Theoretical Computer Science 1 (1976) 199–214. © North-Holland Publishing Company

A DECOMPOSITION THEOREM FOR HYPER-ALGEBRAIC EXTENSIONS OF LANGUAGE FAMILIES*

Jan van LEEUWEN and Derick WOOD

Department of Mathematics, University of Utrecht, Utrecht, The Netherlands
Department of Applied Mathematics, McMaster University, Hamilton, Ontario, Canada

Communicated by Arto Salomaa
Received 30 April 1974
Revised 10 December 1974

Abstract. In modern theories of rewriting structures, hyper-sentential and hyper-algebraic extensions of languages-families have abstracted the imminent features of iterated parallel substitution. After introducing the concept of a (depth-bounded) translation, we show that for each language L hyper-algebraic over a natural family $\mathcal{F}$ there are $\mathcal{F}$-translations Δ, $\bar{\Delta}$ and languages $L_1, ..., L_m$ hyper-sentential over $\mathcal{F}$ such that $L = \bigcup_{i=1}^{m} \Delta^p(L_i) \cup \bar{\Delta}^q(\$)$, for some $p, q \geq 0$.

Two specializations of this result are given, when more assumptions are made about $\mathcal{F}$. These are, firstly, a translation theorem and, secondly, an alphabetic homomorphism theorem for hyper-algebraic extensions (an alphabetic homomorphism is a letter-to-letter or letter-to-ε homomorphism).

1. Introduction

A few years ago new developments in formal language theory focussed on the study of parallel rewriting as a means of language generation. Many of the results obtained are surveyed in Herman and Rozenberg [3] and Rozenberg and Salomaa [8], where also the original motivation from biology is described.

In this paper we study the fundamental aspects of some structural results concerning "tabled" parallel rewriting. The underlying structures were previously studied in van Leeuwen [5–7], Salomaa [9] and Wood [10]. The point of view expressed in these papers is that rewriting systems are considered to be substitutional devices for extending arbitrary families, rather than for the generation of particular families only.

* This work was carried out by the first author under NSF Grant GJ 998, NATO Grant 574 and under the auspices of the Center for Mathematical Methods in the Social, Biological and Health Sciences (SUNY, Buffalo, N.Y.) and by the second author under a National Research Council of Canada Grant No. A-7700.

The structures dealt with in this paper are "levelled trees". Going from one level to the next corresponds to applying one out of a finite set of $\mathcal{F}$-substitutions.

Beginning with an initial language from $\mathcal{F}$ in the first level, the entire collection of words we get level by level is a "language". The new family of languages obtained from $\mathcal{F}$ is said to be the hyper-sentential extension of $\mathcal{F}$. If we only include those words from the various levels which are over a designated "terminal" alphabet, i.e., take canonical restrictions of the languages hyper-sentential over $\mathcal{F}$, then we get a richer family, the hyper-algebraic extension of $\mathcal{F}$. Such extensions correspond to the "sentential form sets" and "(terminal) languages", respectively, in the more classical frame-works.

In this paper we analyse languages in the hyper-algebraic extension of natural families $\mathcal{F}$, and show that in a *well-defined sense* they are composed of finitely many hyper-sentential languages, which means that there is a surprising predictability in taking canonical restrictions. A similar, but stronger, result was recently obtained by Ehrenfeucht and Rozenberg [1–2] for extensions of the family of finite languages but it is here shown to be of a structural rather than combinatorial nature.

2. Preliminaries

For a language L we let Alph (L) be the smallest alphabet such that $L \subseteq$ Alph $(L)^*$.

A collection of languages $\mathcal{F}$ is called a *family* if it contains at least one non-trivial member and is closed under isomorphism. A family $\mathcal{F}$ is called *natural* if it contains all canonical restrictions of its members, i.e., if for each L in $\mathcal{F}$ and alphabet T, $L \cap T^*$ is in $\mathcal{F}$.

For a family $\mathcal{F}$ and alphabets T_1 and T_2 a map $\tau: T_1^* \rightarrow 2^{T_2^*}$ is said to be *an $\mathcal{F}$-substitution* iff it satisfies the following conditions: (i) $\tau(\varepsilon) = \{\varepsilon\}$, (ii) for all a in T_1, $\tau(a)$ is in $\mathcal{F}$, and (iii) for all x, y in T_1^*, $\tau(xy) = \tau(x)\,\tau(y)$.

We now introduce parallel rewriting systems for generating levelled trees in step by step fashion.

Definition. Let $\mathcal{F}$ be a family. *An $\mathcal{F}$-substitution scheme* is a pair $G = (V, \{\tau_1, \ldots, \tau_n\})$, $n > 0$, where V is an alphabet and all τ_i's are $\mathcal{F}$-substitutions over V, such that for all i and for all a in V, $\tau_i(a) \neq \emptyset$.

We write $x \overset{k}{\underset{G}{\Rightarrow}} y$ if y is in $\tau_{i_k} \ldots \tau_{i_1}(x)$ for some choice of $i_1, \ldots, i_k$ in $\{1, \ldots, n\}$. The steps involved can be visualized as iterated replacements (see Fig. 1). Going "downwards" some branches may disappear if the particular τ that is applied leads to the empty word, ε.

Definition. *The hyper-sentential extension of* $\mathcal{F}$ is the family of all languages L such that for some $\mathcal{F}$-substitution scheme G and M in $\mathcal{F}$, $L = \{y: x \overset{k}{\underset{G}{\Rightarrow}} y$ for some x in M and $k \geqslant 0\}$.

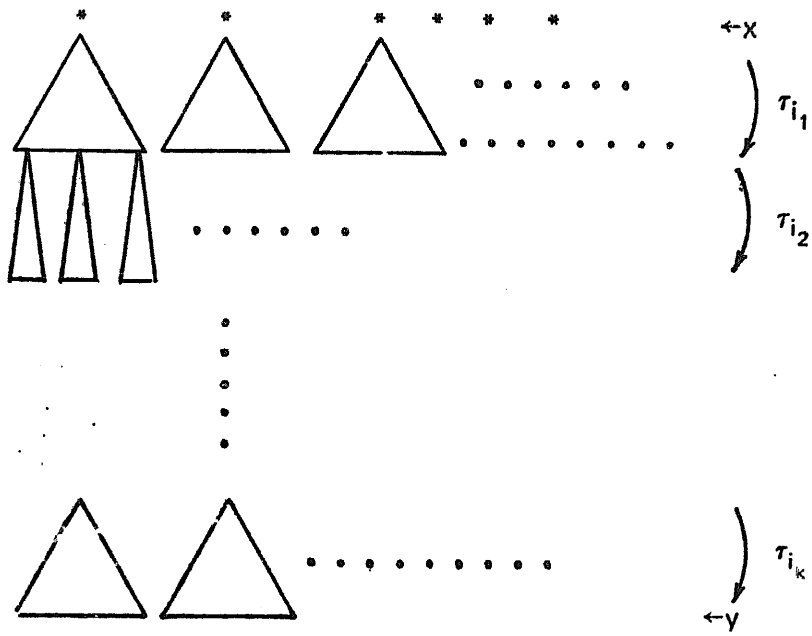


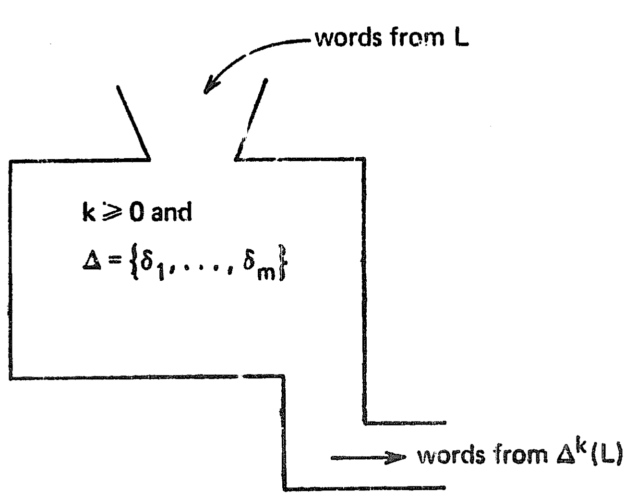
Fig. 1.

Thus, languages in the hyper-sentential extension of $\mathcal{F}$ consist of *everything* that is generated level by level by some $\mathcal{F}$-substitution scheme from initial languages in $\mathcal{F}$.

Definition. *The hyper-algebraic extension of* $\mathcal{F}$ is the smallest natural language family enclosing the hyper-sentential extension of $\mathcal{F}$.

Therefore all languages in an hyper-algebraic extension are of the form $L \cap T^*$ for some hyper-sentential L and alphabet T, and a much richer family results. This has been exemplified in the study of 0L versus E0L-languages, of T0L versus ET0L-languages (all classes of developmental languages see Herman and Rozenberg [3]) and in other studies of "sentential form sets" versus "terminal-restricted languages".

This paper will study to what extent hyper-algebraic languages have a more complex structure than hyper-sentential languages. The remainder of this section is devoted to a tool which we will need in our analysis, and which is of a more automata-theoretic than language-theoretic nature.



Imagine a device which holds an $\mathcal{F}$-substitution scheme $\Delta = (W, \{\delta_1, ..., \delta_m\})$ in its interior, and upon input of some language L over W it processes its words some fixed number of steps, k say, and then outputs the results.

Definition. Given $\mathcal{F}$, let $\overline{\mathcal{F}} = \mathcal{F} \cup \{\$\}$, i.e. the smallest family containing $\mathcal{F}$ and the single symbol sets, let $\Delta = (W, \{\delta_1, ..., \delta_m\})$ be an $\overline{\mathcal{F}}$-substitution scheme and k a fixed integer, $k \geqslant 0$. For a language L over W, $\Delta^k(L) = \{y | x \overset{k}{\Rightarrow} y \text{ for some } x \text{ in } L\}$ is called an *$\mathcal{F}$-translation of L.*

If the family $\mathcal{F}$ is implicitly understood we will just speak of a "translation". In later applications we will frequently wish to translate a language into the terminal words which can be derived from its elements in some fixed number of steps. As an illustration of this use of a translation, we prove a lemma which is of interest in its own right.

Lemma 1 (Selection Lemma). *Let $\mathcal{F}$ be a natural family, G be an $\mathcal{F}$-substitution scheme, L be in $\mathcal{F}$, T be some terminal alphabet and let k, l $\geqslant$ 0. There exist finitely many $L_1, ..., L_m$, m > 0, all isomorphic to canonical restrictions of L, such that* (1) *$\{y$: there exists x in L such that $x \overset{k}{\Rightarrow} y$ and there exists z in T^* such that $y \overset{l}{\Rightarrow} z\}$ is a translation of $\bigcup_{i=1}^{m} L_i$ and* (2) *$\{y: x \overset{k+l}{\Rightarrow} y$ in $T^*\}$ is a translation of $\bigcup_{i=1}^{m} L_i$.*

Proof. Let $G = (V, \{\tau_1, ..., \tau_n\})$. We may assume $k > 0$, otherwise choose $m = 1$ and $L_1 = \{x: x \text{ is in } L \text{ and there exists } z \text{ in } T^* \text{ such that } x \overset{l}{\Rightarrow} z\}$. We consider all possible sequences of substitutions of length k, and select those which lead to y's which can be further expanded into terminal words.

For each σ in $\{\tau_1, ..., \tau_n\}^k$, $\sigma = \tau_{i_k} ... \tau_{i_1}$ and δ in $\{\tau_1, ..., \tau_n\}^l$, define the alphabets $V(\sigma, \delta, j)$, $0 \leqslant j \leqslant k$ as follows:

$$V(\sigma, \delta, k) = \{a: a \text{ in } V \text{ and } \delta(a) \cap T^* \neq \emptyset\}, \text{ and}$$

$$V(\sigma, \delta, j-1) = \{a: a \text{ in } V,\ \tau_{i_j}(a) \cap V(\sigma, \delta, j)^* \neq \emptyset\},\ 1 \leqslant j \leqslant k.$$

Now all words in L over $V(\sigma, \delta, 0)$ can be "translated", when appropriate substitutions are made, into words y which can be further processed to some terminal word. The δ is needed to coordinate the final stage.

For all σ and δ such that $V(\sigma, \delta, 0) \neq \emptyset$, introduce new symbols $[a, \sigma, \delta, j]$, where a is in $V(\sigma, \delta, j)$, $0 \leqslant j < k$. For all x in $V(\sigma, \delta, j)^*$, $x = a_1 ... a_p$, let $x^{(\sigma, \delta, j)}$ denote $[a_1, \sigma, \delta, j] ... [a_p, \sigma, \delta, j]$. Now let $L(\sigma, \delta) = \{x^{(\sigma, \delta, 0)}: x \text{ in } L \cap V(\sigma, \delta, 0)^*\}$, which is easily seen to be isomorphic to a canonical restriction of L. Define the substitution τ as follows:

$$\text{for } 0 \leqslant j < k-1,\ \tau([a, \sigma, \delta, j]) = \{x^{(\sigma, \delta, j+1)}: x \text{ in } \tau_{i_{j+1}}(a) \cap V(\sigma, \delta, j+1)^*\}$$

$$\text{and } \tau([a, \sigma, \delta, k-1]) = \tau_{i_k}(a) \cap V(\sigma, \delta, k)^*.$$

Since $\mathcal{F}$ is a natural family, and therefore closed under canonical restriction and isomorphisms, τ is again an $\mathcal{F}$-substitution. Furthermore $\Delta = (V \cup \bigcup_{\sigma,\delta} \{[a, \sigma, \delta, j]$: a is in $V(\sigma, \delta, j)$ and $0 \leqslant j < k\}$, $\{\tau\})$ is the required translation scheme. Note that there are only finitely many $L(\sigma, \delta)$'s and that only *one* substitution is necessary in the translation scheme, hence the required translation is $\Delta^k(\bigcup_{\sigma,\delta} L(\sigma, \delta))$. The second part of the lemma is proved in a similar way, except that

$$V(\sigma, \delta, k+l) = T,$$

$$V(\sigma, \delta, k+j-1) = \{a: a \text{ is in } V,\ \delta_{i_j}(a) \cap V(\sigma, \delta, k+j)^* \neq \emptyset\},\ 1 \leqslant j \leqslant l,$$

and τ is extended by:

$$\tau([a, \sigma, \delta, k-1]) = \{x^{(\sigma,\delta,k)}: x \text{ in } \tau_{i_k}(a) \cap V(\sigma, \delta, k)^*\};$$

for $0 \leqslant j < l-1$,

$$\tau([a, \sigma, \delta, k+j]) = \{x^{(\sigma,\delta,k+j+1)}: x \text{ in } \delta_{i_{j+1}}(a) \cap V(\sigma, \delta, k+j+1)^*\};$$

and

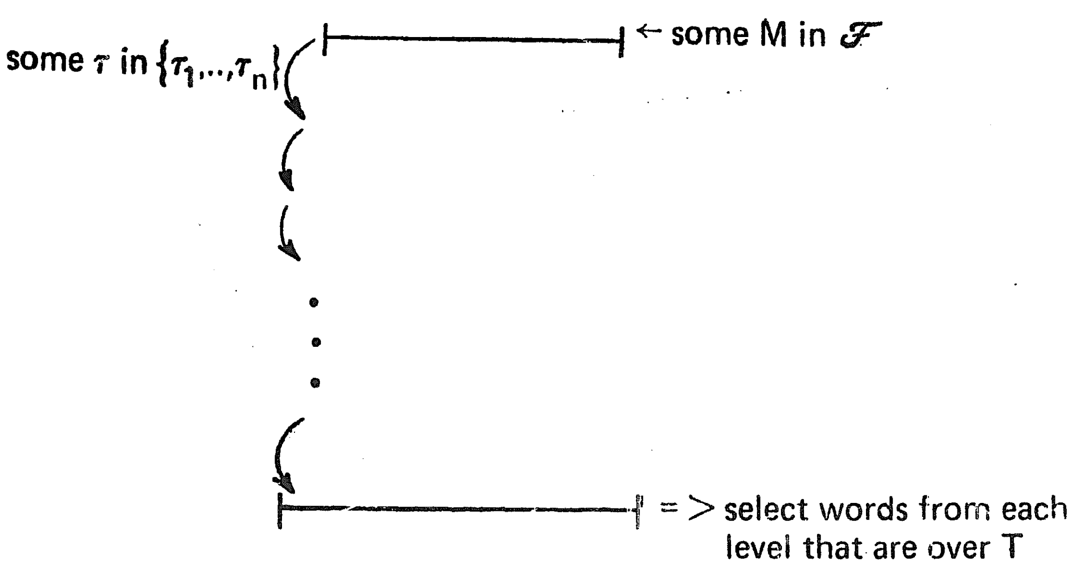
$$\tau([a, \sigma, \delta, k+l-1]) = \delta_{i_l}(a) \cap T^*.$$

The required translation is $\Delta^{k+l}(\bigcup_{\sigma,\delta} L(\sigma, \delta))$.

3. A fundamental theorem

In this Section we will discuss the ideas behind the first few steps of the decomposition theorem.

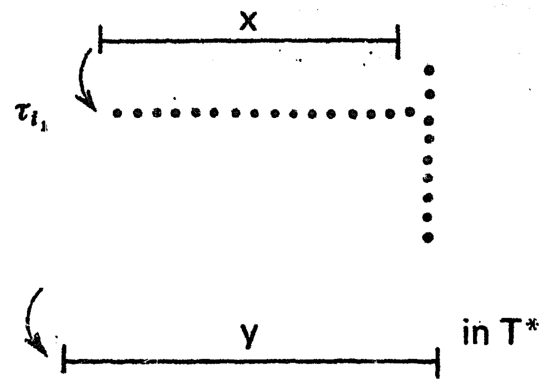
Let $G = (V, \{\tau_1, ..., \tau_n\})$ be an $\mathcal{F}$-substitution scheme, and T an alphabet. We use G to generate a language L hyper-algebraic over $\mathcal{F}$, and wish to decompose L into translations of hyper-sentential languages.



It is required to know precisely where the terminal words will appear, but that is only part of it. We wish to decompose the derivation of a terminal word into "big" steps (realised perhaps by smaller steps), which should be steps in a basic replacement system making up a "hyper-sentential" component.

Hence, in a derivation leading to a terminal word we wish to detect stages ("big" steps) which could have been expanded into a terminal word much earlier, that is, the opposite of a pumping lemma.

Consider some derivation $x \overset{k}{\Rightarrow} y$,



of a word over T. At each level all appearances of some symbol, a say, are replaced, in general, by possibly different words from $\tau_i(a)$ for some i, $x_1, \ldots, x_p$ say. However if at that level each appearance of a is replaced by the same word, x_j say, then the derivation will still give a terminal word, i.e. $x \overset{k}{\Rightarrow} y'$, y' in T^*. Hence, if we are only interested in at what levels terminal words will appear, we may just as well assume this restriction on derivations, and consequently have the steps determined by the alphabet of the intermediate strings.

Slightly generalizing a corresponding concept in Ehrenfeucht and Rozenberg [1], we have:

Definition. For U, T, G, V, let the *spectrum of U with respect to G and T* be Spec $(U, G, T) = \{n$: there exists x such that Alph $(x) = U$ and there exists z in T^* such that, $x \overset{n}{\Rightarrow} z\}$.

As in Wood [10] we obtain:

Theorem 2 (Spectrum theorem). *For all G and U, T, G, V,* Spec (U, G, T) *is ultimately periodic.*

Proof. With the previous reductions we need only consider the behaviour of the sub-alphabets. Going from level to level can easily be modelled by transitions in a (non-deterministic) finite-state automaton. The result follows since all regular languages over a one-letter alphabet form an ultimately periodic set.

Each ultimately periodic set is the union of a finite set and a finite number of arithmetic progressions with equal period. Provided Spec $(U, G, T) \neq \emptyset$, let $N_U = \max$ $(\mathrm{Spec}(U, G, T)) + 1$ if the spectrum is finite, and $\mathrm{Spec}(U, G, T) = \cup \{N_U^{(i)} + \lambda P_U : \lambda \geq 0\}$ $\cup \{n$ in Spec $(U, G, T): n < N_U\}$ for some smallest N_U, otherwise.

Here is the fundamental result.

Theorem 3. *Let G be an $\mathcal{F}$-substitution scheme and $T \subseteq V$ be a terminal alphabet. There exist constants $\alpha \geq 2$ and C, dependent on G and T such that for all $k \geq C$ and* *all* x:

there is a y *in* T^* *such that* $x \overset{k}{\Rightarrow} y$ *iff there is a* z *in* T^* *such that* $x \overset{\alpha k}{\Rightarrow} z$.

Proof. If for all $U \subseteq V$, Spec (U, G, T) is either empty or finite, the theorem follows immediately. Without restriction we therefore assume that some spectra are infinite.

Let $\alpha = 1 + \text{l.c.m.}\,(p_U\colon \text{Spec}\,(U, G, T) \text{ is infinite}\}$, and $C = \max\,\{N_U^{(i)}\colon \text{Spec}\,(U, G, T) \neq \emptyset\}$. Let $k \geqslant C$ be arbitrary.

To prove the theorem we consider two cases.

only if: Suppose $x \overset{k}{\Rightarrow} y$, y in T^*. Since $k \geqslant C$ it follows that Spec (Alph (x), G, T) is infinite, and for some $\lambda \geqslant 0$ and i,

$$k = N^{(i)}_{\text{Alph}(x)} + \lambda \cdot p_{\text{Alph}(x)}\,.$$

Now $\alpha = 1 + \theta \cdot p_{\text{Alph}(x)}$, hence we have:

$$\begin{aligned}\alpha \cdot k &= (1 + \theta p_{\text{Alph}(x)})\,(N^{(i)}_{\text{Alph}(x)} + \lambda p_{\text{Alph}(x)})\\ &= N^{(i)}_{\text{Alph}(x)} + (\lambda + \theta N^{(i)}_{\text{Alph}(x)} + \lambda\theta p_{\text{Alph}(x)})\, p_{\text{Alph}(x)}\end{aligned}$$

and therefore αk is in Spec (Alph (x), G, T).

if: Now suppose that $x \overset{\alpha k}{\Rightarrow} z$, z in T^*. It again follows that Spec (Alph (x), G, T) is infinite, and for some μ and i we have

$$\alpha k = (1 + \theta p_{\text{Alph}(x)})\, k = N^{(i)}_{\text{Alph}(x)} + \mu p_{\text{Alph}(x)}.$$

Write this as

$$k + \theta \cdot k \cdot p_{\text{Alph}(x)} = N^{(i)}_{\text{Alph}(x)} + \mu p_{\text{Alph}(x)},$$

then since $k \geqslant N^{(i)}_{\text{Alph}(x)}$, it follows that

$$\theta \cdot k \cdot p_{\text{Alph}(x)} \leqslant \mu p_{\text{Alph}(x)}.$$

Consequently $\mu - \theta k \geqslant 0$, and

$$k = N^{(i)}_{\text{Alph}(x)} + (\mu - \theta k)\, p_{\text{Alph}(x)}$$

showing that k is in Spec (Alph (x), G, T).

4. The decomposition theorem

Let G be an $\mathcal{F}$-substitution scheme, M in $\mathcal{F}$ be the language to which it is applied, and T be a terminal alphabet. Before formulating the main theorem, let us briefly outline the ideas again.

First we select those words x from M for which Spec (Alph (x), G, T) is infinite. Provided there exist such words, we can choose α and C (depending only on G and T) as in Theorem 3.

Consider a word w in T^* derived in "many" steps, say N steps, such that $N = k(\alpha-1)C+j$, for some j and k, $0 \leqslant j < (\alpha-1)C$ and $k \geqslant 1$ and there is a derivation

$$\begin{array}{l} x_0 = x \text{ in } M \\ \vdots \\ x_j \\ \vdots \\ x_{j+(\alpha-1)C} \\ \vdots \\ x_{j+2(\alpha-1)C} \\ \vdots \\ x_{j+k(\alpha-1)C} = w \text{ in } T^*. \end{array}$$

Since $x_{j+k(\alpha-1)C-\alpha C} = x_{j+(k-1)(\alpha-1)C-C}$ derives a terminal word in αC steps, it derives a terminal word in C steps. Consequently since $x_{j+(k-2)(\alpha-1)C-C}$ derives a terminal word in αC steps, it also derives a terminal word in C steps and so we can continue:

$$\begin{array}{rl} & \vdots \\ & \swarrow x_{j+(k-2)(\alpha-1)C-C} \\ \text{terminal} & \vdots \\ & \swarrow x_{j+(k-1)(\alpha-1)C-C} \\ \text{terminal} & \vdots \\ & \swarrow x_{j+k(\alpha-1)C-C} \\ \text{terminal} & \end{array}$$

all the way up to $x_{j+(\alpha-1)C-C}$.

Thus if we can realize the vertical lines in one "big" step, and the side-branches by a translation, then we have decomposed the system and realised it as the translation of a hyper-sentential language. However there are more details, since, in particular, realizing the translations is a complicated task.

Let us first make the above analysis more precise into an "if and only if" characterization.

Lemma 4. *Let* Spec (Alph (x), G, T) *be infinite, and* $0 \leqslant j < (\alpha-1)C$. *Then* $L(x,j) = \{w \text{ in } T^*: x \overset{N}{\Rightarrow} w \text{ for some } N \equiv j \ (\text{modulo } (\alpha-1)C)\}$ *can be systematically obtained as follows:*

(i) *let* $A = \{y: y \text{ in } T^*, x \overset{j}{\Rightarrow} y\}$, *that is all terminal words which can be derived in j steps, and*

(ii) *consider all sequences* $y_0, y_1, y_2, \ldots,$ *where:* y_0 *is derivable from* x *in* $j+(\alpha-1)C+C$ *steps such that it can lead to a terminal word after an additional C steps, and* y_i *is derivable from* y_{i-1} *in* $(\alpha-1)C$ *steps such that it can lead to a terminal word after an additional C steps and let B be the set of terminal words which can be derived from the* y_i*'s in C steps. Then* $A \cup B = L(x,j)$.

Proof. Assume that the spectrum of x contains infinitely many $N \equiv j$ (modulo $(\alpha-1)\, C$).

The previous analysis has shown that all terminal words in $L\,(x, j)$ will be obtained this way.

We only show here that once a y_0 exists, an infinite sequence of y_i's exists. Using induction, suppose that y_{i-1} exists, deriving a terminal word in C steps. By Theorem 3 it follows that we can derive "another" terminal word from y_{i-1} in αC steps. On the way, after $(\alpha-1)\, C$ steps, we therefore have words from which a terminal word can be derived in $\alpha C-(\alpha-1)\, C = C$ steps. Therefore, if y_0 exists an infinite sequence of y_i's exists.

The lemma holds the clue to breaking up a hyper-algebraic language into hyper-sentential languages.

Theorem 5. *Let $\mathcal{F}$ be a natural family. For each language L hyper-algebraic over $\mathcal{F}$ there exist m, n, translations Δ_i and $\bar{\Delta}_j$, $p_i \geqslant 0$ and $q_j \geqslant 0$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$ and languages $L_1, \ldots, L_m$ hyper-sentential over $\mathcal{F}$ such that*

$$L = \bigcup_{i=1}^{m} \Delta_i^{p_i}(L_i) \cup \bigcup_{j=1}^{n} \bar{\Delta}_j^{q_j}(\$).$$

Proof. Let L be generated from M in $\mathcal{F}$, using the $\mathcal{F}$-substitution scheme $G = (V, \{\tau_1, \ldots, \tau_k\})$ and terminal alphabet T.

Assume first that for all x in M, Spec (Alph (x), G, T) is finite. If all are empty, choose $n = m = 0$. Otherwise there are finitely many $k_1, k_2, \ldots, k_n$, being the levels at which terminal words can appear. For each k_j, use the selection lemma to determine a translation $\bar{\Delta}_j$ and finitely many $M_1, \ldots, M_l$ all over disjoint alphabets but isomorphic to canonical restrictions of M, such that the set of terminal words derivable from M in k_j steps is equal to $\bar{\Delta}_j^{k_j}(\bigcup_{i=1} M_i)$. Since $\mathcal{F}$ is natural, each M_i is in $\mathcal{F}$. Choose a "new" symbol \$, and for each substitution τ in $\bar{\Delta}_j$ introduce $\tau^{(i)}$ $(1 \leqslant i \leqslant l)$ which is like τ but has in addition $\tau^{(i)}(\$) = M_i$. The new scheme is $\bar{\Delta}_j$, and trivially

$$L = \bigcup_{j=1}^{n} \bar{\Delta}_j^{k_j+1}(\$)$$

in this case.

Assume next that for some x in M, Spec (Alph (x), G, T) is infinite. Whatever terminal words are contributed from $\{x\colon \text{Spec (Alph } (x), G, T) \text{ is finite}\}$ (a finite union of languages in $\mathcal{F}$, since $\mathcal{F}$ is natural) can be treated as above.

So we only need to consider the derivations on words x in M with infinite spectrum.

Choose some $U \subseteq V$ such that Spec (U, G, T) is infinite and $M \cap U^* \neq \emptyset$ (U will be denoted as Alph_1 below).

The "construction" below has to be carried out for all such U, but there are only finitely many of them and in the end, collecting all of them together, we will still have a *finite* union of translations.

Choose j, $0 \leqslant j < (\alpha-1) C$. Use the selection lemma to obtain all terminal words derivable in j steps from $M \cap U^*$ (in $\mathcal{F}$) as the finite union of translations of $\$$.

Now assume that the spectrum contains infinitely many $N \equiv j$ (modulo $(\alpha-1) C$). We will use the idea implicit in Lemma 4 to obtain all terminal words derived in such a number of steps.

We will design an $\mathcal{F}$-substitution scheme G' which generates exactly all sequences $y_0, y_1, \ldots$ in such a coded form that there is a translation which works on the y_i' s and intermediate words to get exactly all terminal words required.

The methodology is as follows:

(i) first generate all y_0,

(ii) represent y_i as $[*, \text{Alph}, \delta] \ldots [*, \text{Alph}, \delta]$ where Alph is (a superset of) the alphabet of y_i and δ consists of a sequence of C $\mathcal{F}$-substitutions which can generate terminal words from y_i,

(iii) going from y_i to y_{i+1} in $(\alpha-1) C$ steps is represented by having intermediate words of the form

$$[*, \text{Alph}_1, \text{Alph}_2, \sigma, \delta, t] \ldots$$

where Alph_1 is (a superset of) the alphabet of y_i, Alph_2 is the "guessed" (superset of the) alphabet of y_{i+1}, σ the sequence of $\mathcal{F}$-substitutions of length $(\alpha-1) C$ supposedly realizing the transformation from y_i to y_{i+1}, δ the "guessed" sequence of length C sending y_{i+1} into something terminal and t a counter from 1 to $(\alpha-1) C-1$.

We need the Alph-designations *explicitly*, because the substitutions themselves cannot know what symbols are potentially present, and hence, which replacements to apply.

Clearly not all $(\text{Alph}_1, \text{Alph}_2, \sigma, \delta)$ satisfy the description above. We will however restrict our attention to those which do, and call them *acceptable*. More precisely:

Definition. A quadruple $(\text{Alph}_1, \text{Alph}_2, \sigma, \delta)$ is *acceptable* if σ is of length $(\alpha-1) C$ and can send any word over Alph_1 into a word over Alph_2, and δ is of length C and can send any word over Alph_2 into something terminal.

Note that by Theorem 3, whenever $(\text{Alph}_1, \text{Alph}_2, \sigma, \delta)$ is acceptable, then there are Alph_3, σ' and δ' such that $(\text{Alph}_2, \text{Alph}_3, \sigma', \delta')$ is acceptable.

Letting Q denote $(\text{Alph}_1, \text{Alph}_2, \sigma, \delta)$, an acceptable quadruple, define the alphabets $V(1, Q), \ldots, V((\alpha-1) C-1, Q)$ as follows. Letting

$$V((\alpha-1) C, Q) = \text{Alph}_2,\ V(0, Q) = \text{Alph}_1, \text{and } \sigma = \tau_{i_{(\alpha-1)C}} \ldots \tau_{i_1},$$

then

$$V(j, Q) = \{a \colon \tau_{i_{j+1}}(a) \cap V(j+1, Q)^* \neq \emptyset\},\ 0 \leqslant j < (\alpha-1) C.$$

To perform the initialisation phase (generating the y_0's) acceptable quadruples are introduced where σ is of length $j+(\alpha-1) C-C$, these are denoted by Qs.

Finally, we construct $G(Qs)$ as follows:
symbols:

$$[a, Qs],\ a \text{ in } \mathrm{Alph}_1,$$

$$[a, Qs, t]\ a \text{ in } V(t, Qs),\ 1 \leqslant t < j+(\alpha-1)\,C-C,$$

$$[a, \mathrm{Alph}_1, \delta],\ a \text{ in } \mathrm{Alph}_1,$$

$$[a, Q, t],\ a \text{ in } V(t, Q),\ 1 \leqslant t < (\alpha-1)\,C,$$

for *acceptable quadruples only*, and Alph's which can occur as the "second" coordinate in some acceptable quadruple.

substitutions: there are many substitutions, depending on the choices one may make with a $[a, \mathrm{Alph}_1, \delta]$ symbol. For each map θ: Alph_1's $\rightarrow$ acceptable quadruples Q, beginning with Alph_1, define τ_θ as follows:

$$\tau_\theta([a, Qs]) = [\tau_{i_1}(a) \cap V(1, Qs)^*]^{(Qs,1)}$$

where $[x]^{(Qs,1)}$ means rewrite x as a word consisting of symbols of the form $[a, Qs, 1]$.

$$\tau_\theta([a, Qs, t]) =$$
$$= [\tau_{i_{t+1}}(a) \cap V(t+1, Qs)^*]^{(Qs,t+1)},\quad 1 \leqslant t < j+(\alpha-1)\,C-C-1,$$

$$\tau_\theta([a, Qs, t]) =$$
$$= [\tau_{i_{t+1}}(a) \cap \mathrm{Alph}_2^*]^{(\mathrm{Alph}_2,\delta)},\quad \text{for } t = j+(\alpha-1)\,C-C-1,$$

$$\tau_\theta([a, \mathrm{Alph}, \delta]) =$$
$$= [\tau'_{i_1}(a) \cap V(1, \mathrm{Alph}, \mathrm{Alph}_2, \sigma', \delta')^*]^{(\mathrm{Alph},\mathrm{Alph}_2,\sigma',\delta',1)}$$
$$\text{if } \theta(\mathrm{Alph}) = (\mathrm{Alph}, \mathrm{Alph}_2, \sigma', \delta'),$$

$$\tau_\theta([a, Q, t]) =$$
$$= [\tau_{i_{t+1}}(a) \cap V(t+1, Q)^*]^{(Q,t+1)}\quad \text{if } 1 \leqslant t < (\alpha-1)\,C-1,$$

$$\tau_\theta([a, Q, t]) =$$
$$= [\tau_{i_{t+1}}(a) \cap \mathrm{Alph}_2^*]^{(\mathrm{Alph}_2,\delta)},\quad \text{if } t = (\alpha-1)\,C-1.$$

Because of all our assumptions, *none* of the indicated sets is empty and actually all are renamings of canonical restrictions of $\mathcal{F}$-languages. Hence, each τ_θ is an $\mathcal{F}$-substitution!

Note that $G(Qs)$ starts working on $[M \cap \mathrm{Alph}_1^*]^{(Qs)}$, which is in $\mathcal{F}$, and therefore generates a language hyper-sentential over $\mathcal{F}$. We need to consider all such schemes with $\mathrm{Alph}_1 \subseteq V$ but there are only finitely many of them.

The theorem follows if we can design a translation to send all words of that hyper-sentential language into the terminal words derivable in some C steps.

Note that we have a lot of intermediate words as well, and a number of "dummy" substitutions in the translation are needed to let it all happen in a uniform manner.

Construct the following substitution-scheme $\Delta(Qs)$:

— for all symbols β of $G(Qs)$ introduce new symbols $\beta^{(t)}$, $1 \leqslant t < j+(\alpha-1)C$

— for all symbols $[a, \text{Alph}, \delta]$ define the alphabets $V(1, \text{Alph}, \delta), \ldots, V(C-1, \text{Alph}, \delta)$ as usual

— for each substitution τ from $G(Qs)$ construct the following substitution $\bar{\tau}$ for $\Delta(Qs)$:

$$\bar{\tau}([a, Qs]) = \\ = \tau([a, Qs, 1])^{(1)},$$

$$\bar{\tau}([a, Qs, t]) = \\ = \{[a, Qs, t]^{(1)}\},$$

$$\bar{\tau}([a, Qs, t]^{(i)}) = \\ = \{[a, Qs, t]^{(i+1)}\}, \quad 1 \leqslant i < t,$$

$$\bar{\tau}([a, Qs, t]^{(i)}) = \\ = \tau([a, Qs, t])^{(i+1)}, \quad t \leqslant i < j+(\alpha-1)C-C,$$

$$\bar{\tau}([a, \text{Alph}, \delta]) = \{[a, \text{Alph}, \delta]^{(1)}\},$$

$$\bar{\tau}([a, \text{Alph}, \delta]^{(i)}) = \{[a, \text{Alph}, \delta]^{(i+1)}\}, \quad 1 \leqslant i < j+(\alpha-1)C-C,$$

$$\bar{\tau}([a, Q, t]) = \{[a, Q, t]^{(1)}\},$$

$$\bar{\tau}([a, Q, t]^{(i)}) = \{[a, Q, t]^{(i+1)}\}, \quad 1 \leqslant i < t+j-C,$$

$$\bar{\tau}([a, Q, t]^{(i)}) = \tau([a, Q, t])^{(i+1)}, \; t+j-C \leqslant i < j+(\alpha-1)C-C,$$

$$\bar{\tau}([a, \text{Alph}, \delta]^{(i)}) = \\ = ([\tau_{r_{i-j-(\alpha-1)C+C+1}}(a) \cap V(i-j-(\alpha-1)C+C+1, \text{Alph}, \delta)^*]^{(\text{Alph}, \delta)})^{(i+1)}, \\ j+(\alpha-1)C-C \leqslant i < j+(\alpha-1)C-1,$$

where $\delta = \tau_{r_C} \ldots \tau_{r_1}$,

$$\bar{\tau}([a, \text{Alph}, \delta]^{(i)}) = \tau_{r_C}(a) \cap T^*, \quad \text{for } i = j+(\alpha-1)C-1.$$

(This construction assumes that $j \geqslant C$, but an obvious modification holds when $j < C$).

It follows that the required translation is obtained by iterating $\Delta(Qs)$ exactly $j+(\alpha-1)C$ times. Note that all substitutions are indeed non-empty $\mathcal{F}$-substitutions as required.

Provided all L_i are chosen over disjoint alphabets, and each translation Δ_j has the appropriate number of extra delaying steps added (which is easily done, bearing in mind how we constructed them), we can prove immediately the following stronger form of the above theorem.

Theorem 6 (The decomposition theorem). *Let $\mathcal{F}$ be a natural family. For each language L hyper-algebraic over $\mathcal{F}$ there exist integers m, p, q, translations $\Delta, \bar{\Delta}$, and languages $L_1, ..., L_m$ hyper-sentential over $\mathcal{F}$ such that*

$$L = \bigcup_{i=1}^{m} \Delta^p(L_i) \cup \bar{\Delta}^q(\$), \quad \textit{for some } p, q \geqslant 0.$$

It is this stronger result which we examine in more detail.

5. Applications

In this section we will examine some consequences of the decomposition theorem when we make more assumptions about $\mathcal{F}$.

A first result is obtained if we assume $\mathcal{F}$ to contain all single symbol sets. The hyper-sentential extension of $\mathcal{F}$ in this case is "closed" under union, and the following characterization theorem results.

Theorem 7 (The translation theorem). *Let $\mathcal{F}$ be a natural family containing $\{\$\}$. A language L is hyper-algebraic over $\mathcal{F}$ iff there exists a translation Δ and a language L' hyper-sentential over $\mathcal{F}$ such that $L = \Delta^p(L')$, for some $p \geqslant 0$.*

Proof. *only if*: First note that because $\mathcal{F}$ contains $\{\$\}$, the set $\{\$\}$ is hyper-sentential over $\mathcal{F}$. Hence, the decomposition theorem shows that there exist hyper-sentential languages $L_1, ..., L_m$ and a translation Δ such that $L = \bigcup_{i=1}^{m} \Delta^k(L_i)$.

Make all alphabets disjoint, and it follows that there is a single translation Δ such that $L = \Delta^k(\bigcup_{i=1}^{m} L_i)$.

Design $\tau^{(i)}$ with

$\tau^{(i)}\{\$\}$ = initial language for L_i, and

$\tau^{(i)}$ is identity otherwise, $1 \leqslant i \leqslant m$.

The $\tau^{(i)}$ are clearly all $\mathcal{F}$-substitutions. Now add all substitutions involved in the generation of the (alphabetically disjoint) L_i's. All those substitutions have to be augmented with identity rules for those symbols that are originally foreign for them.

It follows that $\{\$\} \cup \bigcup_{i=1}^{m} L_i$ is hyper-sentential over $\mathcal{F}$.

Now delay Δ one step to let \$ be translated into the start language of the various L_i, and then send them on as usual.

It follows that one single translation on one single hyper-sentential language will realise L.

if: Given $\Delta^p(L') \subseteq T^*, p > 0$ we show that it is hyper-algebraic. Let L' be generated by $G' = (V', \{\tau_1, ..., \tau_m\}), m > 0$ from some initial $\mathcal{F}$-language M', and $\Delta = (V' \cup V, \{\delta_1, ..., \delta_n\}), n > 0$.

Note that the alphabet of Δ must contain the alphabet of G', however V can be empty. We construct G, an $\mathcal{F}$-substitution scheme, by embedding Δ in G', making sure that the δ_i's can only be applied exactly p times to a sentential form generated by the τ_i's.

Let $G = (U, \{\bar{\tau}_1, ..., \bar{\tau}_m, \bar{\delta}_1, ..., \bar{\delta}_n\})$ where $U = V \cup V' \cup \{X\} \cup \{[a, i]: a$ in $V' \cup V, 1 \leqslant i \leqslant p\}$, X does not appear elsewhere, and

(i) for all a in $V \cup V' \cup \{X\}, \bar{\tau}_i(a) = \bar{\delta}_j(a) = \{X\}, 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n$,

(ii) for all $[a, 1]$ in $U, \bar{\tau}_i([a, 1]) = [\tau_i(a)]^{(1)}, 1 \leqslant i \leqslant m$,

(iii) for all $[a,j]$ in $U, \bar{\tau}_i([a,j]) = \{X\}, 1 < j \leqslant p, 1 \leqslant i \leqslant m$,

(iv) for all $[a,j]$ in U, $\bar{\delta}_i([a,j]) = [\delta_i(a)]^{(j+1)}$, $1 \leqslant j < p$, $1 \leqslant i \leqslant n$, and

(v) for all $[a, p]$ in U, $\bar{\delta}_i([a, p]) = \delta_i(a)$, $1 \leqslant i \leqslant n$.

Then $L(G) = \Delta^p(L')$, if $\overline{M}'$ is the initial language of G, where $\overline{M}' \subseteq \{[a, 1]: a \text{ in } V'\}^*$ is isomorphic to M'.

Theorem 8 (The weak coding theorem). *Let $\mathcal{F}$ be a natural family containing all singleton sets. A language L is hyper-algebraic over $\mathcal{F}$ iff there exists an alphabetic homomorphism h such that $L = h(L')$ for some L' hyper-sentential over $\mathcal{F}$.*

Proof. By the translation theorem L is hyper-algebraic over $\mathcal{F}$ iff there exists an $\mathcal{F}$-translation Δ such that $L = \Delta^p L''$ and L'' is hyper-sentential over $\mathcal{F}$. Therefore, the only thing we have to prove is: $\Delta^p L''$ is an $\mathcal{F}$-translation (of a language L'' hyper-sentential over $\mathcal{F}$) iff there is a language L' hyper-sentential over $\mathcal{F}$ and an alphabetic homomorphism h such that $\Delta^p L'' = h(L')$.

if: Trivial, since $\mathcal{F}$ contains $\{\$\}$, and, since it is a natural family, also every single letter set and $\{\varepsilon\}$. Define Δ by $\Delta = (V, \{h\})$, then $\Delta(L'') = h(L')$ where we take $L'' = L'$.

only if: Let L'' be generated from M in $\mathcal{F}$ by $G = (V, \{\delta_1, ..., \delta_n\})$ and let $\Delta = (V, \{\tau_1, ..., \tau_m\})$. Define $G' = (V', \{\delta'_1, ..., \delta'_n, \tau'_1, ..., \tau'_m\})$ as follows. Assume that L contains a word $w = a_1 a_2 ... a_k, k \geqslant 2$. (If this is not the case L is finite and the theorem is trivially true.) Define

$$V' = \bigcup_{i=1}^{p} \{a^{[i]}: a \text{ in } V \cup \bar{V}\} \cup \{[a_i, i]: w = a_1 ... a_k, 1 \leqslant i \leqslant k\} \cup \bar{V} \cup V,$$

where $\bar{V} = \{\bar{a}: a \text{ in } V\}$ and all constituent subsets of V' are disjoint.

To generate L' we start with $\{[a_1, 1] ... [a_k, k]\}$ which is in $\mathcal{F}$ since $\mathcal{F}$ contains all singleton sets.

$$\delta'_i([a_1, 1]) = \tau'_j([a_1, 1]) = \bar{M}^{[1]}$$

$$\delta'_i([a_2, 2]) = \tau'_j([a_2, 2]) = M^{[1]}$$

for $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m$, where

$$M^{[1]} = \{a_1^{[1]} \dots a_r^{[1]}: a_1 \dots a_r \text{ is in } M\}$$
$$\bar{M}^{[1]} = \{\bar{a}_1^{[1]} \dots \bar{a}_r^{[1]}: a_1 \dots a_r \text{ is in } M\}.$$

Hence we obtain at the first level $\bar{M}^{[1]}M^{[1]}$.

Define the δ_i''s and τ_i''s for the remainder of V' as follows:

$$\begin{array}{ll}
\delta_i'(a^{[1]}) = \delta_i(a)^{[1]} & \text{for all } a \text{ in } V, \text{ and } \delta_i'(\cdot) \text{ is identity otherwise, } 1 \leqslant i \leqslant n. \\
\tau_i'(a^{[j]}) = \tau_i(a)^{[j+1]} & \text{for all } a \text{ in } V, \\
\tau_i'(\bar{a}^{[j]}) = \{\bar{a}^{[j+1]}\} & \text{for all } \bar{a} \text{ in } \bar{V},\ 1 \leqslant j \leqslant p-1, \\
\tau_i'(a^{[p]}) = \tau_i(a) & \text{for all } a \text{ in } V, \\
\tau_i'(\bar{a}^{[p]}) = \{\varepsilon\} & \text{for all } \bar{a} \text{ in } V,
\end{array}$$

for all i such that $1 \leqslant i \leqslant m$, and τ_i' is identity otherwise. (Since $\mathcal{F}$ contains all singleton sets we can use identity substitutions and substitutions with $\{\varepsilon\}$).

Define h such that $h([a_i, i]) = a_i$ for $1 \leqslant i \leqslant k$, $h(a) = a$ for all a in V, $h(a^{[j]}) = a$ for all $\bar{a}$ in $\bar{V}$ and $1 \leqslant j \leqslant p$, and $h(a) = \varepsilon$ otherwise. Hence, derivations using δ''s conserve the superscript [1] and simulate the derivations in G while the translations are accomplished by substitutions τ' simulating Δ which end after p applications in words over V. By the definition no parasitic derivations of δ_i''s are possible.

It is easy to see that

$$L' = \{[a_1, 1] \dots [a_k, k]\} \cup \bigcup_{i=0}^{p-1} (\bar{M}^{[i+1]}(\Delta^i L'')^{[i+1]}) \cup \Delta^p L''$$

and therefore $h(L') = \Delta^p L'' = L$.

Much more easy to prove (using a similar method) is:

Theorem 9. *Let $\mathcal{F}$ be a natural family containing $\{\$\}$. A language L is hyper-algebraic over $\mathcal{F}$ iff there exists an alphabetic homomorphism h such that $L - \{\varepsilon\} = h(L') - \{\varepsilon\}$ for some L' hyper-sentential over $\mathcal{F}$.*

Proof. By the translation theorem L is hyper-algebraic over $\mathcal{F}$ iff there exists an $\mathcal{F}$-translation Δ such that $L = \Delta^p L''$ and L'' is hyper-sentential over $\mathcal{F}$. Therefore, the only thing we have to prove is: $\Delta^p L''$ is an $\mathcal{F}$-translation (of a language L'' hyper-sentential over $\mathcal{F}$) iff there is a language L' hyper-sentential over $\mathcal{F}$ and an alphabetic homomorphism h such that $\Delta^p L'' = h(L')$.

if: Trivial, since $\mathcal{F}$ contains $\{\$\}$, and, since it is a natural family, also every single letter set and $\{\varepsilon\}$. Define Δ by $\Delta = (V, \{h\})$, then $\Delta(L'') = h(L')$ where we take $L'' = L'$.

only if: Let $\Delta = (V, \{\delta_1, \dots, \delta_m\})$. Let L'' be generated from M in $\mathcal{F}$ by $G = (V, \{\delta_1, \dots, \delta_n\})$. Define $G' = (V', \{\delta_1', \dots, \delta_n', \tau_1', \dots, \tau_m'\})$ where

$$V' = \bigcup_{i=1}^{p-1} \{a^{[i]}: a \text{ is in } V\} \cup V.$$

Start with the language $M^{[1]} = \{x^{[1]}: x^{[1]} = a_1^{[1]} \dots a_k^{[1]}$ for $a_1 \dots a_k$ in $M\}$ isomorphic with M and define the substitutions of G' as follows.

$$\delta_i'(a^{[1]}) = \delta_i(a)^{[1]} \quad \text{for all } a \text{ in } V,\ 1 \leqslant i \leqslant n; \text{ and identity otherwise.}$$
$$\tau_i'(a^{[j]}) = \tau_i(a)^{[j+1]} \quad \text{for all } a \text{ in } V,\ 1 \leqslant j \leqslant p-1,\ 1 \leqslant i \leqslant m;$$
$$\tau_i'(a^{[p]}) = \tau_i(a) \quad \text{for all } a \text{ in } V,\ 1 \leqslant i \leqslant m; \text{ and identity otherwise.}$$

Define h such that $h(a) = a$ for all a in V and $h(a) = \varepsilon$ otherwise.

Both the translation theorem and the weak coding theorem are remarkable results for parallel rewriting systems. However, in the case that $\mathcal{F} \equiv$ finite languages (clearly satisfying the premises of the weak coding theorem) we only obtain a weakened version of Ehrenfeucht and Rozenberg's results for T0L and 0L languages [1–2].

Acknowledgment

We are very grateful to the referee and to Paul M. B. Vitanyi for correcting many typographical errors in an earlier draft of this paper, and for providing the sharpened form of Theorem 8 and the additional Theorem 9.

References

[1] A. Ehrenfeucht and G. Rozenberg, The equality of E0L languages and codings of 0L languages, Internatl. J. Comput. Math. (to appear).

[2] A. Ehrenfeucht and G. Rozenberg, Nonterminals versus homomorphisms in defining languages for some classes of rewriting systems, Acta Informat. (to appear).

[3] G. Herman and G. Rozenberg, Developmental Systems and Languages (North-Holland, Amsterdam, 1975).

[4] J. van Leeuwen, Pre-set pushdown automata and 0L grammars, UC at Berkeley, CSTR 10 (1973).

[5] J. van Leeuwen, Hyper-AFL's and all pre-set PDA languages are indexed, UC at Berkeley, Memorandum (1973).

[6] J. van Leeuwen, F-iteration languages, UC at Berkeley, Memorandum (1973).

[7] J. van Leeuwen, A Generalisation of Parikh's Theorem in Formal Language Theory, Lecture Notes in Computer Science 14 (Springer, Berlin, 1974) 17–26.

[8] G. Rozenberg and A. Salomaa (Eds.), L-Systems, Lecture Notes in Computer Science 15 (Springer, Berlin, 1974).

[9] A. Salomaa, Macros, iterated substitution and Lindenmayer AFL's, University of Aarhus CSTR-DAIMI PB-18 (1973).

[10] D. Wood, A note on Lindenmayer systems, spectra, Szilard languages and equivalence, Internatl. J. Comput. Information Sci. (to appear).