

## Chapter 1

---

# Some remarks on arbitrary multiple pattern interpretations

C. MARTÍN-VIDE, V. MITRANA

**ABSTRACT.** A word  $w$  is obtained by an arbitrary  $n$ -pattern interpretation of a word  $x$  if there are  $n$  homomorphisms  $h_1, h_2, \dots, h_n$  and a positive integer  $k$  such that  $w = h_{i_1}(x)h_{i_2}(x) \cdots h_{i_k}(x)$  with  $1 \leq i_j \leq n$  for all  $1 \leq j \leq k$ . This arbitrary multiple pattern interpretation of words is naturally extended to languages. We investigate some closure properties of the families of languages obtained by arbitrary multiple pattern interpretations of finite, regular, and context-free languages, respectively. We show that the first of these families forms an infinite hierarchy and give a characterization of the arbitrary multiple pattern interpretation of finite languages. Two concepts of ambiguity and inherent ambiguity of multiple pattern interpretation are defined. It is shown that both properties are decidable for multiple pattern interpretations on finite languages but strong ambiguity is not decidable for multiple pattern interpretations on the class of context-free languages. The paper also contains a series of open problems.

## 1.1 Introduction

In Angluin (1980), a new way for defining a language is considered. Instead of identifying completely a language by generative devices as formal grammars or by recognition devices as automata, sometimes it is useful to consider less strict definitions. In the aforementioned work, the notion of *pattern* is defined as a word containing variables and constants, and then the language defined by a pattern  $\alpha$  consists of all words obtained from  $\alpha$  by substituting a string of constants for each variable. The substitution has to be uniform in the sense that the multiple occurrences of a variable must be replaced with the same string.

In the seminal work of Angluin 1980 the variables have to be replaced with nonempty strings, while in Jiang et al. (1994) substitutions by empty strings are allowed, which makes an essential difference. In Restivo and Salemi (2002), one proposes a generalization of this definition: the distinction between variables and constants is discarded. Given two strings  $x$  and  $w$ , possibly over the same alphabet,

$x$  is a *pattern description* of a string  $w$  ( $w$  is obtained by a *pattern interpretation* of  $x$ ) if  $w$  is the homomorphic image of  $x$ : in other words, there exists a homomorphism  $h$  such that  $w = h(x)$ .

Programming languages can be viewed as pattern interpretations of some languages. For instance, the main non-context-free features of programming languages are the necessity to define labels and the necessity to declare identifiers. The necessity of defining labels can be expressed by pattern interpretation in the following way. A correct program containing labels should be the pattern interpretation of the following general description:

$$(part\_pro)_1(label) : (statement) (part\_pro)_2 goto (label) (part\_pro)_3,$$

where  $(label)$ ,  $(statement)$  and  $(part\_pro)_i$ ,  $i = 1, 2, 3$ , are variables representing labels, statements or other parts of a program, respectively. By interpreting this pattern, we have to substitute the two occurrences of the variable  $(label)$  by the same string of constants, hence observing the semantic restriction regarding labels definition.

In Kudlek et al. (2003) we propose a new pattern interpretation, namely *multiple pattern interpretation*. A word  $w$  is obtained by an arbitrary  $n$ -pattern interpretation of a word  $x$  if there are some homomorphisms  $h_1, h_2, \dots, h_n$ , and  $k \geq 1$ , such that  $w = h_{i_1}(x)h_{i_2}(x) \cdots h_{i_k}(x)$ ,  $1 \leq i_j \leq n$  for all  $1 \leq j \leq k$ . This arbitrary multiple pattern interpretation of words is naturally extended to languages, namely a language  $L$  is the arbitrary multiple pattern interpretation of another language  $E$  if  $L$  contains all words which are obtained by the same arbitrary multiple pattern interpretation of the words in  $E$ .

Multiple pattern interpretations seem to be of basic concern for linguists. Indeed, one may say that each sentence follows a pattern which is an element of a finite or infinite set, that is a language.

Let us first consider the following aspect of language acquisition: two-word utterances in the speech of a two-year old child, in accordance with Owens (2001). To understand them, linguists proposed several strategies actually based on ordered or arbitrary multiple pattern interpretations. First, some words are used without any positional consistency (*agent+action*, *action+object*, *agent+object*): the so-called *grouping pattern* in Brown and Leonard (1986). For example, the child may say “Eat cookie” or “Cookie eat”. Secondly, the utterance is characterized by a consistent word order which reflects patterns heard in adult speech: the so-called *positional associative pattern*, see Braine (1976). A third strategy, called *positional productive pattern* (Braine (1976)), is characterized by consistent word order and creative combinations. That is, children hypothesize a mini-language of patterns (*attribute+entity*, *possessor+possession*, *demonstrative+entity*, *agent+action*, etc.) and then interpret these patterns by words repeatedly heard in specific locations in adult

speech. It has been suggested that positional rules rather than semantic rules are the basis for early multiword utterances (Pine and Lieven (1993)). This strategy applies to adult speech as well, especially for nonnative speakers. For instance, a part of a speech can be constructed by an ordered interpretation of the word:

*article+adjective+noun+verb+pronoun+noun+adverb.*

By interpreting this word through a two-pattern interpretation, one gets the sentences:

*The young man ate his hamburger quickly.*

*A mad racer drove his car recklessly.*

On the other hand, syntactic theory is concerned, unlike traditional grammar, not with just describing specific languages but also with developing a general, universal theory. According to Borsley (1999), this means that other languages are always potentially relevant when one is describing a particular language. Thus, following Chomsky (1965, 1975); Kolb and Mönnich (1999), there exists a non-trivial set of axioms and a learnable extension of it that specify a possible natural language, and every natural language has a theory which is a learnable extension of the initial set. One has to determine a set of primitive blocks, operations which act on these blocks, an initial set and a learning procedure which maps the (primitive blocks of the) initial set onto the (utterances of the) steady state.

Furthermore, the *principles-and-parameters-model* discussed in Vogler (1999) has been established as a grammar formalism, based on the GB theory (Chomsky (1981, 1986)), aimed at describing the syntactical knowledge in a way that gives answers to questions concerning language acquisition and universal properties of languages. Thus, one may assume that the kernel of GB theory consists of a set of principles (= wellformedness conditions) and a way of interpreting them. In this respect, our multiple pattern interpretation of a language may be viewed as a particular case of such a general model.

## 1.2 Basic definitions

Let  $V$  and  $U$  be two alphabets. For a given integer  $n \geq 1$ , we denote by  $\Omega_{n,V,U}$  an  $n$ -tuple  $(h_1, h_2, \dots, h_n)$  of homomorphisms from  $V^*$  to  $U^*$ , and call it an  *$n$ -pattern interpretation*. The subscripts indicating the alphabets will be omitted when the two alphabets are self-understood. A multiple pattern interpretation is said to be *non-erasing* if all its components are non-erasing homomorphisms. For the rest of this paper, if not otherwise stated, all multiple pattern interpretations are non-erasing ones.

The *arbitrary multiple pattern interpretation* of  $L \subseteq V^*$  through  $\Omega_{n,V,U}$  is the language:

$$\Omega_{n,V,U}^*(L) = \{h_{i_1}(w)h_{i_2}(w) \cdots h_{i_r}(w) \mid w \in L, r \geq 1, 1 \leq i_j \leq n, \text{ for all } 1 \leq j \leq r\}.$$

If  $n = 1$ , hence  $\Omega$  consists of a single homomorphism  $h$  from  $V^*$  to  $U^*$ , we write  $h_{V,U}^*(L)$ , or  $h^*(L)$  provided that the alphabets  $V$  and  $U$  are self-understood from the context.

A homomorphism  $h : V^* \rightarrow U^*$  is termed a *letter-to-letter homomorphism* if  $h(a) \in U$  for any  $a \in V$ . A multiple pattern interpretation whose all components are letter-to letter homomorphism is called a multiple pattern letter-to-letter interpretation. The following families of languages are defined:

$$\begin{aligned} \mathbf{HOM}_n^*(\mathbf{X}) &= \{\Omega_{n,V,U}^*(L) \mid \text{for some } n\text{-pattern interpretation } \Omega_{n,V,U} \\ &\quad \text{and } L \in \mathbf{X}\}, \\ \mathbf{LHOM}_n^*(\mathbf{X}) &= \{\Omega_{n,V,U}^*(L) \mid \text{for some } n\text{-pattern letter-to-letter} \\ &\quad \text{interpretation } \Omega_{n,V,U} \text{ and } L \in \mathbf{X}\}, \end{aligned}$$

where  $\mathbf{X} \in \{\mathbf{FIN}, \mathbf{REG}, \mathbf{CF}\}$ . Here **FIN**, **REG**, **CF**, stand for the families of finite, regular, and context-free languages, respectively.

The above definition of a multiple pattern interpretation remembers the definition of a DTOL scheme, a very well-known type of Lindenmayer system. A DTOL scheme may be viewed as a multiple pattern interpretation  $\Omega_{n,V,V}$ , hence the  $n$  homomorphisms are actually endomorphisms on  $V^*$ . However, the way of interpreting a word through a DTOL scheme is different. A word  $w$  is obtained by a the DTOL scheme  $\Omega_{n,V,V}$  interpretation of a word  $x$  if there exists a positive integer  $k$  such that  $w = h_{i_1} \circ h_{i_2} \circ \dots \circ h_{i_k}(x)$ ,  $1 \leq i_j \leq n$ ,  $1 \leq j \leq k$ . Note the main difference: an arbitrary multiple pattern interpretation of a word is defined by a concatenation of the homomorphic images of that word while the interpretation through a DTOL scheme is a composition of the homomorphic images of that word. For more details about Lindenmayer systems the reader is referred to Rozenberg and Salomaa (1980). This makes an essential difference with respect to the families of languages obtained by these interpretation, namely the two families are incomparable. Indeed, the language  $\{a^{3^n} \mid n \geq 0\}$  can be obtained by interpreting the singleton set  $\{a\}$  through the DTOL scheme formed by the homomorphism  $h(a) = aaa$ , but it cannot be the arbitrary multiple pattern interpretation of any language since  $2 \cdot 3^n$  cannot be written as a power of 3. On the other hand, the regular language  $R = \{a^{2^n} \mid n \geq 1\}$  is the arbitrary 1-pattern interpretation of the same singleton set  $\{a\}$  but it cannot be obtained by interpreting any finite language

through a DTOL scheme. Assume the contrary, namely  $R$  is obtained by interpreting a finite language through the DTOL scheme  $\Omega_n = (h_1, h_2, \dots, h_n)$ . Since  $a^{2p}$  with  $p$  an arbitrarily large prime number is in  $R$ , it follows that  $a^{2p} = h_i(a^{2k})$  for some  $1 \leq i \leq n$  and  $k \geq 1$ . This implies that  $k = 1$  and  $h_i(a) = a^p$ . The contradiction follows from the fact that there are many prime numbers.

### 1.3 Some properties of the languages obtained by arbitrary multiple pattern interpretations

Clearly, for any alphabet  $V = \{a_1, a_2, \dots, a_n\}$  we have  $V^* = \Omega_{n, \{a\}, V}^*(\{a\})$ , where  $\Omega_{n, \{a\}, V} = (h_1, h_2, \dots, h_n)$ , each  $h_i$  being defined by  $h_i(a) = a_i$ . It is worth mentioning here a similar fact observed for pattern descriptions (Restivo and Salemi (2002)), namely the “worst” description of any word  $w$  (in the sense that this description gives the least information about the structure of  $w$ ) is the word  $a$ . On the other hand, it is easy to note that any language in  $\mathbf{HOM}_n^*(\mathbf{FIN})$  is either  $\{\varepsilon\}$  or infinite.

Note that if  $L$  is a finite language, then any language obtained by an arbitrary multiple pattern interpretation of  $L$  is a regular language. Indeed, for any  $\Omega_n = (h_1, h_2, \dots, h_n)$

$$\Omega_n^*(L) = \bigcup_{w \in L} \{h_1(w), h_2(w), \dots, h_n(w)\}^+ \quad (*)$$

holds. However, there are regular languages that are not the arbitrary multiple pattern interpretation of any language, finite or not. Such a language is  $a^+b^+$ . This follows immediately from a simple observation: if a word  $w$  lies in a language  $L$  defined by a multiple pattern interpretation of an arbitrary language, then  $ww$  must lie in  $L$ , too. Therefore, the following problem naturally arises: Is it decidable whether or not a given a context-free (regular) language is the arbitrary multiple pattern interpretation of a finite language?

First, we give a characterization of regular languages that can be obtained by an arbitrary multiple pattern interpretation of a finite language.

**Proposition 1.3.1** *A regular language is the arbitrary multiple pattern interpretation of a finite language if and only if it is a finite union of finitely generated semigroups w.r.t. concatenation.*

*Proof*: Clearly, if any arbitrary multiple pattern interpretation of a finite language is a finite union of finitely generated semigroups w.r.t. concatenation, see (\*) above.

Assume now that  $L = \bigcup_{i=1}^k F_i^+ \subseteq U^*$  for some  $k \geq 1$  and finite sets  $F_1, F_2, \dots, F_k$ . Let  $p = \max\{\text{card}(F_i) \mid 1 \leq i \leq k\}$ . It is plain that for each  $1 \leq i \leq k$ , one can find  $F'_i$  such that  $\text{card}(F'_i) = p$  and  $(F'_i)^+ = F_i^+$ . Suppose that  $F'_i = \{x_1^{(i)}, x_2^{(i)}, \dots, x_p^{(i)}\}$ ,  $1 \leq i \leq k$ . We define the alphabet  $V = \{a_1, a_2, \dots, a_k\}$  and the homomorphisms  $h_j : V^* \rightarrow U^*$ ,  $1 \leq j \leq p$ ,  $h_j(a_i) = x_j^{(i)}$ . The equality  $\Omega_p^*(V) = L$ , where  $\Omega_p = (h_1, h_2, \dots, h_p)$ , concludes the proof.  $\square$

Now the aforementioned problem can be reformulated as follows: Is it decidable whether or not a given regular language can be written as a finite union of finitely generated semigroups w.r.t. concatenation? Despite the problem seems to be “classic”, we were not able to find any result regarding this matter either to solve it. The problem is likely decidable for subclasses of regular languages, like slender regular languages (Păun and Salomaa (1995)) but the general case remains *open*.

For the class of context-free languages the problem was solved in Kudlek et al. (2003) by a usual reduction to the Post Correspondence Problem:

**Theorem 1.3.1** *Is it undecidable whether or not a given a context-free language is the arbitrary multiple pattern interpretation of a finite language.*

It is worth mentioning that there are non-context-free languages in  $\text{LHOM}_n^*(\mathbf{REG})$ , for any  $n \geq 1$ . However, for each  $k \geq 1$ , the family  $\text{HOM}_k(\mathbf{CF})$  contains context-sensitive languages only.

**Proposition 1.3.2** *For any  $k \geq 1$ ,  $\text{HOM}_n(\mathbf{CF}) \subset \text{NSPACE}(n)$ .*

*Proof:* Given  $L \subseteq V^*$  (by a context-free grammar or a pushdown automaton) and  $\Omega_{k,V,U} = (h_1, h_2, \dots, h_k)$  for some alphabet  $U$ , it is easy to construct an on-line Turing machine  $M$  with one storage tape which works as follows:

- The read-only input tape contains the string  $w$  of length  $n$  which is to be analyzed.  $M$  guesses a pair  $(x, i)$ , where  $x \in V^*$  is written on the storage tape,  $|x| \leq n$ , and  $1 \leq i \leq k$  such that  $h_i(x)$  is a prefix of  $w$ .

- Then,  $M$  checks whether or not  $x \in L$ . If  $x \in L$ , then  $M$  chooses nondeterministically  $1 \leq j \leq k$  and checks whether or not  $h_j(x)$  is the next factor of  $w$ , and continues in this way until the input string is completely read. When the input string is completely read,  $M$  accepts  $w$ .

- If there is no pair  $(x, i)$  as above, then  $M$  rejects  $w$ . Clearly,  $M$  accepts  $w$  iff  $w \in \Omega_n^*(L)$  and the total space used on the storage tape is bounded by  $|w|$ .  $\square$

It is easy to note that the family  $\mathbf{CF}$  in the above proof can be replaced by the family of context-sensitive languages.

As far as the possibility of having infinite hierarchies of families of languages defined by arbitrary multiple pattern interpretations of finite, regular, or context-free languages is concerned, we state the following partial result:

**Theorem 1.3.2** *Both hierarchies  $\mathbf{HOM}_n^*(\mathbf{FIN}) \subseteq \mathbf{HOM}_{n+1}^*(\mathbf{FIN})$  and  $\mathbf{LHOM}_n^*(\mathbf{FIN}) \subseteq \mathbf{LHOM}_{n+1}^*(\mathbf{FIN})$  are infinite.*

*Proof:* For a given  $n$  we consider the alphabet  $V_n = \{a_1, a_2, \dots, a_n\}$ . It is obvious that  $V_n^+ \in \mathbf{LHOM}_n^*(\mathbf{FIN})$ . Assume now that  $V_n^+ = \Omega_{n-1}^*(L)$ , where  $\Omega_{n-1} = (h_1, h_2, \dots, h_{n-1})$  and  $L$  is a finite language. We take  $z = a_1^{p_1} a_2^{p_2} \dots a_n^{p_n} \in V_n^+$  for arbitrarily large  $p_1, p_2, \dots, p_n$ . Assume that  $z = h_{(i,1)}(x) h_{(i,2)}(x) \dots h_{(i,k)}(x)$  for some  $k \geq 1$ . Since  $L$  is finite and  $p_1, p_2, \dots, p_n$  are arbitrarily large, it follows that there are  $1 \leq j_1, j_2, \dots, j_n \leq k$  such that  $h_{(i,j_t)}(x) \in a_t^+$  for all  $1 \leq t \leq n$ , which is contradictory.  $\square$

We do not know whether any of the hierarchies  $\mathbf{HOM}_n^*(\mathbf{X}) \subseteq \mathbf{HOM}_{n+1}^*(\mathbf{X})$ ,  $\mathbf{LHOM}_n^*(\mathbf{X}) \subseteq \mathbf{LHOM}_{n+1}^*(\mathbf{X})$ ,  $\mathbf{X} \in \{\mathbf{REG}, \mathbf{CF}\}$ , is infinite.

### 1.3.1 Closure properties

**Theorem 1.3.3** *1. For each  $n \geq 1$ , and  $\mathbf{X} \in \{\mathbf{FIN}, \mathbf{REG}, \mathbf{CF}\}$  the family  $\mathbf{HOM}_n^*(\mathbf{X})$  is closed under union and homomorphisms but fails to be closed under concatenation, intersection with regular sets, complement, and set difference.*

*Proof: Union:* Let

$$\begin{aligned} U_1^* \supseteq L_1 &= \overline{\Omega}_n^*(X_1), X_1 \subseteq V_1^*, X_1 \in \mathbf{FIN}, \\ U_2^* \supseteq L_2 &= \tilde{\Omega}_n^*(X_2), X_2 \subseteq V_2^*, X_2 \in \mathbf{FIN}, \\ &\text{where } n \geq 1, \\ \overline{\Omega}_n &= (h_1, h_2, \dots, h_n) \text{ and } \tilde{\Omega}_n = (g_1, g_2, \dots, g_n). \end{aligned}$$

Without loss of generality we may assume that  $V_1$  and  $V_2$  are disjoint. We define  $\Omega_n = (s_1, s_2, \dots, s_n)$ , where each homomorphism  $s_i : (V_1 \cup V_2)^* \rightarrow (U_1 \cup U_2)^*$ ,  $1 \leq i \leq n$ , is defined by

$$s_i(a) = \begin{cases} h_i(a), & \text{if } a \in V_1 \\ g_i(a), & \text{if } a \in V_2 \end{cases}$$

Obviously,  $L_1 \cup L_2 = \Omega_{n+m}^*(X_1 \cup X_2)$ .

*Homomorphisms:* Let  $\Omega_n = (h_1, h_2, \dots, h_n)$  be an  $n$ -pattern interpretation,  $h_i : V^* \rightarrow U^*$  for all  $1 \leq i \leq n$ , and  $g : U^* \rightarrow W^*$  be an arbitrary homomorphism. It is plain that  $\Omega_n^*(L) = \tilde{\Omega}_n^*(L)$  holds for any language  $L \subseteq V^*$  and  $\tilde{\Omega}_n = (g \circ h_1, g \circ h_2, \dots, g \circ h_n)$ .

*Concatenation:* Both languages  $a^+$  and  $b^+$  are multiple pattern interpretation of finite languages, but  $a^+b^+$  cannot be the multiple pattern interpretation of any language.

*Intersection with regular sets:* It follows immediate from the fact that  $V^*$  is a multiple pattern interpretation of a finite language whereas there are regular languages which cannot be obtained by a multiple pattern interpretation of any language.

*Complement and set difference:* We take the language  $L = \{a^{2n} \mid n \geq 1\} = h^*(\{a\})$ , where  $h(a) = aa$ . But  $\overline{L} = \{a\}^* \setminus L$  cannot be the arbitrary multiple pattern interpretation of any language since if  $w \in \overline{L}$ , then  $ww$  is also in  $\overline{L}$ , hence both  $|w|$  and  $|ww|$  must be odd, which is contradictory.  $\square$

We do not know whether or not a family  $\mathbf{HOM}_n(\mathbf{X})$  as above is closed under Kleene closure, but the next result may be interpreted as follows: The Kleene closure of an arbitrary multiple pattern interpretation loses, in some cases, information about the structure imposed by the pattern interpretation. Formally,

**Theorem 1.3.4** *Let  $\mathbf{F}$  be a family of languages closed under homomorphisms and union. Then, the Kleene closure of any arbitrary multiple pattern interpretation of a language in  $\mathbf{F}$  is in  $\mathbf{F}$ .*

*Proof :* Let  $L \subseteq V^*$  be a language in  $\mathbf{F}$  and  $\Omega_n = (h_1, h_2, \dots, h_n)$  be an  $n$ -pattern interpretation, for some  $n \geq 1$  and  $h_i : V^* \rightarrow U^*$ ,  $1 \leq i \leq n$ . We construct the new alphabets  $V_i = \{a_i \mid a \in V\}$  and define the letter-to-letter homomorphisms  $c_i : V^* \rightarrow V_i^*$ ,  $c_i(a) = a_i$ ,  $1 \leq i \leq n$ . We now consider the language

$$R = \left( \bigcup_{i=1}^n c_i(L) \right)^*$$

which is still in  $\mathbf{F}$ , and the homomorphism  $g : (\cup_{i=1}^n V_i)^* \rightarrow U^*$ , defined by  $g(a_i) = h_i(a)$  for all  $1 \leq i \leq n$ , and  $a \in V$ . We claim that

$$g(R) = (\Omega_n^*(L))^*$$

holds. Indeed, if  $z = g(y) \in g(R)$ , then  $y = x_1x_2 \dots x_p$  with  $x_j = c_{(i,j)}(z_j)$ ,  $z_j \in L$ ,  $1 \leq j \leq p$ . But

$$z = g \circ c_{(i,1)}(z_1) \dots g \circ c_{(i,p)}(z_p) = h_{(i,1)}(z_1) \dots h_{(i,p)}(z_p) \in (\Omega_n^*(L))^*.$$

Conversely, if  $y = z_1z_2 \dots z_r \in (\Omega_n^*(L))^*$ , then there are the strings  $x_1, x_2, \dots, x_r$  in  $L$  and the positive integers  $k_1, k_2, \dots, k_r$  such that  $z_i = h_{(i,1)}(x_i) \dots h_{(i,k_i)}(x_i)$ ,  $1 \leq (i, 1), (i_2), \dots, (i, k_i) \leq n$ ,  $1 \leq i \leq r$ . Hence

$$y = g(c_{(1,1)}(x_1) \dots c_{(1,k_1)}(x_1)c_{(2,1)}(x_2) \dots c_{(2,k_2)}(x_2) \dots c_{(r,1)}(x_r) \dots c_{(r,k_r)}(x_r)),$$

therefore  $y \in g(R)$ . □

**Corollary 1.3.1** *The Kleene closure of any arbitrary multiple pattern interpretation of a regular (context-free) language is regular (context-free).*

### 1.3.2 Ambiguity

Given an  $n$ -pattern interpretation  $\Omega_n = (h_1, h_2, \dots, h_n)$  and a language  $L$ , we say that  $\Omega_n$  is *weakly ambiguous* on  $L$  if there exists a word  $x \in L$  such that

$$h_{i_1}(x)h_{i_2}(x) \cdots h_{i_k}(x) = h_{j_1}(x)h_{j_2}(x) \cdots h_{j_p}(x),$$

holds for some  $k, p \geq 1$ .

Given an  $n$ -pattern interpretation  $\Omega_n = (h_1, h_2, \dots, h_n)$  and a language  $L$ , we say that  $\Omega_n$  is *strongly ambiguous* on  $L$  if there exist two different words  $x$  and  $y$  in  $L$  such that:

$$h_{i_1}(y)h_{i_2}(y) \cdots h_{i_k}(y) = h_{j_1}(x)h_{j_2}(x) \cdots h_{j_p}(x),$$

holds for some  $k, p \geq 1$ .

If  $\Omega_n$  is weakly/strongly ambiguous on any language  $L$  in a family of languages  $F$ , then  $\Omega_n$  is said to be *inherently weak/strong ambiguous* on  $F$ .

**Theorem 1.3.5** *It is decidable whether or not an arbitrary multiple pattern interpretation is weakly/strongly ambiguous on a finite language  $L$ .*

*Proof* : First we discuss how the strong ambiguity can be algorithmically checked. Let  $L = \{x_1, x_2, \dots, x_k\} \subseteq V^*$  and  $\Omega_{n, V, U}$  an arbitrary multiple pattern interpretation. We set  $L_i = L \setminus \{x_i\}$  for any  $1 \leq i \leq k$ . It is plain that  $\Omega_n$  is not strongly ambiguous on  $L$  if and only if  $\Omega_n^*(L_i) \cap \Omega_n^*(\{x_i\}) = \emptyset$  for all  $1 \leq i \leq k$ . Since  $\Omega_n^*(L_i) \cap \Omega_n^*(\{x_i\})$  is a regular language which can be effectively constructed and the emptiness problem is decidable for regular languages, we are done.

Let  $h_i(x_j) = w_{(i,j)}$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq k$ ; clearly  $\Omega_n$  is not weakly ambiguous if and only if for each  $1 \leq i \leq n$  the following two conditions are satisfied:

- (i)  $w_{(i,j)} \neq w_{(i,r)}$ ,  $1 \leq j \neq r \leq k$ ,
- (ii)  $\{w_{(i,1)}, w_{(i,2)}, \dots, w_{(i,k)}\}$  is a code, or equivalently the semigroup  $\{w_{(i,1)}, w_{(i,2)}, \dots, w_{(i,k)}\}^+$  is free.

Obviously, the first condition can be algorithmically checked while the second condition can be checked by Sardinas-Paterson algorithm (Berstel and Perrin

(1984)) for testing injectivity of the homomorphism  $f : \{a_1, a_2, \dots, a_k\}^* \rightarrow U^*$  defined by  $f(a_j) = w_{(i,j)}$ ,  $1 \leq j \leq k$ . It is known that, provided (i) holds, (ii) is satisfied if and only if  $f$  is injective (see, Shyr and Thierrin (1977)).  $\square$

**Theorem 1.3.6** *Let  $\mathbf{F}$  be a family of languages having the following two properties:*

1. *It is effectively closed under union, letter-to-letter homomorphisms and concatenation with symbols.*
2. *The problem “Given two languages  $L_1, L_2 \in \mathbf{F}$ , is  $L_1 \cap L_2$  empty?” is undecidable.*

*Then, given an  $n$ -pattern interpretation  $\Omega_n$ ,  $n \geq 1$ , and a language  $L \in \mathbf{F}$ , one cannot algorithmically decide whether or not  $\Omega_n$  is strongly ambiguous on  $L$ .*

*Proof :* Let  $L_1 \subseteq V_1^*$  and  $L_2 \subseteq V_2^*$  be two arbitrary languages in  $\mathbf{F}$ . We construct the letter-to-letter homomorphism  $g : V_2^* \rightarrow U^*$ , where  $U = \{X_a \mid a \in V_2\}$  such that  $U \cap V_1 = \emptyset$ , defined by  $g(a) = X_a$  for each  $a \in V_2$ , and then consider the language

$$L = \{\$\}L_1\{\#\} \cup \{\$\}g(L_2)\{\#\},$$

where  $\$$  and  $\#$  are two new symbols. Now we take the homomorphism  $h : (V_1 \cup U \cup \{\$, \#\})^* \rightarrow (V_1 \cup V_2 \cup \{\$, \#\})^*$ , defined by  $h(a) = a$  for  $a \in V_1$ ,  $h(X_b) = b$  for all  $b \in V_2$ , and  $h(\$) = \$$ ,  $h(\#) = \#$ .

Clearly,  $h$  is strongly ambiguous on  $L$  if and only if  $L_1 \cap L_2 \neq \emptyset$ . Indeed, if  $w \in L_1 \cap L_2$ , then  $\$g(w)\# \in \{\$\}g(L_2)\{\#\} \subseteq L$  and  $\$g(w)\#$  is different than  $\$w\#$ . But  $h(\$w\#) = h(\$g(w)\#)$ , hence  $h$  is strongly ambiguous on  $L$ .

Conversely, if  $h$  is strongly ambiguous on  $L$ , then there are  $x, y \in L$ ,  $x \neq y$ , such that  $h^n(x) = h^m(y)$  for some positive integers  $n, m$ . As  $h^k(z)$  contains exactly  $k$  occurrences of  $\$$  for any  $k \geq 1$  and  $z \in L$ , it follows that  $n = m$ . By the definition of  $h$ , one cannot have both strings either from  $\{\$\}L_1\{\#\}$  or from  $\{\$\}g(L_2)\{\#\}$ . Assume that  $x = \$z\#$ , with  $z \in L_1$ , and  $y = \$g(w)\#$ , with  $w \in L_2$  (the other case may be treated similarly). For  $h^n(x) = h^n(y)$ , it follows that  $(\$z\#)^n = (\$w\#)^n$ , hence  $z = w$ , that is  $L_1 \cap L_2 \neq \emptyset$ .  $\square$

Since the family of context-free languages has all properties above, we get:

**Corollary 1.3.2** *Given an  $n$ -pattern interpretation  $\Omega_n$ ,  $n \geq 1$ , and a context-free language  $L$ , one cannot algorithmically decide whether or not  $\Omega_n$  is strongly ambiguous on  $L$ .*

The decidability status of the weak ambiguity remains open.

## 1.4 Conclusion and further work

We have investigated some properties of the families of languages obtained by arbitrary multiple pattern interpretations of finite, regular, and context-free languages. Some closure properties, most of them being negative results, of these families were presented. In spite of the fact that a characterization of the arbitrary multiple pattern interpretation of finite languages was given the problem of deciding whether or not a regular language is such a language remained open. Two concepts of ambiguity and inherent ambiguity of multiple pattern interpretation were defined. It was shown that both properties were decidable for multiple pattern interpretations on finite languages but strong ambiguity was not decidable for multiple pattern interpretations on the class of context-free languages.

We finish with a brief discussion about some directions for further research. A multiple pattern interpretation is said to be *inherently weakly/strongly ambiguous* on a family of languages  $\mathbf{X}$  if it is weakly/strongly ambiguous on any language in  $\mathbf{X}$ . Clearly, there exist multiple pattern interpretations which are inherently weakly/strongly ambiguous on a given family  $\mathbf{X}$ ; it suffices to take all the homomorphic images as being power of a common word. A language  $L \in \mathbf{HOM}_n^*(\mathbf{X})$  is said to be inherently weakly/strongly ambiguous if for any multiple pattern interpretation  $\Omega_n$  such that  $\Omega_n^*(E) = L$ , for some  $E \in \mathbf{X}$ , then  $\Omega_n$  is weakly/strongly ambiguous on  $E$ . Note that it is not obligatory  $\Omega_n$  be inherently ambiguous on  $\mathbf{X}$ . We hope to return to this topic in a further work.

## Bibliography

- Angluin, D. (1980) Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62.
- Berstel, J. and Perrin, D. (1984) *The Theory of Codes*. Academic Press, New York.
- Borsley, R. (1999) *Syntactic Theory. A Unified Approach*. Edward Arnold, London.
- Braine, M. (1976) Children's first word combinations. *Monographs of the Society for Research in Child Development*, 41 (Serial NO. 164).
- Brown, B. and Leonard, L. (1986) Lexical influences on children's early positional patterns. *Journal of Child Language*, 13:219–229.
- Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA.
- Chomsky, N. (1975) *The Logical Structure of Linguistic Theory*. Plenum, New York.

*Some remarks on arbitrary multiple pattern interpretations: C. Martín-Vide, V. Mitrana/12*

Chomsky, N. (1981) *Lectures on Government and Binding*. Foris, Dordrecht.

Chomsky, N. (1986) *Knowledge of Language. Its Nature, Origin and Use*. Praeger, New York.

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

Kolb, H.P. and Mönnich, U. (1999) Introduction. In *The Mathematics of Syntactic Structure*, H.P. Kolb and U. Mönnich, (eds.). Mouton de Gruyter, Berlin.

Kudlek, M., Martín-Vide, C., and Mitrana, V. (2003) Multiple pattern interpretations. *GRAMMARS*, 5:223–238.

Owens, R.E. (2001) *Language Development: An Introduction*. Allyn and Bacon, Boston.

Păun, G. and Salomaa, A. (1995) Thin and slender languages. *Discrete Appl. Math.*, 61:257–270.

Pine, J.M. and Lieven, E.V.M. (1993) Reanalysing rote-learned phrases: individual differences in the transition to multiword speech. *Journal of Child Language*, 20:551–571.

Restivo, A. and Salemi, S. (2002) Words and patterns. *Lecture Notes in Computer Science 2295*. Springer, Berlin, 117–129.

Rozenberg, G. and Salomaa, A. (1980) *The Mathematical Theory of L Systems*. Academic Press, New York.

Rozenberg, G. and Salomaa, A. (eds.) (1997) *Handbook of Formal Languages*. 3 vols. Springer, Berlin.

Shyr, H.J. and Thierrin, G. (1977) Languages and MOL schemes. *R.A.I.R.O. Informatique Theorique/Theoretical Computer Science*, 1,4: 293–301.

Vogler, H. (1999) Principle languages and principle-based parsing. In *The Mathematics of Syntactic Structure*, H.P. Kolb and U. Mönnich, (eds.). Mouton de Gruyter, Berlin, 83–111.