A Negative Result on Inductive Inference of Extended Pattern Languages

Daniel Reidenbach

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

Abstract. The question of learnability of the class of extended pattern languages is considered to be one of the eldest and outstanding open problems in inductive inference of formal languages. This paper provides an appropriate answer presenting a subclass – the terminal-free extended pattern languages – that is not learnable in the limit. In order to achieve this result we will have to limit the respective alphabet of terminal symbols to exactly two letters.
In addition we will focus on the impact of ambiguity of pattern languages on inductive inference of terminal-free extended pattern languages. The conventional view on nondeterminism in patterns inspired by formal language theory is transformed into an approach that meets the requirements of inductive inference. These studies will lead to some useful learnability criteria for classes of terminal-free extended pattern languages.

1 Introduction

The analysis of learnability of formal languages – originating in [Gol67] – is one of the main subjects in inductive inference. Meanwhile there exist some powerful criteria on language identification in the limit (like in [Ang80], [Wri89] and [BCJ99]) deriving from these studies. Contrary to the discouraging findings concerning *super-finite* classes of languages, learnability of the class of *pattern languages* was shown by Angluin ([Ang79] and [Ang80]). In the sequel there has been a variety of additional studies (e.g. in [LW91], [WZ94]) concerning complexity of learning algorithms, consequences of different input data, and so on.

Pattern languages in the sense of Angluin disallow any empty substitution of variables. The question whether the class of *extended pattern languages* – tolerating empty substitutions and also known as *erasing pattern languages* or *E-pattern languages* – is learnable has proven to be more complicated than that of "standard" pattern languages. It has been the focus of attention since 1982 when Shinohara was the first to deal with extended pattern languages and it was characterized as "one of the outstanding open problems of inductive inference" by Mitchell in 1998 (cf. [Mit98]). Up to the present there are only two non-trivial subclasses of extended pattern languages known to be learnable, both of them

restricting the occurrences of variables. In detail, the class of extended pattern languages where the patterns contain at most m distinct variables ([Wri89]) and the class of *quasi-regular pattern languages*, with every variable occurring exactly m times (first shown in [Shi82a] for $m = 1$, the general case shown by Mitchell in [Mit98]), can be mentioned. Mitchell also pointed out that the full class of extended pattern languages is learnable in the limit if the respective alphabet of terminal symbols is infinite or singular. The research on extended pattern languages within the scope of formal language theory was initiated in [JKS+94] and led among others to some interesting findings concerning the decidability of inclusion.

In order to take an undisguised look at the difficulties leading to the restrictions in the approaches of Shinohara, Wright and Mitchell we will focus in the following sections on terminal-free extended pattern languages. The main result of this paper will state that the class of extended pattern languages – and also that of terminal-free extended pattern languages – is not learnable in the limit if the respective terminal alphabet consists of exactly two letters. Subsequent to this the impact of nondeterminism of pattern languages on the questions of learning theory will be analysed, but first a sufficiently precise definition of the concepts to be used will be given.

2 Preliminaries

Let Σ be an alphabet of *terminal* symbols and $X = \{x_1, x_2, x_3, \cdots\}$ an infinite alphabet of *variables*, $\Sigma \cap X = \emptyset$. If we are talking just of an *alphabet* we mean an alphabet of terminals. If A is an arbitrary alphabet then A^+ denotes the set of all non-empty words over A and A^* the set of all (empty and non-empty) words over A. We will use lower case letters from the end of the Latin alphabet in order to name words of terminal symbols. We designate the *empty* word as e. $|\cdot|$ denotes the size of an alphabet or the length of a word, respectively, and $|w|_{\mathtt{a}}$ the frequency of a letter $\mathtt{a}$ in a word w. A *pattern* is a word over $\Sigma \cup X$, a *terminal-free pattern* is a word over X; naming patterns we will use lower case letters from the beginning of the Greek alphabet. $\mathrm{var}(\alpha)$ denotes the set of all variables of a pattern α.

A *substitution* is a morphism $\sigma : (\Sigma \cup X)^* \longrightarrow \Sigma^*$ such that $\sigma(\mathtt{a}) = \mathtt{a}$ for all $\mathtt{a} \in \Sigma$. An *inverse substitution* is a morphism $\overline{\sigma} : \Sigma^* \longrightarrow X^*$. The *extended pattern language* of a pattern α is defined as

$$L_\Sigma(\alpha) := \{w \in \Sigma^* \mid \exists \sigma : \ \sigma(\alpha) = w\}.$$

If there is no need to give emphasis to the concrete shape of Σ we denote the extended pattern language of a pattern α simply as $L(\alpha)$. Each function $t : \mathbb{N} \longrightarrow \Sigma^*$ satisfying $\{t(n) \mid n \geq 0\} = L(\alpha)$ is called a *text* for $L(\alpha)$.

Following [Mit98] we designate a pattern α as *succinct* if and only if for all patterns β

$$L(\beta) = L(\alpha) \implies |\alpha| \leq |\beta|.$$

According to the studies of Mateescu and Salomaa ([MS94]) we denote a word w as *ambiguous* (in respect of a pattern α) if and only if there exist two substitutions σ and σ' such that $\sigma(x_i) \neq \sigma'(x_i)$ for some $x_i \in \mathrm{var}(\alpha)$, but $\sigma(\alpha) = w = \sigma'(\alpha)$. We call a word *unambiguous* (in respect of a pattern α) if it is not ambiguous.

In [JSSY95] it is shown that the inclusion of two arbitrary extended pattern languages is not decidable. Fortunately this fact does not hold true for terminal-free extended pattern languages. As this is of great importance for the following studies we now cite a respective theorem of [JSSY95]:

Fact 1. *Let $\alpha, \beta \in X^*$ be two arbitrarily given terminal-free patterns. Then $L(\beta) \subseteq L(\alpha)$ if and only if there exists a morphism $\phi : X^* \longrightarrow X^*$ such that $\phi(\alpha) = \beta$.*

We investigate the identification of extended pattern languages in Gold's learning model (cf. [Gol67]), so we have to agree on the corresponding notions. Let S be any total computable function reading initial segments of texts and returning patterns. Each such function is called a *strategy*. If α is a pattern and t a text for $L(\alpha)$ we say that S *identifies* $L(\alpha)$ *from* t, iff the sequence of patterns returned by S, when reading t, converges to a pattern β, such that $L(\beta) = L(\alpha)$. Any set PAT* of extended pattern languages is *learnable in the limit* (or: *inferrable from positive data*) iff there is a strategy S identifying each language $L \in$ PAT* from any corresponding text.

The learnability characterization used in this paper originates from Angluin. In fact it combines Condition 1 and Theorem 1 of [Ang80]:

Fact 2. *An arbitrary subclass* PAT* *of extended pattern languages is inferrable from positive data iff there exists an effective procedure that enumerates for every pattern α with $L(\alpha) \in$ PAT* a set T_α such that*

- *$T_\alpha \subseteq L(\alpha)$,*
- *T_α is finite, and*
- *$T_\alpha \not\subseteq L(\beta)$ for all $L(\beta) \in$ PAT* with $L(\beta) \subset L(\alpha)$.*

T_α is called a telltale (in respect of α and PAT**).*

The second learnability criterion we will use also derives from Angluin (combining Condition 2, Condition 4 and Corollary 3 of [Ang80]):

Fact 3. *Let* PAT* *be an arbitrary subclass* PAT* *of extended pattern languages such that for two languages $L_1, L_2 \in$ PAT* inclusion is decidable. Then* PAT* *is inferrable from positive data if there exists for every pattern α with $L(\alpha) \in$ PAT* a set T_α such that*

- *$T_\alpha \subseteq L(\alpha)$,*
- *T_α is finite, and*
- *$T_\alpha \not\subseteq L(\beta)$ for all $L(\beta) \in$ PAT* with $L(\beta) \subset L(\alpha)$.*

3 A Non-Learnable Subclass of Extended Pattern Languages

In this section we will present a specific and simply structured extended pattern language that does not have a telltale. This fact entails the conclusion that the class of extended pattern languages is not learnable in the limit.

Our argumentation will lead to conflicting presumptions; on the one hand elements with such a feature seem to be quite frequent among the set of all patterns, on the other hand we will have to limit the used alphabet to exactly two letters, turning the examined class of languages into a rather peculiar one.

To begin with we will name a well known type of pattern that is as useful for our line of reasoning as it is inconvenient for the needs of inductive inference:

Definition 1 (Passe-partout). *Let α be a pattern and $W \subset L(\alpha)$ a finite set of words. Let β be a pattern, such that*

- *$W \subseteq L(\beta)$ and*
- *$L(\beta) \subset L(\alpha)$.*

We then say that β is a passe-partout (for α and W).

Note that if there exists a passe-partout for a pattern α and a set of words W, then W is not a telltale for $L(\alpha)$.

Now we will present the crucial lemma of this section:

Lemma 1. *Let $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ be an alphabet and*

$$\alpha := x_1 x_1 \; x_2 x_2 \; x_3 x_3$$

a pattern. Then for any finite $W \subset L_\Sigma(\alpha)$ there exists a terminal-free passe-partout.

Proof. If W is empty the above statement is trivially true. Given an arbitrary non-empty $W = \{w_1, w_2, \cdots, w_n\} \subset L(\alpha)$, the following procedure constructs a passe-partout β:

As an inverse substitution we define for every w_i a morphism $\overline{\sigma}_i : \Sigma^* \longrightarrow X^*$ by

$$\overline{\sigma}_i(\mathtt{c}) := \begin{cases} x_{2i-1} & , \quad \mathtt{c} = \mathtt{a}, \\ x_{2i} & , \quad \mathtt{c} = \mathtt{b}. \end{cases}$$

As $W \subset L(\alpha)$, for every w_i, $1 \leq i \leq n$, there exists a substitution σ_i satisfying $\sigma_i(\alpha) = w_i$. Constructing a set of $3n$ strings $\gamma_{i,k} \in X^*$ we now will identify the necessary elements of β.

Case (i) $\sigma_i(x_3) = v_1\,\mathtt{a}\,v_2,\ \ v_1, v_2 \in \{\mathtt{b}\}^* \ \wedge\ \sigma_i(x_1), \sigma_i(x_2) \in \{\mathtt{b}\}^*$ or
$\sigma_i(x_3) = v_1\,\mathtt{b}\,v_2,\ \ v_1, v_2 \in \{\mathtt{a}\}^* \ \wedge\ \sigma_i(x_1), \sigma_i(x_2) \in \{\mathtt{a}\}^*$.
Thus, $\sigma_i(x_3)$ contains a letter exactly once and w_i contains this letter exactly twice. In this case we define

$$\begin{aligned}\gamma_{i,1} &:= \overline{\sigma}_i\,(\sigma_i(x_1)\,\sigma_i(x_2))\,,\\ \gamma_{i,2} &:= \overline{\sigma}_i\,(\sigma_i(x_3))\,,\\ \gamma_{i,3} &:= e.\end{aligned}$$

Note that in (i) w_i is ambiguous, so that the above definition provides a pattern $\gamma_i := \gamma_{i,1}\,\gamma_{i,2}\,\gamma_{i,3}$ with $w_i \in L(\gamma_i)$.

Case (ii) $\sigma_i(x_3)$ is empty or w_i contains every letter of $\sigma_i(x_3)$ at least four times.
In this case we simply define

$$\gamma_{i,k} := \overline{\sigma}_i\,(\sigma_i(x_k))\,,\ 1 \le k \le 3.$$

Obviously (ii) also provides a pattern $\gamma_i := \gamma_{i,1}\,\gamma_{i,2}\,\gamma_{i,3}$ with $w_i \in L(\gamma_i)$.

Combining the fragments of all γ_i in an appropriate manner we now compose the resulting pattern of the procedure:

$$\beta := \underbrace{\gamma_{1,1}\,\gamma_{2,1}\,\cdots\,\gamma_{n,1}}_{\sim x_1}\ \underbrace{\gamma_{1,1}\,\gamma_{2,1}\,\cdots\,\gamma_{n,1}}_{\sim x_1}\ \underbrace{\gamma_{1,2}\,\gamma_{2,2}\,\cdots\,\gamma_{n,2}}_{\sim x_2}\ \underbrace{\gamma_{1,2}\,\gamma_{2,2}\,\cdots\,\gamma_{n,2}}_{\sim x_2}$$
$$\underbrace{\gamma_{1,3}\,\gamma_{2,3}\,\cdots\,\gamma_{n,3}}_{\sim x_3}\ \underbrace{\gamma_{1,3}\,\gamma_{2,3}\,\cdots\,\gamma_{n,3}}_{\sim x_3}.$$

In order to conclude the proof we now show that β indeed is a passe-partout for α and W:

1. We define a substitution $\sigma_i' : X^* \longrightarrow \Sigma^*$ by

$$\sigma_i'(x_j) := \begin{cases} \mathtt{a} & , \quad j = 2i-1,\\ \mathtt{b} & , \quad j = 2i,\\ e & , \quad \text{else.}\end{cases}$$

   Obviously $\sigma_i'(\beta) = w_i$, and thus $W \subseteq L(\beta)$.
2. α and β are both terminal-free patterns, and due to the above depicted shape of these patterns there exists a morphism $\phi : X^* \longrightarrow X^*$ with $\phi(\alpha) = \beta$. Thus, $L(\beta)$ is a subset of $L(\alpha)$ (according to the inclusion criterion in [JSSY95] described in Fact 1).
   On the other hand there exists no morphism $\psi : X^* \longrightarrow X^*$ with $\psi(\beta) = \alpha$, as $\gamma_{i,3} \neq \delta_1\,x_j\,\delta_2$, $1 \le i \le n$, if $x_j \notin \mathrm{var}(\delta_k)$, $1 \le k \le 2$, and $x_j \notin \mathrm{var}(\gamma_{i,l})$, $1 \le l \le 2$. Because of this fact

$$\beta \neq \cdots x_p \cdots x_p \cdots x_q \cdots x_q \cdots x_r \cdots x_r \cdots$$

   if there are no other occurrences of at least one of these variables in β.
   $\Longrightarrow\ L(\beta) \subset L(\alpha)$ □

Obviously every variable of α occurs exactly twice. Ironically its language thus belongs to a class that – according to [Mit98] – is learnable in the limit. Nevertheless, the findings of Mitchell and Lemma 1 are consistent as β not necessarily is quasi-regular. Accordingly the quasi-regular pattern languages to some extent are not learnable on account of their shape as such but as they do not include all possible passe-partouts.

In addition we want to point out that ambiguity of words plays a major role in the proof of Lemma 1: a telltale for an extended pattern language has to include words with a substitution of variables containing a unique letter – as taken into consideration in case (i). If the alphabet consists of just two letters these specific words may turn out to be ambiguous, leading to a decisive loss of significance. We will revert to this aspect in the next section.

Referring to Angluin the impact of Lemma 1 on inductive inference can be stated with little effort:

Theorem 1. *The class of terminal-free extended pattern languages is not inferrable from positive data if the respective alphabet consists of exactly two letters.*

Proof. Let Σ be an alphabet, $|\Sigma| = 2$. Lemma 1 provides a terminal-free pattern α, such that for any finite set $W \subset L_\Sigma(\alpha)$ there exists a passe-partout. Therefore the class of terminal-free extended pattern languages does not satisfy Angluin's Condition 1, and according to Theorem 1 of [Ang80] it is not inferrable from positive data (as presented in Fact 2). □

Theorem 1 entails the finding, that all positive results on inductive inference of extended pattern languages cited in the introduction follow the only practicable course: any learnable (sub-)class of these languages has to be provided with some restrictions on the shape of the variables or the alphabet of terminal symbols.

Finally we now explicitly will formulate the trivial conclusion of Theorem 1:

Corollary 1. *The class of extended pattern languages is not inferrable from positive data if the respective alphabet consists of exactly two letters.*

As explained in section 2 we investigate in this paper the standard learning model of inductive inference regarding a class of languages as learnable if, roughly speaking, for every of its elements a syntactical convergence of hypotheses can be achieved. A second, widely analysed model – known as *behaviorally correct* learning or BC-learning – replaces this aspect by the weaker claim of a semantic convergence (cf. [CL82], [OW82] and concerning the inference of functions [Bar74]). According to [BCJ96] Angluin's Condition 2 (cf. [Ang80]) fully

characterizes any BC-learnable class of languages. Hence it is obvious that the class of extended pattern languages (and that of terminal-free extended pattern languages as well) is not BC-learnable, too, since there does not *exist* any telltale for a terminal-free extended pattern language (as shown in Lemma 1). In addition we want to point out that – due to the decidability of inclusion – every result in Gold's model concerning the learnability of terminal-free extended pattern languages directly can be interpreted as a statement on BC-learning.

4 The Importance of Unambiguous Words

As already mentioned above ambiguity of words is the core of the construction of a passe-partout in the proof of Lemma 1. In this section we will return to that point with a more general view.

Nondeterminism of pattern languages has been examined by Mateescu and Salomaa in [MS94]. Within the scope of learning theory however it seems to be useful to focus on a slightly different aspect of nondeterminism. Instead of searching for the maximum ambiguity of a pattern and its words we conjecture that it is beneficial to analyse the minimum ambiguity of the words of extended pattern languages. Being more precisely we suggest to pose the question, whether there exist certain unambiguous words in every terminal-free extended pattern language. This approach is inspired by the special needs of inductive inference concerning the analysis of subset relations of languages.

The following Theorem – providing a criterion for the learnability of terminal-free extended pattern languages – will specify our intention:

Theorem 2. *Let Σ be an alphabet. Let $\mathrm{Pat}^{\star}_{\mathrm{tf}}$ be a set of terminal-free patterns and $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ the corresponding class of extended pattern languages. If for any $\alpha \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$ there exists a finite set of substitutions $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$ such that*

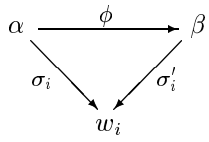
1. *for every i, $1 \leq i \leq n$, $\sigma_i(\alpha)$ is unambiguous in respect of α and*
2. *for every $x_j \in \mathrm{var}(\alpha)$ there exists an i, $1 \leq i \leq n$, and an $\mathtt{a} \in \Sigma$ such that $|\sigma_i(x_j)|_{\mathtt{a}} = 1$ and $|\sigma_i(\alpha)|_{\mathtt{a}} = |\alpha|_{x_j}$*

then $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ is inferrable from positive data.

Proof. Given $\alpha \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$, we define a set T_α of words over Σ by

$$T_\alpha := \{w_i \mid 1 \leq i \leq n \ \wedge \ \sigma_i(\alpha) = w_i\}.$$

We now will show that T_α is a telltale for $L(\alpha)$ in respect of $\mathrm{PAT}^{\star}_{\mathrm{tf}}$. For that purpose assume $T_\alpha \subseteq L(\beta) \subseteq L(\alpha)$ for some $\beta \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$. Then – according to the inclusion criterion of [JSSY95] described in Fact 1 – there exists a morphism $\phi : X^* \longrightarrow X^*$ such that $\phi(\alpha) = \beta$. Moreover, there exists a second set of substitutions $\{\sigma'_1, \sigma'_2, \cdots, \sigma'_n\}$ with $\sigma'_i(\beta) = w_i$ for all i, $1 \leq i \leq n$. Consequently, the following diagram illustrates the relationship of α, β, and w_i for every i, $1 \leq i \leq n$:



Hence it is obvious that

$$\sigma_i(x_j) = \sigma_i'(\phi(x_j)) \text{ for all } x_j \in \mathrm{var}(\alpha) \text{ and } 1 \leq i \leq n$$

as every w_i is unambiguous in respect of α. Because of this fact (and because of condition 2) $\phi(x_j)$ must have the following shape:

$$\phi(x_j) = \gamma_1\, x_{j_{\mathtt{a}}}\, \gamma_2 \text{ for all } x_j \in \mathrm{var}(\alpha)$$

with $\gamma_1, \gamma_2 \in X^*$ and $|\beta|_{x_{j_{\mathtt{a}}}} = |\alpha|_{x_j}$. Thus, the morphism $\psi : X^* \longrightarrow X^*$ defined by

$$\psi(x_k) := \begin{cases} x_j & , \quad k = j_{\mathtt{a}}, \\ e & , \quad \text{else} \end{cases}$$

leads to $\psi(\beta) = \alpha$ and – according to Fact 1 – $L(\beta) = L(\alpha)$. Consequently, $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ satisfies the conditions of Fact 3 as T_α is a telltale for $L(\alpha)$ in respect of $\mathrm{PAT}^{\star}_{\mathrm{tf}}$.
$\Longrightarrow$ $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ is inferrable from positive data. □

As a consequence of Theorem 2 of [MS94] – dealing with changing degrees of ambiguity of the same extended pattern language depending on the question of whether the respective pattern is succinct or not – we consider it as vital to restrict the search for unambiguous words on a set of succinct patterns. In addition we may see the results of the previous section as a hint that – depending on the concrete shape of the class of pattern languages to be examined – an alphabet of at least three letters is necessary in order to construct a set of unambiguous words.

The following example demonstrates the way how Theorem 2 might be used:

Example 1. Following Shinohara (cf. [Shi82b]) we define:

Definition 2 (Terminal-free non-cross patterns). *A pattern α is called a terminal-free non-cross pattern if and only if it satisfies*

$$\alpha = x_1^{e_1}\, x_2^{e_2}\, x_3^{e_3}\, \cdots x_n^{e_n}$$

for some n and numbers $e_1, e_2, \cdots, e_n$ with $n \geq 1$ and $e_i \geq 1$, $1 \leq i \leq n$.

We denote an extended pattern language $L(\alpha)$ as terminal-free extended non-cross pattern language if and only if α is a terminal-free non-cross pattern.

We state without proof that the class of terminal-free extended non-cross pattern languages is inferrable from positive data for any finite alphabet Σ with at least two letters. To this end let α be an arbitrarily chosen terminal-free non-cross pattern and $\{\mathtt{a}, \mathtt{b}\} \subseteq \Sigma$. Given the substitution σ by

$$\sigma(x_j) := \mathtt{a}\,\mathtt{b}^j$$

the set T_α defined by

$$T_\alpha := \begin{cases} \{\mathtt{a}\} & , \quad \exists\, i : 1 \leq i \leq n \wedge e_i = 1, \\ \{\sigma(\alpha)\}. & , \quad \text{else} \end{cases}$$

is a telltale for α. Note that the absence of possible passe-partouts among the terminal-free non-cross patterns again is the *conditio sine qua non* for this conclusion. As – according to [Mit98] – the full class of extended pattern languages is learnable in the limit if the repective alphabet is infinite or consists of just one letter, the above statement implies the learnability of the class of terminal-free extended non-cross pattern languages for any alphabet.

For the purpose of this paper however our example looks different: Let again α be a terminal-free non-cross pattern and Σ a finite alphabet, $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \subseteq \Sigma$. Then the set of substitutions $\{\sigma_1^{\mathrm{nc}}, \sigma_2^{\mathrm{nc}}, \cdots, \sigma_n^{\mathrm{nc}}\}$ given by

$$\sigma_j^{\mathrm{nc}}(i_k) := \begin{cases} \mathtt{b}^{f_i}\mathtt{b}^{f_i} & , \quad j \neq i \wedge i \text{ is odd}, \\ \mathtt{c}^{f_i}\mathtt{c}^{f_i} & , \quad j \neq i \wedge i \text{ is even}, \\ \mathtt{b}^{f_i}\,\mathtt{a}\,\mathtt{b}^{f_i} & , \quad j = i \wedge i \text{ is odd}, \\ \mathtt{c}^{f_i}\,\mathtt{a}\,\mathtt{c}^{f_i} & , \quad j = i \wedge i \text{ is even}, \end{cases}$$

$1 \leq i \leq n$ and $1 \leq j \leq n$, with

$$f_i := \begin{cases} 1 & , \quad i = 1, \\ \prod_{j=1}^{i-1} e_j & , \quad i > 1 \end{cases}$$

satisfies the conditions of Theorem 2 if $L(\alpha) \neq L(x_1)$. $L(x_1)$ is learnable in the limit using the set $\{\mathtt{a}\}$ as a telltale. Thus, choosing this approach leads to the conclusion that the class of terminal-free extended non-cross pattern languages is inferrable from positive data if Σ is finite and consists of at least three letters.

Consequently, in the present example Theorem 2 does not lead to an optimal result. Thus, we suggest to examine the use Theorem 1 for those classes of terminal-free extended pattern languages that turn out to be not learnable in the limit if the corresponding alphabet consists of two letters, such as the full class of terminal-free extended pattern languages (cf. Theorem 1).

5 Diffuse Words

In Theorem 2 we use unambiguous words in order to guarantee fixed spheres of responsibility for every variable when generating a word. We now will present an – in a sense – weaker claim leading to a comparable result concerning the existence of a telltale in a terminal-free extended pattern language.

We will start with a precise explanation of the concept to be used:

Definition 3 (Diffuse). *Let α be a pattern, $|\alpha| := n$, and σ a substitution. If $m \leq n$ let $\alpha^m = x_{i_1}\, x_{i_2} \cdots x_{i_m}$ be the initial segment of length m of α. Let ϵ be the smallest natural number such that for every substitution σ' with $\sigma'(\alpha) = \sigma(\alpha)$ and for every m, $1 \leq m \leq n$,*

$$|\sigma(\alpha^m)| - \epsilon \;\leq\; |\sigma'(\alpha^m)| \;\leq\; |\sigma(\alpha^m)| + \epsilon.$$

If $|\sigma(x_i)| \geq 2\epsilon + 1$ for all $x_i \in \mathrm{var}(\alpha)$ then we call the word $\sigma(\alpha)$ diffuse (of degree ϵ)(in respect of α).

Thus, a diffuse word contains certain letters that – regarding all possible substitutions – have to be generated by distinct variables. Note that all words being diffuse of degree 0 are unambiguous but not every unambiguous word necessarily has to be diffuse of degree 0 (because of the condition $|\sigma(x_i)| \geq 1$ for all x_i).

The following example illustrates Definition 3:

Example 2. We define a pattern α by

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

Obviously α is terminal-free and succinct. We examine the substitution σ given by

$$\sigma(x_1) := \mathtt{baa},\ \sigma(x_2) := \mathtt{aba},\ \sigma(x_3) := \mathtt{bba},\ \sigma(x_4) := \mathtt{bbb}.$$

There exists only one different substitution σ' such that $\sigma'(\alpha) = \sigma(\alpha)$, σ' given by

$$\sigma'(x_1) = \mathtt{ba},\ \sigma'(x_2) = \mathtt{aaba},\ \sigma'(x_3) = \mathtt{bb},\ \sigma'(x_4) = \mathtt{abbb}.$$

Taking a look at the resulting word of both substitutions

$$\overbrace{\mathtt{b\,a\,a\,a\,b\,a\,b\,b\,a\,b\,b\,b\,b\,a\,a\,b\,b\,b\,b\,b\,a\,a\,b\,a}}^{\sigma(\alpha)}$$
$$\underbrace{\mathtt{b\,a\,a\,a\,b\,a\,b\,b\,a\,b\,b\,b\,b\,a\,a\,b\,b\,b\,b\,b\,a\,a\,b\,a}}_{\sigma'(\alpha)}$$

it is obvious that $\sigma(\alpha)$ is diffuse of degree 1.

The learning criterion for terminal-free extendend pattern languages based on diffuse words reads as follows:

Theorem 3. *Let Σ be an alphabet. Let $\mathrm{Pat}^{\star}_{\mathrm{tf}}$ be a set of terminal-free patterns and $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ the corresponding class of extended pattern languages. If for every $\alpha \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$ there exist natural numbers $\epsilon_1, \epsilon_2, \cdots, \epsilon_n \geq 0$ and a finite set of substitutions $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$ such that*

1. *for every i, $1 \leq i \leq n$, $\sigma_i(\alpha)$ is diffuse of degree ϵ_i in respect of α and*
2. *for every $x_j \in \mathrm{var}(\alpha)$ there exists an i, $1 \leq i \leq n$, such that*

$$\sigma_i(x_j) = v_{i_1}\, v_{i_2}\, \mathtt{a}\, v_{i_3}\, v_{i_4}$$

for a letter $\mathtt{a} \in \Sigma$ and some $v_{i_1}, v_{i_2}, v_{i_3}, v_{i_4} \in \Sigma^*$, $|v_{i_1}| = \epsilon_i = |v_{i_4}|$, such that $|\sigma_i(\alpha)|_{\mathtt{a}} = |\alpha|_{x_j}$

then $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ is inferrable from positive data.

Proof. Given $\alpha \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$, we define a set T_α of words over Σ by

$$T_\alpha := \{w_i \mid 1 \leq i \leq n \ \wedge\ \sigma_i(\alpha) = w_i\}.$$

We now will show that T_α is a telltale for $L(\alpha)$ in respect of $\mathrm{PAT}^{\star}_{\mathrm{tf}}$. For that purpose assume $T_\alpha \subseteq L(\beta) \subseteq L(\alpha)$ for some $\beta \in \mathrm{Pat}^{\star}_{\mathrm{tf}}$. Then – according to Fact 1 – there exists a morphism $\phi : X^* \longrightarrow X^*$ such that $\phi(\alpha) = \beta$. Moreover, there exists a second set of substitutions $\{\sigma'_1, \sigma'_2, \cdots, \sigma'_n\}$ with $\sigma'_i(\beta) = w_i$ for all i, $1 \leq i \leq n$.

Because every w_i, $1 \leq i \leq n$, is diffuse in respect of α and because of condition 2 we may conclude that for every $x_j \in \mathrm{var}(\alpha)$ there exists a σ'_i such that

$$\sigma'_i(\phi(x_j)) = u_1\, v_{i_2}\, \mathtt{a}\, v_{i_3}\, u_2$$

for a letter $\mathtt{a} \in \Sigma$, some words $u_1, u_2 \in \Sigma^*$ and v_{i_2}, v_{i_3} deriving from $\sigma_i(x_j)$. In addition it is obvious that $|\sigma'_i(\beta)|_{\mathtt{a}} = |\alpha|_{x_j}$ can be stated for this σ'_i. Therefore – like in the proof of Theorem 2 – ϕ must have the shape

$$\phi(x_j) = \gamma_1\, x_{j_{\mathtt{a}}}\, \gamma_2 \text{ for all } x_j \in \mathrm{var}(\alpha)$$

with $\gamma_1, \gamma_2 \in X^*$ and $|\beta|_{x_{j_{\mathtt{a}}}} = |\alpha|_{x_j}$. Thus, the morphism $\psi : X^* \longrightarrow X^*$ defined by

$$\psi(x_k) := \begin{cases} x_j & , \quad k = j_{\mathtt{a}}, \\ e & , \quad \text{else} \end{cases}$$

leads to $\psi(\beta) = \alpha$ and $L(\beta) = L(\alpha)$. Consequently, $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ satisfies the conditions of Fact 3 as T_α is a telltale for $L(\alpha)$ in respect of $\mathrm{PAT}^{\star}_{\mathrm{tf}}$.
$\Longrightarrow$ $\mathrm{PAT}^{\star}_{\mathrm{tf}}$ is inferrable from positive data. □

Note that Example 1 is valid for Theorem 3 as well, since all of the words generated by the given substitutions are diffuse of degree 0 in respect of any terminal-free non-cross pattern. Additionally, we generally suggest to restrict the search for substitutions satisfying the conditions of Theorem 3 on succinct patterns and an alphabet of at least three letters.

6 Concluding Remarks

Since we focus in the present paper on terminal-free patterns it seems worth mentioning that most aspects of the previous two sections may also be expressed using the terms of Post's correspondence problem (as it is revealed by Example 2).

Finally we presume that for every succinct terminal-free pattern there exists a set of substitutions satisfying the conditions of Theorem 2 (or Theorem 3, respectively), if the corresponding alphabet consists of at least three letters. Thus, we conjecture (referring to Theorem 1 and the results from [Mit98]) that the class of terminal-free extended pattern languages is inferrable from positive data if and only if the respective alphabet does not consist of exactly two letters.

Acknowledgements

The results of this paper are part of a diploma thesis at the University of Kaiserslautern. The author wishes to thank his supervisor Sandra Zilles for her extraordinary support, Jochen Nessel for some useful hints, and Rolf Wiehagen for his inspiring introduction to inductive inference and continuous advice.

References

[Ang79] D. Angluin. Finding patterns common to a set of strings. In *Proceedings, 11th Annual ACM Symposium on Theory of Computing*, pages 130–141, 1979.

[Ang80] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[Bar74] J. Barzdin. Two theorems on the limiting synthesis of functions. *Theory of Algorithms and Programs, Latvian State University, Riga*, 210:82–88, 1974.

[BCJ96] G.R. Baliga, J. Case, and S. Jain. Synthesizing enumeration techniques for language learning. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 169–180, 1996.

[BCJ99] G.R. Baliga, J. Case, and S. Jain. The synthesis of language learners. *Information and Computation*, 152:16–43, 1999.

[CL82] J. Case and L. Lynes. Machine inductive inference and language identification. In *Lecture Notes in Computer Science*, volume 140, pages 107–115. Proceeedings of the 9th International Colloquium on Automata, Languages and Programming, 1982.

[Gol67] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[JKS+94] T. Jiang, E. Kinber, A. Salomaa, K. Salomaa, and S. Yu. Pattern languages with and without erasing. *Intern. J. Computer Math.*, 50:147–163, 1994.

[JSSY95] T. Jiang, A. Salomaa, K. Salomaa, and S. Yu. Decision problems for patterns. *Journal of Computer and System Science*, 50:53–63, 1995.

[LW91] S. Lange and R. Wiehagen. Polynomial-time inference of arbitrary pattern languages. *New Generation Computing*, 8:361–370, 1991.

[Mit98] A.R. Mitchell. Learnability of a subclass of extended pattern languages. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 64–71, 1998.

[MS94] A. Mateescu and A. Salomaa. Nondeterminism in patterns. In *Lecture Notes in Computer Science*, volume 775, pages 661–668. STACS 94, 11th Annual Symposium on Theoretical Aspects of Computer Science, 1994.

[OW82] D. Osherson and S. Weinstein. Criteria of language learning. *Information and Control*, 52:123–138, 1982.

[Shi82a] T. Shinohara. Polynomial time inference of extended regular pattern languages. In *Lecture Notes in Computer Science*, volume 147, pages 115–127. RIMS Symposia on Software Science and Engineering, Proceedings, Kyoto, 1982.

[Shi82b] T. Shinohara. Polynomial time inference of pattern languages and its application. In *Proceedings of the 7th IBM Symposium on Mathematical Foundations of Computer Science*, pages 191–209, 1982.

[Wri89] K. Wright. Identification of unions of languages drawn from an identifiable class. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pages 328–333, 1989.

[WZ94] R. Wiehagen and T. Zeugmann. Ignoring data may be the only way to learn efficiently. *Journal of Experimental and Theoretical Artificial Intelligence*, 6:131–144, 1994.