Unambiguous Morphic Images of Strings

Dominik D. Freydenberger, Daniel Reidenbach*,**, and Johannes C. Schneider

Fachbereich Informatik, Technische Universität Kaiserslautern,
Postfach 3049, 67653 Kaiserslautern, Germany
{d_freyde,reidenba,j_schne}@informatik.uni-kl.de

Abstract. Motivated by the research on pattern languages, we study a fundamental combinatorial question on morphisms in free semigroups: With regard to any string α over some alphabet we ask for the existence of a morphism σ such that $\sigma(\alpha)$ is unambiguous, i.e. there is no morphism ρ with $\rho \neq \sigma$ and $\rho(\alpha) = \sigma(\alpha)$. Our main result shows that a rich and natural class of strings is provided with unambiguous morphic images.

1 Introduction

In the past decades a lot of effort has been spent on investigating the properties of those *morphisms* which map a string over some alphabet Σ onto a string over a second alphabet Σ' (cf., e.g., Lothaire [9], Choffrut and Karhumäki [2], Harju and Karhumäki [3]). In this context, many problems only arise if Σ contains more symbols than Σ', and therefore—in order to address these difficulties as precisely as possible—we assume, for the remainder of our paper, $\Sigma = \mathbb{N}$ and $\Sigma' = \{\mathtt{a}, \mathtt{b}\}$. Consequently, we regard the set of morphisms mapping the strings in an infinitely generated free semigroup to the strings in a free monoid with two generators. According to the closely related research on *pattern languages* (cf. Mateescu and Salomaa [10]) we call an element of $\mathbb{N}^+$ a *pattern* and an element of $\{\mathtt{a}, \mathtt{b}\}^*$ a *word*. We separate all symbols in a pattern by a dot (see, e.g., the example pattern α' below) so as to avoid any confusion.

Quite a number of the basic questions to be asked on suchlike mappings deals with the problem of finding a morphism which, in spite of the resulting alphabet reduction, preserves the structure of its input string as far as possible; this is a manifest topic, e.g., in the theory of codes (cf. Jürgensen and Konstantinidis [5]). Even though any answer to this question strongly depends on the formal definition of what is considered to be a "structure-preserving" morphism, from a very intuitive point of view, one surely would agree that, for instance, the shape of the pattern $\alpha' = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 1 \cdot 4 \cdot 3 \cdot 2$ is not adequately reflected by its morphic image $w_1 = \mathtt{a}^{10}$. Obviously, for such a task, a code—that, in our sense, is nothing but an *injective* morphism—is a more appropriate choice: If we apply the morphism $\sigma'(i) = \mathtt{a}\,\mathtt{b}^i$, $i \in \mathbb{N}$, to α' then we receive the word $w_2 = \sigma'(\alpha) = \mathtt{a}\,\mathtt{b}\,\mathtt{a}\,\mathtt{b}^2\,\mathtt{a}\,\mathtt{b}^3\,\mathtt{a}\,\mathtt{b}^4\,\mathtt{a}\,\mathtt{b}\,\mathtt{a}\,\mathtt{b}^4\,\mathtt{a}\,\mathtt{b}^3\,\mathtt{a}\,\mathtt{b}^2$ which, due to the distinct lengths

* Corresponding author
** Partly supported by Deutsche Forschungsgemeinschaft (DFG), Grant Wi 1638/1-3

of subwords over the single symbol $\mathtt{b}$, that allow the definition of an inverse morphism, seems to describe its preimage quite accurately.

However, in those settings where we are confronted with a variety of morphic images of one and the same pattern α—such as in *inductive inference of pattern languages* (cf., e.g., Angluin [1]), that deals with the inferrability of an (unknown) pattern from the set of *all* of its morphic images—the mere injectivity of morphisms can turn out to be insufficient for reflecting the shape of α. Surprisingly, with regard to these problems, a second property of a morphic image w demonstrably is much more important, namely its *ambiguity* (cf. Reidenbach [13]), i.e. the question whether there are at least two morphisms σ, ρ such that, for some symbol i in α, $\sigma(i) \neq \rho(i)$, but nevertheless $\sigma(\alpha) = w = \rho(\alpha)$. Returning to our example it can easily be seen that w_2 is ambiguous with respect to α' as it can, e.g., also be generated by the morphism ρ' with $\rho'(1) = \mathtt{a\,b\,a\,b}^2$, $\rho'(2) = \varepsilon$, $\rho'(3) = \mathtt{a\,b}^3\,\mathtt{a\,b}^2$ and $\rho'(4) = \mathtt{b}^2$, where ε is the empty word:

$$\overbrace{\mathtt{a\,b}}^{\sigma'(1)}\,\overbrace{\mathtt{a\,b\,b}}^{\sigma'(2)}\,\overbrace{\mathtt{a\,b\,b\,b}}^{\sigma'(3)}\,\overbrace{\mathtt{a\,b\,b\,b\,b}}^{\sigma'(4)}\,\overbrace{\mathtt{a\,b}}^{\sigma'(1)}\,\overbrace{\mathtt{a\,b\,b\,b\,b}}^{\sigma'(4)}\,\overbrace{\mathtt{a\,b\,b\,b}}^{\sigma'(3)}\,\overbrace{\mathtt{a\,b\,b}}^{\sigma'(2)}$$

$$\underbrace{\mathtt{a\,b\,a\,b\,b}}_{\rho'(1)}\,\underbrace{\mathtt{a\,b\,b\,b\,a\,b\,b}}_{\rho'(3)}\,\underbrace{\mathtt{b\,b}}_{\rho'(4)}\,\underbrace{\mathtt{a\,b\,a\,b\,b}}_{\rho'(1)}\,\underbrace{\mathtt{b\,b}}_{\rho'(4)}\,\underbrace{\mathtt{a\,b\,b\,b\,a\,b\,b}}_{\rho'(3)}$$

Consequently, w_2 does not adequately substantiate the existence of the symbol 2 in α' since this symbol is not needed for generating w_2; thus, from that point of view and in spite of its injectivity, we do not consider σ' to meet our vague yet well-founded requirements for a structure-preserving morphism. But even if we restrict our examination to nonerasing morphisms, i.e. if we use the free semigroup $\{\mathtt{a}, \mathtt{b}\}^+$ instead of $\{\mathtt{a}, \mathtt{b}\}^*$ as value set of the morphisms, the multitude of potential generating morphisms blurs the evidence of α in w_2.

Unfortunately, this ambiguity of words is a frequent property of many patterns, and, effortlessly, examples can be given for which there is no morphism at all leading to an unambiguous word; on the other hand, it is by no means obvious for which patterns there exist such structure-preserving morphic images. In the present paper, we examine this combinatorial problem of intrinsic interest systematically. To this end, we concentrate on the ambiguity of those words that are images of injective morphisms, and we explicitly distinguish between the general case where the set of all morphisms $\rho_{\mathrm{E}} : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ is considered and the restricted case that focuses on nonerasing morphisms $\rho_{\mathrm{NE}} : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$. Our paper is organised as follows: After some brief formal definitions we collect a number of rather evident preliminary results before we show that a rich and natural class of patterns is characterised by the ability of morphically generating unambiguous words. This main result answers a question posed in [12].

Obviously, our work shows some connections to *equality sets* (cf., e.g., Harju and Karhumäki [3], Lipponen and Păun [8]): If, for some pattern α, we find a morphism σ such that $\sigma(\alpha)$ is unambiguous then α is a "non-solution" to the Post Correspondence Problem for σ and any other morphism ρ. Finally, it seems worth mentioning that, in a sense, our work complements a research that has been initiated by Mateescu and Salomaa [11]. As explained above we show that, for *every* pattern in some class, there exists *at least one* word that has exactly one generating morphism, whereas, in a more general context, [11] examines the question whether, for an arbitrary upper bound $n \in \mathbb{N}$, there exists *at least one* pattern such that *each* of its morphic images has at most n distinct generating morphisms. In our restricted setting, for all patterns with occurrences of at least two different symbols, this question has a trivial answer in the negative.

2 Definitions and Basic Notes

We begin the formal part of this paper with a number of basic definitions. Major parts of our terminology are adopted from the research on pattern languages (cf. Mateescu and Solomaa [10]). Additionally, for notions not explained explicitly, we refer the reader to Choffrut and Karhumäki [2].

Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let Σ be an *alphabet*, i.e. an enumerable set of symbols. We regard two different alphabets: $\mathbb{N}$ and $\{\mathtt{a}, \mathtt{b}\}$ with $\mathtt{a} \neq \mathtt{b}$. Henceforth we call any symbol in $\mathbb{N}$ a *variable* and any symbol in $\{\mathtt{a}, \mathtt{b}\}$ a *letter*. A *string (over Σ)* is a finite sequence of symbols from Σ. For the *concatenation* of two strings $w_1, w_2 \in \Sigma^*$ we write $w_1 \cdot w_2$ or simply $w_1 w_2$. The string that results from the n-fold concatenation of a string w occasionally is denoted by w^n. $|\, x \,|$ stands for the size of a set x or the length of a string x, respectively. We denote the *empty string* by ε, i.e. $|\varepsilon| = 0$. In order to distinguish between a string over $\mathbb{N}$ and a string over $\{\mathtt{a}, \mathtt{b}\}$, we call the former a *pattern* and the latter a *word*. We name patterns with lower case letters from the beginning of the Greek alphabet such as α, β, γ. With regard to an arbitrary pattern α, $\mathrm{var}(\alpha)$ denotes the set of all variables occurring in α. For every alphabet Σ, Σ^* is the set of all (empty and non-empty) strings over Σ, and $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$. We say that a string $v \in \Sigma^*$ is a *substring* of a string $w \in \Sigma^*$ if and only if, for some $u_1, u_2 \in \Sigma^*$, $w = u_1 v\, u_2$. Subject to the concrete alphabet considered, we call a substring a *subword* or *subpattern*. Additionally, we use the notions $w = \ldots v \ldots$ if v is a substring of w, $w = v \ldots$ if v is a *prefix* of w, and $w = \ldots v$ if v is a *suffix* of w. $|w|_v$ denotes the number of occurrences of a substring v in a string w. We do not use this notion for substrings with overlapping occurrences.

Since we deal with free semigroups, a *morphism* σ is a mapping that is compatible with the concatenation, i.e. for patterns $\alpha, \beta \in \mathbb{N}^+$, a morphism $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ satisfies $\sigma(\alpha \cdot \beta) = \sigma(\alpha) \cdot \sigma(\beta)$. Hence, a morphism is fully explained as soon as it is declared for all variables in $\mathbb{N}$. Note that we restrict ourselves to total morphisms, even though we normally declare a morphism only for those variables explicitly that, in the respective context, are relevant.

Let $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ be a morphism. Then σ is called *nonerasing* provided that, for every $i \in \mathbb{N}$, $\sigma(i) \neq \varepsilon$. Note that σ necessarily is nonerasing if it is injective. For any pattern $\alpha \in \mathbb{N}^+$ with $\sigma(\alpha) \neq \varepsilon$, we call $\sigma(\alpha)$ *weakly unambiguous (with respect to α)* if there is no nonerasing morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ such that $\rho(\alpha) = \sigma(\alpha)$ and, for some $i \in \mathrm{var}(\alpha)$, $\rho(i) \neq \sigma(i)$. If, in addition, there is no (arbitrary) morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ with $\rho(\alpha) = \sigma(\alpha)$ and $\rho(i) \neq \sigma(i)$ for some $i \in \mathrm{var}(\alpha)$, then $\sigma(\alpha)$ is called *(strongly) unambiguous (with respect to α)*. Obviously, if $\sigma(\alpha)$ is strongly unambiguous then it is weakly unambiguous as well. Finally, $\sigma(\alpha)$ is *ambiguous (with respect to* α*)* if and only if it is not weakly unambiguous.

As mentioned above, our subject is closely related to pattern languages (cf., e.g., Mateescu and Salomaa [10]), and therefore we consider it useful to provide an adequate background for some explanatory remarks. The *pattern language* of a pattern is the set of all of its possible morphic images in some fixed free monoid Σ^* (in our case $\Sigma = \{\mathtt{a}, \mathtt{b}\}$). More precisely and with regard to any $\alpha \in \mathbb{N}^+$, we distinguish between its *E-pattern language* $L_{\mathrm{E}}(\alpha) = \{\sigma(\alpha) \mid \sigma : \mathbb{N}^+ \longrightarrow \Sigma^*\}$ and its *NE-pattern language* $L_{\mathrm{NE}}(\alpha) = \{\sigma(\alpha) \mid \sigma : \mathbb{N}^+ \longrightarrow \Sigma^+\}$. Note that this definition implies that the full *class of E-(resp. NE-)pattern languages*, i.e. the set $\{L_{\mathrm{E}}(\alpha) \mid \alpha \in \mathbb{N}^+\}$ resp. $\{L_{\mathrm{NE}}(\alpha) \mid \alpha \in \mathbb{N}^+\}$, considered in this paper merely covers a special case which, in literature, usually is referred to as *terminal-free* (or: *pure*) pattern languages. This is due to the fact that, contrary to our view, a pattern commonly is seen as a string in $(\mathbb{N} \cup \Sigma)^+$ and not just in $\mathbb{N}^+$.

We conclude the definitions in this section with a crucial partition of the set of all patterns subject to the following criterion:

Definition 1. *We call any* $\alpha \in \mathbb{N}^+$ succinct *if and only if there exists no decomposition* $\alpha = \beta_0\,\gamma_1\,\beta_1\,\gamma_2\,\beta_2\ \ldots\ \beta_{n-1}\,\gamma_n\,\beta_n$ *with* $n \geq 1$, $\beta_k \in \mathbb{N}^*$ *and* $\gamma_k \in \mathbb{N}^+$, $k \leq n$, *such that*

1. *for every* k, $1 \leq k \leq n$, $|\gamma_k| \geq 2$,
2. *for every* k, $1 \leq k \leq n$, *and for every* k', $0 \leq k' \leq n$, $\mathrm{var}(\gamma_k) \cap \mathrm{var}(\beta_{k'}) = \emptyset$,
3. *for every* k, $1 \leq k \leq n$, *there exists an* $i_k \in \mathrm{var}(\gamma_k)$ *such that* $|\gamma_k|_{i_k} = 1$ *and, for every* k', $1 \leq k' \leq n$, *if* $i_k \in \mathrm{var}(\gamma_{k'})$ *then* $\gamma_k = \gamma_{k'}$.

We call $\alpha \in \mathbb{N}^+$ prolix *if and only if it is not succinct.*

Example 1. Obviously, any pattern α, $|\alpha| \geq 2$, necessarily is prolix if there is a variable $i \in \mathbb{N}$ such that $|\alpha|_i = 1$. Our initial example $\alpha' = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 1 \cdot 4 \cdot 3 \cdot 2$ and the pattern $\alpha_1 = 1 \cdot 1$ are succinct, whereas $\alpha_2 = 1 \cdot 2 \cdot 3 \cdot 3 \cdot 1 \cdot 2 \cdot 3$ is prolix with $\beta_0 = \varepsilon$, $\gamma_1 = 1 \cdot 2$, $\beta_1 = 3 \cdot 3$, $\gamma_2 = 1 \cdot 2$ and $\beta_2 = 3$. Note that this obligatory decomposition of a prolix pattern does not have to be unique. Additional and more complex examples can be found in the subsequent sections and in [13].

According to Reidenbach [13] the succinct patterns are the shortest generators for their respective E-pattern language—this explains the terms "succinct" and "prolix". In other words, for every succinct pattern α and for every pattern β, if $L_{\mathrm{E}}(\alpha) = L_{\mathrm{E}}(\beta)$ then $|\alpha| \leq |\beta|$. Consequently, the class of E-pattern languages equals the set $\{L_{\mathrm{E}}(\alpha) \mid \alpha \in \mathbb{N}^+,\ \alpha \text{ is succinct}\}$ although the set of all patterns is a proper superset of the set of all succinct patterns.

In addition to this view, the set of prolix patterns exactly corresponds to the class of finite *fixed points* of nontrivial morphisms, i.e. for every prolix pattern α there exists a morphism $\phi : \mathbb{N}^* \longrightarrow \mathbb{N}^*$ such that, for an $i \in \mathrm{var}(\alpha)$, $\phi(i) \neq i$ and yet $\phi(\alpha) = \alpha$ (cf., e.g., Head [4], Levé and Richomme [7]).

Finally note that all results in this paper hold for morphisms to arbitrary finitely generated free monoids with three or more generators instead of $\{\mathtt{a}, \mathtt{b}\}^*$ as well. With regard to the positive results, this follows by definition, and the proofs of the negative results can be adapted with little effort.

3 Weakly Unambiguous Words

We begin our examination with some momentuous statements on weakly unambiguous morphic images. The first is an evident yet strong positive result:

Proposition 1. *There is a nonerasing morphism* $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ *such that, for every* $\alpha \in \mathbb{N}^+$, $\sigma(\alpha)$ *is weakly unambiguous.*

Proof. For every $i \in \mathbb{N}$, let $|\sigma(i)| = 1$. Then, for every $\alpha \in \mathbb{N}^+$, $|\sigma(\alpha)| = |\alpha|$ and, consequently, $\sigma(\alpha)$ is weakly unambiguous. □

In Proposition 1 we fully restrict ourselves to nonerasing morphisms, and therefore the view applied therein exactly corresponds to the concept of NE-pattern languages. Indeed, the weak unambiguity of the words referred to in the proof is of major importance for inductive inference of NE-pattern languages: Due to the fact given in Proposition 1, for every NE-pattern language L a pattern α with $L = L_{\mathrm{NE}}(\alpha)$ can be inferred from the set of all of the shortest words in this language (shown by Lange and Wiehagen [6]). With regard to E-pattern languages, however, this is provably impossible since, in general, these words are not strongly unambiguous (cf. Reidenbach [12]). Consequently, in respect of pattern inference, the unambiguity of certain words—which are not generated by an injective morphism—is surprisingly powerful.

For the main goal of our approach (see Section 1), however, injectivity of morphisms is vital. Unfortunately, for those morphisms the outcome significantly differs from Proposition 1:

Theorem 1. *There is no injective morphism* $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ *such that, for every* $\alpha \in \mathbb{N}^+$, $\sigma(\alpha)$ *is weakly unambiguous.*

Proof. Assume to the contrary that there is such a morphism σ. Since σ is injective, $\sigma(\alpha) \neq \sigma(\beta)$ for every $\alpha \neq \beta$. In particular, this implies $\sigma(i) \neq \sigma(i')$ for every $i, i' \in \mathbb{N}$ with $i \neq i'$. Hence, there must be a $j \in \mathbb{N}$ with $\sigma(j) = w_1 w_2$ for some $w_1, w_2 \in \{\mathtt{a}, \mathtt{b}\}^+$. Now, for an arbitrary $j' \neq j$, let $\alpha := j \cdot j'$. Then, for the morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ given by $\rho(j) := w_1$ and $\rho(j') := w_2\, \sigma(j')$, $\rho(\alpha) = \sigma(\alpha)$, and, thus, $\sigma(\alpha)$ is ambiguous. This contradicts the assumption. □

Obviously, Theorem 1 includes the analogous result for strong unambiguity. Consequently, there is no single injective morphism which, when applied to arbitrary patterns, leads to unambiguous words. Thus, two natural questions arise from Theorem 1: Is there a significant subclass of all patterns for which the opposite of Theorem 1 holds true? Is there at least a way to find for every pattern an individual injective morphism that leads to an unambiguous word? In the following section we examine these questions with regard to strong unambiguity.

4 Strongly Unambiguous Words

Bearing the consequences of Theorem 1 in mind the present section deals with strongly unambiguous words. We begin with the observation that the example pattern α in the proof of Theorem 1 is prolix. However, if we focus on succinct patterns then the analogue turns out to be true; as we now ask for strong unambiguity we even can prove the opposite of Proposition 1:

Proposition 2. *There is no nonerasing morphism* $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ *such that, for every succinct* $\alpha \in \mathbb{N}^+$*,* $\sigma(\alpha)$ *is strongly unambiguous.*

Proof. Assume to the contrary that there is such a morphism. Then there exist $j, j' \in \mathbb{N}$, $j \neq j'$, and an $\mathtt{c} \in \{\mathtt{a}, \mathtt{b}\}$ with $\sigma(j) = v\,\mathtt{c}$ and $\sigma(j') = v'\,\mathtt{c}$, $v, v' \in \Sigma^*$. For some $k, k' \in \mathbb{N}$, $k \neq k'$, $j \neq k \neq j'$ and $j \neq k' \neq j'$, we then regard the pattern $\alpha := j \cdot k \cdot j \cdot k' \cdot j' \cdot k \cdot j' \cdot k'$. Obviously, α is succinct. Now consider the morphism ρ, given by $\rho(j) := v$, $\rho(j') := v'$, $\rho(k) := \mathtt{c}\,\sigma(k)$, $\rho(k') := \mathtt{c}\,\sigma(k')$. Then evidently $\sigma(\alpha) = \rho(\alpha)$, but, e.g., $\sigma(j) \neq \rho(j)$. This is a contradiction. □

Consequently, for every succinct pattern, it is necessary to give an individual injective morphism that leads to a strongly unambiguous word—provided that such a morphism exists. The hope for a positive answer to this question is supported by the following fact whereby, for many patterns, even completely inappropriate looking morphisms generate a strongly unambiguous word.

Proposition 3. *For every nonerasing morphism* $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ *there exists a succinct* $\alpha \in \mathbb{N}^+$*,* $|\operatorname{var}(\alpha)| \geq 2$*, such that* $\sigma(\alpha)$ *is strongly unambiguous.*

Proof. Since our argumentation solely deals with the length of the morphic images of the variables, we can utilise the following fact on linear combinations:

Claim 1. For all $p, q \in \mathbb{N}$ *there exist* $r, s \in \mathbb{N}$*,* $r, s \geq 2$ *such that there are no* $p', q' \in \mathbb{N}_0 \setminus \{p, q\}$ *satisfying* $rp + sq = rp' + sq'$*.*

With $r > q$, $s > p$, and $\gcd(r, s) = 1$, Claim 1 can be proven with a bit of effort.

Now, for some $i, j \in \mathbb{N}$, $i \neq j$, let $p := |\sigma(i)|$, $q := |\sigma(j)|$. Furthermore, let $\alpha := i^r \cdot j^s$ with r, s derived from Claim 1. Obviously, α is succinct. Assume to the contrary that there is a morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ with $\rho(\alpha) = \sigma(\alpha)$ and, for some $k \in \{i, j\}$, $\rho(k) \neq \sigma(k)$. Then ρ must satisfy $|\rho(i)| \neq |\sigma(i)|$, $|\rho(j)| \neq |\sigma(j)|$ and $|\rho(\alpha)| = |\sigma(\alpha)|$. Consequently, with $p' := |\rho(i)|$, $q' := |\rho(j)|$, $rp + sq = |\sigma(\alpha)| = |\rho(\alpha)| = rp' + sq'$. This contradicts Claim 1. □

Before we go further into this matter of strongly unambiguous morphic images for succinct patterns (see Theorem 3), we turn our attention to prolix patterns. Here we can easily give a definite answer, which, alternatively, can be seen as a consequence of the fact that every prolix pattern is a fixed point of some nontrivial morphism (cf. Section 2):

Theorem 2. *For any prolix* $\alpha \in \mathbb{N}^+$ *and for any nonerasing morphism* $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$*,* $\sigma(\alpha)$ *is not strongly unambiguous.*

Proof. Assume to the contrary that there are a prolix pattern α and a nonerasing morphism σ such that $\sigma(\alpha)$ is strongly unambiguous. Then, as α is prolix, there exists a decomposition $\alpha = \beta_0\,\gamma_1\,\beta_1\,\gamma_2\,\beta_2\ \ldots\ \beta_{n-1}\,\gamma_n\,\beta_n$ satisfying the conditions of Definition 1. With regard to this decomposition and for every k, $1 \leq k \leq n$, let i_k be the smallest $i \in \mathrm{var}(\gamma_k)$ such that $|\gamma_k|_{i_k} = 1$ and, for every k', $1 \leq k' \leq n$, if $i_k \in \mathrm{var}(\gamma_{k'})$, then $\gamma_k = \gamma_{k'}$. By definition, for every γ_k, $1 \leq k \leq n$, such an i_k exists, and, for every $\beta_{k'}$, $0 \leq k' \leq n$, $i_k \notin \mathrm{var}(\beta_{k'})$. Now we define ρ as follows: For all k, $1 \leq k \leq n$, $\rho(i_k) := \sigma(\gamma_k)$, for all $i \in \mathrm{var}(\gamma_k) \setminus \{i_k\}$, $\rho(i) := \varepsilon$, and, for all k, $0 \leq k \leq n$ and for all $i \in \mathrm{var}(\beta_k)$, $\rho(i) := \sigma(i)$. Then $\sigma(\alpha) = \rho(\alpha)$, but, since σ is nonerasing and every γ_k contains at least two variables, there is an $i \in \mathrm{var}(\alpha)$ with $\sigma(i) \neq \rho(i)$. Thus, $\sigma(\alpha)$ is not strongly unambiguous. □

Thus, for every prolix pattern there is no strongly unambiguous word at all—at least as long as we restrict ourselves to the images of nonerasing morphisms. If this requirement is omitted then we face a fairly intricate situation:

Example 2. Let $\alpha_1 := 1 \cdot 1 \cdot 2 \cdot 2 \cdot 3 \cdot 4 \cdot 3 \cdot 4$, $\alpha_2 := 1 \cdot 2 \cdot 2 \cdot 1 \cdot 3 \cdot 4 \cdot 3 \cdot 4$, and $\beta_1 := 1 \cdot 2 \cdot 3 \cdot 3 \cdot 1 \cdot 2$, $\beta_2 := 1 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 1 \cdot 2 \cdot 2$. The patterns are prolix. For α_1 and β_1 there is no morphism σ such that $\sigma(\alpha_1)$ or $\sigma(\beta_1)$ are unambiguous. Contrary to this, for α_2 and β_2 there exist suitable words such as `abba` and `baab`.

As shown in Example 2, there are prolix patterns for which we can unambiguously map certain subpatterns onto strings in $\{\mathtt{a}, \mathtt{b}\}^*$, whereas for different, quite similar appearing patterns this is impossible. Furthermore, these subpatterns can consist of parts of some β_k as well as parts of some γ_k in the required decomposition of the patterns (cf. Definition 1). We now briefly discuss this phenomenon, and we begin with a criterion which covers both prolix and succinct patterns:

Condition 1. *A pattern $\alpha \in \mathbb{N}^+$ satisfies Condition 1 if and only if there exists an $i \in \mathrm{var}(\alpha)$ such that, for $n = |\mathrm{var}(\alpha)| - 1$, for all $j_1, j_2, \ldots, j_n \in \mathrm{var}(\alpha) \setminus \{i\}$ and for all $k_1, k_2, \ldots, k_n \in \mathbb{N}_0$, $|\alpha|_i \neq k_1|\alpha|_{j_1} + k_2|\alpha|_{j_2} + \ldots + k_n|\alpha|_{j_n}$.*

For those patterns satisfying Condition 1 we can give a positive result:

Proposition 4. *For every $\alpha \in \mathbb{N}^+$ satisfying Condition 1 there exists a morphism $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ such that $\sigma(\alpha)$ is strongly unambiguous.*

Proof. With $\sigma(i) := \mathtt{a}$ (i as defined in Condition 1) and, for all $j \in \mathbb{N}$ with $j \neq i$, $\sigma(j) := \varepsilon$, Proposition 4 follows immediately. □

For prolix patterns with exactly two different variables, Condition 1 even characterises the subclass for which there are strongly unambiguous words:

Proposition 5. *Let $\alpha \in \mathbb{N}^+$ be prolix, $\mathrm{var}(\alpha) := \{i, j\}$. Then there exists a morphism $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ such that $\sigma(\alpha)$ is strongly unambiguous if and only if $|\alpha|_i \neq |\alpha|_j$.*

Proof. For the *if* part, w. l. o. g. assume $|\alpha|_i < |\alpha|_j$. Then the existence of a morphism σ such that $\sigma(\alpha)$ is strongly unambiguous is guaranteed by Proposition 4. We proceed with the *only if* part: Let $|\alpha|_i = |\alpha|_j$. Then, since α is prolix, α can only be of the form $(i \cdot j)^n$ (or $(j \cdot i)^n$), $n \in \mathbb{N}$. Thus, there is no strongly unambiguous morphic image for α. □

Note that Propositions 4 and 5 utilise a morphism that is non-empty for a single variable only. In general, of course, one might wish to find a morphism that assigns non-empty words to a preferably large number of variables in a prolix pattern and, nevertheless, leads to a strongly unambiguous word (cf. Example 2). However, as soon as the number of variables to be mapped onto non-empty words exceeds the number of letters in the target alphabet, we consider it an extraordinarily challenging problem to find reasonably strong criteria.

We now return to the remaining crucial question of this paper left open after Propositions 2 and 3 and Theorem 2, namely the existence of *injective* morphisms generating strongly unambiguous words for *succinct* patterns. Particularly the proof of Proposition 2 suggests that a finitely generated free monoid might not be rich enough to include strongly unambiguous morphic images for all succinct patterns. On the other hand, the proof of the comprehensive negative result for prolix patterns (cf. Theorem 2) strongly utilises the properties of these patterns as declared in Definition 1, and, indeed, our main result (to be proven in Section 4.1) shows the opposite of Theorem 2 to be true for succinct patterns:

Theorem 3. *For every succinct $\alpha \in \mathbb{N}^+$, there is an injective morphism $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ such that $\sigma(\alpha)$ is strongly unambiguous.*

Consequently, for every succinct string in an infinitely generated free semigroup there is a morphic image in a free monoid with two generators that—in accordance with our requirements explained in Section 1—sufficiently preserves its structure. With regard to pattern languages, Theorem 3 proves that in every E-pattern language there is an unambiguous word with respect to any shortest generating pattern. For a restatement of our main result in terms of equality sets or fixed points of morphisms, see the notes in Section 1 or Section 2, respectively.

Finally, we can use Theorems 2 and 3 for a characterisation of succinctness:

Corollary 1. *Let $\alpha \in \mathbb{N}^+$. Then α is succinct if and only if there exists an injective morphism $\sigma : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^+$ such that $\sigma(\alpha)$ is strongly unambiguous.*

4.1 Proof of Theorem 3

We begin this section with a procedure which, for every succinct pattern, constructs a morphism that generates a strongly unambiguous word:

Definition 2. *Let $\alpha \in \mathbb{N}^+$. For every $j \in \mathrm{var}(\alpha)$, consider the following sets: $L_j := \{k \mid \alpha = \ldots \cdot k \cdot j \cdot \ldots\}$ and $R_j := \{k \mid \alpha = \ldots \cdot j \cdot k \cdot \ldots\}$. Thus, L_j consists of all "left neighbours" of j in α and R_j of all "right neighbours". With these sets, construct two relations $\sim_l$ and $\sim_r$ on $\mathrm{var}(\alpha)$: For all $k, k' \in \mathrm{var}(\alpha)$*

- *$k \sim_l k'$ if and only if there are $j_1, j_2, \ldots, j_t \in \mathrm{var}(\alpha)$, $t \geq 1$, such that*
  1. *$L_{j_1} \cap L_{j_2} \neq \emptyset$, $L_{j_2} \cap L_{j_3} \neq \emptyset$, …, $L_{j_{t-1}} \cap L_{j_t} \neq \emptyset$ and*
  2. *$k \in L_{j_1}$ and $k' \in L_{j_t}$.*
- *$k \sim_r k'$ if and only if there are $j_1, j_2, \ldots, j_s \in \mathrm{var}(\alpha)$, $s \geq 1$, such that*
  1. *$R_{j_1} \cap R_{j_2} \neq \emptyset$, $R_{j_2} \cap R_{j_3} \neq \emptyset$, …, $R_{j_{s-1}} \cap R_{j_s} \neq \emptyset$ and*
  2. *$k \in R_{j_1}$ and $k' \in R_{j_s}$.*

Evidently, $\sim_l$ and $\sim_r$ are equivalence relations, and, for every $k \in \text{var}(\alpha)$, there exist equivalence classes $L^\sim$ and $R^\sim$ with $k \in L^\sim$ and $k \in R^\sim$. Given in arbitrary order each, let $L_1^\sim, L_2^\sim, \ldots, L_p^\sim$ be all equivalence classes resulting from $\sim_l$ and $R_1^\sim, R_2^\sim, \ldots, R_q^\sim$ all equivalence classes resulting from $\sim_r$. Consequently, $L_1^\sim \cup L_2^\sim \cup \ldots \cup L_p^\sim$ and $R_1^\sim \cup R_2^\sim \cup \ldots \cup R_q^\sim$ are two disjoint partitions of $\text{var}(\alpha)$ induced by $\sim_l$ and $\sim_r$. Then, for $i \in \{1, 2, \ldots, p\}$, $i' \in \{1, 2, \ldots, q\}$ and for every $k \in \text{var}(\alpha)$, the morphism $\sigma_\alpha^{\text{su}}$ is given by

$$\sigma_\alpha^{\text{su}}(k) := \begin{cases} \mathtt{a}\,\mathtt{b}^{3k}\,\mathtt{a}\ \mathtt{a}\,\mathtt{b}^{3k+1}\,\mathtt{a}\ \mathtt{a}\,\mathtt{b}^{3k+2}\,\mathtt{a}\,, & \nexists\, i : k = \min L_i^\sim \wedge \nexists\, i' : k = \min R_{i'}^\sim, \\ \mathtt{b}\,\mathtt{a}^{3k}\,\mathtt{b}\ \mathtt{a}\,\mathtt{b}^{3k+1}\,\mathtt{a}\ \mathtt{a}\,\mathtt{b}^{3k+2}\,\mathtt{a}\,, & \nexists\, i : k = \min L_i^\sim \wedge \exists\, i' : k = \min R_{i'}^\sim, \\ \mathtt{a}\,\mathtt{b}^{3k}\,\mathtt{a}\ \mathtt{a}\,\mathtt{b}^{3k+1}\,\mathtt{a}\ \mathtt{b}\,\mathtt{a}^{3k+2}\,\mathtt{b}\,, & \exists\, i : k = \min L_i^\sim \wedge \nexists\, i' : k = \min R_{i'}^\sim, \\ \mathtt{b}\,\mathtt{a}^{3k}\,\mathtt{b}\ \mathtt{a}\,\mathtt{b}^{3k+1}\,\mathtt{a}\ \mathtt{b}\,\mathtt{a}^{3k+2}\,\mathtt{b}\,, & \exists\, i : k = \min L_i^\sim \wedge \exists\, i' : k = \min R_{i'}^\sim. \end{cases}$$

Obviously, for every $\alpha \in \mathbb{N}^+$, $\sigma_\alpha^{\text{su}}$ is injective.

As an illustration of Definition 2 we now identify $\sigma_\alpha^{\text{su}}$ for an example pattern:

Example 3. Let $\alpha := 1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 4 \cdot 3 \cdot 1 \cdot 2 \cdot 3 \cdot 2 \cdot 4$. Evidently, α is succinct (cf. Definition 1). Then $L_1 = \{3\}$, $L_2 = \{1, 3\}$, $L_3 = \{2, 4\}$, $L_4 = \{2\}$ and $R_1 = \{2\}$, $R_2 = \{3, 4\}$, $R_3 = \{1, 2\}$, $R_4 = \{3\}$. This leads to $L_1^\sim = \{1, 3\}$, $L_2^\sim = \{2, 4\}$ and $R_1^\sim = \{1, 2\}$, $R_2^\sim = \{3, 4\}$, and, thus, $\sigma_\alpha^{\text{su}}(1) = \mathtt{b} \ldots \mathtt{b}$, $\sigma_\alpha^{\text{su}}(2) = \mathtt{a} \ldots \mathtt{b}$, $\sigma_\alpha^{\text{su}}(3) = \mathtt{b} \ldots \mathtt{a}$ and $\sigma_\alpha^{\text{su}}(4) = \mathtt{a} \ldots \mathtt{a}$.

Note that, for the pattern in Example 3, the injective morphism $\sigma'(i) = \mathtt{a}\,\mathtt{b}^i$, $i \in \mathbb{N}$, generates an unambiguous word as well, and even the non-injective morphism σ'' given by $\sigma''(2) := \mathtt{b}$, $\sigma''(1) = \sigma''(3) = \sigma''(4) := \mathtt{a}$ has this property. Additionally, for every pattern α satisfying, for some $n \geq 1$ and $p_1, p_2, \ldots, p_n \geq 2$, $\alpha = 1^{p_1} \cdot 2^{p_2} \cdot \ldots \cdot n^{p_n}$, $\sigma'(\alpha)$ is known to be strongly unambiguous (cf. Reidenbach [12]). Thus, $\sigma_\alpha^{\text{su}}$ is "sufficient" for generating an unambiguous word (as to be shown in the subsequent lemmata), but, in general, it is not "necessary" since there can be significantly shorter words with the desired property. Contrary to this, for our initial example α', it is obviously necessary to give a morphism which is more sophisticated than σ' (cf. Section 1).

As the underlying principles of both Definition 2 and the subsequent lemmata are fairly complex we now briefly discuss the line of reasoning in this section: The basic idea for Definition 2 is derived from the proof of Proposition 2. Therein, we can observe that, for the abstract example pattern, the ambiguity of the regarded word is caused by the fact that, for all of the left neighbours of some variables (i.e., in terms of Definition 2, for some L_i), the morphic images end with the same letter. We call an L_i *(morphically) homogeneous* (*with respect to* a morphism σ) if it shows such a property. Thus, it seems reasonable to choose a morphism such that in each L_i with $|L_i| \geq 2$ there are two variables whose morphic images end with different letters (or, in other words, convert L_i into a *(morphically) heterogeneous* set), but this idea may lead to conflicting assignments:

Example 4. Let $\alpha := 1 \cdot 2 \cdot 3 \cdot 2 \cdot 1 \cdot 3 \cdot 1$. Thus, $L_1 = \{2, 3\}$, $L_2 = \{1, 3\}$ and $L_3 = \{1, 2\}$. Then, for a binary alphabet, there is no morphism σ such that, at a time, L_1, L_2 and L_3 are morphically heterogeneous with respect to σ.

Fortunately, a thorough combinatorial consideration shows that it suffices to guarantee heterogeneity of each $L_i^{\sim}$ (cf. Lemma 2); as these sets are disjoint, this avoids any contradictory assignments. Note that these statements analogously hold for R_i and $R_i^{\sim}$ (regarding the first letter of the morphic images of the variables in these sets instead of the last one).

Before we formally analyse the consequences of morphic heterogeneity we now address the injectivity of $\sigma_\alpha^{\mathrm{su}}$. Of course, according to our goal of finding a structure-preserving morphic image, we have to choose injective morphisms; in addition, however, we can observe that non-injectivity can cause ambiguity:

Example 5. Let $\alpha := 1\cdot 2\cdot 1\cdot 3\cdot 2\cdot 1\cdot 4\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 6\cdot 5\cdot 6\cdot 7\cdot 1\cdot 7\cdot 2$. This pattern is succinct, and $L_1^{\sim} = \mathrm{var}(\alpha)$, $R_1^{\sim} = \mathrm{var}(\alpha)$. Then, for $L_1^{\sim}$ and $R_1^{\sim}$, the non-injective morphism σ given by $\sigma(1) := \mathtt{b}$ and $\sigma(i) := \mathtt{a}$, $i \in \mathrm{var}(\alpha) \setminus \{1\}$ leads to the desired heterogeneity. Nevertheless, there is a morphism ρ with $\rho(\alpha) = \sigma(\alpha)$ and $\rho \neq \sigma$, namely $\rho(4) := \varepsilon$, $\rho(5) := \mathtt{a\,a}$ and $\rho(i) := \sigma(i)$, $i \in \mathrm{var}(\alpha) \setminus \{4, 5\}$.

The injectivity of $\sigma_\alpha^{\mathrm{su}}$ is brought about by the assignment of three unique *segments* $\mathtt{c\,d}^m\,\mathtt{c}$, $\mathtt{c}, \mathtt{d} \in \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{c} \neq \mathtt{d}$, $m \in \mathbb{N}$, to each variable. This allows to prove the following phenomenon, which, in a similar way, has been examined in [13]:

Lemma 1. *Let $\alpha \in \mathbb{N}^+$ be succinct. Then, for every morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ with $\rho(\alpha) = \sigma_\alpha^{\mathrm{su}}(\alpha)$ and for every $i \in \mathrm{var}(\alpha)$, $\rho(i) = \ldots\, \mathtt{g}\ \mathtt{a\,b}^{3i+1}\,\mathtt{a}\ \mathtt{h}\, \ldots$, $\mathtt{g}, \mathtt{h} \in \{\mathtt{a}, \mathtt{b}\}$.*

Lemma 1 requires an extensive reasoning. Due to space constraints, we omit the proof and refer the reader to Lemma 1 in [13], which can give a rough idea of it.

We conjecture that strong unambiguity can also be ensured by a morphism which, for every variable $i \in \mathbb{N}$, assigns only the first and the last segment of $\sigma_\alpha^{\mathrm{su}}(i)$ to i, but, in this case, we expect the proof of the equivalent to the following lemma to be significantly more difficult.

As explained above, we now conclude the proof of our main result with the examination of the use of morphic heterogeneity:

Lemma 2. *Let $\alpha \in \mathbb{N}^+$ be succinct. If, for every morphism $\rho : \mathbb{N}^+ \longrightarrow \{\mathtt{a}, \mathtt{b}\}^*$ with $\rho(\alpha) = \sigma_\alpha^{\mathrm{su}}(\alpha)$ and for every $i \in \mathrm{var}(\alpha)$, $\rho(i) = \ldots\ \mathtt{a\,b}^{3i+1}\,\mathtt{a}\ \ldots$ then $\sigma_\alpha^{\mathrm{su}}(\alpha)$ is strongly unambiguous.*

Proof. If $|\,\mathrm{var}(\alpha)| = 1$ then every morphic image of α is strongly unambiguous, and therefore, in this case, Lemma 2 holds trivially. Hence, let $|\,\mathrm{var}(\alpha)| \geq 2$. We start the proof with a small remark that is needed at several stages of the proof:

Claim 1. For every $i \in \mathrm{var}(\alpha)$, $\mathtt{c}, \mathtt{d} \in \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{c} \neq \mathtt{d}$, and $y \in \{0, 1, 2\}$, $\rho(i) \neq \ldots\ \mathtt{c\,d}^{3i+y}\,\mathtt{c}\ \ldots\ \mathtt{c\,d}^{3i+y}\,\mathtt{c}\ \ldots$.

Proof (Claim 1). Claim 1 directly follows from the precondition $\rho(\alpha) = \sigma_\alpha^{\mathrm{su}}(\alpha)$ since, obviously, $|\,\sigma_\alpha^{\mathrm{su}}(\alpha)|_{\mathtt{c\,d}^{3i+y}\,\mathtt{c}} = |\alpha|_i$. □(*Claim 1*)

Now assume to the contrary that there is a morphism ρ with $\rho(\alpha) = \sigma_\alpha^{\mathrm{su}}(\alpha)$ such that, for every $i \in \mathrm{var}(\alpha)$, $\rho(i) = \ldots\ \mathtt{a\,b}^{3i+1}\,\mathtt{a}\ \ldots$ and, for some $i' \in \mathrm{var}(\alpha)$, $\rho(i') \neq \sigma_\alpha^{\mathrm{su}}(i')$. Then there necessarily is a $j \in \mathrm{var}(\alpha)$ such that, for some $\mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}, \mathtt{h} \in \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{c} \neq \mathtt{d}$, $\mathtt{e} \neq \mathtt{f}$,

(a) $\rho(j) = \dots \texttt{g c d}^{3j} \texttt{c a b}^{3j+1} \texttt{a} \dots$ or
(b) $\rho(j) = \dots \texttt{a b}^{3j+1} \texttt{a e f}^{3j+2} \texttt{e h} \dots$.

We restrict the following reasoning to case (a) since an analogous argumentation can be applied to case (b) (using $\sim_r$ instead of $\sim_l$): Note that, due to $|\operatorname{var}(\alpha)| \geq 2$ and the succinctness of α, $|\alpha|_j \geq 2$ and therefore $L_j \neq \emptyset$ (for the definition of L_j, see Definition 2). Hence, let k be an arbitrary variable in L_j. Consequently, for any $\texttt{c}, \texttt{d} \in \{\texttt{a}, \texttt{b}\}$, $\texttt{c} \neq \texttt{d}$, $\rho(k) \neq \dots \texttt{c d}^{3k+2} \texttt{c}$.

Case 1: $\alpha = j \dots$.
Then we can directly follow from Claim 1: $\sigma^{\mathrm{su}}_\alpha(\alpha) = \texttt{e f}^{3j} \texttt{e} \dots \neq \rho(\alpha) = \dots \texttt{g c d}^{3j} \texttt{c} \dots$, with $\texttt{c}, \texttt{d}, \texttt{e}, \texttt{f}, \texttt{g} \in \{\texttt{a}, \texttt{b}\}$, $\texttt{c} \neq \texttt{d}$, $\texttt{e} \neq \texttt{f}$. This contradicts the condition $\sigma^{\mathrm{su}}_\alpha(\alpha) = \rho(\alpha)$.

Case 2: $\alpha = \dots k$.
Then $\sigma^{\mathrm{su}}_\alpha(\alpha) = \dots \texttt{c d}^{3k+2} \texttt{c} \neq \rho(\alpha)$ since $\rho(k) \neq \dots \texttt{c d}^{3k+2} \texttt{c}$ for $\texttt{c}, \texttt{d} \in \{\texttt{a}, \texttt{b}\}$, $\texttt{c} \neq \texttt{d}$. This again contradicts the condition $\sigma^{\mathrm{su}}_\alpha(\alpha) = \rho(\alpha)$.

Case 3: $\alpha \neq j \dots$ and $\alpha \neq \dots k$.
For the equivalence classes $L^\sim_1, L^\sim_2, \dots, L^\sim_p$ derived from the construction of $\sigma^{\mathrm{su}}_\alpha$, let $\iota \in \{1, 2, \dots, p\}$ with $L_j \subseteq L^\sim_\iota$. Then, since all $L^\sim_1, L^\sim_2, \dots, L^\sim_p$ are pairwise disjoint, this ι is unique. Now we can collect a number of facts that facilitate the argumentation in Case 3. The first holds as α is succinct:

Claim 2. If $\alpha \neq j \dots$ and $\alpha \neq \dots k$ then $|L^\sim_\iota| \geq 2$.

Proof (Claim 2). If $|L_j| \geq 2$ then Claim 2 holds trivially. Hence, let $L_j = \{k\}$. Then, for every occurrence of j in α, the conditions $\alpha \neq j \dots$ and $|\alpha|_j \geq 2$ lead to $\alpha = \dots k \cdot j \dots$. Thus, due to the succinctness of α, there are some $j_1, j_2, \dots, j_m \in \operatorname{var}(\alpha)$, $m \geq 2$, with $\alpha = \dots k \cdot j_r \dots$, $1 \leq r \leq m$. Additionally, because of $\alpha \neq \dots k$, there must be an $s \in \{1, 2, \dots, m\}$ and a $\bar{k} \in \operatorname{var}(\alpha)$, $\bar{k} \neq k$, with $\alpha = \dots \bar{k} \cdot j_s \dots$, since, otherwise, α would either be prolix or start with a j_r, $r \in \{1, 2, \dots, m\}$, leading to the same argumentation as in Case 1. Consequently, $L_{j_s} \supseteq \{k, \bar{k}\}$ and therefore $L_j \subset \{k, \bar{k}\} \subseteq L_{j_s} \subseteq L^\sim_\iota$. □(*Claim 2*)

Now, for every $L^\sim$ among $L^\sim_1, L^\sim_2, \dots, L^\sim_p$, the next fact follows by definition since these equivalence classes are composed by union of *non-disjoint* sets (cf. Definition 2 and, e.g., Example 3):

Claim 3. If $|L^\sim| \geq 2$ then, for every $\hat{k} \in L^\sim$, there is an $L_{\hat{\jmath}} \subseteq L^\sim$ with $|L_{\hat{\jmath}}| \geq 2$ and $\hat{k} \in L_{\hat{\jmath}}$.

We conclude the list of preliminary claims with the following one, that deals with a crucial phenomenon which is reflected in the transitivity of $\sim_l$:

Claim 4. For every $\hat{k} \in L^\sim_\iota$ and any $\texttt{e}, \texttt{f} \in \{\texttt{a}, \texttt{b}\}$, $\texttt{e} \neq \texttt{f}$, $\rho(\hat{k}) \neq \dots \texttt{e f}^{3\hat{k}+2} \texttt{e} \dots$.

Proof (Claim 4). With regard to any $\hat{k}' \in L_j \subseteq L^\sim_\iota$, Claim 4 holds because of the precondition $\rho(i) = \dots \texttt{a b}^{3i+1} \texttt{a} \dots$, $i \in \operatorname{var}(\alpha)$, because of Claim 1 and the fact that, for some $\texttt{c}, \texttt{d}, \texttt{e}, \texttt{f} \in \{\texttt{a}, \texttt{b}\}$, $\texttt{c} \neq \texttt{d}$, $\texttt{e} \neq \texttt{f}$, $\rho(j) = \dots \texttt{e c d}^{3j} \texttt{c a b}^{3j+1} \texttt{a} \dots$ and

$$\rho(\hat{k}' \cdot j) = \dots \texttt{a b}^{3\hat{k}'+1} \texttt{a e f}^{3\hat{k}'+2} \texttt{e c d}^{3j} \texttt{c a b}^{3j+1} \texttt{a} \dots .$$

We now regard all $\hat{k}'' \in L^{\sim}_{\iota}$ for which there is an $L_{j'}$ with $\hat{k}', \hat{k}'' \in L_{j'}$ (recall that $\hat{k}' \in L_j$). Then—since Claim 4 is satisfied for $\hat{k}'$ and, consequently, $\rho(j') = \ldots\ \mathtt{e}\ \mathtt{c}\,\mathtt{d}^{3j'}\ \mathtt{c}\ \mathtt{a}\,\mathtt{b}^{3j'+1}\,\mathtt{a}\ \ldots$, $\mathtt{c}, \mathtt{d}, \mathtt{e} \in \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{c} \neq \mathtt{d}$—Claim 4 holds for these $\hat{k}''$ as well. Now we proceed to all $\hat{k}''' \in L^{\sim}_{\iota}$ for which there is an $L_{j''}$ with $\hat{k}'', \hat{k}''' \in L_{j''}$ (recall that $\hat{k}'' \in L_{j'}$). Then, as Claim 4 is satisfied for $\hat{k}''$, Claim 4 holds for all $\hat{k}'''$ and so on. Consequently, according to the construction of $L^{\sim}_{\iota}$ (cf. Definition 2) Claim 4 holds for every $\hat{k} \in L^{\sim}_{\iota}$. □(*Claim 4*)

We now can conclude our argumentation on Case 3: According to Claim 2, $|L^{\sim}_{\iota}| \geq 2$. Let $k_{\sharp} := \min L^{\sim}_{\iota}$; then, due to Claim 3, there is an $j_{\sharp} \in \mathrm{var}(\alpha)$ with $k_{\sharp} \in L_{j_{\sharp}}$ and $|L_{j_{\sharp}}| \geq 2$. Consequently, let $\bar{k}_{\sharp} \in \mathrm{var}(\alpha)$ with $k_{\sharp} \neq \bar{k}_{\sharp}$ and $\{k_{\sharp}, \bar{k}_{\sharp}\} \subseteq L_{j_{\sharp}}$. Then, because of $k_{\sharp} = \min L^{\sim}_{\iota}$ and $\bar{k}_{\sharp} \in L^{\sim}_{\iota}$, $\sigma^{\mathrm{su}}_{\alpha}(k_{\sharp}) = \ldots\ \mathtt{b}$ and $\sigma^{\mathrm{su}}_{\alpha}(\bar{k}_{\sharp}) = \ldots\ \mathtt{a}$. Referring to the condition $\rho(i) = \ldots\ \mathtt{a}\,\mathtt{b}^{3i+1}\,\mathtt{a}\ \ldots$, $i \in \mathrm{var}(\alpha)$, to Claim 1 and to Claim 4, these different endings of $\sigma^{\mathrm{su}}_{\alpha}(k_{\sharp})$ and $\sigma^{\mathrm{su}}_{\alpha}(\bar{k}_{\sharp})$ imply

$$\ldots\ \boxed{\mathtt{b}}\ \mathtt{c}\,\mathtt{d}^{3j_{\sharp}}\ \mathtt{c}\ \mathtt{a}\,\mathtt{b}^{3j_{\sharp}+1}\,\mathtt{a}\ \ldots = \rho(j_{\sharp}) = \ldots\ \boxed{\mathtt{a}}\ \mathtt{c}\,\mathtt{d}^{3j_{\sharp}}\ \mathtt{c}\ \mathtt{a}\,\mathtt{b}^{3j_{\sharp}+1}\,\mathtt{a}\ \ldots\ ,$$

for some $\mathtt{c}, \mathtt{d} \in \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{c} \neq \mathtt{d}$. This contradicts $\mathtt{a} \neq \mathtt{b}$. □

With Lemma 1 and Lemma 2, Theorem 3 follows immediately.

Acknowledgements. The authors are indebted to Frank Hechler and the referees for numerous useful comments and, in particular, for pointing out the analogy between prolix patterns and fixed points of morphisms.

References

1. D. Angluin. Finding patterns common to a set of strings. *J. Comput. Syst. Sci.*, 21:46–62, 1980.
2. C. Choffrut and J. Karhumäki. Combinatorics of words. In [14].
3. T. Harju and J. Karhumäki. Morphisms. In [14].
4. T. Head. Fixed languages and the adult languages of 0L schemes. *Intern. J. of Computer Math.*, 10:103–107, 1981.
5. H. Jürgensen and S. Konstantinidis. Codes. In [14].
6. S. Lange and R. Wiehagen. Polynomial-time inference of arbitrary pattern languages. *New Generation Comput.*, 8:361–370, 1991.
7. F. Levé and G. Richomme. On a conjecture about finite fixed points of morphisms. *Theor. Comp. Sci.*, to appear.
8. M. Lipponen and G. Păun. Strongly prime PCP words. *Discrete Appl. Math.*, 63:193–197, 1995.
9. M. Lothaire. *Combinatorics on Words.* Addison-Wesley, Reading, MA, 1983.
10. A. Mateescu and A. Salomaa. Patterns. In [14].
11. A. Mateescu and A. Salomaa. Finite degrees of ambiguity in pattern languages. *RAIRO Inform. Théor. Appl.*, 28(3–4):233–253, 1994.
12. D. Reidenbach. A non-learnable class of E-pattern languages. *Theor. Comp. Sci.*, to appear.
13. D. Reidenbach. A discontinuity in pattern inference. In *Proc. STACS 2004*, volume 2996 of *LNCS*, pages 129–140, 2004.
14. G. Rozenberg and A. Salomaa. *Handbook of Formal Languages*, volume 1. Springer, Berlin, 1997.