Existence and nonexistence of descriptive patterns

Existence and nonexistence of descriptive patterns[☆]

Dominik D. Freydenberger[a], Daniel Reidenbach[b,*]

[a] *Institut für Informatik, Goethe-Universität, Postfach 111932, D-60054 Frankfurt am Main, Germany*
[b] *Department of Computer Science, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom*

Abstract

In the present paper, we study the existence of descriptive patterns, i. e. patterns that cover all words in a given set through morphisms and that are optimal in terms of revealing commonalities of these words. Our main result shows that if patterns may be mapped to words by arbitrary morphisms, then there exist infinite sets of words that do not have a descriptive pattern. This answers a question posed by Jiang, Kinber, Salomaa, Salomaa and Yu (*International Journal of Computer Mathematics* 50, 1994). Since the problem of whether a pattern is descriptive depends on the inclusion relation of so-called pattern languages, our technical considerations lead to a number of deep insights into the inclusion problem for and the topology of the class of terminal-free E-pattern languages.

Keywords: Pattern languages, Descriptive patterns

1. On Patterns Descriptive of a Set of Strings

A *pattern* is a finite string that consists of variables taken from an alphabet X and terminal symbols taken from an alphabet Σ. For any pattern α and any word w over Σ, α is said to cover w if w can be obtained from α by substituting the variables with appropriate strings of terminal symbols. Whenever α contains several occurrences of the same variable, the substitution of variables needs to be "uniform", i. e. each of the occurrences must be replaced with the same word over Σ. Therefore, and more formally, such a substitution is simply a terminal-preserving morphism $\sigma : (\Sigma \cup X)^* \to \Sigma^*$, i. e. a morphism that satisfies $\sigma(a) = a$ for every terminal symbol a in the pattern. For instance, the pattern $\alpha := xy\mathtt{b}x\mathtt{a}$ (where x, y are variables and $\mathtt{a}, \mathtt{b}$ are terminal symbols) covers the word $w_1 := \mathtt{abababa}$ since there is a substitution σ, given by $\sigma(x) := \mathtt{ab}$ and

[☆] A preliminary version of this paper was presented at the conference DLT 2009.
[*] Corresponding author.
Email addresses: freydenberger@em.uni-frankfurt.de (Dominik D. Freydenberger), D.Reidenbach@lboro.ac.uk (Daniel Reidenbach)

$\sigma(y) := \mathtt{a}$, satisfying $\sigma(\alpha) = w$. In contrast to this, α does not cover, e. g., $w_2 := \mathtt{bbbbaa}$.

Due to the simplicity of the concepts involved, the above described notion of a pattern is studied in a variety of fields of research. The present paper mainly deals with two quite closely related approaches: Firstly, a pattern α over $\Sigma \cup X$ can be regarded as a generator of a formal language $L(\alpha)$, the so-called *pattern language*, which simply comprises all words in Σ^* that can be obtained from the pattern by arbitrary substitutions. Secondly, for any given finite or infinite language S, patterns can be used to approximate S; i. e., a pattern α is sought that is *consistent* with S (which means that α covers all words in S or alternatively, in terms of pattern languages, $L(\alpha) \supseteq S$). The latter concept is motivated by the fact that if a pattern is consistent with a language S, then this pattern reveals a common structure of the strings in S. Hence, and since they are compact devices that can be easily read and interpreted by humans, patterns can be very helpful when commonalities of data represented by strings are analysed.

The characteristics of pattern languages have been intensively studied in the past decades. Therefore, quite a number of basic properties of pattern languages, e. g. regarding the usual decision problems for classes of formal languages, are known (cf. the surveys by Mateescu and Salomaa [10] and Salomaa [14] and our recent paper [6]). Furthermore, pattern languages have been a focus of interest of inductive inference from the very beginning, investigating whether it is possible to infer a pattern from the words in its pattern language (see Ng and Shinohara [11]). It is quite remarkable that many of the corresponding results in language theory and inductive inference differ for the two main types of pattern languages that are normally considered, namely the *NE*-pattern language of a pattern (introduced by Angluin [1]), which merely consists of those words in Σ^* that can be obtained from the pattern by *nonerasing* substitutions (i. e. substitutions that do not replace any variables with the empty word), and the *E*-pattern language (established by Shinohara [15]), which additionally comprises those words that can be derived from the pattern by substituting the empty word for arbitrary variables.

The problem of finding a consistent pattern for an arbitrary set S of strings is often referred to as *(string) pattern discovery*, and many of its applications are derived from tasks in bioinformatics (cf. Brazma et al. [2]). In contrast to the inductive inference approach to pattern languages, where a pattern shall be inferred that exactly describes the given language, string pattern discovery faces the problem that S can typically have many consistent patterns showing very different characteristics. For instance, both

$$\begin{aligned} \alpha_1 &:= xyxyx \text{ and} \\ \alpha_2 &:= x\mathtt{ab}y \end{aligned}$$

are consistent with the language

$$S_0 := \{ \begin{array}{l} \texttt{ababa}, \\ \texttt{ababbababbab}, \\ \texttt{babab}\}, \end{array}$$

and the pattern $\alpha_0 := x$ is consistent with every set of strings, anyway. Hence, the algorithms of string pattern discovery require an underlying notion of the quality of a pattern in order to determine what patterns to strive for. With regard to the above example set and patterns, it seems quite likely that one might not be interested in a procedure outputting α_0 when reading S_0. Concerning α_1 and α_2, however, it is, a priori, by no means evident which of them to prefer. Thus, the definition of the quality of a pattern might often depend on the field of application where string pattern discovery is conducted. In addition to this, it is a worthwhile goal to develop *generic* notions of quality of consistent patterns that can inform the design of pattern discovery algorithms.

In this regard, the *descriptiveness* of patterns is a well-known and plausible concept, that is also used within the scope of inductive inference (cf. Ng and Shinohara [11]). A pattern δ is said to be descriptive of a given set S of strings if there is no pattern α satisfying $L(\delta) \supset L(\alpha) \supseteq S$. Intuitively, this means that if δ is descriptive of S, then no consistent pattern for S provides a strictly closer match than δ. Thus, although δ does not need to be unique (as to be further discussed below), it is guaranteed that it is one of the most accurate approximations of S that can be provided by patterns. While descriptiveness is unquestionably an appropriate notion of quality of consistent patterns, it leads to major technical challenges, as its application requires insights into the inclusion problem for pattern languages, which is known to be undecidable in the general case and still combinatorially involved for some major natural subclasses where it is decidable. This aspect is crucial to the subsequent formal parts of our paper.

Since the definition of a descriptive pattern is based on the concept of pattern languages, the question of whether NE- or E-pattern languages are chosen can have a significant impact on the descriptiveness of a pattern. This is reflected by the terminology we use: we call a pattern δ an NE-descriptive pattern if it is descriptive in terms of its NE-pattern language and the NE-pattern languages of all patterns in $(\Sigma \cup X)^+$; accordingly, we call δ E-descriptive if its descriptiveness is based on interpreting all patterns as generators of E-pattern languages. In order to illustrate these terms, we now briefly discuss the descriptiveness of the example patterns introduced above (though the full verification of our corresponding claims is not always straightforward and might require certain tools to be introduced later). If we deal with S_0 and the patterns in the context of NE-pattern languages, then it can be stated that both α_1 and α_2 are NE-descriptive of S_0, since no NE-pattern languages can comprise S_0 and, at the same time, be a proper sublanguage of the NE-pattern languages of α_1 or α_2. If we study S_0 in terms of E-pattern languages, it turns out that α_1 is also E-descriptive of S_0, i. e. there is no pattern generating an E-pattern language

that is consistent with S_0 and strictly included in the E-pattern language of α_1. However, the second NE-descriptive example pattern α_2 is not E-descriptive of S_0, since the E-pattern language generated by

$$\alpha_3 := x\texttt{abab}y$$

is a proper sublanguage of the E-pattern language of α_2 and comprises S_0. The pattern α_3, in turn, is even E-descriptive of S_0, but not NE-descriptive, since it is not consistent with S_0 if we disallow empty substitutions. Exactly the same holds for $\alpha_4 := x\texttt{baba}y$, which also is consistent with S_0 if we allow the empty substitution of variables, generates an E-pattern language that is strictly included in the E-pattern language of α_2 and is E-descriptive, but not NE-descriptive of S_0.

The present paper examines the basic underlying problem of descriptive pattern discovery, namely the *existence* of such patterns; this means that we study the question of whether or not, for a given language S, there is a pattern that is descriptive of S. To this end, four different cases can be considered: NE-descriptive patterns of finite languages, NE-descriptive patterns of infinite languages, E-descriptive patterns of finite languages and E-descriptive patterns of infinite languages. The problem of the existence of the former three types of descriptive patterns is either trivial or has already been solved in previous publications. We therefore largely study the latter case, and our corresponding main result answers a question posed by Jiang, Kinber, Salomaa, Salomaa and Yu [7]. Our technical considerations do not only provide insights into the actual topic of our paper, but – due to the definition of descriptive patterns – also reveal vital phenomena related to the inclusion of E-pattern languages and, hence, the topology of class of terminal-free E-pattern languages. Due to the way the inclusion of terminal-free E-pattern languages is characterised, this implies that we have to deal with combinatorial properties of morphisms in free monoids. Furthermore, crucial parts of our reasoning are based on infinite *unions* of pattern languages, which means that our paper shows additional connections to so-called multi-pattern languages (cf. Dumitrescu et al. [3]). While [3] features unions of pattern languages where the generating patterns form a context-free language, our work is essentially based on multi-pattern languages where the underlying set of patterns – apart from an infinite variable alphabet we have to use – is defined similarly to an HD0L language (see Kari et al. [9]).

A preliminary version [5] of this paper was presented at the conference DLT 2009.

2. Basic Definitions and Preparatory Technical Considerations

This paper is largely self-contained. For notations not explicitly defined, Rozenberg and Salomaa [13] can be consulted.

Let $\mathbb{N} := \{0, 1, 2, 3, \ldots\}$ and, for every $k \geq 0$, $\mathbb{N}_k := \{n \in \mathbb{N} \mid n \geq k\}$. The symbols $\subseteq$, $\subset$, $\supseteq$ and $\supset$ refer to subset, proper subset, superset and proper superset relation, respectively. The symbol ∞ stands for infinity. For an arbitrary

alphabet A, a *string* (over A) is a finite sequence of symbols from A, and λ stands for the *empty string*. The symbol A^+ denotes the set of all nonempty strings over A, and $A^* := A^+ \cup \{\lambda\}$. For the *concatenation* of two strings w_1, w_2 we write $w_1 \cdot w_2$ or simply $w_1 w_2$. We say that a string $v \in A^*$ is a *factor* of a string $w \in A^*$ if there are $u_1, u_2 \in A^*$ such that $w = u_1 v u_2$. The notation $|K|$ stands for the size of a set K or the length of a string K; the term $|w|_a$ refers to the number of occurrences of the symbol a in the string w. For any $w \in \Sigma^*$ and any $n \in \mathbb{N}$, w^n denotes the *n-fold concatenation of w*, with $w^0 := \lambda$.

For any alphabets A, B, a *morphism* is a function $h : A^* \to B^*$ that satisfies $h(vw) = h(v)h(w)$ for all $v, w \in A^*$. Given morphisms $g : A^* \to B^*$ and $h : B^* \to C^*$ (for alphabets A, B, C), their *composition* $h \circ g$ is defined by $(h \circ g)(w) := h(g(w))$ for all $w \in A^*$. For every morphism $h : A^* \to A^*$ and every $n \geq 0$, h^n denotes the *n-fold iteration of h*, i. e., $h^{n+1} := h \circ h^n$, where h^0 is the identity on A^*.

A morphism $h : A^* \to B^*$ is said to be *nonerasing* if $h(a) \neq \lambda$ for all $a \in A$. For any string $w \in C^*$, where $C \subseteq A$ and $|w|_a \geq 1$ for every $a \in C$, the morphism $h : A^* \to B^*$ is called a *renaming (of w)* if $h : C^* \to B^*$ is injective and $|h(a)| = 1$ for every $a \in C$.

Let Σ be a (finite or infinite) alphabet of so-called *terminal symbols* (or: *letters*) and X an infinite set of *variables* with $\Sigma \cap X = \emptyset$. We normally assume $\{\mathtt{a}, \mathtt{b}, \ldots\} \subseteq \Sigma$ and $\{y, z, x_0, x_1, x_2 \ldots\} \subseteq X$. A *pattern* is a string over $\Sigma \cup X$, a *terminal-free pattern* is a string over X and a *word* is a string over Σ. The set of all patterns over $\Sigma \cup X$ is denoted by Pat_Σ. For any pattern α, we refer to the set of variables in α as $\mathrm{var}(\alpha)$.

A morphism $\sigma : (\Sigma \cup X)^* \to (\Sigma \cup X)^*$ is called *terminal-preserving* if $\sigma(a) = a$ for every $a \in \Sigma$. A terminal-preserving morphism $\sigma : (\Sigma \cup X)^* \to \Sigma^*$ is called a *substitution*. Let $S \subseteq \Sigma^*$; then we say that a pattern α is *consistent with S* if, for every $w \in S$, there exists a substitution σ satisfying $\sigma(\alpha) = w$.

Intuitively, the pattern language of a pattern α is the maximum set of words α is consistent with. Formally, we consider two types of pattern languages, depending on whether we restrict ourselves to nonerasing substitutions: the *NE-pattern language* $L_{\mathrm{NE},\Sigma}(\alpha)$ of a pattern $\alpha \in \mathrm{Pat}_\Sigma$ is given by

$$L_{\mathrm{NE},\Sigma}(\alpha) := \{\sigma(\alpha) \mid \sigma : (\Sigma \cup X)^* \to \Sigma^* \text{ is a nonerasing substitution}\},$$

and the *E-pattern language* $L_{\mathrm{E},\Sigma}(\alpha)$ of α is given by

$$L_{\mathrm{E},\Sigma}(\alpha) := \{\sigma(\alpha) \mid \sigma : (\Sigma \cup X)^* \to \Sigma^* \text{ is a substitution}\}.$$

The term *pattern language* refers to any of the definitions introduced above. We call a pattern language *terminal-free* if it is generated by a terminal-free pattern.

We now can introduce our terminology on the main topic of this paper, namely the descriptiveness of a pattern. For any alphabet Σ and any language $S \subseteq \Sigma^*$, a pattern $\delta \in \mathrm{Pat}_\Sigma$ is said to be *NE-descriptive (of S)* provided that $L_{\mathrm{NE},\Sigma}(\delta) \supseteq S$ and, for every $\alpha \in \mathrm{Pat}_\Sigma$ with $L_{\mathrm{NE},\Sigma}(\alpha) \supseteq S$, $L_{\mathrm{NE},\Sigma}(\alpha) \not\subset$

$L_{\mathrm{NE},\Sigma}(\delta)$. Analogously, δ is called *E-descriptive (of S)* if $L_{\mathrm{E},\Sigma}(\delta) \supseteq S$ and, for every $\alpha \in \mathrm{Pat}_\Sigma$ with $L_{\mathrm{E},\Sigma}(\alpha) \supseteq S$, $L_{\mathrm{E},\Sigma}(\alpha) \not\subset L_{\mathrm{E},\Sigma}(\delta)$.

Obviously, the definition of a descriptive pattern is based on the inclusion of pattern languages, which is an undecidable problem for both the full class of NE-pattern languages and the full class of E-pattern languages (cf. Jiang et al. [8], Freydenberger and Reidenbach [6]). A significant part of our subsequent technical considerations, however, can be restricted to terminal-free E-pattern languages, and here the inclusion problem is known to be decidable. This directly results from the following characterisation:

Theorem 1 (Jiang et al. [8]). *Let Σ be an alphabet, $|\Sigma| \geq 2$, and let $\alpha, \beta \in X^+$ be terminal-free patterns. Then $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ if and only if there exists a morphism $h : X^* \to X^*$ satisfying $h(\beta) = \alpha$.*

While Theorem 1 is a powerful tool when dealing with the inclusion of terminal-free E-pattern languages, the examination of the descriptiveness of a pattern requires insights into *proper* inclusion relations, and therefore we use some further combinatorial results on morphisms in free monoids to give a more convenient criterion that can replace the use of Theorem 1.

In accordance with Reidenbach and Schneider [12], we designate a terminal-free pattern $\alpha \in X^+$ as *morphically imprimitive* if there is a pattern $\beta \in X^*$ satisfying the following conditions: $|\beta| < |\alpha|$ and there are morphisms $g, h : X^* \to X^*$ such that $g(\alpha) = \beta$ and $h(\beta) = \alpha$. Otherwise, α is *morphically primitive*. Let $\alpha \in X^+$ be morphically primitive. A morphism $h : X^* \to X^*$ is said to be an *imprimitivity morphism (for α)* provided that $|h(\alpha)| > |\alpha|$ and there is a morphism $g : X^* \to X^*$ satisfying $(g \circ h)(\alpha) = \alpha$. Referring to these concepts, we now can give a characterisation of certain proper inclusion relations between terminal-free E-pattern languages:

Lemma 2. *Let Σ be an alphabet, $|\Sigma| \geq 2$, $\alpha \in X^+$ a morphically primitive pattern and $h : X^* \to X^*$ a morphism. Then $L_{\mathrm{E},\Sigma}(h(\alpha)) \subset L_{\mathrm{E},\Sigma}(\alpha)$ if and only if h is neither an imprimitivity morphism for α nor a renaming of α.*

Proof. We firstly consider the *if* direction: If h is neither an imprimitivity morphism for α nor a renaming of α, then $|h(\alpha)| < |\alpha|$ or there is no morphism g mapping $h(\alpha)$ to α. In the latter case, due to Theorem 1, $L_{\mathrm{E},\Sigma}(h(\alpha)) \not\supseteq L_{\mathrm{E},\Sigma}(\alpha)$. In the former case, if there is a morphism g mapping $h(\alpha)$ to α, then α is not morphically primitive, which contradicts the condition of the lemma. Hence, there is no such morphism, and this again implies $L_{\mathrm{E},\Sigma}(h(\alpha)) \not\supseteq L_{\mathrm{E},\Sigma}(\alpha)$. Since Theorem 1 shows that $L_{\mathrm{E},\Sigma}(h(\alpha)) \subseteq L_{\mathrm{E},\Sigma}(\alpha)$, we have $L_{\mathrm{E},\Sigma}(h(\alpha)) \subset L_{\mathrm{E},\Sigma}(\alpha)$.

We proceed with the *only if* direction: If $L_{\mathrm{E},\Sigma}(h(\alpha)) \subset L_{\mathrm{E},\Sigma}(\alpha)$, then there is no morphism mapping $h(\alpha)$ to α. However, the definition of an imprimitivity morphism mapping α to some pattern β implies the existence of a morphism mapping β to α again. The same trivially holds for any renaming of α. Thus, h is neither an imprimitivity morphism for α nor a renaming of α. □

The question of whether a given morphism is an imprimitivity morphism for a pattern can be easily answered using the following insight:

Theorem 3 (Reidenbach and Schneider [12]). *Let $\alpha \in X^+$ be a morphically primitive pattern. Then a morphism $h : X^* \to X^*$ is an imprimitivity morphism for α if and only if*

1. *for every $x \in \mathrm{var}(\alpha)$, there exists an $x_h \in \mathrm{var}(h(x))$ such that $|h(x)|_{x_h} = 1$ and $|h(y)|_{x_h} = 0$ for every $y \in \mathrm{var}(\alpha) \setminus \{x\}$, and*
2. *there exists an $x \in \mathrm{var}(\alpha)$ with $|h(x)| \geq 2$.*

Evidently, Lemma 2 can only be applied if there is a tool for checking whether a terminal-free pattern is morphically primitive. This is provided by the following characterisation:

Theorem 4 (Reidenbach and Schneider [12]). *A pattern $\alpha \in X^+$ is morphically primitive if and only if there is no factorisation*

$$\alpha = \beta_0\, \gamma_1\, \beta_1\, \gamma_2\, \beta_2\, \ldots\, \beta_{n-1}\, \gamma_n\, \beta_n$$

with $n \geq 1$, $\beta_k \in X^*$ and $\gamma_k \in X^+$, $k \leq n$, such that

1. *$|\gamma_k| \geq 2$ for every k, $1 \leq k \leq n$,*
2. *$\mathrm{var}(\beta_0 \ldots \beta_n) \cap \mathrm{var}(\gamma_1 \ldots \gamma_n) = \emptyset$,*
3. *for every k, $1 \leq k \leq n$, there exists an $x_k \in \mathrm{var}(\gamma_k)$ such that $|\gamma_k|_{x_k} = 1$ and, for every k', $1 \leq k' \leq n$, if $x_k \in \mathrm{var}(\gamma_{k'})$ then $\gamma_k = \gamma_{k'}$.*

Thus, with Lemma 2, Theorem 3 and Theorem 4 we now have an appropriate tool for deciding on particular proper inclusion relations between terminal-free E-pattern languages.

3. Descriptive Patterns and Infinite Strictly Decreasing Chains of Pattern Languages

Before we state and prove the main results of our paper, we discuss some simple yet enlightening observations that establish a connection between descriptiveness of patterns and infinite strictly decreasing chains of pattern languages over some fixed alphabet, i. e. sequences $(L_i)_{i \in \mathbb{N}}$ of pattern languages satisfying, for every $j \in \mathbb{N}$, $L_j \supset L_{j+1}$. This aspect is already briefly mentioned by Jiang et al. [7].

Since, by definition, a descriptive pattern generates a smallest pattern language comprising a language S, S does not have a descriptive pattern if and only if no pattern language L satisfying $L \supseteq S$ is smallest. Hence, the existence of a descriptive pattern essentially depends on the existence of a pattern language that is not contained in an infinite strictly decreasing chain:

Observation 5. *Let Σ be an alphabet and $S \subseteq \Sigma^*$ a language. Then there is no pattern that is NE-descriptive (or E-descriptive) of S if and only if, for every pattern α with $L_{\mathrm{NE},\Sigma}(\alpha) \supseteq S$ (or $L_{\mathrm{E},\Sigma}(\alpha) \supseteq S$, respectively) there is*

- *a sequence of patterns $\alpha_i \in \mathrm{Pat}_\Sigma$, $i \in \mathbb{N}$, satisfying, for every $j \in \mathbb{N}$,*

  - $L_{\mathrm{NE},\Sigma}(\alpha_j) \supset L_{\mathrm{NE},\Sigma}(\alpha_{j+1})$ *(or* $L_{\mathrm{E},\Sigma}(\alpha_j) \supset L_{\mathrm{E},\Sigma}(\alpha_{j+1})$*, respectively) and*
  - $L_{\mathrm{NE},\Sigma}(\alpha_j) \supseteq S$ *(or* $L_{\mathrm{E},\Sigma}(\alpha_j) \supseteq S$*, respectively)*

  and

- *an* $n \in \mathbb{N}$ *with* $L_{\mathrm{NE},\Sigma}(\alpha_n) = L_{\mathrm{NE},\Sigma}(\alpha)$ *(or* $L_{\mathrm{E},\Sigma}(\alpha_n) = L_{\mathrm{E},\Sigma}(\alpha)$*, respectively).*

Proof. Directly from the definition of an NE-descriptive (or E-descriptive) pattern. □

Consequently, the question of whether there is a descriptive pattern for a language S requires insights into the inclusion problem for pattern languages. As partly stated in Section 2, this problem is undecidable in the general case, but it is decidable for the class of terminal-free E-pattern languages (though combinatorially complex and, according to Ehrenfeucht and Rozenberg [4], NP-complete).

In order to illustrate and substantiate Observation 5 and as a reference for further considerations in Section 4, we now give some examples of strictly decreasing chains of pattern languages. We begin with a sequence of patterns that has almost identical properties for both NE- and E-pattern languages:

Example 6. Let Σ be any alphabet. For every $i \in \mathbb{N}$, we define $\alpha_i := x_1^{2^i}$, i. e. $\alpha_0 = x_1$, $\alpha_1 = x_1^2$, $\alpha_2 = x_1^4$, $\alpha_3 = x_1^8$ and so on. It can be easily seen that, for every $j \in \mathbb{N}$, there is a morphism $h : \{x_1\}^+ \to \{x_1\}^+$, defined by $h(x_1) := x_1^2$, satisfying $h(\alpha_j) = \alpha_{j+1}$. Since, for both NE- and E-pattern languages, the existence of such a morphism is a sufficient condition for an inclusion relation (cf. Lemma 3.1 by Angluin [1] and Theorem 2.3 by Jiang et al. [7], respectively), $L_{\mathrm{NE},\Sigma}(\alpha_j) \supseteq L_{\mathrm{NE},\Sigma}(\alpha_{j+1})$ and $L_{\mathrm{E},\Sigma}(\alpha_j) \supseteq L_{\mathrm{E},\Sigma}(\alpha_{j+1})$ are satisfied. In the given example, it is evident that all inclusions of NE-pattern languages are strict. The same holds for the inclusion of E-pattern languages; alternatively, for all but unary alphabets Σ, it is directly proven by Lemma 2 (using Theorem 3 and Theorem 4) given in Section 2. Hence, the sequence of α_i leads to an infinite strictly decreasing chain for NE-pattern languages as well as for E-pattern languages. Nevertheless, the sequence of patterns is irrelevant in the context of Observation 5, as the sets $S_{\mathrm{NE}} := \bigcap_{i=0}^{\infty} L_{\mathrm{NE},\Sigma}(\alpha_i)$ and $S_{\mathrm{E}} := \bigcap_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\alpha_i)$, i. e. those languages all patterns are consistent with, satisfy $S_{\mathrm{NE}} = \emptyset$ and $S_{\mathrm{E}} = \{\lambda\}$.

Our next example looks quite similar to Example 6, but here a difference between NE- and E-pattern languages can be noted:

Example 7. Let Σ be an alphabet with $|\Sigma| \geq 2$. For every $i \in \mathbb{N}$, we define $\alpha_i := x_1^{2^i} y^2$, i. e. $\alpha_0 = x_1 y^2$, $\alpha_1 = x_1^2 y^2$, $\alpha_2 = x_1^4 y^2$, $\alpha_3 = x_1^8 y^2$ and so on. Referring to the same facts as mentioned in Example 6, it can be shown that the patterns again define one infinite strictly decreasing chain of NE-pattern languages and another one of E-pattern languages. However, while

the set $S_{\mathrm{NE}} := \bigcap_{i=0}^{\infty} L_{\mathrm{NE},\Sigma}(\alpha_i)$ again is empty, $S_{\mathrm{E}} := \bigcap_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\alpha_i)$ now equals $L_{\mathrm{E},\Sigma}(y^2)$. Hence, we have a chain of E-pattern languages that are all consistent with a nontrivial language. Nevertheless, $L_{\mathrm{E},\Sigma}(y^2)$ obviously has a descriptive pattern, namely $\delta := y^2$, and this of course holds for all infinite sequences of patterns where S_{E} equals an E-pattern language. Consequently, the existence of a single infinite strictly decreasing chain of E-pattern languages L_i satisfying, for every $i \in \mathbb{N}$, $L_i \supseteq S$, does not mean that there is no E-descriptive pattern for S. Furthermore, it is worth mentioning that we can replace S_{E} with a finite language and still preserve the above described properties of the α_i and δ. For $\Sigma \supseteq \{\mathtt{a}, \mathtt{b}\}$, this is demonstrated, e. g., by the language $S := \{\mathtt{aa}, \mathtt{bb}\}$, which satisfies, for every $i \in \mathbb{N}$, $S \subseteq L_{\mathrm{E},\Sigma}(\alpha_i)$ and has the E-descriptive pattern δ.

Our final example presents a special phenomenon of E-pattern languages, namely the existence of bi-infinite strictly decreasing/increasing chains of such languages:

Example 8. Let Σ be an alphabet with $|\Sigma| \geq 2$. For every $i \in \mathbb{Z}$, we define

$$\alpha_i := \begin{cases} x_1^{2^{-i}} & \text{if } i \text{ is negative,} \\ x_1^2 x_2^2 \ \ldots \ x_{i+2}^2 & \text{else.} \end{cases}$$

Hence, for example, from $i = -3$ to $i = 2$ the patterns read $\alpha_{-3} = x_1^8$, $\alpha_{-2} = x_1^4$, $\alpha_{-1} = x_1^2$, $\alpha_0 = x_1^2 x_2^2$, $\alpha_1 = x_1^2 x_2^2 x_3^2$, and $\alpha_2 = x_1^2 x_2^2 x_3^2 x_4^2$. Using Theorem 4, it is easy to show that all patterns are morphically primitive. Theorem 3 demonstrates that all morphisms mapping an α_k to an α_j, $j < k$, are not imprimitivity morphisms. Therefore we can conclude from Lemma 2 that $L_{\mathrm{E},\Sigma}(\alpha_j) \subset L_{\mathrm{E},\Sigma}(\alpha_k)$ if and only if $j < k$. For the given patterns, $S_{\mathrm{E}} := \bigcap_{i=-\infty}^{\infty} L_{\mathrm{E},\Sigma}(\alpha_i)$ equals $\{\lambda\}$, but if we define, for every $i \in \mathbb{Z}$, $\alpha'_i := y^2 \alpha_i$, then these α'_i generate a bi-infinite strictly decreasing/increasing chain of E-pattern languages where $S_{\mathrm{E}} := \bigcap_{i=-\infty}^{\infty} L_{\mathrm{E},\Sigma}(\alpha'_i) = L_{\mathrm{E},\Sigma}(y^2)$ is an E-pattern language.

Note that the example patterns given above are terminal-free merely for the sake of convenience. They can be effortlessly turned into certain patterns containing terminal symbols and still showing equivalent properties.

4. The Existence of Descriptive Patterns

In the present chapter we study the existence of patterns that are descriptive of sets S of strings. According to our remarks in Section 1, four main cases can be considered, depending on whether S is finite or infinite and whether NE- or E-descriptiveness is examined. We focus on the existence of E-descriptive patterns for infinite languages since, for the other three cases, answers are absolutely straightforward or directly or indirectly provided by Angluin [1] and Jiang et al. [7]. In order to give a comprehensive description and further explain some of our formal concepts and statements we nevertheless also briefly describe the known or trivial cases.

Using Observation 5, the question of the existence of *NE-descriptive* patterns can be easily answered for all types of languages S. We begin with the case of a *finite* S. Here, it is primarily necessary to observe that a word w can only be covered by a pattern α through nonerasing substitutions if α is not longer than w. Hence, for any finite alphabet Σ and any word over Σ, there are only finitely many NE-pattern languages over Σ covering this word; this property of a class of languages is commonly referred to as *finite thickness* (cf. Wright [17]). Quite obviously, the same holds for infinite alphabets Σ, since the number of different terminal symbols that can occur in patterns covering w is limited by the number of different terminal symbols in w. With regard to infinite sequences of patterns (generating languages that all differ from each other) over a fixed alphabet, this means that none of them can contain infinitely many patterns that cover, e. g., the shortest word in a given finite set of strings. This immediately shows that, for every finite S, there exists an NE-descriptive pattern:

Proposition 9 (Angluin [1]). *Let Σ be an alphabet and $S \subseteq \Sigma^+$ a finite language. Then there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is NE-descriptive of S.*

Note that Angluin [1] does not explicitly state Proposition 9, but directly studies more challenging questions by introducing a procedure computing an NE-descriptive pattern for any finite language S and examining the computational complexity of the problem of finding such patterns for finite languages.

With regard to NE-descriptive patterns for *infinite* languages S, the same reasoning as for finite languages S leads to the analogous result:

Proposition 10. *Let Σ be an alphabet and $S \subseteq \Sigma^+$ an infinite language. Then there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is NE-descriptive of S.*

Proof. Directly from Observation 5 and the finite thickness of the class of NE-pattern languages. □

A closer look at the underlying reasoning proving Propositions 9 and 10 reveals that it does not need to consider whether any infinite sequence of patterns leads to an infinite strictly decreasing chain of NE-pattern languages (as featured by Observation 5), but can be completely based on the concept of finite thickness. If we nevertheless wish to examine the properties of such chains, then we can easily observe that, for every sequence of patterns α_i, $i \in \mathbb{N}$, with $L_{\mathrm{NE},\Sigma}(\alpha_i) \supset L_{\mathrm{NE},\Sigma}(\alpha_{i+1})$, the set $S_{\mathrm{NE}} := \bigcap_{i=0}^{\infty} L_{\mathrm{NE},\Sigma}(\alpha_i)$ necessarily is empty. Hence, Examples 6 and 7 illustrate the only option possible.

With regard to *E-descriptiveness*, the situation is more complex. As shown by Examples 7 and 8, the class of E-pattern languages does not have finite thickness and there are even finite and infinite languages that are contained in all E-pattern languages of an infinite strictly decreasing chain. Nevertheless, it is known that every nontrivial *finite* language has an E-descriptive pattern:

Theorem 11 (Jiang et al. [7]). *Let Σ be an alphabet and $S \subseteq \Sigma^*$ a finite language, $S \neq \{\lambda\}$. Then there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is E-descriptive of S.*

The proof for Theorem 11 given by Jiang et al. [7] demonstrates that for every finite language S an upper bound n can be given such that, for every pattern α consistent with S, there exists a pattern β satisfying $|\beta| \leq n$ and $S \subseteq L_{\mathrm{E},\Sigma}(\beta) \subseteq L_{\mathrm{E},\Sigma}(\alpha)$. So if, for any finite S, there is a sequence of patterns α_i, $i \in \mathbb{N}$, leading to an infinite strictly decreasing chain of E-pattern languages comprising S – which implies that there is no upper bound for the length of the α_i – then all but finitely many of these patterns need to have variables that are not required for generating the words in S. This phenomenon is illustrated by Example 7, where only the subpattern y^2 of all patterns is necessary in order to map the patterns to the words in S_{E}.

In the proof for Theorem 11, the upper bound n equals the sum of the lengths of the words in S. Thus, this method cannot be adopted when investigating the existence of E-descriptive patterns for *infinite* sets of words. In fact, as to be demonstrated below, we here need to consider two subcases depending on the number of different letters occurring in the words of S. If the underlying alphabet is unary, then the descriptiveness of a pattern is related to the inclusion relation of E-pattern languages over this unary alphabet. The structure of such E-pattern languages, however, is significantly simpler than that of E-pattern languages over larger alphabets; in particular, the full class of these languages is a specific subclass of the regular languages (namely the linear unary languages). Therefore it can be shown that, for every sequence of patterns $(\alpha_i)_{i\in\mathbb{N}}$ leading to an infinite strictly decreasing chain of E-pattern languages over a unary alphabet, the language $S_{\mathrm{E}} := \bigcap_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\alpha_i)$ is finite. Referring to Observation 5, this directly leads to the following result:

Theorem 12. *Let Σ be an alphabet, $|\Sigma| = 1$, and $S \subseteq \Sigma^*$ an infinite language. Then there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is E-descriptive of S.*

The proof for Theorem 12 is given in Section 5.1.

In contrast to this, Example 7 demonstrates that, for alphabets with at least two letters, there is an infinite strictly decreasing chain of E-pattern languages such that the intersection of all these languages is infinite. Since this intersection is an E-pattern language, Example 7 can nevertheless not be used to establish a result that differs from those given for the other cases. In order to answer the question of whether this holds true for all such chains, we now consider a more sophisticated infinite sequence of patterns, that is defined as follows:

Definition 13. We define the pattern

$$\alpha_0 := y^2 z^2$$

and the morphism $\phi : X^* \to X^*$ (note that we assume $X \supseteq \{y, z, x_0, x_1, x_2 \ldots\}$) through, for every $i \in \mathbb{N}$,

$$\begin{aligned} \phi(x_i) &:= x_{i+1}, \\ \phi(y) &:= y^2 x_1, \\ \phi(z) &:= x_1 z^2. \end{aligned}$$

Then, for every $i \in \mathbb{N}$, the pattern α_{i+1} is given by

$$\alpha_{i+1} := \phi(\alpha_i) = \phi^i(\alpha_0).$$

This means that, for example,

$$\begin{aligned}
\alpha_1 &= y^2x_1\, y^2x_1\, x_1z^2\, x_1z^2,\\
\alpha_2 &= (y^2x_1y^2x_1x_2)\,(y^2x_1y^2x_1x_2)\,(x_2x_1\, z^2x_1z^2)\,(x_2x_1z^2x_1z^2),\\
\alpha_3 &= (y^2x_1y^2x_1x_2\; y^2x_1y^2x_1x_2\; x_3)\,(y^2x_1y^2x_1x_2\; y^2x_1y^2x_1x_2\; x_3)\\
&\quad\; (x_3\; x_2x_1z^2x_1z^2\; x_2x_1z^2x_1z^2)\,(x_3\; x_2x_1z^2x_1z^2\; x_2x_1z^2x_1z^2).
\end{aligned}$$

It can be shown that this sequence $(\alpha_i)_{i\in\mathbb{N}}$ defines an infinite strictly decreasing chain of E-pattern languages. Furthermore, if we define the morphism $\psi : X^* \to X^*$ through $\psi(x_i) := x_i$ and $\psi(y) := \psi(z) := x_0$, then, for every alphabet Σ with $|\Sigma| \geq 2$, $L_\Sigma := \bigcup_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\psi(\alpha_i))$ satisfies $L_\Sigma \subseteq \bigcap_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\alpha_i)$. As a side note, it is worth mentioning that L_Σ is a multi-pattern language (cf. Dumitrescu et al. [3]) where the set $\{\psi(\alpha_i) \mid i \in \mathbb{N}\}$ of generating patterns is defined similarly to an HD0L language (albeit over an infinite alphabet of variables); such a concept has not been considered by previous literature. Finally, it can be demonstrated that the sequence $(\alpha_i)_{i\in\mathbb{N}}$ has a very particular property, since for every pattern γ with $L_{\mathrm{E},\Sigma}(\gamma) \supseteq L_\Sigma$ there exists an α_i satisfying $L_{\mathrm{E},\Sigma}(\gamma) \supseteq L_{\mathrm{E},\Sigma}(\alpha_i)$. Referring to Observation 5, this implies the main result of our paper:

Theorem 14. *For every alphabet Σ with $|\Sigma| \geq 2$ there is an infinite language $L_\Sigma \subset \Sigma^*$ that has no E-descriptive pattern.*

The proof for Theorem 14 is given in Section 5.2.

Consequently, when searching for descriptive patterns, the case of E-descriptive patterns of infinite languages over alphabets of at least two letters is the only one where the existence of such patterns is not always guaranteed. This directly answers a question posed by Jiang et al. [7].

Finally, it can be shown that, while the proof of Theorem 14 is based on the particular shape of the infinite union L_Σ of E-pattern languages described above, L_Σ can be replaced by a language L_Σ^t which, for every pattern $\psi(\alpha_i)$, $i \geq 0$, contains just a single word. In order to describe this insight more precisely, we have to introduce the following concept:

Definition 15. A language L is called *properly thin* if, for every $n \geq 0$, L contains at most one word of length n.

Referring to this definition, we can strengthen Theorem 14 as follows:

Corollary 16. *For every alphabet Σ with $|\Sigma| \geq 2$, there is an infinite properly thin language $L_\Sigma^t \subset \Sigma^*$ that has no E-descriptive pattern.*

The proof for Corollary 16 is given in Section 5.3.

5. Proof of the Major Theorems

The present section contains the proofs of the major theorems given in this paper.

5.1. Proof of Theorem 12

Before we give the actual proof of Theorem 12, we introduce some concepts that are only relevant to this section.

To begin with, we extend the operations addition, subtraction, multiplication and division from the natural numbers to operations on natural numbers with sets of natural numbers in the canonical way; i. e., for $\star \in \{+, -, \cdot, /\}$ and $M \subseteq \mathbb{N}$, $b \in \mathbb{N}$ let $M \star b := \{m \star b \mid m \in M\}$. Note that in all cases where we use division or subtraction, the results will always be natural numbers; furthermore, we make free use of the commutativity of multiplication and addition and write $b + M$ or $b \cdot M$ instead of $M + b$ or $M \cdot b$, respectively. For any (possibly infinite) $M \subseteq \mathbb{N}$, let $\gcd(M)$ denote the *greatest common divisor* of all elements of M.

Let $n \geq 1$ and $M = \{m_1, \ldots, m_n\} \subset \mathbb{N}_1$. We define the *linear hull of* M as $\mathrm{lin}(M) := \{m \mid m = k_1 m_1 + \ldots + k_n m_n \text{ for some } k_1, \ldots, k_n \in \mathbb{N}\}$, and $\mathrm{lin}(\emptyset) := \{0\}$.

It is obvious that every unary language L is isomorphic to its Parikh set $P(L) := \{|w| \mid w \in L\} \subseteq \mathbb{N}$. We say that a unary language L is *linear* if there is a $b \geq 0$ and a finite set $G \subset \mathbb{N}$ such that $P(L) = b + \mathrm{lin}(G)$. This allows us to state the following observation on unary pattern languages:

Proposition 17. *A unary language is linear if and only if it is a pattern language.*

Proof. Let $\Sigma = \{\mathtt{a}\}$. We begin with the *if* direction. Let $\alpha \in \mathrm{Pat}_\Sigma$ with $\mathrm{var}(\alpha) = \{x_1, \ldots, x_n\}$ for some $n \geq 0$. Let $b := |\alpha|_\mathtt{a}$ and, for $1 \leq i \leq n$, $g_i := |\alpha|_{x_i}$; furthermore, we define $\beta := \mathtt{a}^b x_1^{g_1} \ldots x_n^{g_n}$. As Σ is unary, $L_{\mathrm{E},\Sigma}(\alpha) = L_{\mathrm{E},\Sigma}(\beta)$ holds, and it is easy to see that $P(L_{\mathrm{E},\Sigma}(\beta)) = b + \mathrm{lin}(\{g_1, \ldots, g_n\})$.

Conversely, if some language $L \subseteq \Sigma^*$ is linear, then there exist a $b \geq 0$ and a finite set $G = \{g_1, \ldots, g_n\} \supset \mathbb{N}$ (with $n \geq 0$) satisfying $P(L) = b + \mathrm{lin}(G)$. If we define β as above, $P(L_{\mathrm{E},\Sigma}(\beta)) = b + \mathrm{lin}(G) = P(L)$ leads to $L_{\mathrm{E},\Sigma}(\beta) = L$. □

Also, note this important fact on linear hulls:

Lemma 18. *For every finite $M \subset \mathbb{N}$, there exists an $n \geq 1$ with $\mathrm{lin}(M) \supseteq \gcd(M) \cdot \mathbb{N}_n$.*

Proof. The case of $\gcd(M) = 1$ is well known, a proof can be found in Chapter 3.15 of Wilf [16]. If $\gcd(M) > 1$, let $M' := M/\gcd(M)$. Then, as $\gcd(M') = 1$, there is an $n \geq 1$ such that $\mathrm{lin}(M') \supseteq \mathbb{N}_n$, and therefore, $\mathrm{lin}(M) = \gcd(M) \cdot \mathrm{lin}(M') \supseteq \gcd(M) \cdot \mathbb{N}_n$. □

Now that all necessary tools have been introduced, we are ready for the proof of Theorem 12:

PROOF. Let $\Sigma := \{\mathtt{a}\}$. Furthermore, let

$$\begin{aligned} b &:= \min(P(S)), \\ P'_S &:= P(S) - b, \\ g &:= \gcd(P'_S), \\ P''_S &:= P'_S/g \end{aligned}$$

and $\alpha := \mathtt{a}^b x_1^g$. It is easy to verify that $L_{\mathrm{E},\Sigma}(\alpha) \supseteq S$, $P(L_{\mathrm{E},\Sigma}(\alpha)) = b + g \cdot \mathbb{N}$ and $P(S) = b + g \cdot P''_S$. Although α is not necessarily E-descriptive of S, we shall see that there is always only a finite number of pattern languages between $L_{\mathrm{E},\Sigma}(\alpha)$ and S.

Since Σ is unary, we have, for every pattern $\beta \in \mathrm{Pat}_\Sigma$ with $L_{\mathrm{E},\Sigma}(\alpha) \supset L_{\mathrm{E},\Sigma}(\beta) \supseteq S$,

$$P(L_{\mathrm{E},\Sigma}(\alpha)) \supset P(L_{\mathrm{E},\Sigma}(\beta)) \supseteq P(S).$$

This, in turn, is equivalent to

$$b + g \cdot \mathbb{N} \supset P(L_{\mathrm{E},\Sigma}(\beta)) \supseteq b + g \cdot P''_S.$$

Due to this relation and Proposition 17, we can conclude with some effort that there is a finite $G_\beta \supset \mathbb{N}$ with $P(L_{\mathrm{E},\Sigma}(\beta)) = b + g \cdot \mathrm{lin}(G_\beta)$. Therefore,

$$b + g \cdot \mathbb{N} \supset b + g \cdot \mathrm{lin}(G_\beta) \supseteq b + g \cdot P''_S,$$

which is equivalent to

$$\mathbb{N} \supset \mathrm{lin}(G_\beta) \supseteq P''_S.$$

As $\gcd(P''_S) = 1$, there is a finite $C_S \subset P''_S$ with $\gcd(C_S) = 1$. We observe that

$$\mathrm{lin}(G_\beta) \supseteq P''_S \supset C_S,$$

and, as C_S is a finite subset of $\mathrm{lin}(G_\beta)$,

$$\mathrm{lin}(G_\beta) \supseteq \mathrm{lin}(C_S).$$

Due to Lemma 18, there is an $n \geq 0$ such that $\mathrm{lin}(C_S) \supseteq \mathbb{N}_n$, and thus, $\mathrm{lin}(G_\beta) \supseteq \mathbb{N}_n$, which leads to $P(L_{\mathrm{E},\Sigma}(\beta)) \supseteq b + g \cdot \mathbb{N}_n$.

Now, assume that there is an infinite sequence $(\beta_i)_{i \geq 0}$ over Pat_Σ such that $L_{\mathrm{E},\Sigma}(\alpha) \supset L_{\mathrm{E},\Sigma}(\beta_i) \supset L_{\mathrm{E},\Sigma}(\beta_{i+1}) \supset S$ for every $i \geq 0$. Then there is an infinite sequence $(G_{\beta_i})_{i \geq 0}$ of finite subsets of $\mathbb{N}$ with, for every $i \geq 0$, $P(L_{\mathrm{E},\Sigma}(\beta_i)) = b + g \cdot \mathrm{lin}(G_{\beta_i})$ and $\mathrm{lin}(G_{\beta_i}) \supset \mathrm{lin}(G_{\beta_{i+1}}) \supset \mathbb{N}_n$. As $\mathbb{N}_n$ is cofinite, such an infinite sequence cannot exist – therefore, due to Observation 5, there must be some pattern that is E-descriptive of S. □

5.2. Proof of Theorem 14

In order to prove Theorem 14, we define L_Σ through the infinite sequence of patterns α_i, $i \in \mathbb{N}$, given by Definition 13 in such a way that the words of L_Σ are structurally so close to the patterns α_i that, for every pattern $\delta \in \mathrm{Pat}_\Sigma$

with $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_\Sigma$, there is a $j \in \mathbb{N}$ with $L_{\mathrm{E},\Sigma}(\delta) \supset L_{\mathrm{E},\Sigma}(\alpha_j) \supset L_\Sigma$. Thus, regardless of how closely $L_{\mathrm{E},\Sigma}(\delta)$ approximates L_Σ, there is always an α_j that provides a better description of L_Σ.

Before we define L_Σ, we wish to support our subsequent reasoning by stating that, for sufficiently large $i \in \mathbb{N}$, the patterns α_i introduced in Definition 13 can be given as follows:

$$\begin{aligned}\alpha_i &= (((((\ldots((((y^2x_1)^2x_2)^2x_3)^2x_4)^2\ldots x_{i-4})^2x_{i-3})^2x_{i-2})^2x_{i-1})^2x_i)^2\\ &\quad (x_i(x_{i-1}(x_{i-2}(x_{i-3}(x_{i-4}\ldots(x_4(x_3(x_2(x_1z^2)^2)^2)^2)^2\ldots)^2)^2)^2)^2)^2.\end{aligned}$$

As briefly introduced in Section 4, L_Σ is an infinite union of E-pattern languages. The corresponding patterns are derived from the patterns α_i by a morphism $\psi : X^* \to X^*$, defined through

$$\begin{aligned}\psi(x_i) &:= x_i,\\ \psi(y) := \psi(z) &:= x_0.\end{aligned}$$

Applying ψ to the patterns α_i, we receive an infinite sequence of patterns $(\beta_i)_{i\geq 0}$; i. e., we define, for every $i \geq 0$, $\beta_i := \psi(\alpha_i)$. As the rather simple structure of ψ suggests, any pattern β_j, $j \in \mathbb{N}$, is structurally very close to the patterns α_j, since the only difference is that both y and z are replaced by the variable x_0:

$$\begin{aligned}\beta_0 &= x_0^2\,x_0^2,\\ \beta_1 &= x_0^2x_1\,x_0^2x_1\,x_1x_0^2\,x_1x_0^2,\\ \beta_2 &= (x_0^2x_1x_0^2x_1x_2)\,(x_0^2x_1x_0^2x_1x_2)\,(x_2x_1\,x_0^2x_1x_0^2)\,(x_2x_1x_0^2x_1x_0^2),\\ \beta_3 &= (x_0^2x_1x_0^2x_1x_2\,x_0^2x_1x_0^2x_1x_2\,x_3)\,(x_0^2x_1x_0^2x_1x_2\,x_0^2x_1x_0^2x_1x_2\,x_3)\\ &\quad (x_3\,x_2x_1x_0^2x_1x_0^2\,x_2x_1x_0^2x_1x_0^2)\,(x_3\,x_2x_1x_0^2x_1x_0^2\,x_2x_1x_0^2x_1x_0^2),\\ &\vdots\\ \beta_i &= (((((\ldots((((x_0^2x_1)^2x_2)^2x_3)^2x_4)^2\ldots x_{i-4})^2x_{i-3})^2x_{i-2})^2x_{i-1})^2x_i)^2\\ &\quad (x_i(x_{i-1}(x_{i-2}(x_{i-3}(x_{i-4}\ldots(x_4(x_3(x_2(x_1x_0^2)^2)^2)^2\ldots)^2)^2)^2)^2)^2.\end{aligned}$$

Finally, for any alphabet Σ with $|\Sigma| \geq 2$, we define $L_\Sigma := \bigcup_{i=0}^{\infty} L_{\mathrm{E},\Sigma}(\beta_i)$.

The relation between the patterns β_i can again be expressed by a morphism, namely $\mu : X^* \to X^*$ given by

$$\begin{aligned}\mu(x_i) &:= \begin{cases}\lambda & \text{if } i = 0,\\ x_{i-1} & \text{if } i > 0,\end{cases}\\ \mu(y) &:= y,\\ \mu(z) &:= z.\end{aligned}$$

Figuratively speaking, the morphism μ permits us to move downward in the sequence $(\beta_i)_{i\geq 0}$ (note that μ is given for the variables y and z due to technical reasons arising later). This is illustrated by Figure 1 and further substantiated by the following lemma:

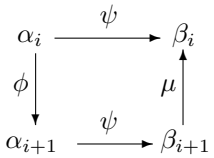

Figure 1: Morphic relations between the elements of the sequences $(\alpha_i)_{i\geq 0}$ and $(\beta_i)_{i\geq 0}$.

Lemma 19. *For all $i, j \geq 0$, $\mu^j(\beta_{i+j}) = \beta_i$.*

Proof. If $j = 0$, the claim is trivially true. We now consider $j = 1$. By definition, $\mu(\beta_{i+1}) = (\mu \circ \psi \circ \phi)(\alpha_i)$. The morphism $\mu \circ \psi \circ \phi : X^* \to X^*$ works as follows:

$$\begin{aligned}(\mu \circ \psi \circ \phi)(x) &= \begin{cases}(\mu \circ \psi)(x_{k+1}) & \text{if } x = x_k,\\ (\mu \circ \psi)((y)^2 x_1) & \text{if } x = y,\\ (\mu \circ \psi)(x_1 (z)^2) & \text{if } x = z\end{cases}\\ &= \begin{cases}\mu(x_{k+1}) & \text{if } x = x_k,\\ \mu((x_0)^2 x_1) & \text{if } x = y,\\ \mu(x_1 (x_0)^2) & \text{if } x = z\end{cases}\\ &= \begin{cases}x_k & \text{if } x = x_k,\\ x_0 & \text{if } x = y \text{ or } x = z\end{cases}\\ &= \psi(x).\end{aligned}$$

Therefore, $\mu(\beta_{i+1}) = \psi(\alpha_i) = \beta_i$. For all larger values of j, the claim holds by induction. □

Referring to Theorem 1, Figure 1 already illustrates certain inclusion relations between the languages generated by the patterns α_i and β_j, $i, j \in \mathbb{N}$. The following lemma shows that these inclusions are proper, which in particular means that the patterns in $(\alpha_i)_{i\geq 0}$ lead to a strictly decreasing chain of E-pattern languages (as featured by Observation 5). Additionally, the lemma describes the relation of the given E-pattern languages to L_Σ. A summary of selected inclusion relations is provided by Figure 2.

Lemma 20. *For every $i \in \mathbb{N}$, the following statements hold:*

1. $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\alpha_{i+1}) \supset L_\Sigma$,
2. $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\beta_i)$,
3. $L_{\mathrm{E},\Sigma}(\beta_i) \subset L_{\mathrm{E},\Sigma}(\beta_{i+1}) \subset L_\Sigma$.

Proof. For every $i \geq 0$, the proper inclusion relations $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\alpha_{i+1})$, $L_{\mathrm{E},\Sigma}(\beta_{i+1}) \supset L_{\mathrm{E},\Sigma}(\beta_i)$ and $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\beta_i)$ follow from Lemma 2: By

definition, $\alpha_{i+1} = \phi(\alpha_i)$ and $\beta_i = \psi(\alpha_i)$, and, due to Lemma 19, $\beta_i = \mu(\beta_{i+1})$. Furthermore, the following claim holds true:

Claim. For every $i \in \mathbb{N}$, the patterns α_i and β_i are morphically primitive.

Proof of Claim. According to Theorem 4, every morphically imprimitive pattern γ must – among other requirements that need to be satisfied – contain at least one variable that, for each of its occurrences in γ, has the same left neighbours or the same right neighbours. More formally, there must be an $x \in \mathrm{var}(\gamma)$ such that there exists a factorisation

$$\gamma = \widehat{\gamma}_1 \, \chi_{x,L} \, x \, \chi_{x,R} \, \widehat{\gamma}_2 \, \chi_{x,L} \, x \, \chi_{x,R} \, \widehat{\gamma}_3 \, \cdots \, \widehat{\gamma}_{n-1} \, \chi_{x,L} \, x \, \chi_{x,R} \, \widehat{\gamma}_n$$

with $n \geq 2$, $\chi_{x,L}, \chi_{x,R}, \widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_n \in X^* \setminus \{x\}$ and $\chi_{x,L} \neq \lambda$ or $\chi_{x,R} \neq \lambda$.

If we now consider any pattern α_i, $i \in \mathbb{N}$, then neither y nor z nor x_i can have that property, because they have squared occurrences. More precisely, for $x \in \{y, z, x_i\}$, $\alpha_i = \ldots xx \ldots$, which due to $\chi_{x,L}, \chi_{x,R} \in X^* \setminus \{x\}$ implies $\chi_{x,L} = \lambda$ and $\chi_{x,R} = \lambda$. For every $x_j \in \mathrm{var}(\alpha_i) \setminus \{y, z, x_i\}$, $\alpha_i = \ldots x_j x_{j+1} \ldots$ and $\alpha_i = \ldots x_j y \ldots$, and this means that $\chi_{x_j,R} = \lambda$. Furthermore, for every such x_j, α_i satisfies $\alpha_i = \ldots x_{j+1} x_j \ldots$ and $\alpha_i = \ldots z x_j \ldots$, and this implies $\chi_{x_j,L} = \lambda$. In other words, there is no variable in α_i that, for each of its occurrences, has the same left neighbours or the same right neighbours. Consequently, α_i is morphically primitive.

If we substitute x_0 for y and z in the above reasoning, then it shows that every β_i, $i \in \mathbb{N}$, is morphically primitive, too. This proves the correctness of the Claim. □*(Claim)*

Finally, according to Theorem 3, ϕ, ψ and μ are not imprimitivity morphisms for the patterns they are applied to; by definition, none of the morphisms in question is a renaming of any of the patterns involved. Thus, all conditions of Lemma 2 are satisfied, and this directly proves the correctness of our initial statement. In addition to this, these inclusion relations immediately imply $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\beta_j)$ for all $i, j \geq 0$.

For every $i \geq 0$, the inclusion $L_\Sigma \supseteq L_{\mathrm{E},\Sigma}(\beta_i)$ follows from the definition of L_Σ, which in turn immediately leads to $L_\Sigma \neq L_{\mathrm{E},\Sigma}(\beta_i)$, as otherwise $L_{\mathrm{E},\Sigma}(\beta_{i+1}) \supset L_{\mathrm{E},\Sigma}(\beta_i)$ would not be satisfied.

By definition, for every $w \in L_\Sigma$, there is an $i \geq 0$ with $w \in L_{\mathrm{E},\Sigma}(\beta_i)$; and therefore, $w \in L_{\mathrm{E},\Sigma}(\alpha_j)$ for every $j \geq 0$, which implies $L_{\mathrm{E},\Sigma}(\alpha_j) \supseteq L_\Sigma$. Finally, $L_{\mathrm{E},\Sigma}(\alpha_j) = L_\Sigma$ would contradict $L_{\mathrm{E},\Sigma}(\alpha_j) \supset L_{\mathrm{E},\Sigma}(\alpha_{j+1}) \supseteq L_\Sigma$. □

Regarding the possible existence of a pattern δ that is E-descriptive of L_Σ, the language $L_{\mathrm{E},\Sigma}(\delta)$ must, by definition, not be a superlanguage of any of the E-pattern languages in the strictly decreasing chain established by Lemma 20. More precisely, for every pattern $\delta \in \mathrm{Pat}_\Sigma$, if there is an $i \geq 0$ with $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\alpha_i)$, we have

$$L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\alpha_{i+1}) \supset L_\Sigma,$$

which leads to the following lemma:

$$\begin{array}{ccc}
L_{\mathrm{E},\Sigma}(\alpha_0) & \supset & L_{\mathrm{E},\Sigma}(\beta_0) \\
\cup & & \cap \\
L_{\mathrm{E},\Sigma}(\alpha_1) & \supset & L_{\mathrm{E},\Sigma}(\beta_1) \\
\cup & & \cap \\
L_{\mathrm{E},\Sigma}(\alpha_2) & \supset & L_{\mathrm{E},\Sigma}(\beta_2) \\
\cup & & \cap \\
L_{\mathrm{E},\Sigma}(\alpha_3) & \supset & L_{\mathrm{E},\Sigma}(\beta_3) \\
\cup & & \cap \\
L_{\mathrm{E},\Sigma}(\alpha_4) & \supset & L_{\mathrm{E},\Sigma}(\beta_4) \\
\cup & & \cap \\
\vdots & & \vdots
\end{array}$$

Figure 2: Inclusion relations between the E-pattern languages of α_i and β_j, $i, j \geq 0$.

Lemma 21. *If $\delta \in \mathrm{Pat}_\Sigma$ and $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\alpha_i)$ for some $i \geq 0$, then δ is not E-descriptive of L_Σ.*

Therefore, although the language that is generated by a pattern that is E-descriptive of L_Σ (if any) has to contain every language $L_{\mathrm{E},\Sigma}(\beta_i)$, it may not contain any single language $L_{\mathrm{E},\Sigma}(\alpha_i)$. The main idea of our construction is that this requirement is inherently contradictory, as we shall see that whenever a pattern δ can generate every language $L_{\mathrm{E},\Sigma}(\beta_i)$, then δ can generate almost all of the languages $L_{\mathrm{E},\Sigma}(\alpha_i)$ as well.

We now assume to the contrary that there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is E-descriptive of L_Σ. As $\lambda \in L_\Sigma \subseteq L_{\mathrm{E},\Sigma}(\delta)$, δ cannot contain any terminals. Therefore, Theorem 1 permits us to describe all relevant inclusion relations through morphisms.

According to Theorem 1, for every $i \in \mathbb{N}$, there is a morphism $\theta_i : X^* \to X^*$ such that $\theta_i(\delta) = \beta_i$, since $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\beta_i)$ holds by definition. We now choose an infinite sequence of morphisms $(\theta_i)_{i \geq 0}$ such that for every $i \geq 0$,

1. $\theta_i(\delta) = \beta_i$, and
2. θ_i erases as many variables of δ as possible; i. e., for every morphism ρ with $\rho(\delta) = \theta_i(\delta) = \beta_i$,

$$|\{x \in \mathrm{var}(\delta) \mid \rho(x) = \lambda\}| \leq |\{x \in \mathrm{var}(\delta) \mid \theta_i(x) = \lambda\}|.$$

Such a sequence must exist, as $\mathrm{var}(\delta)$ is finite. Furthermore, we choose integers m, n such that θ_m and θ_{m+n} erase exactly the same variables of δ; i. e., for all $x \in \mathrm{var}(\delta)$, $\theta_m(x) = \lambda$ if and only if $\theta_{m+n}(x) = \lambda$. Again, this is possible due to $\mathrm{var}(\delta)$ being finite. Due to technical reasons and without loss of generality, we assume $m, n \geq 2$.

As we shall see, this choice allows us to modify θ_{m+n} in such a way that the resulting morphism maps δ to α_{m+1}, which (according to Lemma 21) leads to the desired contradiction. Our modification mostly targets those variables

in $\mathrm{var}(\delta)$ that contain occurrences of x_{n-1} in their images under θ_{m+n}. To this end, we define

$$\widehat{X} := \{x \in \mathrm{var}(\delta) \mid x_{n-1} \in \mathrm{var}(\theta_{m+n}(x))\},$$
$$\widehat{X}_L := \{x \in \widehat{X} \mid \theta_{m+n}(x) \text{ contains } x_{n-2}x_{n-1},\ x_{n-1}x_n \text{ or } x_{n-1}x_0 \text{ as a factor}\},$$
$$\widehat{X}_R := \{x \in \widehat{X} \mid \theta_{m+n}(x) \text{ contains } x_{n-1}x_{n-2},\ x_nx_{n-1} \text{ or } x_0x_{n-1} \text{ as a factor}\}.$$

In order to construct a well-defined morphism, we need to show that $\widehat{X}_R$ and $\widehat{X}_L$ form a partition of $\widehat{X}$; as we shall see, $\widehat{X}_L$ contains exactly those variables that are mapped to occurrences of x_{n-1} in the left side of β_{m+n}, while $\widehat{X}_R$ contains those variables that are mapped to occurrences on the right side. Then we can use these variables as "anchors" for a modification of θ_{m+n} that permits us to obtain α_{n+1} from δ.

Our corresponding reasoning is based on the following insight:

Lemma 22. *For every $x \in \mathrm{var}(\delta)$, if $\theta_{m+n}(x)$ contains a variable x_i with $i < n$, then $\theta_{m+n}(x)$ also contains a variable x_j with $j \geq n$.*

Proof. To begin with, recall that $\theta_{m+n}(\delta) = \beta_{m+n}$ and $(\mu^n \circ \theta_{m+n})(\delta) = \beta_m$ (cf. Lemma 19). Assume to the contrary that there is an $x \in \mathrm{var}(\delta)$ such that $\mathrm{var}(\theta_{m+n}(x)) \neq \emptyset$ and $\mathrm{var}(\theta_{m+n}(x)) \subseteq \{x_0, \ldots, x_{n-1}\}$. Note that for all $n \geq 0$,

$$\mu^n(x_i) = \begin{cases} \lambda & i < n, \\ x_{i-n} & i \geq n. \end{cases}$$

Therefore, $\mu^n(x_i) = \lambda$ if and only if $i < n$; and thus $(\mu^n \circ \theta_{m+n})(x) = \lambda$.

Moreover, for every $y \in \mathrm{var}(\delta)$, if $\theta_{m+n}(y) = \lambda$, then $(\mu^n \circ \theta_{m+n})(y) = \lambda$. But θ_m and θ_{m+n} erase exactly the same variables of δ. Thus, although $\mu^n \circ \theta_{m+n}$ erases more variables than θ_m, $(\mu^n \circ \theta_{m+n})(\delta) = \beta_m = \theta_m(\delta)$ holds, which is a contradiction to the second criterion in our choice of $(\theta_i)_{i \geq 0}$. □

Note that this implies that, for all $x \in \widehat{X}$, $|\theta_{m+n}(x)| \geq 2$. Now we can prove that $\widehat{X}_L$ and $\widehat{X}_R$ form a partition of $\widehat{X}$:

Lemma 23. *$\widehat{X}_L \cup \widehat{X}_R = \widehat{X}$ and $\widehat{X}_L \cap \widehat{X}_R = \emptyset$.*

Proof. To see that $\widehat{X}_L \cup \widehat{X}_R = \widehat{X}$ must hold, recall the shape of β_{m+n}:

$$\begin{aligned}\beta_{m+n} =& ((\ldots(((\ldots((x_0)^2x_1)^2\ldots x_{n-2})^2x_{n-1})^2x_n)^2\ldots)^2x_{m+n})^2 \\ & (x_{m+n}(\ldots(x_n(x_{n-1}(x_{n-2}\ldots(x_1(x_0)^2)^2\ldots)^2)^2)^2\ldots)^2)^2.\end{aligned}$$

Due to Lemma 22, $|\theta_{m+n}(x)| \geq 2$ for each $x \in \widehat{X}$. Thus, every $\theta_{m+n}(x)$ contains not only an occurrence of x_{n-1}, but at least one left or right neighbour. If some occurrence of x_{n-1} lies in the left half of β_{m+n}, its left neighbour is always an occurrence of x_{n-2} (recall that $n \geq 2$), and its right neighbour is either x_n or x_0. On the other hand, if it lies in the right half of β_{m+n}, its right neighbour is

always x_{n-2}, and its left neighbour is either x_0 or x_n. Thus, if some $x \in \widehat{X}$ is mapped to an occurrence of x_{n-1} in the left half of β_{m+n}, $\theta_{m+n}(x)$ contains a factor $x_{n-2}x_{n-1}$, $x_{n-1}x_n$ or $x_{n-1}x_0$, and $x \in \widehat{X}_L$. Likewise, if it is mapped to an occurrence in the right half, $\theta_{m+n}(x)$ contains $x_{n-1}x_{n-2}$, x_nx_{n-1} or x_0x_{n-1}, and $x \in \widehat{X}_R$. Therefore, $\widehat{X}_L \cup \widehat{X}_R = \widehat{X}$.

In order to prove disjointness, we make another structural observation: We can safely assume that every variable in δ occurs at least twice – otherwise $L_{\mathrm{E},\Sigma}(\delta) = \Sigma^* \supset L_{\mathrm{E},\Sigma}(\alpha_0)$ would hold, and δ would not be E-descriptive of L_Σ according to Lemma 21. Thus, there is no variable x such that $x_{m+n}x_{m+n}$ is a factor of $\theta_{m+n}(x)$, as $x_{m+n}x_{m+n}$ occurs only once in β_{m+n}. This means that $x_{m+n}x_{m+n}$ forms an insurmountable barrier: For every occurrence of a variable from $\mathrm{var}(\delta)$, its image under θ_{m+n} lies either in the left or the right half of β_{m+n}. But if this image is longer than a single letter, the images of all occurrences of this variable must be mapped to the same side of β_{m+n}. According to Lemma 22, this is true for all variables of $\widehat{X}$. Therefore, for every $x \in \widehat{X}$, either $x \in \widehat{X}_L$ or $x \in \widehat{X}_R$ holds, which implies $\widehat{X}_L \cap \widehat{X}_R = \emptyset$. □

This permits us to define a morphism $\rho : X^* \to X^*$ through

$$\rho(x) := \begin{cases} (\widehat{\rho}_L \circ \theta_{m+n})(x) & \text{if } x \in \widehat{X}_L, \\ (\widehat{\rho}_R \circ \theta_{m+n})(x) & \text{if } x \in \widehat{X}_R, \\ \theta_{m+n}(x) & \text{otherwise,} \end{cases}$$

where the morphisms $\widehat{\rho}_L, \widehat{\rho}_R : X^* \to X^*$ are given by

$$\widehat{\rho}_L(x) := \begin{cases} y & \text{if } x = x_{n-1}, \\ x & \text{otherwise,} \end{cases} \qquad \widehat{\rho}_R(x) := \begin{cases} z & \text{if } x = x_{n-1}, \\ x & \text{otherwise.} \end{cases}$$

According to Lemma 23, the morphism ρ is well-defined, and, as to be proven next, $(\mu^{n-1} \circ \rho)(\delta) = \alpha_{m+1}$. Applying ρ to δ leads to

$$\begin{aligned} \rho(\delta) =& ((\dots(((\dots((x_0)^2x_1)^2\dots x_{n-2})^2y)^2x_n)^2\dots)^2x_{m+n})^2 \\ & (x_{m+n}(\dots(x_n(z(x_{n-2}\dots(x_1(x_0)^2)^2\dots)^2)^2)^2\dots)^2)^2, \end{aligned}$$

and, as $(m+n)-(n-1) = m+1$ and $\mu(x_i) = \lambda$ for every $i \leq n-2$, we obtain

$$\begin{aligned} (\mu^{n-1} \circ \rho)(\delta) &= ((\dots((y)^2x_1)^2\dots)^2x_{m+1})^2(x_{m+1}(\dots(x_1(z)^2)^2\dots)^2)^2 \\ &= \alpha_{m+1}. \end{aligned}$$

The morphism $\mu^{n-1} \circ \rho$ maps δ to α_{m+1}, and, thus, Theorem 1 immediately leads to $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\alpha_{m+1})$. Therefore, due to Lemma 21, the pattern δ cannot be E-descriptive of L_Σ. This contradiction concludes the proof of Theorem 14. □

5.3. Proof of Corollary 16

Our proof of Corollary 16 is based on the following technical lemma, that is given by Jiang et al. [8] in the context of their proof of Theorem 1:

Lemma 24 (Jiang et al. [8]). *Let Σ be an alphabet, $\Sigma \supseteq \{\mathtt{a}, \mathtt{b}\}$, and let $\alpha, \beta \in X^+$ be terminal-free patterns, $k := |\beta|$. Let the morphism $\tau_k : X^* \to X^*$ be given by, for every $i \in \mathbb{N}$,*

$$\tau_k(x_i) := \mathtt{ab}^{ki+1}\mathtt{a}\,\mathtt{ab}^{ki+2}\mathtt{a}\,\ldots\,\mathtt{ab}^{ki+k-1}\mathtt{a}\,\mathtt{ab}^{ki+k}\mathtt{a}.$$

Then $\tau_k(\alpha) \in L_{\mathrm{E},\Sigma}(\beta)$ if and only if there exists a morphism $h : X^* \to X^*$ satisfying $h(\beta) = \alpha$.

Furthermore, we wish to point out that the patterns α_i and β_i, $i \in \mathbb{N}$, referred to in the present section are defined in Definition 13 and Section 5.2, respectively.

We prove Corollary 16 by giving a thin language $L^t_\Sigma \subset L_\Sigma$ such that for every $\delta \in \mathrm{Pat}_\Sigma$ with $L_{\mathrm{E},\Sigma}(\delta) \supseteq L^t_\Sigma$ and for infinitely many $i \geq 0$, there is a morphism $\theta_i : X^* \to X^*$ with $\theta_i(\delta) = \beta_i$. Then for every such δ, there is a $j \geq 0$ with $L_{\mathrm{E},\Sigma}(\delta) \supset L_{\mathrm{E},\Sigma}(\alpha_j) \supset L^t_\Sigma$.

Proof. Let $\mathtt{a}, \mathtt{b} \in \Sigma$ with $\mathtt{a} \neq \mathtt{b}$. For every $n \geq 1$, we define a substitution $\tau_n : X^* \to \Sigma^*$ by

$$\tau_n(x_i) := \mathtt{ab}^{ni+1}\mathtt{a}\,\mathtt{ab}^{ni+2}\mathtt{a}\,\ldots\,\mathtt{ab}^{ni+n-1}\mathtt{a}\,\mathtt{ab}^{ni+n}\mathtt{a},$$

and we assume that τ_0 denotes the constant λ-function. We then define

$$L^t_\Sigma := \bigcup_{n \geq 0}^{\infty} \tau_n(\beta_n).$$

It is easy to see that L^t_Σ is properly thin, as for every $n \geq 0$, $|\tau_n(\beta_n)| < |\tau_{n+1}(\beta_{n+1})|$.

We assume to the contrary that there is a pattern $\delta \in \mathrm{Pat}_\Sigma$ that is E-descriptive of L^t_Σ. First note that – since $L_{\mathrm{E},\Sigma}(\alpha_i) \supset L_{\mathrm{E},\Sigma}(\alpha_{i+1}) \supset L_\Sigma \supset L^t_\Sigma$ for every $i \geq 0$ (see Lemma 20) – there is no $j \in \mathbb{N}$ with $L_{\mathrm{E},\Sigma}(\delta) \supseteq L_{\mathrm{E},\Sigma}(\alpha_j)$ (as described by Lemma 21). Furthermore, as $\tau_0(\beta_0) = \lambda$, $\lambda \in L^t_\Sigma \subseteq L_{\mathrm{E},\Sigma}(\delta)$ holds, and therefore δ must be terminal-free.

According to Lemma 24, for every $\delta \in X^+$ and every $n \geq |\delta|$, $\tau_n(\beta_n) \in L_{\mathrm{E},\Sigma}(\delta)$ if and only if there is a morphism $\theta_n : X^* \to X^*$ such that $\theta_n(\delta) = \beta_n$. Furthermore, for every $m < n$ and the morphism μ introduced in Section 5.2, $(\mu^{n-m} \circ \theta_n)(\delta) = \mu^{n-m}(\beta_n) = \beta_{n-(m-n)} = \beta_m$ holds. Thus, there is an infinite sequence $(\theta_i)_{i \geq 0}$ with $\theta_i(\delta) = \beta_i$ for all $i \geq 0$, which allows us to construct a morphism that maps δ to some α_j just as in the proof for Theorem 14. Thus, $L_{\mathrm{E},\Sigma}(\delta) \supset L_{\mathrm{E},\Sigma}(\alpha_j) \supset L^t_\Sigma$, and this contradicts our assumption of δ being E-descriptive of L^t_Σ. □

6. Conclusions and Further Directions of Research

In the present paper, we have studied the existence and nonexistence of patterns that are descriptive of a set of strings. We have explained that this question is related to the existence of infinite strictly decreasing chains of pattern

languages. Our main result has demonstrated that there exist infinite languages over alphabets of at least two letters that do not have an E-descriptive pattern.

This insight leads to the question of characteristic criteria describing infinite languages without an E-descriptive pattern. Our main proof has given one example of such languages, namely a particular infinite union of E-pattern languages. Although we have demonstrated that an infinite properly thin language can be substituted for this union, we anticipate that only very special languages (and very special infinite strictly decreasing chains of E-pattern languages) can be used for such a proof. Thus, we expect the nonexistence of E-descriptive patterns to be a rare phenomenon. In addition to the said criteria, we consider it worthwhile to further investigate the existence of efficient procedures finding descriptive patterns of given languages (for those cases where descriptive patterns exist). So far, this question has only been answered for NE-descriptive patterns of finite languages (see Angluin [1]), demonstrating that no such procedure can have polynomial runtime (provided that P$\neq$NP). We feel that a more pleasant result might be possible for E-descriptive patterns.

References

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

[2] Brazma, A., Jonassen, I., Eidhammer, I., Gilbert, D., 1998. Approaches to the automatic discovery of patterns in biosequences. Journal of Computational Biology 5, 279–305.

[3] Dumitrescu, S., Păun, G., Salomaa, A., 1996. Languages associated to finite and infinite sets of patterns. Revue Roumaine de Mathématiques Pures et Appliquées 41, 607–625.

[4] Ehrenfeucht, A., Rozenberg, G., 1979. Finding a homomorphism between two words is NP-complete. Information Processing Letters 9, 86–88.

[5] Freydenberger, D., Reidenbach, D., 2009. Existence and nonexistence of descriptive patterns. In: Proc. 13th International Conference on Developments in Language Theory, DLT 2009. Vol. 5583 of Lecture Notes in Computer Science. pp. 228–239.

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

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

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

[9] Kari, L., Rozenberg, G., Salomaa, A., 1997. L systems. In: Rozenberg, G., Salomaa, A. (Eds.), Handbook of Formal Languages. Vol. 1. Springer, Ch. 5, pp. 253–328.

[10] Mateescu, A., Salomaa, A., 1997. Patterns. In: Rozenberg, G., Salomaa, A. (Eds.), Handbook of Formal Languages. Vol. 1. Springer, Ch. 4.6, pp. 230–242.

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

[12] Reidenbach, D., Schneider, J., 2009. Morphically primitive words. Theoretical Computer Science 410, 2148–2161.

[13] Rozenberg, G., Salomaa, A., 1997. Handbook of Formal Languages. Vol. 1. Springer, Berlin.

[14] Salomaa, K., 2004. Patterns. In: Martin-Vide, C., Mitrana, V., Păun, G. (Eds.), Formal Languages and Applications. No. 148 in Studies in Fuzziness and Soft Computing. Springer, pp. 367–379.

[15] Shinohara, T., 1982. Polynomial time inference of extended regular pattern languages. In: Proc. RIMS Symposia on Software Science and Engineering, Kyoto. Vol. 147 of Lecture Notes in Computer Science. pp. 115–127.

[16] Wilf, H., 1994. generatingfunctionology, 2nd Edition. Academic Press, New York.

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