x_4^2 x_5^2 \text{ and}\\
\gamma \;:=\;& x_1 \mathtt{a} x_2 x_3^2 x_4^2 x_5^2 x_6^2 x_7 \mathtt{b} x_8 \mathtt{a} x_2 x_9^2 x_4^2 x_5^2 x_{10}^2 x_7 \mathtt{b}\\
& x_{11} \mathtt{c} x_{12} x_{13}^2 x_{14}^2 x_{15}^2 x_{16}^2 x_{17} \mathtt{d} x_{18} \mathtt{c} x_{12} x_{19}^2 x_{14}^2 x_{15}^2 x_{20}^2 x_{17} \mathtt{d}\\
& x_{21} x_{14}^2 x_{15}^2 x_{14}^2 x_{15}^2 x_{14}^2 x_{15}^2 x_{22} x_4^2 x_5^2 x_4^2 x_5^2 x_4^2 x_5^2.
\end{aligned}$$

Then $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$, $L_{\mathrm{E},\Sigma}(\beta) \not\subseteq L_{\mathrm{E},\Sigma}(\alpha)$ *and* $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.

Proof. In order to illustrate the structure of the patterns α, β and γ, we wish to point out that there exist simple terminal-preserving morphisms ϕ and ψ satisfying $\phi(\gamma) = \alpha$ and $\psi(\gamma) = \beta$. The morphism ϕ is defined by $\phi(x_i) := x_i$ for $i \in \{1, 2, 3, 4\}$, $\phi(x_5) := \varepsilon$, and $\phi(x_i) := x_{i-1}$ for all other variables in γ. In other words, α can be generated from γ by deleting the variable x_5 (and converting the resulting pattern into canonical form). Similarly, γ can be turned into β by deleting the variable x_{15}. Hence, formally, the morphism ψ is given by $\psi(x_i) := x_i$ for $i \in \{1, 2, \ldots, 14\}$, $\psi(x_{15}) := \varepsilon$, and $\psi(x_i) := x_{i-1}$ for the other variables in γ. According to Jiang et al. [9], the existence of ϕ and ψ implies that $L_{\mathrm{E},\Sigma}(\alpha)$ and $L_{\mathrm{E},\Sigma}(\beta)$ are sublanguages of $L_{\mathrm{E},\Sigma}(\gamma)$. Hence, $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\gamma)$.

The next step of our proof establishes that $L_{\mathrm{E},\Sigma}(\gamma)$ is included in $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$. Thus, we now prove that, for every substitution σ, there exists a substitution τ that satisfies $\tau(\alpha) = \sigma(\gamma)$ or $\tau(\beta) = \sigma(\gamma)$. Intuitively, this is possible since a substitution is necessarily ambiguous with respect to γ if it maps both x_5 and x_{15} in γ to a nonempty word.

Formally, we consider the nine cases listed below. Cases 1 – 4 lead to a substitution τ such that $\tau(\alpha) = \sigma(\gamma)$, and in Cases 5 – 9 the substitution τ satisfies $\tau(\beta) = \sigma(\gamma)$. Note that in Case 9 we could alternatively give a substitution that would show $\sigma(\gamma) \in L_{\mathrm{E},\Sigma}(\alpha)$.

<u>Case 1</u> $\sigma(x_4 x_5)$ contains the letter $\mathtt{a}$:

Define
$$\begin{aligned}
\tau(x_1) &:= \sigma(x_1 \mathtt{a} x_2 x_3^2)[\sigma(x_4^2 x_5^2)/\mathtt{a}\langle 1\rangle],\\
\tau(x_7) &:= \sigma(x_8 \mathtt{a} x_2 x_9^2)[\sigma(x_4^2 x_5^2)/\mathtt{a}\langle 1\rangle],\\
\tau(x_2) &:= [\mathtt{a}\langle 1\rangle\backslash\sigma(x_4^2 x_5^2)],\\
\tau(x_{21}) &:= \sigma(x_{22} x_4^2 x_5^2 x_4^2 x_5^2 x_4^2 x_5^2),\\
\tau(x) &:= \varepsilon,\ x \in \{x_3, x_4, x_8\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3, x_4, x_7, x_8, x_{21}\}.
\end{aligned}$$

<u>Case 2</u> $\sigma(x_4 x_5)$ does not contain the letter $\mathtt{a}$, but it contains the letter $\mathtt{b}$:

Define
$$\begin{aligned}
\tau(x_7) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^2 x_5^2 x_6^2 x_7 \mathtt{b} x_8)],\\
\tau(x_{10}) &:= [\mathtt{b}\langle 1\rangle\backslash\sigma(x_4^2 x_5^2 x_{10}^2 x_7 \mathtt{b} x_{11})],\\
\tau(x_6) &:= [\sigma(x_4^2 x_5^2)/\mathtt{b}\langle 1\rangle],\\
\tau(x_{21}) &:= \sigma(x_{22} x_4^2 x_5^2 x_4^2 x_5^2 x_4^2 x_5^2),\\
\tau(x) &:= \varepsilon,\ x \in \{x_4, x_5, x_9\},\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_9, x_{10}, x_{21}\}.
\end{aligned}$$

<u>Case 3</u> $\sigma(x_4 x_5) \in \{\mathtt{c}\}^*$:

Define
$$\begin{aligned}
\tau(x_4) &:= \sigma(x_4 x_5),\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3, x_4\}.
\end{aligned}$$

<u>Case 4</u> $\sigma(x_4 x_5) \in \{\mathtt{d}\}^+$:

Define
$$\begin{aligned}
\tau(x_4) &:= \sigma(x_4 x_5),\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \mathrm{var}(\alpha) \setminus \{x_1, x_2, x_3, x_4\}.
\end{aligned}$$

<u>Case 5</u> $\sigma(x_4 x_5) \in \{\mathtt{c}, \mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$ and $\sigma(x_{14} x_{15})$ contains the letter $\mathtt{c}$:

$$\begin{aligned}
\text{Define}\quad \tau(x_{11}) &:= \sigma(x_{11}\mathtt{c}x_{12}x_{13}^2)[\sigma(x_{14}^2x_{15}^2)/\mathtt{c}\langle 1\rangle],\\
\tau(x_{17}) &:= \sigma(x_{18}\mathtt{c}x_{12}x_{19}^2)[\sigma(x_{14}^2x_{15}^2)/\mathtt{c}\langle 1\rangle],\\
\tau(x_{12}) &:= [\mathtt{c}\langle 1\rangle\backslash\sigma(x_{14}^2x_{15}^2)],\\
\tau(x_{20}) &:= \sigma(x_{21}x_{14}^2x_{15}^2x_{14}^2x_{15}^2x_{14}^2x_{15}^2),\\
\tau(x) &:= \varepsilon,\ x \in \{x_{13}, x_{14}, x_{18}\},\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \{x_{15}, x_{16}, x_{19}, x_{21}\}.
\end{aligned}$$

Case 6 $\sigma(x_4x_5) \in \{\mathtt{c},\mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$, and $\sigma(x_{14}x_{15})$ does not contain the letter $\mathtt{c}$, but the letter $\mathtt{d}$:

$$\begin{aligned}
\text{Define}\quad \tau(x_{17}) &:= [\mathtt{d}\langle 1\rangle\backslash\sigma(x_{14}^2x_{15}^2x_{16}^2x_{17}\mathtt{d}x_{18})],\\
\tau(x_{20}) &:= [\mathtt{d}\langle 1\rangle\backslash\sigma(x_{14}^2x_{15}^2x_{20}^2x_{17}\mathtt{b}x_{21}x_{14}^2x_{15}^2x_{14}^2x_{15}^2x_{14}^2x_{15}^2)],\\
\tau(x_{16}) &:= [\sigma(x_{14}^2x_{15}^2)/\mathtt{d}\langle 1\rangle],\\
\tau(x) &:= \varepsilon,\ x \in \{x_{14}, x_{15}, x_{19}\},\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \{x_{18}, x_{21}\}.
\end{aligned}$$

Case 7 $\sigma(x_4x_5) \in \{\mathtt{c},\mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$, and $\sigma(x_{14}x_{15}) \in \{\mathtt{a}\}^*$:

$$\begin{aligned}
\text{Define}\quad \tau(x_{14}) &:= \sigma(x_{14}x_{15}),\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \{x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}\}.
\end{aligned}$$

Case 8 $\sigma(x_4x_5) \in \{\mathtt{c},\mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$, and $\sigma(x_{14}x_{15}) \in \{\mathtt{b}\}^+$:

$$\begin{aligned}
\text{Define}\quad \tau(x_{14}) &:= \sigma(x_{14}x_{15}),\\
\tau(x) &:= \sigma(x),\ x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}\},\\
\tau(x_i) &:= \sigma(x_{i+1}),\ x_i \in \{x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}\}.
\end{aligned}$$

Case 9 $\sigma(x_4x_5) \in \{\mathtt{c},\mathtt{d}\}^+ \setminus (\{\mathtt{c}\}^+ \cup \{\mathtt{d}\}^+)$, and $\sigma(x_{14}x_{15}) \in \{\mathtt{a},\mathtt{b}\}^+ \setminus (\{\mathtt{a}\}^+ \cup \{\mathtt{b}\}^+)$:

This implies that $\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)$, which is a factor of $\sigma(\gamma)$, contains two occurrences of the factor $\mathtt{ab}$ followed by two occurrences of the factor $\mathtt{cd}$.

$$\begin{aligned}
\text{Define}\quad \tau(x_1) &:= \sigma([\gamma/x_{21}\langle 1\rangle]x_{21})[\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)/\mathtt{ab}\langle 1\rangle],\\
\tau(x_8) &:= [\mathtt{ab}\langle 1\rangle\backslash\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)/\mathtt{ab}\langle 2\rangle],\\
\tau(x_{11}) &:= [\mathtt{ab}\langle 2\rangle\backslash\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)/\mathtt{cd}\langle 1\rangle],\\
\tau(x_{17}) &:= [\mathtt{cd}\langle 1\rangle\backslash\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)/\mathtt{cd}\langle 2\rangle],\\
\tau(x_{20}) &:= [\mathtt{cd}\langle 2\rangle\backslash\sigma((x_{14}^2x_{15}^2)^3x_{22}(x_4^2x_5^2)^3)],\\
\tau(x) &:= \varepsilon,\ x \in \mathrm{var}(\beta) \setminus \{x_1, x_8, x_{11}, x_{17}, x_{20}\}.
\end{aligned}$$

Hence, for every $w \in L_{\mathrm{E},\Sigma}(\gamma)$, we can give a substitution τ that yields $w \in L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$. Referring to the first paragraph of this proof, this implies that $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$.

We conclude this proof by showing that $L_{\mathrm{E},\Sigma}(\alpha)$ and $L_{\mathrm{E},\Sigma}(\beta)$ are incomparable, and we start with $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$. To this end, we consider the following substitution:

$$\sigma_1(x_i) := \begin{cases} \mathtt{b} & , \quad i \in \{12, 15\}\,, \\ \mathtt{ab} & , \quad i = 13\,, \\ \mathtt{ab}^2 & , \quad i = 14\,, \\ \mathtt{a} & , \quad i \in \{18, 19\}\,, \\ \varepsilon & , \quad \text{otherwise}\,. \end{cases}$$

Thus,

$$\sigma_1(\alpha) = \mathtt{abab}\,\underline{\mathtt{c}}\,\mathtt{b}^2\,(\mathtt{ab})^2\,(\mathtt{ab}^2)^2\,\mathtt{b}^2\,\underline{\mathtt{d}}\,\underline{\mathtt{c}}\,\mathtt{a}^2\,(\mathtt{ab})^2\,(\mathtt{ab}^2)^2\,\mathtt{a}^2\,\underline{\mathtt{d}}\,((\mathtt{ab})^2(\mathtt{ab}^2)^2)^3.$$

Note that this word has only two occurrences of the letter $\mathtt{c}$ and two occurrences of the letter $\mathtt{d}$. These are underlined in the above representation of the word, for ease of reference.

If we now assume to the contrary that there exists a substitution τ_1 satisfying $\tau_1(\beta) = \sigma_1(\alpha)$, then we can immediately observe that $\tau_1(x_{14}) = \varepsilon$, since $\sigma_1(\alpha)$ does not contain a factor v^6 to the right of the rightmost occurrence of $\mathtt{d}$, while β has an occurrence of x_{14}^6 to the right of the second occurrence of $\mathtt{d}$. Similarly, $\tau_1(x_{12}) = \tau_1(x_{16}) = \varepsilon$, since immediately to the right of the first occurrence of $\mathtt{c}$ and immediately to the left of the first occurrence of $\mathtt{d}$ in $\sigma_1(\alpha)$ we have the letter $\mathtt{b}$, while $\mathtt{a}$ is the letter to the right of the second occurrence of $\mathtt{c}$ and to the left of the second occurrence of $\mathtt{d}$. Thus, $\tau_1(x_{13}^2 x_{15}^2)$ needs to equal $w_1 := \mathtt{b}^2\,(\mathtt{ab})^2\,(\mathtt{ab}^2)^2\,\mathtt{b}^2$. This is a contradiction, because w_1 does not consist of two consecutive squares. Thus, $\sigma_1(\alpha) \notin L_{\mathrm{E},\Sigma}(\beta)$, and therefore $L_{\mathrm{E},\Sigma}(\alpha) \not\subseteq L_{\mathrm{E},\Sigma}(\beta)$.

The final part of our proof, demonstrating that $L_{\mathrm{E},\Sigma}(\beta) \not\subseteq L_{\mathrm{E},\Sigma}(\alpha)$, is very similar to our reasoning in the previous paragraph. We now make use of the substitution σ_2 given by

$$\sigma_2(x_i) := \begin{cases} \mathtt{d} & , \quad i \in \{3, 6\}\,, \\ \mathtt{cd} & , \quad i = 4\,, \\ \mathtt{cd}^2 & , \quad i = 5\,, \\ \mathtt{c} & , \quad i \in \{9, 10\}\,, \\ \varepsilon & , \quad \text{otherwise}\,, \end{cases}$$

and we consider

$$\sigma_2(\beta) = \underline{\mathtt{a}}\,\mathtt{d}^2\,(\mathtt{cd})^2\,(\mathtt{cd}^2)^2\,\mathtt{d}^2\,\underline{\mathtt{b}}\,\underline{\mathtt{a}}\,\mathtt{c}^2\,(\mathtt{cd})^2\,(\mathtt{cd}^2)^2\,\mathtt{c}^2\,\underline{\mathtt{b}}\,\mathtt{cdcd}\,((\mathtt{cd})^2(\mathtt{cd}^2)^2)^3,$$

where the sole occurrences of $\mathtt{a}$ and of $\mathtt{b}$ are underlined.

Using a similar argument as above, we can see that if there exists a substitution τ_2 satisfying $\tau_2(\alpha) = \sigma_2(\beta)$, then $\tau_2(x_2) = \tau_2(x_4) = \tau_2(x_6) = \varepsilon$. Hence, $\tau_2(x_3^2 x_5^2)$ should equal $\mathtt{c}^2\,(\mathtt{cd})^2\,(\mathtt{cd}^2)^2\,\mathtt{c}^2$, but this is not possible. Consequently, $\sigma_2(\beta) \notin L_{\mathrm{E},\Sigma}(\alpha)$, and therefore $L_{\mathrm{E},\Sigma}(\beta)$ is not a subset of $L_{\mathrm{E},\Sigma}(\alpha)$. $\square$

We are not able to give equivalent examples for larger alphabets, and we expect the question of their existence to be a complex and important problem. This is because the above examples depend on the ambiguity of terminal-preserving morphisms, which is a phenomenon that underpins many properties of pattern languages. Similar constructions to those in Propositions 25, 26, and 27 have been used to disprove longstanding conjectures on inductive inference (see Reidenbach [15, 17]) and the equivalence problem (see Reidenbach [16]) for E-pattern languages over alphabets of up to 4 letters and, similarly, it has so far not been possible to expand those techniques to arbitrary alphabets. Our examples, thus, suggest a close link between the problem in the current section and the two most important open problems for E-pattern languages over alphabets with at least 5 letters, and we expect that substantial progress on any one of them will require combinatorial insights that will allow the others to be solved as well.

For all $Z, Z' \in \{\mathrm{E}, \mathrm{NE}\}$, we have seen example patterns α and β such that $L_{Z,\Sigma}(\alpha) \cup L_{Z,\Sigma}(\beta)$ is a Z'-pattern language. We shall now try to generalise these examples in order to obtain characterisations of such pairs of patterns.

For the case $Z = Z' = \mathrm{E}$, we are only able to state a necessary condition for $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$ that, unfortunately, is not very strong:

Theorem 28. *Let Σ be an arbitrary alphabet, and let α, β and γ be patterns with $L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$. Furthermore, let w_α, w_β and w_γ be the residuals of α, β and γ, respectively. Then $w_\gamma = w_\alpha$ and w_γ is a subsequence of w_β or $w_\gamma = w_\beta$ and w_γ is a subsequence of w_α.*

Proof. We first note that $|w_\alpha| \geq |w_\gamma|$ and $|w_\beta| \geq |w_\gamma|$, since otherwise $w_\alpha \notin L_{\mathrm{E},\Sigma}(\gamma)$ or $w_\beta \notin L_{\mathrm{E},\Sigma}(\gamma)$, respectively. If $|w_\alpha| > |w_\gamma|$ and $|w_\beta| > |w_\gamma|$, then $w_\gamma \notin L_{\mathrm{E},\Sigma}(\alpha) \cup L_{\mathrm{E},\Sigma}(\beta)$, which is a contradiction and, thus, $|w_\alpha| = |w_\gamma|$ or $|w_\beta| = |w_\gamma|$ holds. If $|w_\alpha| = |w_\gamma|$, then $w_\alpha = w_\gamma$, since otherwise $w_\alpha \notin L_{\mathrm{E},\Sigma}(\gamma)$. Furthermore, since $w_\beta \in L_{\mathrm{E},\Sigma}(\gamma)$, w_γ is a subsequence of w_β. Analogously, from the assumption that $|w_\beta| = |w_\gamma|$ we can conclude that $w_\beta = w_\gamma$ and that w_γ is a subsequence of w_α. □

In view of the fact that the examples of Propositions 25, 26 and 27 are rather complicated, we expect that a full characterisation for the case $Z = Z' = \mathrm{E}$ is difficult to obtain.

For the case $Z = Z' = \mathrm{NE}$, we can present a strong necessary condition that, similarly to Theorem 8, strengthens a result by Shinohara [20]:

Theorem 29. *Let Σ be an alphabet with $\{\mathtt{a}, \mathtt{b}\} \subseteq \Sigma$ and let α, β and γ be patterns. If $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$, then one of the following three statements is true:*

- $L_{\mathrm{NE},\Sigma}(\alpha) \subseteq L_{\mathrm{NE},\Sigma}(\beta)$ *and* $\beta = \gamma$.
- $L_{\mathrm{NE},\Sigma}(\beta) \subseteq L_{\mathrm{NE},\Sigma}(\alpha)$ *and* $\alpha = \gamma$.

- $|\Sigma| = 2$ *and*

$$\alpha = \delta_0\, \mathtt{a}\, \delta_1\, \mathtt{a}\, \delta_2 \dots \delta_{m-1}\, \mathtt{a}\, \delta_m\,,$$
$$\beta = \delta_0\, \mathtt{b}\, \delta_1\, \mathtt{b}\, \delta_2 \dots \delta_{m-1}\, \mathtt{b}\, \delta_m\,,$$
$$\gamma = \delta_0\, x\, \delta_1\, x\, \delta_2 \dots \delta_{m-1}\, x\, \delta_m\,,$$

where $m \geq 1$*,* $\delta_i \in (X \cup \Sigma)^*$*,* $0 \leq i \leq m$*.*

Proof. In this proof, we use the following notation: in order to refer to the symbol at a certain position j, $1 \leq j \leq n$, in a word $w = a_1 \cdot a_2 \cdot \dots \cdot a_n$, $a_i \in \Sigma$, $1 \leq i \leq n$, we use $w[j] := a_j$.

We assume that $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{NE},\Sigma}(\gamma)$, but neither $L_{\mathrm{NE},\Sigma}(\alpha) \subseteq L_{\mathrm{NE},\Sigma}(\beta)$ and $\beta = \gamma$ nor $L_{\mathrm{NE},\Sigma}(\beta) \subseteq L_{\mathrm{NE},\Sigma}(\alpha)$ and $\alpha = \gamma$ holds.

If $|\alpha| < |\beta|$, then $|\gamma| = |\alpha|$, since $|\gamma| < |\alpha|$ or $|\gamma| > |\alpha|$ means that the shortest words of $L_{\mathrm{NE},\Sigma}(\gamma)$ cannot be obtained from α or the shortest words of $L_{\mathrm{NE},\Sigma}(\alpha)$ cannot be obtained from γ, respectively. Obviously, $\alpha = \gamma$ implies $L_{\mathrm{NE},\Sigma}(\beta) \subseteq L_{\mathrm{NE},\Sigma}(\alpha)$; thus, we can assume that $\alpha \neq \gamma$, which is a contradiction, since this means that the sets of shortest words of $L_{\mathrm{NE},\Sigma}(\alpha)$ and $L_{\mathrm{NE},\Sigma}(\gamma)$ are different. In the same way, the assumption $|\beta| < |\alpha|$ leads to a contradiction and therefore we can conclude that $|\alpha| = |\beta| = |\gamma|$.

We call α and β *equivalent with respect to the occurrences of terminal symbols* if, for every i, $1 \leq i \leq |\alpha|$, $\alpha[i] \in \Sigma$ or $\beta[i] \in \Sigma$ implies $\alpha[i] = \beta[i]$.

Claim 1: The patterns α and β are not equivalent with respect to the occurrences of terminal symbols.

Proof of Claim 1: We assume that α and β are equivalent with respect to the occurrences of terminal symbols, i. e., for every i, $1 \leq i \leq |\alpha|$, if $\alpha[i] \in \Sigma$ or $\beta[i] \in \Sigma$, then $\alpha[i] = \beta[i]$. We note that if

- for all i, j, $1 \leq i < j \leq |\alpha|$, $\alpha[i] = \alpha[j]$ implies $\beta[i] = \beta[j]$ or,
- for all i, j, $1 \leq i < j \leq |\beta|$, $\beta[i] = \beta[j]$ implies $\alpha[i] = \alpha[j]$,

then α can be mapped to β by a nonerasing substitution or β can be mapped to α by a nonerasing substitution, respectively, which implies $L_{\mathrm{NE},\Sigma}(\beta) \subseteq L_{\mathrm{NE},\Sigma}(\alpha)$ or $L_{\mathrm{NE},\Sigma}(\alpha) \subseteq L_{\mathrm{NE},\Sigma}(\beta)$, respectively. Hence, we can assume that this is not the case, which implies that there must be positions i, j, i', j', $1 \leq i, j, i', j' \leq |\alpha|$, such that $\alpha[i] = \alpha[j]$, $\beta[i] \neq \beta[j]$, $\alpha[i'] \neq \alpha[j']$ and $\beta[i'] = \beta[j']$. This means that $\gamma[i], \gamma[j], \gamma[i'], \gamma[j'] \in \mathrm{var}(\gamma)$ with $\gamma[i] \neq \gamma[j]$ and $\gamma[i'] \neq \gamma[j']$; thus, we can obtain a shortest word v from γ with $v[i] \neq v[j]$ and $v[i'] \neq v[j']$. Since all shortest words $v' \in L_{\mathrm{NE},\Sigma}(\alpha)$ satisfy $v'[i] = v'[j]$ and all shortest words $v'' \in L_{\mathrm{NE},\Sigma}(\beta)$ satisfy $v''[i'] = v'[j']$, $v \notin L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$, which is a contradiction. Consequently, α and β are not equivalent with respect to the occurrences of terminal symbols. $\square$(Claim 1)

If α and β are not equivalent with respect to the occurrences of terminal symbols, then there exists a position i, $1 \leq i \leq |\alpha|$, such that

- $\alpha[i] \in \Sigma$ and $\beta[i] \in \text{var}(\beta)$,
- $\alpha[i] \in \text{var}(\alpha)$ and $\beta[i] \in \Sigma$ or
- $\alpha[i], \beta[i] \in \Sigma$ and $\alpha[i] \neq \beta[i]$.

Next, we show that there does not exist an i, $1 \leq i \leq |\alpha|$, such that $\alpha[i] \in \Sigma$ and $\beta[i] \in \text{var}(\beta)$ or $\beta[i] \in \Sigma$ and $\alpha[i] \in \text{var}(\alpha)$.

Claim 2: For every i, $1 \leq i \leq |\alpha|$, $\alpha[i] \in \Sigma$ if and only if $\beta[i] \in \Sigma$.

Proof of Claim 2: We assume to the contrary that for some position p, $1 \leq p \leq |\alpha|$, $\alpha[p] = \mathtt{a} \in \Sigma$ and $\beta[p] = x \in \text{var}(\beta)$.

We assume that there exists a position q, $1 \leq q \leq |\beta|$, of β such that $\beta[q] \in \Sigma$. If $\alpha[q] \in \text{var}(\alpha)$, then $\gamma[p], \gamma[q] \in \text{var}(\gamma)$ with $\gamma[p] \neq \gamma[q]$, which implies that in $L_{\text{NE},\Sigma}(\gamma)$ there is a shortest word v with $v[p] \neq \alpha[p]$ and $v[q] \neq \beta[q]$. This is a contradiction, since v is neither in $L_{\text{NE},\Sigma}(\alpha)$ nor in $L_{\text{NE},\Sigma}(\beta)$. If, on the other hand, $\alpha[q] \in \Sigma$ and $\alpha[q] \neq \beta[q]$, then we obtain a contradiction in the same way. Hence, for every i, $1 \leq i \leq |\beta|$, if $\beta[i] \in \Sigma$, then $\alpha[i] = \beta[i]$.

We recall that $\beta[p] = x$ and we define $p_1, p_2, \ldots, p_m$ to be exactly the positions that satisfy $\beta[p_i] = x$, $1 \leq i \leq m$. If there exists an i, $1 \leq i \leq m$, with $\alpha[p_i] \neq \mathtt{a}$, then $\gamma[p] \neq \gamma[p_i]$ and $\gamma[p], \gamma[p_i] \in \text{var}(\gamma)$. Therefore, we can obtain a shortest word v from γ with $v[p] \neq v[p_i]$ and $v[p] \neq \mathtt{a}$, which is neither in $L_{\text{NE},\Sigma}(\alpha)$ nor $L_{\text{NE},\Sigma}(\beta)$. Consequently, $\alpha[p_i] = \mathtt{a}$, $1 \leq i \leq m$.

Now we assume that there exist positions q, q', $1 \leq q < q' \leq |\beta|$, such that $\beta[q], \beta[q'] \in \text{var}(\beta)$, $\beta[q] = \beta[q']$ and $\alpha[q] \neq \alpha[q']$. As demonstrated above, $\beta[q] = \beta[q'] = x$ implies $\alpha[q] = \alpha[q'] = \mathtt{a}$; thus, we can conclude that $\beta[q] \neq x$ and $\beta[q'] \neq x$. Therefore $\gamma[q], \gamma[q'], \gamma[p] \in \text{var}(\gamma)$, $\gamma[q] \neq \gamma[q']$, since $\alpha[q] \neq \alpha[q']$, and $\gamma[p] \notin \{\gamma[q], \gamma[q']\}$, since in $L_{\text{NE},\Sigma}(\beta)$ there is a shortest word v with $v[p] = \mathtt{b}$ and $v[q] = v[q'] = \mathtt{a}$. This implies that we can obtain a shortest word v from γ by substituting $\gamma[q]$ and $\gamma[q']$ by different symbols and $\gamma[p]$ by a symbol different from $\mathtt{a}$. This is a contradiction, since v is neither in $L_{\text{NE},\Sigma}(\alpha)$ nor in $L_{\text{NE},\Sigma}(\beta)$.

Consequently, for all i, j, $1 \leq i < j \leq |\beta|$, $\beta[i], \beta[j] \in \text{var}(\beta)$ and $\beta[i] = \beta[j]$ implies $\alpha[i] = \alpha[j]$, and, as demonstrated above, for every i, $1 \leq i \leq |\beta|$, $\beta[i] \in \Sigma$ implies $\beta[i] = \alpha[i]$. This means that there exists a nonerasing substitution that maps β to α and, thus, $L_{\text{NE},\Sigma}(\alpha) \subseteq L_{\text{NE},\Sigma}(\beta)$, a contradiction.

Analogously, we can show that if there exists an i, $1 \leq i \leq |\beta|$, such that $\beta[i] \in \Sigma$ and $\alpha[i] \in \text{var}(\alpha)$, then $L_{\text{NE},\Sigma}(\beta) \subseteq L_{\text{NE},\Sigma}(\alpha)$. Hence, for every i, $1 \leq i \leq |\alpha|$, $\alpha[i] \in \Sigma$ if and only if $\beta[i] \in \Sigma$. □(Claim 2)

The two claims from above show that α and β are not equivalent with respect to the occurrences of terminal symbols and, for every i, $1 \leq i \leq |\alpha|$, $\alpha[i] \in \Sigma$ if and only if $\beta[i] \in \Sigma$. Thus, there is at least one position r, $1 \leq r \leq |\alpha|$, with $\alpha[r] = \mathtt{a}$ and $\beta[r] = \mathtt{b}$. This implies that $\gamma[r] \in \text{var}(\gamma)$ and if $|\Sigma| \geq 3$, then we can substitute $\gamma[r]$ by a symbol that is different from both $\mathtt{a}$ and $\mathtt{b}$, which is a contradiction. Thus, $|\Sigma| = 2$. If there exists a position r' with $\alpha[r'] = \mathtt{b}$ and $\beta[r'] = \mathtt{a}$, then $\gamma[r], \gamma[r'] \in \text{var}(\gamma)$ and, thus, we can obtain a shortest word v from γ with $v[r] = v[r'] = \mathtt{a}$, which is neither in $L_{\text{NE},\Sigma}(\alpha)$

nor in $L_{\mathrm{NE},\Sigma}(\beta)$. So there are some positions $r_1, r_2, \dots, r_m$ with $m \geq 1$ and $1 \leq r_i \leq |\alpha|$, $1 \leq i \leq m$, such that $\alpha[r_i] = \mathtt{a}$ and $\beta[r_i] = \mathtt{b}$ and for all other positions $j \in \{1, 2, \dots, |\alpha|\} \setminus \{r_1, r_2, \dots, r_m\}$, $\alpha[j] \in \Sigma$ or $\beta[j] \in \Sigma$ implies $\alpha[j] = \beta[j]$.

In order to conclude the proof, we show that these positions r_i, $1 \leq i \leq m$, are in fact the only positions where α and β differ. To this end, we assume that there are positions i, j, $1 \leq i < j \leq |\alpha|$, with $\alpha[i], \alpha[j] \in \mathrm{var}(\alpha)$ (which particularly means that $\beta[i], \beta[j] \in \mathrm{var}(\beta)$), $\alpha[i] = \alpha[j]$ and $\beta[i] \neq \beta[j]$. We note that this implies $\gamma[i], \gamma[j] \in \mathrm{var}(\gamma)$ with $\gamma[i] \neq \gamma[j]$ and $\gamma[r_1] \notin \{\gamma[i], \gamma[j]\}$ since otherwise we cannot obtain shortest words v from γ with $v[r_1] \notin \{v[i], v[j]\}$. Thus, there is a shortest word v in $L_{\mathrm{NE},\Sigma}(\gamma)$ with $v[i] \neq v[j]$ and $v[r_1] = \mathtt{a}$. This is a contradiction, since v is neither in $L_{\mathrm{NE},\Sigma}(\alpha)$ nor in $L_{\mathrm{NE},\Sigma}(\beta)$. The assumption that there are positions i, j, $1 \leq i < j \leq |\alpha|$, with $\beta[i], \beta[j] \in \mathrm{var}(\beta)$, $\beta[i] = \beta[j]$ and $\alpha[i] \neq \alpha[j]$ leads to a contradiction in the same way. Hence, such positions do not exist, which means that α and β are identical up to the position $r_1, r_2, \dots, r_m$. More precisely,

$$\alpha = \delta_0 \, \mathtt{a} \, \delta_1 \, \mathtt{a} \, \delta_2 \dots \delta_{m-1} \, \mathtt{a} \, \delta_m \,,$$
$$\beta = \delta_0 \, \mathtt{b} \, \delta_1 \, \mathtt{b} \, \delta_2 \dots \delta_{m-1} \, \mathtt{b} \, \delta_m \,,$$

where $\delta_i \in (X \cup \Sigma)^*$, $0 \leq i \leq m$. Furthermore, this implies that

$$\gamma = \delta_0 \, x \, \delta_1 \, x \, \delta_2 \dots \delta_{m-1} \, x \, \delta_m \,,$$

where $x \notin \bigcup_{i=0}^{m} \mathrm{var}(\delta_i)$. □

It remains to consider the cases $Z = \mathrm{NE}$, $Z' = \mathrm{E}$ and $Z = \mathrm{E}$, $Z' = \mathrm{NE}$, for which we have full characterisations. Before we prove these characterisations, we recall that Jiang et al. show in [9] that, for every pattern α, we can construct finite sets of patterns Γ and Δ such that $L_{\mathrm{E},\Sigma}(\alpha) = \bigcup_{\beta \in \Gamma} L_{\mathrm{NE},\Sigma}(\beta)$ and $L_{\mathrm{NE},\Sigma}(\alpha) = \bigcup_{\beta \in \Delta} L_{\mathrm{E},\Sigma}(\beta)$. More precisely, Γ is the set of all patterns that can be obtained from α by erasing some (possibly none) of the variables and Δ contains all pattern that can be obtained from α by substituting each $x \in \mathrm{var}(\alpha)$ by bx, for some $b \in \Sigma$. We note that the examples $L_{\mathrm{NE},\Sigma}(\mathtt{abc}) \cup L_{\mathrm{NE},\Sigma}(\mathtt{a}x\mathtt{b}x\mathtt{c}x) = L_{\mathrm{E},\Sigma}(\mathtt{a}x\mathtt{b}x\mathtt{c}x)$ and $L_{\mathrm{E},\{\mathtt{a},\mathtt{b}\}}(\mathtt{aa}x) \cup L_{\mathrm{E},\{\mathtt{a},\mathtt{b}\}}(\mathtt{ab}x) = L_{\mathrm{NE},\{\mathtt{a},\mathtt{b}\}}(\mathtt{a}x)$ of Proposition 24 are applications of exactly this construction.

The characterisation for the case $Z = \mathrm{NE}$, $Z' = \mathrm{E}$ follows from the fact that we can prove that if we restrict ourselves to unions of only two pattern languages, then $L_{\mathrm{E},\Sigma}(\alpha) = \bigcup_{\beta \in \Gamma} L_{\mathrm{NE},\Sigma}(\beta)$ is the only possible way to describe an E-pattern language by NE-pattern languages.

Theorem 30. *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let α, β and γ be patterns. Then $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$ if and only if $\alpha \in \Sigma^+$ and $\beta = \gamma = u_1 \, x^{j_1} \, u_2 \, x^{j_2} \dots x^{j_m} \, u_{m+1}$, $j_i \in \mathbb{N}_0$, $1 \leq i \leq m$, such that $u_1 \, u_2 \dots u_{m+1} = \alpha$.*

Proof. The *if* direction follows trivially from the observation that if γ is a one-variable pattern with residual w, then $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(\gamma) \cup \{w\}$.

In order to prove the *only if* direction, let α, β and γ be patterns with $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta) = L_{\mathrm{E},\Sigma}(\gamma)$. If γ is terminal-free, then $\varepsilon \in L_{\mathrm{E},\Sigma}(\gamma)$, which is a contradiction, since $\varepsilon \notin L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$. Hence, we can assume that γ has a non-empty residual w. Since w is the unique shortest word in $L_{\mathrm{E},\Sigma}(\gamma)$, it is also the unique shortest word in $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$, which implies that α or β must equal w. This is due to the fact that if α or β is shorter than w, then in $L_{\mathrm{NE},\Sigma}(\alpha)$ or $L_{\mathrm{NE},\Sigma}(\beta)$, respectively, there exists a word shorter than w, if both α and β are longer than w, then $w \notin L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$, and if $|\alpha| = |w|$ or $|\beta| = |w|$ with $\alpha \neq w$ or $\beta \neq w$, respectively, then, since $|\Sigma| \geq 2$, in $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$ there exists a word different from w, but with the same length. Without loss of generality, we assume that $\alpha = w$. If $|\beta| = |\alpha| = |w|$, then, as explained above, $\beta = w$, which implies $\alpha = \beta = \gamma = w$ and, thus, the condition of the theorem is satisfied. In the following, we consider the case $|\alpha| < |\beta|$.

We note that in $L_{\mathrm{NE},\Sigma}(\alpha) \cup L_{\mathrm{NE},\Sigma}(\beta)$, there is no word v with $|w| < |v| < |\beta|$, since $L_{\mathrm{NE},\Sigma}(\alpha) = \{w\}$ and every word in $L_{\mathrm{NE},\Sigma}(\beta)$ has a length of at least $|\beta|$. Thus, for every $x \in \mathrm{var}(\gamma)$, $|\gamma|_x \geq |\beta| - |w|$, since if there exists an $x \in \mathrm{var}(\gamma)$ with $|\gamma|_x < |\beta| - |w|$, then erasing all variables in γ except x yields a pattern γ' with $|w| < |\gamma'| < |\beta|$, from which we could derive a word v with $|w| < |v| < |\beta|$. On the other hand, there exists at least one variable $x \in \mathrm{var}(\gamma)$ with $|\gamma|_x = |\beta| - |w|$, since otherwise γ cannot produce the shortest words in $L_{\mathrm{NE},\Sigma}(\beta)$. Consequently, all the shortest words in $L_{\mathrm{NE},\Sigma}(\beta)$ can be obtained from γ by erasing all variables except some variable x with $|\gamma|_x = |\beta| - |w|$, which instead is substituted by a single symbol. We emphasise here that the choice of the variable x with $|\gamma|_x = |\beta| - |w|$ that is not erased depends on the actual shortest word to be produced. More precisely, for every shortest word v in $L_{\mathrm{NE},\Sigma}(\beta)$, there exists a $\gamma' = u_1 x u_2 x \ldots x u_{k+1}$, where $w = u_1 u_2 \ldots u_{k+1}$, $k = |\beta| - |w|$ and v can be obtained from γ' by substituting all occurrences of x by a single symbol $b \in \Sigma$, i.e., $v = u_1 b u_2 b \ldots b u_{k+1}$. For the following argument, let v be some fixed shortest word from $L_{\mathrm{NE},\Sigma}(\beta)$ and let γ' be the corresponding pattern as described above. Since v is a shortest word from $L_{\mathrm{NE},\Sigma}(\beta)$, it can be obtained by substituting every variable of β by a single symbol. Therefore, for every i, $1 \leq i \leq |v|$, the symbol at position i in v either corresponds to an occurrence of a terminal symbol at position i in β or to an occurrence of a variable at position i in β. If, for some i, $1 \leq i \leq |v|$, the symbol b at position i in v corresponds to a terminal symbol in β, then all shortest words in $L_{\mathrm{NE},\Sigma}(\beta)$ have an occurrence of symbol b at position i, which implies that at position i in γ' there must also be an occurrence of b. Since the residual of γ' is w, this directly implies that w', the residual of β, is a substring of w. If $w' \neq w$, i.e., $|w'| < |w|$, then there must exist a symbol $b \in \Sigma$, such that $|w'|_b < |w|_b$. This is a contradiction, since by substituting all variables in β by some words that do not contain occurrences of b, we can produce a word that does not contain w as a substring and therefore it is not a word in $L_{\mathrm{NE},\Sigma}(\gamma)$. Thus, we can conclude that $w' = w$. This, in particular, implies that all the shortest words v in $L_{\mathrm{NE},\Sigma}(\beta)$ contain exactly the symbols of w and $|\beta| - |w|$ occurrences of *the same* symbol $b \in \Sigma$, since, as explained above, v is obtained from γ by erasing all variables but one, which is substituted by a single symbol. This directly implies that β must contain $|\beta| - |w|$ occurrences of the same variable and, therefore, is a one-variable pattern of the form $\beta = u_1 x^{j_1} u_2 x^{j_2} \dots x^{j_m} u_{m+1}$, where $w = u_1 u_2 \dots u_m$ and $j_i \geq 1$, $1 \leq i \leq m$.

It remains to show that $\beta = \gamma$. To this end, we observe that if $|\operatorname{var}(\gamma)| \geq 2$, then in $L_{\mathrm{E},\Sigma}(\gamma)$ there are words that cannot be obtained from β (e. g., all words constructed by erasing all variables in γ except two, which are substituted by different single symbols). Furthermore, $|\operatorname{var}(\gamma)| = 1$ implies that $L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(\gamma) \cup \{w\}$. Hence,

$$L_{\mathrm{E},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(\gamma) \cup \{w\} = L_{\mathrm{NE},\Sigma}(\beta) \cup \{w\} = L_{\mathrm{NE},\Sigma}(\beta) \cup L_{\mathrm{NE},\Sigma}(\alpha)$$

and since $w \notin L_{\mathrm{NE},\Sigma}(\gamma) \cup L_{\mathrm{NE},\Sigma}(\beta)$, the equality $L_{\mathrm{NE},\Sigma}(\gamma) = L_{\mathrm{NE},\Sigma}(\beta)$ follows, which implies $\gamma = \beta$. $\square$

With respect to the case $Z = \mathrm{E}$, $Z' = \mathrm{NE}$, we can even present a characterisation for the situation $L_{\mathrm{NE},\Sigma}(\alpha) = \bigcup_{i=1}^{k} L_{\mathrm{E},\Sigma}(\beta_i)$ with $k \leq |\Sigma|$. It shall be explained later on that this characterisation is a generalisation of the construction given by Jiang et al. [9].

Theorem 31. *Let $\ell \geq 2$ and let Σ be an alphabet with $\{\mathtt{a}_1, \mathtt{a}_2, \dots, \mathtt{a}_\ell\} \subseteq \Sigma$. Furthermore, let $\alpha_1, \alpha_2, \dots, \alpha_\ell$ and γ be patterns with $L_{\mathrm{E},\Sigma}(\alpha_i) \neq L_{\mathrm{E},\Sigma}(\alpha_j)$, $1 \leq i < j \leq \ell$. Then $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i) = L_{\mathrm{NE},\Sigma}(\gamma)$ if and only if, for some permutation π of $(1, 2, \dots, \ell)$,*

- $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \dots, \mathtt{a}_\ell\}$,
- $\gamma = u_1\, x\, u_2\, x\, u_3 \dots u_k\, x\, u_{k+1}$, $k \geq 1$, $u_i \in \Sigma^*$, $1 \leq i \leq k+1$, *and,*
- *for every i, $1 \leq i \leq \ell$,*

  $$\alpha_i = u_1\, \alpha'_i\, \mathtt{a}_{\pi(i)}\, \alpha''_i\, u_2\, \alpha'_i\, \mathtt{a}_{\pi(i)}\, \alpha''_i\, u_3 \dots u_k\, \alpha'_i\, \mathtt{a}_{\pi(i)}\, \alpha''_i\, u_{k+1}\,,$$

  where $\alpha'_i, \alpha''_i \in X^*$,
- *for every i, $1 \leq i \leq \ell$, there exists a $y_i \in \operatorname{var}(\alpha_i)$ with $|\alpha_i|_{y_i} = k$ and*
  - $|\alpha'_i|_{y_i} = 1$ *for all i, $1 \leq i \leq \ell$, or*
  - $|\alpha''_i|_{y_i} = 1$ *for all i, $1 \leq i \leq \ell$.*

Proof. We first show the *if* direction. To this end, we assume that the conditions of the theorem hold and, in particular, we assume that $|\alpha'_i|_{y_i} = 1$ for all i, $1 \leq i \leq \ell$ (the case where $|\alpha''_i|_{y_i} = 1$, $1 \leq i \leq \ell$, can be dealt with analogously). For every i, $1 \leq i \leq \ell$, every word $u_1 v \mathtt{a}_i u_2 v \mathtt{a}_i u_3 \dots u_k v \mathtt{a}_i u_{k+1} \in L_{\mathrm{NE},\Sigma}(\gamma)$ can be obtained from α_i by substituting y_i by v and erasing all other variables, which shows $L_{\mathrm{NE},\Sigma}(\gamma) \subseteq \bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$. Furthermore, every word in $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$ is of the form $u_1 v u_2 v u_3 \dots u_k v u_{k+1}$, where v is a non-empty word and, thus, it is in $L_{\mathrm{NE},\Sigma}(\gamma)$. This proves $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i) \subseteq L_{\mathrm{NE},\Sigma}(\gamma)$, which concludes the *if* direction.

We now show the *only if* direction. Throughout our argument, we assume that the permutation π satisfies $\pi(i) = i$ for every i, $1 \leq i \leq \ell$, and therefore write i instead of $\pi(i)$. The proof for any other permutation is identical.

We assume that $L_{\mathrm{E},\Sigma}(\alpha_i) \neq L_{\mathrm{E},\Sigma}(\alpha_j)$, $1 \leq i < j \leq \ell$, and $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i) = L_{\mathrm{NE},\Sigma}(\gamma)$. We first prove the following claim.

Claim 1: $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_\ell\}$ and, for some $k \geq 1$, there are $u_i \in \Sigma^*$, $1 \leq i \leq k+1$, such that

$$\begin{aligned}\gamma &= u_1\, x\, u_2\, x\, u_3 \ldots u_k\, x\, u_{k+1}\,,\\ \alpha_i &= u_1\, \alpha'_i\, \mathtt{a}_i\, \pi_{i,u_2}\, \mathtt{a}_i\, \pi_{i,u_3} \ldots \pi_{i,u_k}\, \mathtt{a}_i\, \alpha''_i\, u_{k+1}, 1 \leq i \leq \ell\,,\end{aligned}$$

where, for every i, $1 \leq i \leq \ell$, and j, $2 \leq j \leq k$, $\alpha'_i, \alpha''_i \in X^*$ and π_{i,u_j} is a pattern with residual u_j.

Proof of Claim 1: For every i, $1 \leq i \leq \ell$, α_i contains at least one terminal symbol, since otherwise $\varepsilon \in \bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$, which is a contradiction as $\varepsilon \notin L_{\mathrm{NE},\Sigma}(\gamma)$. Furthermore, we note that γ must contain at least one variable, since otherwise $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i) = L_{\mathrm{NE},\Sigma}(\gamma) = \{\gamma\}$, which contradicts $L_{\mathrm{E},\Sigma}(\alpha_i) \neq L_{\mathrm{E},\Sigma}(\alpha_j)$, $1 \leq i < j \leq \ell$. For every i, $1 \leq i \leq \ell$, let w_i be the residual of α_i and let Φ be the set of shortest words of $\bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$. Our next goal is to show that $\Phi = \{w_i \mid 1 \leq i \leq \ell\}$. If there is a word $u \in \Phi \setminus \{w_i \mid 1 \leq i \leq \ell\}$, then, for some i, $1 \leq i \leq \ell$, u can be obtained from α_i by substituting a variable with a non-empty word, which implies that $|w_i| < |u|$ and therefore $u \notin \Phi$, a contradiction. Thus, $\Phi \subseteq \{w_i \mid 1 \leq i \leq \ell\}$, which, in particular, means $|\Phi| \leq \ell$. Since at least ℓ shortest words can be obtained from γ, by substituting a variable by the symbols $\mathtt{a}_1$, $\mathtt{a}_2$, …, $\mathtt{a}_\ell$, respectively, we can conclude that $|\Phi| = \ell$. Thus, $\Phi = \{w_i \mid 1 \leq i \leq \ell\}$ with $|\Phi| = \ell$. Since $|\operatorname{var}(\gamma)| > 1$ or $|\Sigma| > \ell$ implies that $L_{\mathrm{NE},\Sigma}(\gamma)$ contains strictly more than ℓ shortest words, $|\operatorname{var}(\gamma)| = 1$ and $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_\ell\}$ is implied. More precisely, γ is of the form

$$\gamma = u_1\, x\, u_2\, x\, u_3 \ldots u_k\, x\, u_{k+1}\,,$$

where $k \in \mathbb{N}_0$, and $u_i \in \Sigma^*$, $1 \leq i \leq k+1$. Since $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_\ell\}$, the ℓ unique shortest words in $L_{\mathrm{NE},\Sigma}(\gamma)$, i.e., the words w_i, $1 \leq i \leq \ell$, can be obtained from γ by substituting x by $\mathtt{a}_i$, $1 \leq i \leq \ell$, respectively. Hence, without loss of generality,

$$w_i = u_1\, \mathtt{a}_i\, u_2\, \mathtt{a}_i\, u_3 \ldots u_k\, \mathtt{a}_i\, u_{k+1}, 1 \leq i \leq \ell\,.$$

Since, for every i, $1 \leq i \leq \ell$, w_i is the residual of α_i and since all words in $L_{\mathrm{NE},\Sigma}(\gamma) = \bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$ have u_1 as a prefix and u_{k+1} as a suffix (due to the structure of γ), we can conclude that the α_i are of the following form:

$$\alpha_i = u_1\, \alpha'_i\, \mathtt{a}_i\, \pi_{i,u_2}\, \mathtt{a}_i\, \pi_{i,u_3} \ldots \pi_{i,u_k}\, \mathtt{a}_i\, \alpha''_i\, u_{k+1}, 1 \leq i \leq \ell\,,$$

where, for every i, $1 \leq i \leq \ell$, and j, $2 \leq j \leq k$, $\alpha'_i, \alpha''_i \in X^*$ and π_{i,u_j} is a pattern with residual u_j. □(Claim 1)

Next, we show that, for every i and j, $1 \leq i \leq \ell$, $2 \leq j \leq k$, π_{i,u_j} contains u_j as a (non-scattered) factor, i. e., $\pi_{i,u_j} = \psi\, u_j\, \psi'$, where ψ and ψ' are terminal-free patterns. To this end, we first prove another claim:

Claim 2: Let i, $1 \leq i \leq \ell$, let $z \in \mathrm{var}(\alpha_i)$ and let $\widehat{\alpha}_i$ be the pattern obtained by erasing all variables of α_i except z. Then

$$\widehat{\alpha}_i = u_1\, z^n \mathtt{a}_i z^{n'}\, u_2\, z^n \mathtt{a}_i z^{n'}\, u_3 \ldots u_k\, z^n \mathtt{a}_i z^{n'} u_{k+1}\,,$$

where $n, n' \in \mathbb{N}_0$ with $|\alpha|_z = k(n + n')$.

Proof of Claim 2: We can construct a word v by erasing all variables from α_i except z, which is substituted by $\mathtt{a}_{i'}$, where $1 \leq i' \leq \ell$, $i \neq i'$. Obviously, $v \in L_{\mathrm{E},\Sigma}(\alpha_i)$ and therefore also $v \in L_{\mathrm{NE},\Sigma}(\gamma)$. This means that it is possible to obtain v from γ by substituting x of γ by a non-empty word u. Furthermore, since the residual of γ is $u_1 u_2 \ldots u_{k+1}$, the residual of α is $u_1 \mathtt{a}_i u_2 \mathtt{a}_i \ldots \mathtt{a}_i u_{k+1}$ and there are k occurrences of x in γ, we can conclude that $|u|_{\mathtt{a}_i} = 1$ and $|u|_{\mathtt{a}_{i'}} = \frac{|\alpha_i|_z}{k}$. More precisely, $u = \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'}$ with $k(n + n') = |\alpha_i|_z$. Hence,

$$v = u_1\, \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'}\, u_2\, \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'}\, u_3 \ldots u_k\, \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'} u_{k+1}\,.$$

Now let p, $1 \leq p \leq |v|$, be an arbitrary position of v that corresponds to a symbol in one of the $\mathtt{a}_{i'}^n$ or $\mathtt{a}_{i'}^{n'}$ factors, i. e., there exists a q, $1 \leq q \leq k$, such that $p = |u_1 \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'} \ldots \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'} u_q| + r$, for some r with $1 \leq r \leq n$, or $p = |u_1 \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'} \ldots \mathtt{a}_{i'}^n \mathtt{a}_i \mathtt{a}_{i'}^{n'} u_q \mathtt{a}_{i'}^n \mathtt{a}_i| + r$, for some r with $1 \leq r \leq n'$. If, after erasing all variables from α_i except z, there is an occurrence of symbol $\mathtt{a}_{i'}$ at position p, then, for every word v' that is obtained from α by erasing all variables except z, which is substituted by a single symbol, $v'[p] = \mathtt{a}_{i'}$ must be satisfied. We shall assume that $\mathtt{a}_{i'}$ at position p is, in fact, not produced by an occurrence of variable z. We can now obtain a word v' from α by erasing all variables except z, which is substituted by the symbol $\mathtt{a}_i$. It must be possible to obtain this word v' from γ as well, which is only possible by substituting variable x by the word $\mathtt{a}_i^{n+n'+1}$; thus,

$$v' = u_1\, \mathtt{a}_i^n \mathtt{a}_i\, \mathtt{a}_i^{n'}\, u_2\, \mathtt{a}_i^n \mathtt{a}_i\, \mathtt{a}_i^{n'}\, u_3 \ldots u_k\, \mathtt{a}_i^n \mathtt{a}_i\, \mathtt{a}_i^{n'}\,.$$

It can be easily verified that $v'[p] = \mathtt{a}_i$, which is a contradiction to the observations from above. Therefore, such positions p cannot exist and every occurrence of a symbol $\mathtt{a}_{i'}$ in one of the $\mathtt{a}_{i'}^n$ or $\mathtt{a}_{i'}^{n'}$ factors is produced by an occurrence of variable z. This implies that the pattern $\widehat{\alpha}_i$ obtained from α_i by erasing all variables except z is of the form

$$\alpha' = u_1\, z^n \mathtt{a}_i z^{n'}\, u_2\, z^n \mathtt{a}_i z^{n'}\, u_3 \ldots u_k\, z^n \mathtt{a}_i z^{n'} u_{k+1}\,.$$

□(Claim 2)

We note that Claim 2 particularly implies that, for every i, j, $1 \leq i \leq \ell$, $2 \leq j \leq k$, and for every $z \in \mathrm{var}(\alpha_i)$, every occurrence of z in π_{i,u_j} is either to the left of the first terminal symbol or to the right of the last terminal symbol that belongs to the residual u_j of π_{i,u_j}. Consequently, for every i, $1 \leq i \leq \ell$, α_i is of the following form:

$$\alpha_i = u_1\, \delta_{i,1}\, \mathtt{a}_i\, \delta'_{i,1}\, u_2\, \delta_{i,2}\, \mathtt{a}_i\, \delta'_{i,2}\, u_3\, \delta_{i,3} \ldots \delta'_{i,k-1}\, u_k\, \delta_{i,k}\, \mathtt{a}_i\, \delta'_{i,k}\, u_{k+1}\,,$$

where $\delta_{i,j}, \delta'_{i,j} \in X^*$, $1 \leq j \leq k$. We note that from Claim 2, we can further conclude that every variable $z \in \mathrm{var}(\alpha)$ has the same number of occurrences in the factors between the terminal words u_j, $1 \leq j \leq k$, i.e., for every i, $1 \leq i \leq \ell$, $|\delta_{i,1}\mathtt{a}\delta'_{i,1}|_z = |\delta_{i,2}\mathtt{a}\delta'_{i,2}|_z = \ldots = |\delta_{i,k}\mathtt{a}\delta'_{i,k}|_z$, which particularly implies $|\delta_{i,1}\mathtt{a}\delta'_{i,1}| = |\delta_{i,2}\mathtt{a}\delta'_{i,2}| = \ldots = |\delta_{i,k}\mathtt{a}\delta'_{i,k}|$. In the following, we shall show that these factors are actually identical. To this end, let i, $1 \leq i \leq \ell$ and let σ be an arbitrary substitution for α_i, which substitutes every variable by exactly one symbol. Since $\sigma(\alpha_i) \in L_{\mathrm{NE},\Sigma}(\gamma)$, there is a substitution τ with $\tau(\gamma) = \sigma(\alpha_i)$. Furthermore, since σ substitutes every variable by a single symbol and since $|\delta_{i,1}\mathtt{a}_i\delta'_{i,1}| = |\delta_{i,2}\mathtt{a}_i\delta'_{i,2}| = \ldots = |\delta_{i,k}\mathtt{a}_i\delta'_{i,k}|$, $|\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1})| = |\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1})| = \ldots = |\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1})|$ is implied. In particular, this means that $|\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1})| = \frac{|\sigma(\alpha_i)| - |u_1 u_2 \ldots u_k|}{k}$ and since $|\tau(x)| = \frac{|\tau(\gamma)| - |u_1 u_2 \ldots u_k|}{k}$ clearly holds, we conclude that $|\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1})| = |\sigma(\delta_{i,2}\mathtt{a}_i\delta'_{i,2})| = \ldots = |\sigma(\delta_{i,k}\mathtt{a}_i\delta'_{i,k})| = |\tau(x)|$. Since

$$\begin{array}{rllllllll} \sigma(\alpha_i) = & u_1 & \sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1}) & u_2 & \sigma(\delta_{i,2}\mathtt{a}_i\delta'_{i,2}) & \ldots & \sigma(\delta_{i,k}\mathtt{a}_i\delta'_{i,k}) & u_{k+1} \\ = & u_1 & \tau(x) & u_2 & \tau(x) & \ldots & \tau(x) & u_{k+1}\,, \end{array}$$

it follows that $\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1}) = \sigma(\delta_{i,2}\mathtt{a}_i\delta'_{i,2}) = \ldots = \sigma(\delta_{i,k}\mathtt{a}_i\delta'_{i,k}) = \tau(x)$. So for every substitution σ for α_i that substitutes every variable by a single symbol, $\sigma(\delta_{i,1}\mathtt{a}_i\delta'_{i,1}) = \sigma(\delta_{i,2}\mathtt{a}_i\delta'_{i,2}) = \ldots = \sigma(\delta_{k,1}\mathtt{a}_i\delta'_{k,1})$ is implied, which is only possible if $\delta_{i,1}\mathtt{a}_i\delta'_{i,1} = \delta_{i,2}\mathtt{a}_i\delta'_{i,2} = \ldots = \delta_{i,k}\mathtt{a}_i\delta'_{i,k}$. Thus,

$$\alpha_i = u_1\, \alpha'_i\, \mathtt{a}_i\, \alpha''_i\, u_2\, \alpha'_i\, \mathtt{a}_i\, \alpha''_i\, u_3 \ldots u_k\, \alpha'_i\, \mathtt{a}_i\, \alpha''_i\, u_{k+1}\,,$$

where $\alpha'_i, \alpha''_i \in X^*$.

It remains to show that there exist variables $y_i \in \mathrm{var}(\alpha_i)$, $1 \leq i \leq \ell$, with $|\alpha_i|_{y_i} = k$ and such that $|\alpha'_i|_{y_i} = 1$ for every i, $1 \leq i \leq \ell$, or $|\alpha''_i|_{y_i} = 1$ for every i, $1 \leq i \leq \ell$. To this end, for every i, $1 \leq i \leq \ell$, we define the following properties:

- $P_{i,l}$: There exists a variable $y_i \in \mathrm{var}(\alpha)$ with $|\alpha_i|_{y_i} = k$ and $|\alpha'_i|_{y_i} = 1$.
- $P_{i,r}$: There exists a variable $y_i \in \mathrm{var}(\alpha)$ with $|\alpha_i|_{y_i} = k$ and $|\alpha''_i|_{y_i} = 1$.

We first show that $P_{1,l}$ or $P_{1,r}$ is satisfied. Let v_1 be the word obtained from γ by substituting x by $\mathtt{a}_1\mathtt{a}_1$. Obviously, $v_1 \notin \bigcup_{i=2}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$, and, since $v_1 \in \bigcup_{i=1}^{\ell} L_{\mathrm{E},\Sigma}(\alpha_i)$, this implies that $v_1 \in L_{\mathrm{E},\Sigma}(\alpha_1)$. The word v_1 can only be obtained from α_1 by erasing all variables except some variable y_i with k occurrences, which is substituted by $\mathtt{a}_1$. Hence, property $P_{1,l}$ or $P_{1,r}$ is satisfied. Next, we assume that $P_{1,l}$ is not satisfied, which implies that $P_{1,r}$ is satisfied. Furthermore, let $v_2, v_3, \ldots, v_\ell$ be the words obtained from γ by substituting x by $\mathtt{a}_2\mathtt{a}_1, \mathtt{a}_3\mathtt{a}_1, \dots, \mathtt{a}_k\mathtt{a}_1$, respectively. Since $P_{1,l}$ is not satisfied, $\{v_2, v_3, \dots, v_k\} \cap L_{\mathrm{E},\Sigma}(\alpha_1) = \emptyset$. Furthermore, for every i, j, $2 \le i, j \le \ell$, $i \neq j$, $|v_i|_{\mathtt{a}_j} < |\alpha_j|_{\mathtt{a}_j}$, which implies that $v_i \notin L_{\mathrm{E},\Sigma}(\alpha_j)$. Consequently, for every i, $2 \le i \le \ell$, $v_i \in L_{\mathrm{E},\Sigma}(\alpha_i)$ and v_i can only be obtained from α_i by erasing all variables except some variable y_i with k occurrences, which is substituted by $\mathtt{a}_1$. Hence, properties $P_{i,r}$, $1 \le i \le \ell$, are satisfied. Analogously, the assumption that $P_{1,r}$ is not satisfied leads to the situation that properties $P_{i,l}$, $1 \le i \le \ell$, are satisfied. Consequently, there exist variables $y_i \in \mathrm{var}(\alpha_i)$, $1 \le i \le \ell$, such that $|\alpha'_i|_{y_i} = 1$ for every i, $1 \le i \le \ell$, or $|\alpha''_i|_{y_i} = 1$ for every i, $1 \le i \le \ell$. This concludes the proof. □

If we apply the construction of Jiang et al. [9] to a one-variable pattern γ, then we obtain patterns α_i, $1 \le i \le |\Sigma|$, that satisfy the conditions of the patterns in the statement of Theorem 31. More precisely, this corresponds to the special case where $\alpha'_i\alpha''_i = y_i$, $1 \le i \le |\Sigma|$. Moreover, it can be easily verified that if γ and patterns α_i, $1 \le i \le |\Sigma|$, satisfy the conditions of the statement of Theorem 31, then, depending on whether $|\alpha'_i|_{y_i} = 1$ for all i, $1 \le i \le |\Sigma|$, or $|\alpha''_i|_{y_i} = 1$ for all i, $1 \le i \le |\Sigma|$, we can obtain patterns β_i from the patterns α_i by replacing $\alpha'_i\mathtt{a}_i\alpha''_i$ by $y_i\mathtt{a}_i$ or by $\mathtt{a}_i y_i$, respectively, and $\bigcup_{i=1}^{|\Sigma|} L_{\mathrm{E},\Sigma}(\beta_i) = L_{\mathrm{NE},\Sigma}(\gamma)$ still holds. Furthermore, the patterns β_i are exactly the patterns that are obtained if we apply the construction of Jiang et al. [9].

5. Summary

In the present paper we have investigated a variety of closure properties of standard classes of pattern languages, showing that they are not closed under most of the typical operations applied to classes of languages. Table 1 contains references to a selection of our results. Entries highlighted with the symbol † refer to statements that provide characteristic conditions, and references in brackets () indicate that an insight into the full class of E-pattern languages indirectly follows from our reasoning on the subclass of terminal-free E-pattern languages.

Further results of our paper that are not included in Table 1 consider closure properties of classes of pattern languages over unary alphabets (see Propositions 2, 5, 9, 18, 21, 23), and the relation between E-pattern languages and finite unions of NE-pattern languages and vice versa (see Propositions 3 and 24, and Theorems 30 and 31).

Acknowledgments

The authors wish to thank the anonymous referees of this paper and of its conference version [4] for their numerous helpful remarks, which, among other things, have yielded stronger versions of Propositions 19 and 20. Furthermore, Kai Salomaa's suggestion to extend this paper to pattern languages over unary alphabets is gratefully acknowledged, and reflected by several results in this paper.

Operation	nePAT_Σ	ePAT_Σ	$\mathrm{nePAT}_{\mathrm{tf},\Sigma}$	$\mathrm{ePAT}_{\mathrm{tf},\Sigma}$
Union	Thm 1, Prop. 24, Thm 29	Prop. 25, Prop. 26, Prop. 27, Thm 28	Prop. 2, Thm 8†, Cor. 10†	Prop. 2, Thm 4†, Cor. 10†
Intersection	Thm 1, Sect. 4.2	Sect. 4.2	Prop. 11	Prop. 16, Cor. 17
Complementation	Thm 1, Prop. 22†	Prop. 22†	Prop. 18†	Prop. 18†
Morphisms, inverse morphisms	Thm 1	(Prop. 19)	Prop. 19	Prop. 19
Kleene plus, Kleene star	Thm 1	(Prop. 20)	Prop. 20	Prop. 20

Table 1: The notations nePAT_Σ, ePAT_Σ, $\mathrm{nePAT}_{\mathrm{tf},\Sigma}$ and $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ stand for the classes of general NE-pattern languages, general E-pattern languages, terminal-free NE-pattern languages and terminal-free E-pattern languages, respectively, over some finite alphabet Σ with $|\Sigma| \geq 2$. Further remarks on the table are given in Section 5.

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

[2] H. Bayer. Allgemeine Eigenschaften von Patternsprachen. Projektarbeit, Fachbereich Informatik, Universität Kaiserslautern, 2007. In German.

[3] K. Culik II and J. Karhumäki. On the equality sets for homomorphisms on free monoids with two generators. *Theoretical Informatics and Applications (RAIRO)*, 14:349–369, 1980.

[4] J. D. Day, D. Reidenbach, and M. L. Schmid. Closure properties of pattern languages. In *Proc. 18th International Conference on Developments in Language Theory, DLT 2014*, volume 8633 of *Lecture Notes in Computer Science*, pages 279–290, 2014.

[5] H. Fernau and M. L. Schmid. Pattern matching with variables: A multivariate complexity analysis. *Information and Computation*, 242:287–305, 2015.

[6] H. Fernau, M. L. Schmid, and Y. Villanger. On the parameterised complexity of string morphism problems. In *Proc. 33rd IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2013*, volume 24 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 55–66, 2013.

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

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

[9] 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.

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

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

[12] M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading, MA, 1983.

[13] R. C. Lyndon and M. P. Schützenberger. The equation $a^m = b^n c^p$ in a free group. *Michigan Mathematical Journal*, 9:289–298, 1962.

[14] A. Marron and K. Ko. Identification of pattern languages from examples and queries. *Information and Computation*, 74:91–112, 1987.

[15] D. Reidenbach. A non-learnable class of E-pattern languages. *Theoretical Computer Science*, 350:91–102, 2006.

[16] D. Reidenbach. An examination of Ohlebusch and Ukkonen's conjecture on the equivalence problem for E-pattern languages. *Journal of Automata, Languages and Combinatorics*, 12:407–426, 2007.

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

[18] D. Reidenbach and M.L.Schmid. Patterns with bounded treewidth. *Information and Computation*, 239:87–99, 2014.

[19] D. Reidenbach and M.L.Schmid. Regular and context-free pattern languages over small alphabets. *Theoretical Computer Science*, 518:80–95, 2014.

[20] T. Shinohara. Inferring unions of two pattern languages. *Bulletin of Informatics and Cybernetics*, 20:83–88, 1983.

[21] T. Shinohara and H. Arimura. Inductive inference of unbounded unions of pattern languages from positive data. *Theoretical Computer Science*, 241:191–209, 2000.

[22] K. Wright. Identification of unions of languages drawn from an identifiable class. In *Proc. 2nd Annual Workshop on Computational Learning Theory, COLT 1989*, pages 328–333, 1989.