JOURNAL OF COMPUTER AND SYSTEM SCIENCES 55, 477–488 (1997)
ARTICLE NO. SS971497

The Set of Minimal Words of a Context-free Language is Context-free

Jean Berstel*

Institut Gaspard Monge, Université Marne-la-Vallée, 2, rue de la Butte Verte, F-93166 Noisy-le-Grand Cedex

and

L. Boasson

LITP, Université Denis Diderot, 2, place Jussieu, F-75251 Paris Cedex 05, France

Received February 10, 1995; revised June 17, 1996

Let A be a finite, totally ordered alphabet, and let $\preccurlyeq$ be the lexicographic ordering on A^*. Let X be a subset of A^*. The *language of minimal words* of X is the subset of X composed of the lexicographically minimal word of X for each length: $\mathrm{Min}(X) = \{x \in X \mid \forall w \in X, |w| = |x| \Rightarrow x \preccurlyeq w\}$. The aim of this paper is to prove that if L is a context-free language, then the language $\mathrm{Min}(L)$ context-free.

INTRODUCTION

Let A be a finite, totally ordered alphabet, and let $\preccurlyeq$ be the lexicographic ordering on A^*. Let X be a subset of A^*. The *language of minimal words* of X is the subset of X composed of the lexicographically minimal word of X for each length:

$$\mathrm{Min}(X) = \{x \in X \mid \forall w \in X,\ |w| = |x| \Rightarrow x \preccurlyeq w\}.$$

The aim of this paper is to prove the following theorem:

THEOREM. *Let L be a context-free language. The language* $\mathrm{Min}(L)$ *is context-free.*

This result extends to context-free languages a well-known result for regular languages (see, e.g., [5]). According to [4], the same result holds for context-sensitive languages but was open in the context-free case. We do not know about a simple proof of the theorem; our proof is "brute-force" in the sense that it proceeds by inspection and transformation of grammars.

The proof is in four steps. The first (Section 3) considers the unique infinite word in the adherence of a context-free language that is minimal for the lexicographic order. We show (Theorem 3.2) that this word is ultimately periodic. This result is used in the next section in order to classify the nonterminals of a context-free grammar into three categories:

— a nonterminal is *covered* if it generates at least one word that is lexicographically smaller than the minimal infinite word of the adherence,

— a nonterminal is *dense* if it generates infinitely many words that are prefixes of the minimal infinite word of the adherence,

— a nonterminal is *decreasing* if it is neither covered nor dense.

The second step of the proof (Section 4) is a reduction step. We show, using the classification of nonterminals, that the general problem reduces to the particular case of linear languages (Theorem 4.1). The third step (section 5) shows that, for a linear language, its minimal words are exactly those of some linear and bounded language (Theorem 5.1). In the last step, we solve the problem directly for bounded linear languages (Theorem 6.1).

We observe that the present proof uses some variety of concepts from formal language theory which are rarely used together: infinite words and adherences [3], iteration theorems and linear languages [1, 6, 8], rational transductions [2, 5], combinatorics on words [7]. At present, we do not see how to derive a simpler proof that avoids the use of these notions.

1. NOTATIONS

A *context-free grammar* $G = (V, S, P)$ over an alphabet A is composed of finite alphabet V of *variables* or *nonterminals* disjoint from A, a disinguished nonterminal S called the *axiom* and a finite set $P \subset V \times (V \cup A)^*$ of *productions* or *derivation rules*. Letters in A are called *terminal letters*.

* Travail soutenu par le PRC "Mathématiques et Informatique" et par ESPRIT BRA Working Group 6317—ASMICS 2. Correspondence author. E-mail: Jean.Berstel@univ-mlv.fr.

0022-0000/97 $25.00

Given words $u, v \in (V \cup A)^*$, we write $u \to v$ whenever there exist factorizations $u = xXy$, $v = x\alpha y$, with (X, α) a production. A *derivation* from u to v is a sequence $(u_0, u_1, ..., u_k)$ of words in $(V \cup A)^*$ such that $u_{i-1} \to u_i$ for $i = 1, ..., k$, and $u = u_0$, $v = u_k$. If this holds, we write $u \xrightarrow{*} v$. The *language generated* by a variable X in grammar G is the set

$$L_G(X) = \{w \in A^* \mid X \xrightarrow{*} w\}.$$

The language *generated by the grammar* is the language $L_G(S)$, where S is the axiom.

In the sequel, all alphabets are finite and totally ordered. Let A be an alphabet. We denote by A^* the set of finite words over A, by $A^{\mathbb{N}}$ the set of infinite words, and we set $A^\infty = A^* \cup A^{\mathbb{N}}$.

A word u is a *prefix* of a word v if there exists a word w such that $v = uw$. We then write

$$u \sqsubseteq v.$$

If u and v are infinite words, then $u \sqsubseteq v$ implies $u = v$. For $x \in A^\infty$, we denote by $\mathrm{Pref}(x)$ the set of *finite* words that are prefixes of x. For $X \subset A^\infty$, we set $\mathrm{Pref}(X) = \bigcup_{x \in X} \mathrm{Pref}(x)$.

A word u is *strongly smaller* than a word v if there exist a finite word w and two letters a, b with $a < b$ such that wa is a prefix of u and wb is a prefix of v. In this case, we write

$$u \ll v.$$

Observe that $u \ll v$ implies $ux \ll vy$ for all words x, y. The *lexicographic order* is defined, for $x, y \in A^\infty$ by

$$x \leqslant y \quad \text{if and only if} \quad x \sqsubseteq y \quad \text{or} \quad x \ll y.$$

It is a total order. For infinite words, $x \leqslant y$ iff $x = y$ or $x \ll y$. Let

$$w = a_1 a_2 \cdots a_n \qquad (a_i \in A)$$

be a finite word. The set

$$\mathrm{Sous}(w) = \{u \in A^* \mid u \ll w\}$$

of words that are strongly smaller than w is

$$\mathrm{Sous}(w) = (A_1 \cup a_1 A_2 \cup \cdots \cup a_1 \cdots a_{n-1} A_n)\, A^*$$

with $A_i = \{b \in A \mid b < a_i\}$. It is a regular languange. Set similarly, for an infinite word $\mathbf{x}$,

$$\mathrm{Sous}(\mathbf{x}) = \{w \in A^* \mid w \ll \mathbf{x}\}.$$

Observe the following:

LEMMA 1.1. *If the word* $\mathbf{x}$ *is ultimately periodic, then* $\mathrm{Sous}(\mathbf{x})$ *is a regular language.*

Proof. Let $s, t \in A^*$ with $t \neq \varepsilon$ be two words such that

$$\mathbf{x} = st^\omega.$$

Then

$$\mathrm{Sous}(\mathbf{x}) = \mathrm{Sous}(s) \cup st^* \,\mathrm{Sous}(t),$$

showing that this language is regular. ∎

Let us remark that the set $\mathrm{Sous}(\mathbf{x})$ can effectively be computed as soon as the description of $\mathbf{x}$ as an ultimately periodic word is effective.

2. ADHERENCE

Let $X \subset A^*$. The *adherence* of X is the set of those infinite words whose prefixes are prefixes of words in X:

$$\mathrm{Adh}(X) = \{\mathbf{x} \in A^{\mathbb{N}} \mid \mathrm{Pref}(\mathbf{x}) \subset \mathrm{Pref}(X)\}.$$

The adherence of a set X is empty if and only if X is finite (recall that the alphabet A is finite).

If A^∞ is equipped with the usual topology, the *closure* of $X \subset A^*$ is the set $\bar{X}$ defined by

$$\bar{X} = X \cup \mathrm{Adh}(X).$$

In particular, a convergent sequence of elements in $\bar{X}$ converges to an element in $\bar{X}$.

The *center* of X is the set of prefixes of the adherence of X:

$$\mathrm{Centre}(X) = \mathrm{Pref}(\mathrm{Adh}(X)).$$

It is easily seen that

$$\mathrm{Adh}(\mathrm{Centre}(X)) = \mathrm{Adh}(X),$$
$$\mathrm{Centre}(\mathrm{Adh}(X)) = \mathrm{Centre}(X).$$

For context-free languages, one can given an explicit description of the adherence and of the center. For this, consider a context-free language $L \subset A^*$ and a grammar $G = (V, S, P)$ in reduced Greibach normal form that generates L.

An infinite left derivation is a sequence

$$u_0, u_1, ..., u_n, ...$$

of (finite) words such that $u_0 = S$ and

$$u_n \xrightarrow[g]{} u_{n+1}$$

for all $n \geqslant 0$. By definition, this means that for all $n \geqslant 0$, one has

$$u_n = x_n X_n \gamma_n$$

for appropriate words $x_n \in A^*$, $X_n \in V$, and $\gamma_n \in V^*$. Moreover, since this is a left derivation, one has $x_n \sqsubseteq x_{n+1}$, and since the grammar is in Greibach normal form, x_n tends to infinity. The sequence x_n defines an infinite word $\mathbf{x}$ by the property that all words x_n are prefixes of $\mathbf{x}$. We then write

$$S \xrightarrow{\omega} \mathbf{x}.$$

Since $\mathrm{Pref}(\mathbf{x}) = \mathrm{Pref}(\{x_n \mid n \geqslant 0\})$ and $\mathrm{Pref}(\{x_n \mid n \geqslant 0\}) \subset \mathrm{Pref}(L)$ because the grammar is reduced, one has

$$\mathbf{x} \in \mathrm{Adh}(L).$$

Conversely, if $\mathbf{x} \in \mathrm{Adh}(L)$, one can prove that $S \xrightarrow{\omega} \mathbf{x}$. In other terms,

PROPOSITION 2.1 [3]. *Let L be a context-free language over A generated by a reduced grammar $G = (V, S, P)$ in Greibach normal form. Then*

$$\mathrm{Adh}(L) = \{\mathbf{x} \in A^{\mathbb{N}} \mid S \xrightarrow{\omega} \mathbf{x}\}.$$

3. MINIMUM WORD OF THE ADHERENCE

LEMMA 3.1. *Let X be an infinite subset of A^*. Then* $\mathrm{Adh}(X)$ *contains a minimum element for the lexicographic order.*

This property results from general considerations about the topology of A^∞. In view of its simplicity, we give an elementary proof.

Proof. Let $a_1 \in A$ be the smallest letter such that

$$\mathrm{Adh}(X) \cap a_1 A^{\mathbb{N}} \neq \varnothing$$

and more generally, assuming that $a_1, a_2, ..., a_{i-1}$ are already known, let a_i be the smallest letter such that

$$\mathrm{Adh}(X) \cap a_1 a_2 \cdots a_{i-1} a_i A^{\mathbb{N}} \neq \varnothing.$$

This defines an infinite word

$$\mathbf{x} = a_1 a_2 \cdots a_i \cdots.$$

One has

$$\mathbf{x} \in \mathrm{Adh}(X)$$

since

$$\mathrm{Pref}(\mathbf{x}) = \{a_1 \cdots a_i \mid i \geqslant 0\} \subset \mathrm{Pref}(\mathrm{Adh}(X)) \subset \mathrm{Pref}(X).$$

Moreover, if $\mathbf{y} \in \mathrm{Adh}(X)$, then by construction $\mathbf{x} \preccurlyeq \mathbf{y}$. ∎

It is convenient to call *minimum word* of the adherence of X the infinite minimum word of $\mathrm{Adh}(X)$.

THEOREM 3.2. *Let L be an infinite context-free language over A. The minimum word of the adherence of L is ultimately periodic.*

Proof. Let $G = (V, S, P)$ be a reduced grammar in Greibach normal form generating L, and let $\mathbf{z}$ be the minimum word in the adherence of L. Consider an infinite left derivation,

$$S \xrightarrow{\omega} \mathbf{z}.$$

There exist two finite words $s, t \in A^*$, with $t \neq \varepsilon$, a variable $T \in V$, and two words $\alpha, \beta \in (V \cup A)^*$ such that

$$S \xrightarrow{*} sT\alpha \xrightarrow{*} stT\beta\alpha \xrightarrow{\omega} st\mathbf{x} = \mathbf{z}$$

for some infinite word $\mathbf{x}$. This shows that $s\mathbf{x}$ and st^ω belong to $\mathrm{Adh}(L)$, and by the minimality of $\mathbf{z}$, one has

$$st\mathbf{x} \preccurlyeq s\mathbf{x}, \qquad st\mathbf{x} \preccurlyeq st^\omega,$$

whence, simplifying by s,

$$t\mathbf{x} \preccurlyeq \mathbf{x} \preccurlyeq t^\omega.$$

When you iterate the first inequality, you get

$$t^n\mathbf{x} \preccurlyeq \mathbf{x} \qquad (n \geqslant 0);$$

thus $\mathbf{x} = t^\omega$, and consequently, $\mathbf{z} = st^\omega$. ∎

PROPOSITION 3.3. *Given an infinite context-free language L, the minimum word of the adherence of L is effectively computable.*

Proof. Let $G = (V, S, P)$ be a grammar in Greibach normal form generating L. One may assume that every nonterminal of G generate an infinite language.

Let h be an integer, and consider the set of left derivations of length h for which there are words $s \in A^*$, $t \in A^+$, $\alpha, \beta \in (V \cup A^*)$ and a nonterminal $T \in V$ such that

$$S \xrightarrow{*} sT\alpha \xrightarrow{+} stT\beta\alpha \qquad (*)$$

For each derivation of this type, one has

$$st^\omega \in \mathrm{Adh}(L)$$

and st^ω is the minimum word in the adherence of L if and only if

$$\text{Centre}(L) \cap \text{Sous}(st^\omega) = \varnothing. \qquad (**)$$

Since the center of L is effectively computable and since $\text{Sous}(st^\omega)$ is a regular language, this equality is decidable.

In order to compute the minimum word, it suffices to enumerate the left derivations of length h satisfying $(*)$ successively for $h = 1, 2, \ldots$ until one finds a couple of words s, t satisfying the condition $(**)$. Since this couple exists, it will be found in finite time. ▮

COROLLARY 3.4. *Let* $\mathbf{z}$ *be the minimum word of the adherence of an infinite language* L. *If* $L = \text{Centre}(L)$, *then* $\text{Min}(L) = \text{Pref}(\mathbf{z})$ *is a regular language.* ▮

4. LANGUAGE OF MINIMAL WORDS

Let X be a subset of A^*. The *language of minimal words of* X is the subset of X formed, for each length, of the lexicographically minimal word of X:

$$\text{Min}(X) = \{x \in X \mid \forall w \in X, |w| = |x| \Rightarrow x \leqslant w\}.$$

The aim of this section is to prove the following reduction theorem.

THEOREM 4.1. *Let* L *be a context-free language. There exists a linear context-free language* M *such that* $\text{Min}(L) = \text{Min}(M)$.

We begin by some preliminaries. Two languages L and M will be called *length-disjoint* if the set of length of words in L and the set of length of words in M are disjoint.

We first prove a technical lemma.

LEMMA 4.2. *For any infinity context-free language* L, *there exists a context-free grammar such that*

(1) *Every variable generates an infinite language.*

(2) *There exists an integer* p *such that for every nonterminal* S, *there exists an integer* n_S *satisfying*

$$\{|u| \mid u \in L_S\} = n_S + \mathbb{N} \cdot p.$$

(3) *There exists an integer* k *such that the set of words in* L *of length greater than* k *is a length-disjoint finite union of languages generated by variables of the grammar.*

Observe that we do not claim that the language L is generated by such a grammar. Condition (3) only states that L is a length-disjoint finite union of languages satisfying (2).

Proof. (1) is well-known (see, e.g., [6]).

(2) and (3) Let G be a context-free grammar generating L. According to Parikh's theorem (see, e.g., [1, 6]), for each variable S, there exist a finite set F_S of integers, an integer p_S and a finite number of integers $n_{S,i}$ such that

$$\{|u| \mid \in L_S\} = F_S \cup \bigcup_i (n_{S,i} + \mathbb{N} \cdot p_S).$$

Let K be the union of the F_S over all variables S, and let k be the maximum in K. Let next p be the lcm of the integers p_S. Clearly, by changing the set of integers $n_{S,i}$, we may rewrite the above equation as

$$\{|u| \mid u \in L_S\} = F_S \cup \bigcup_i (n_{S,i} + \mathbb{N} \cdot p).$$

We will now get rid of words of length at most k. For this, we transform the grammar in the following way:

— In each rule $X \to \alpha$, replace each variable S in α by the set $S + L_S^{\leqslant k}$, where $L_S^{\leqslant k} = \{w \in L_S \mid |w| \leqslant k\}$.

— In the resulting grammar, remove all terminal rules $Y \to u$ with $|u| \leqslant k$.

The grammar H obtained in this way satisfies:

— Each variable S generates, in H, the language $L_S \backslash L_S^{\leqslant k}$.

— If G is in Greibach normal form, then H is in weak Greibach normal form (i.e., every right-hand side starts with a terminal letter).

Consequently, up to the words of length at most k in L_S, we may now assume that

$$\{|u| \mid u \in L_S\} = \bigcup_i (n_{S,i} + \mathbb{N} \cdot p).$$

In the last step, we transform such a grammar to get rid of the union. For this, we define new variables S_i for $i \leqslant \max(n_{S,i})$ designed to generate the words in L_S of lengths $n_{S,i} + \mathbb{N} \cdot p$. This is done in the usual way by replacing each production

$$X \to u_0 X^{(1)} u_1 \cdots X^{(m)} u_m$$

by the set

$$X_i \to u_0 X_{i_1}^{(1)} u_1 \cdots X_{i_m}^{(m)} u_m$$

for all $(i, i_1, \ldots, i_m)$ such that

$$i \equiv |u_0| + i_1 + |u_1| + \cdots + i_m + |u_m| \pmod{p}.$$

By construction, the language generated by the variables S_i are length-disjoint for each S. Their union is exactly the original language L_S. Hence, (2) and (3) are proved. ▮

Clearly, given two length-disjoint languages L and M, one has $\mathrm{Min}(L \cup M) = \mathrm{Min}(L) \cup \mathrm{Min}(M)$. Hence, we may admit the following.

Assumption A. From now on, we assume that the language under consideration is generated by a grammar satifying the conditions of Lemma 4.2.

4.1. *Covers*

Let L be a language over A. A *cover* of L is a word $x \in L$ such that $x \ll w$ for every long enough word $w \in L$. For instance, every word in the language $L = \{b^n a^n \mid n > 0\}$ is a cover of L. On the contrary, the language $L = \{a^n b^n \mid n > 0\}$ has no cover.

PROPOSITION 4.3. *Let L be an infinite language and let $\mathbf{z}$ be the minimum word of the adherence of L. A word $x \in L$ is a cover of L if and only if $x \ll \mathbf{z}$.*

Proof. Let $x \in L$ be such that $x \ll \mathbf{z}$. Set (see Fig. 1)

$$E = \{w \mid w \ll x\}, \qquad F = \{w \mid x \sqsubseteq w\}, \qquad T = E \cup F.$$

Every word w in T satisfies $w \ll \mathbf{z}$. The set T is finite, since otherwise $\mathrm{Adh}(T)$ would be nonempty and would contain a word strictly smaller than $\mathbf{z}$ which is impossible. Since every word w is comparable to x in the lexicographic order, it follows $x \ll w$ for every word $w \in L \backslash T$ that is longer than x, showing that x is cover of L.

Conversely, let x be a cover of L. Then x is in L and $x \ll w$ for every long enough word $w \in L$. Let y be the prefix of $\mathbf{z}$ of length $|x|$. There exist infinitely many words u such that $yu \in L$, and since x is a cover, one has $x \ll yu$ for u long enough. But then $x \ll y$ and consequently $x \ll \mathbf{z}$. ▮

Let G be a context-free grammar, and let S be a nonterminal generating an infinite-context-free language L. Denote by $\mathbf{z}$ the minimum word of its adherence. We distinguish the three following situations:

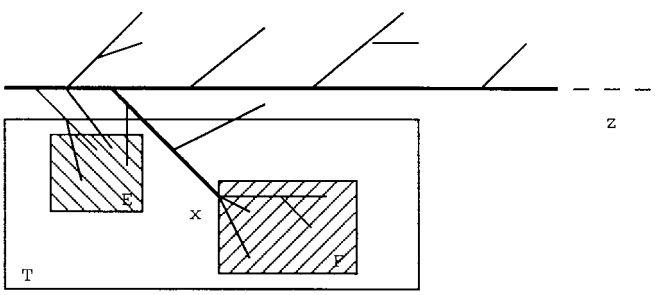

FIG. 1. Illustration of the proof of Proposition 4.3.

- L has a cover. This happens if and only if $L \cap \mathrm{Sous}(\mathbf{z}) \neq \varnothing$; we say that the nonterminal S is *covered*;
- L has no cover, and $L \cap \mathrm{Pref}(\mathbf{z})$ is infinite; in this case, the nonterminal S is called *dense*;
- L has no cover and $L \cap \mathrm{Pref}(\mathbf{z})$ is finite; in this case, S is *decreasing*.

Since $\mathbf{z}$ is ultimately periodic and computable, the two sets $\mathrm{Sous}(\mathbf{z})$ and $\mathrm{Pref}(\mathbf{z})$ are regular. Thus one can effectively test which case arises for a nonterminal.

The distinction into three cases can be explained intuitively as follows. In order to derive, from some sentential form SuT, a word of sufficiently great length l that is minimal among the words of this length, one must choose a minimal word of length l_S generated by S and a minimal word of length l_T generated by T, for integers satisfying $l_S + |u| + l_T = l$. The problem resides in the choice of l_S and l_T. Intuitively, if S is a covered nonterminal, one can choose l_S to be the length of a cover, and l_T will be big. On the contrary, if S is decreasing, one chooses l_S to the big and l_T to be small. Finally, the case of a dense nonterminal S is more delicate; as we shall see, the choice of l_S depends on the nature of the word u and of the language L_T. We shall give the necessary lemmas to make these ideas precise.

EXAMPLE 1. Consider the grammar

$$S \to baSaa$$
$$S \to b.$$

The minimum word of the adherence is $\mathbf{z} = (ba)^\omega$. Every word of the language excepted b is a cover. The smallest cover is *babaa*. The nonterminal is covered.

EXAMPLE 2. Consider the grammar

$$S \to abS$$
$$S \to abSbb$$
$$S \to a.$$

Here, $\mathbf{z} = (ab)^\omega$, and the prefixes of $\mathbf{z}$ generated by S are the words of $(ab)^* a$. The nonterminal is dense.

EXAMPLE 3. Consider the grammar

$$S \to bSa$$
$$S \to aSb$$
$$S \to a.$$

Here, $\mathbf{z} = a^\omega$. The nonterminal is decreasing.

4.2. *Decreasing Nonterminals*

Let L be an infinite context-free language and let $\mathbf{z} = st^\omega$ be the minimum word of its adherence. The language $L \cap st^*A^*$ is infinite. For each word $w \in L \cap st^*A^*$, we denote by $\rho(w)$, and call *rank* of w, the greatest integer n such that st^n is a prefix of w. Thus, every word w in $L \cap st^*A^*$ of length at least $|s|$ can be written as

$$w = st^{\rho(w)}h \qquad (*)$$

for some word h such that t is not a prefix of h. Since $\mathbf{z}$ belong to the adherence of L, the set of ranks of words in L, denoted by $Rg(L)$, is infinite.

If the nonterminal generating the language L is decreasing, the word h of equation $(*)$ satisfies $t \ll h$, provided w is long enough.

Lemma 4.4. *If L is generated by a decreasing nonterminal, there exist two integers N_0 and M_0 such that, for any integer $n \geqslant N_0$ that is in $Rg(L)$ and for any word $v \in L \cap st^*A^*$ of minimal length with rank $\rho(v) = n$, there exists a word $w \in L \cap st^*A^*$ satisfying*:

(i) $|v| \leqslant |w| \leqslant |v| + M_0$

(ii) $\rho(v) < \rho(w)$, *whence* $w \ll v$.

Proof. We start by showing how to compute ρ. Let f be the function from A^* into $\{x, y\}^*$, where x, y are two letters, defined as

$$f(w) = \begin{cases} \varnothing & \text{if} \quad w \notin st^*A^* \\ x^n y^p & \text{if} \quad w = st^n h \quad \text{and} \quad n = \rho(w), \quad |h| = p. \end{cases}$$

It is easy to see that f is a *rational* function [2, 5]: let a and x be two new letters not in A, and set $C = A \cup \{a, x\}$. Consider the morphism ϕ from C^* into A^* that is the identity on A, and such that $\phi(a) = s$ and $\phi(x) = t$. Further, observe that the language $K = \{u \in A^* \mid t \ll u\}$ is regular. If $w = st^n h$ and $n = \rho(w)$, then $\phi^{-1}(w) \cap ax^*K = ax^n h$. Let ψ be the morphism from C^* into $\{x, y\}^*$ that erases a, does not change x, and maps letters in A onto y. Then $\psi(ax^n h) = x^n y^{|h|}$ and $f = \phi^{-1} \circ \bigcap ax^*K \circ \psi$.

Consequently, the language $B = f(L) = f(L \cap st^*A^*) \subset x^*y^*$ is a bounded context-free language over $\{x, y\}$, in bijection with $L \cap st^*A^*$ by f. The language B is infinite, and the set of $n \in \mathbb{N}$ is such that $x^n y^p \in B$ for some p is infinite.

Let N_0 be the constant of some iteration lemma for B, let $n \in Rg(L)$ with $n \geqslant N_0$, and let $v \in L \cap st^*A^*$ be of rank n and of minimal length among the words of rank n in $L \cap st^*A^*$. Then $v = st^n h$ for a word h of minimal length, and $t \ll h$. One has $f(v) = x^n y^p$, where $p = |h|$. By the iteration lemma, there exist two integers k and l, with $k, l \geqslant 0$, $0 < k + l < N_0$ such that $x^{n+ki} y^{p+li} \in B$ for $i \geqslant -1$. It follows that $k > 0$, since otherwise $l > 0$ and $x^n y^{p-l} \in B$, contrary to the minimality of p.

Thus $x^{n+k} y^{p+l} \in B$. Let $w \in L \cap st^*A^*$ such that $f(w) = x^{n+k} y^{p+l}$. Then $|v| \leqslant |w| = |v| + k\,|t| + l < |v| + M_0$, where, for instance, $M_0 = N_0(|t| + 1)$. Furthermore, $\rho(w) = n + k > \rho(v)$, whence $w \ll v$. ∎

Example 4. Consider the language

$$L = \{ba^{2n}b^{2n} \mid n \geqslant 0\} \cup \{ba^n b^{2n+1} \mid n \geqslant 1\}.$$

The minimum word in the adherence is $\mathbf{z} = ba^\omega$. The lemma holds for $M_0 = 4$. Consider indeed a word $v \in L$.

If $v = ba^{2n}b^{2n}$, the word $w = ba^{2(n+1)}b^{2(n+1)}$, of length $|v| + 4$ is convenient.

If $v = ba^n b^{2n+1}$, one may choose $w = ba^{n+1}b^{2n+3}$, or also a word of the form $w = ba^{n+i}b^{n+i}$, provided that $n + i$ is even, that $i > 0$, and that the word is not too long. Taking $i = \lceil (n + 3k)/2 \rceil$, an elementary computation shows that $|v| < |w| \leqslant |v| + 3$.

4.3. *Lemmas*

We now establish three lemmas describing how, according to their type, the variables are used in derivations of words in $\mathrm{Min}(L)$. In this section, $u, v, \ldots$ are terminal words, $S, T, \ldots$ are variables, and γ, δ denote words in $(A \cup V)^*$.

Lemma 4.5. *Let S be a covered variable, let $h \in L_S$ be a cover of L_S, and let $n_0 > |h|$ be an integer such that $h \ll h'$ for all $h' \in L_S$ of length greater than n_0. If*

$$S_0 \xrightarrow{*} uSvT\delta \xrightarrow{*} x = uyvzw \in \mathrm{Min}(L),$$

with $S \xrightarrow{*} y$ and $T \xrightarrow{*} z$, then $|y| \leqslant n_0$.

Proof. Otherwise, $|y| > n_0$, and then $h \ll y$. Let then z' be a word in L_T of length $|z'| = |z| + |y| - |h|$. Such a word exists because $|y| > |h|$ and $|y| \equiv |h| \pmod p$ (Recall our assumption (A)). Then, $uSvT\delta \xrightarrow{*} uhvz'w$. Now this new word has the same length as x, and it is strongly smaller than x. Contradiction. ∎

Let now S be a dense variable. Denote K_S the language formed by the empty word and by the prefixes of $\mathbf{z}_S$ which belong to L_S:

$$K_S = \{\varepsilon\} \cup (L_S \cap \mathrm{Pref}(\mathbf{z}_S)).$$

This language is infinite and it is rational. Thus there exists an integer d_S such that, for all integer n, there exists $h \in K_S$ satisfying $|h| \leqslant n < |h| + d_S$.

Denote by R_S the set

$$R_S = K_S\, \mathrm{Pref}(A^{d_S}) \cap L_S.$$

This is the set of words in L_S obtained from K_S by adding a suffix of bounded length. This set is rational; first, observe that the set K_S is contained in Pref($\mathbf{z}_S$). Since $\mathbf{z}_S$ is ultimately periodic, it is of the form $\mathbf{z}_S = st^\omega$. Consequently, Pref($\mathbf{z}_S$) is, up to a finite set, a finite union of languages of the form st^*p, where p ranges over the prefixes of t. The language K_S is, up to a finite set, contained in this finite union. The language K_S Pref(A^{d_S}) is, up to a finite set, contained in a finite union of languages of the form st^*q, where q ranges over the *finite* set Pref(t) Pref(A^{d_S}). It then follows that the context-free language R_S is also contained in this finite union of regular sets of the form st^*q. This implies that R_S itself is rational.

LEMMA 4.6. *Let S be a dense variable. If*

$$S_0 \xrightarrow{*} uSvT\delta \xrightarrow{*} x = uyvzw \in \mathrm{Min}(L),$$

with $S \xrightarrow{*} y$ and $T \xrightarrow{*} z$, then y belongs to the rational language R_S.

Proof. There exists a word h in K_S such that $|h| \leqslant |y| < |h| + d_S$. Let y' be the longest prefix of y which is in K_S. If $|y'| \geqslant |h|$, then h is a prefix of y and y belongs to K_S Pref(A^{d_S}), and therefore to R_S. Hence, assume that $|y'| < |h|$. It follows that $h \ll y$, since no word in L_S is strongly smaller than a prefix of $\mathbf{z}_S$ because S is not covered. Let then z' be a word in L_T of length $|z'| = |z| + |y| - |h|$. Such a word does exist for the same reason as before. But now $S_0 \xrightarrow{*} uhvz'w$ and $uhvz'w \ll uyvzw$ by construction; this contradicts the minimality of x. ∎

LEMMA 4.7. *Let S be a decreasing variable. There exists a constant Δ such that, if*

$$S_0 \xrightarrow{*} uS\gamma T\delta \xrightarrow{*} x = uyvzw \in \mathrm{Min}(L)$$

with $\gamma \xrightarrow{*} v$, $S \xrightarrow{*} y$ and $T \xrightarrow{*} z$, and if y is long enough, then $|z| \leqslant \Delta$.

Proof. Indeed, if y is long enough, there exists a word y' in L_S such that $|y'| \leqslant |y|$ and $\rho(y') \geqslant N_0$, with N_0 the constant of Lemma 4.4. We may now assume that $\rho(y) \geqslant N_0$. If not, we replace y by y' and z by a word $z' \in L_T$ of length $|z'| = |z| + |y| - |y'| > |z|$, and we obtain a word in L_S having the same length than x and which is strongly smaller than x.

Let $n = \rho(y)$ be the rank of y. By Lemma 4.4, there exists a word y'' in L_S such that $|y'| \leqslant |y''| \leqslant |y'| + M_0$ and $\rho(y'') > \rho(y')$. Moreover, $|y'| \leqslant |y|$. We check now that $|y| < |y''|$. In fact, if $|y''| \leqslant |y|$, we replace, just as previously, y by y'' in the word x and z by a word $z' \in L_T$ of length $|z'| = |z| + |y| - |y''| \geqslant |z|$, and we get a word in L_S having the same length than x and which is strongly smaller than x.

So, we have $|y'| \leqslant |y| < |y''| < |y'| + M_0 \leqslant |y| + M_0$. If z is long enough (for instance, $|z| \geqslant n_T + M_0 \cdot p$, where p and n_T are the constants of Lemma 4.2), there exists a word $z' \in L_T$ such that $|z'| = |z| + |y| - |y''|$, because $|z| - |z'| = |y| - |y''| \leqslant M_0$. Since $y'' \ll y$, we get $uy''vz'w \ll uyvzw$. This proves the lemma. ∎

4.4. *End of the Proof*

Proof of the theorem. Each rule of the grammar is replaced by a rational set of linear rules. Let

$$S_0 \xrightarrow{*} x$$

with $x \in \mathrm{Min}(L)$. Any variable used in the derivation generates a minimal word of its length; otherwise, it is replaced by a lexicographically smaller one of the same length, and we get a word lexicographically smaller than x. Let then

$$S \to u_0 S_1 u_1 \cdots S_k u_k$$

be a rule used in the derivation of S_0 in x. By Lemma 4.5, each covered variable which is not S_k is derived in a word of bounded length. Similarly, each dense variable S_i which is not S_k is derived in a word of the language R_{S_i}. Finally, by Lemma 4.7, each decreasing variable is derived either in a short enough word or in a long word. In the latter case, all the variables following it are derived in short words.

More formally, starting with the original grammar, we construct incrementally a new grammar H by repeating the following process. Let

$$S \to u_0 S_1 u_1 \cdots S_k u_k$$

be a rule. If $k \leqslant 1$, it is just copied. In the other cases,

— if S_1 is covered, it is replaced by the finite set of words in L_{S_1} of length bounded by the length of a cover (according to Lemma 4.5).

— if S_1 is dense, it is replaced by the rational set R_{S_1} (according to Lemma 4.6).

— if S_1 is decreasing, it is replaced by the following two sets of rules:

$$S \to u_0 S_1 u_1 F_2 u_2 \cdots F_k u_k, \tag{1}$$

where F_2, ..., F_k are the finite sets of words in L_{S_2}, ..., L_{S_k} of length bounded by the constant Δ given by Lemma 4.7;

$$S \to u_0 F_1 u_1 S_2 u_2 \cdots S_k u_k, \tag{2}$$

where F_1 is the finite set of words in L_{S_1} so short that Lemma 4.7 does not apply.

This process terminates when all rules contain at most one variable.

Due to dense variables, in the resulting linear grammar H, the right-hand sides of variables may be rational sets. This grammar can be viewed as the result of the substitution of rational languages into an ordinary linear grammar. Thus, the language M generated is the result of a rational substitution into a linear language. It is well known (see [2]) that such a language is itself linear. By construction, $\mathrm{Min}(M) = \mathrm{Min}(L)$. Since M is linear, Theorem 4.1 is proved. ▮

5. REDUCTION TO A BOUNDED LANGUAGE

In this section, we prove the following result.

THEOREM 5.1. *Given any linear language L, there exists a linear bounded language L' such that* $\mathrm{Min}(L) = \mathrm{Min}(L')$.

Let $G = (V, S_0, P)$ be a linear grammar. Given a rule $p \in P$, we denote $\alpha(p)$ its left member and $\omega(p)$ the (unique) variable of the right member if the rule is nonterminal, setting $\omega(p) = \varnothing$ if the rule is terminal.

Consider P as a new alphabet; define $\Delta = \{pq \in P^2 \mid \omega(q) \neq \alpha(p)\}$, and let

$$D = P^+ \backslash P^* \, \Delta P^*.$$

We call *derivations* the words of the rational language D. If $d = p_1 \cdots p_n$ is in D, we set $\alpha(d) = \alpha(p_1)$ and $\omega(d) = \omega(p_n)$. Finally, we denote $u \xrightarrow[d]{*} v$ the fact that u is derived into v by the successive use of the rules $p_1, ..., p_n$.

A *loop* is an element $d \in D$ such that $\alpha(d) = \omega(d)$. We denote by B_S the set of the loops such that $\alpha(d) = \omega(d) = S$.

A derivation is *elementary* if no proper factor of d is a loop. The set of elementary derivations is finite.

LEMMA 5.2. *Let $h \in \mathrm{Min}(L)$. If there exists two loops $p, q \in B_S$, and two derivations d, d' such that*

$$S_0 \xrightarrow[dpqpd']{*} h;$$

then

$$S_0 \xrightarrow[dppqd']{*} h, \qquad \textit{and} \qquad S_0 \xrightarrow[dqppd']{*} h.$$

Proof. There exist words $u, v, x, z, y, s, t \in A^*$ such that

$$S_0 \xrightarrow[d]{*} sSt, \qquad S \xrightarrow[p]{*} uSv, \qquad S \xrightarrow[q]{*} xSy, \quad S \xrightarrow[d]{*} z$$

so that

$$S_0 \xrightarrow{*} h = suxuzvyvt.$$

It follows that

$$S_0 \xrightarrow[dppqd']{*} suuxzyvvt, \qquad S_0 \xrightarrow[dqppd']{*} sxuuzvvyt.$$

Since h is minimal, on one hand we have $suxuzvyt \preccurlyeq suuxzyvvt$, and on the other hand, we have $suxuzvyvt \preccurlyeq sxuuzvvyt$. Hence, $xu \preccurlyeq ux$ and $ux \preccurlyeq xu$; whence $xu = ux$, implying also that $vy = yv$, which in turn shows that these three words are equal. ▮

Let B be the set of elementary loops and let R be the set formed of the empty word and of the elementary derivations which are not loops. Any derivation can be decomposed in a product of elementary derivations. We set

$$M = D \cap \bigcup \{r_0 b_1^* r_1 b_2^* \cdots r_{n-1} b_n^* r_n\},$$

where the union is over all the $r_1, ..., r_n \in R$ and over all the $b_1, ..., b_n \in B$ which are *pairwise different*. The language M is then rational and bounded.

PROPOSITION 5.3. *Let $h \in \mathrm{Min}(L)$. There exists $d \in M$ such that $S_0 \xrightarrow[d]{*} h$.*

Proof. Let d be a derivation sequence that generates h from S_0. We can tranform d into a derivation where all occurrences of a same elementary loop are consecutive. Indeed, if there exists a factorization $d = spqpt$ for an elementary loop p, then q is also a loop. By the lemma, the derivation $sppqt$ also derives h from S_0. If $p_1, ..., p_n$ are the elementary loops that are factors of d, we then may transform d in a derivation

$$d' = r_0 p_1^{k_1} r_1 \cdots r_{n_1} p_n^{k_n} r_n,$$

where $k_1, ..., k_n$ are positive integers, and $r_0, ..., r_n$ are elementary derivations or the empty word. It follows that $d' \in M$. ▮

Proof of the theorem. It is easy to construct, given a grammar G for L, a linear grammar generating the language L' formed of those words that can be derived in G by a derivation belonging to the set M just defined above. Now, by the proposition, we have $\mathrm{Min}(L) \subset \mathrm{Min}(L')$, and since $L' \subset L$, we get the equality $\mathrm{Min}(L) = \mathrm{Min}(L')$. Obviously, L' is a linear bounded language. ▮

6. THE CASE OF A BOUNDED LANGUAGE

In this section, we prove the following result.

THEOREM 6.1. *For any context-free linear bounded language L, the language* $\mathrm{Min}(L)$ *is context-free.*

It is convenient to introduce the following notation: given two words u, v and a language L, we set

$$\langle u, L, v\rangle = \bigcup_{n\geqslant 0} u^n L v^n.$$

Note that

$$\mathrm{Min}(\langle u, L, v\rangle) = \mathrm{Min}(\langle u, \mathrm{Min}(L), v\rangle).$$

A linear language B is *bounded of degree k* if it is of the form

$$B = \{\alpha_1 u_1^{n_1}\alpha_2 u_2^{n_2}\cdots\alpha_k u_k^{n_k}\beta v_k^{n_k}\alpha'_k v_{k-1}^{n_{k-1}}\cdots v_1^{n_1}\alpha'_1 \mid n_1, ..., n_k \geqslant 0\}$$

with $(u_1, v_1), ..., (u_k, v_k) \neq (\varepsilon, \varepsilon)$. In other words, B is (linear) bounded of degree k if it can be written as

$$B = \alpha\langle u, B', v\rangle\ \alpha'$$

with B' of degree $k-1$.

The proof of the theorem is a consequence of three lemmas that we state now.

LEMMA 6.2. *If B is a finite union of languages of degree* 1, *then* Min(B) *is context-free.*

LEMMA 6.3. *If B is bounded of degree* 2, *then* Min(B) = Min(B'), *where B' is a union of bounded languages of degree* 1.

LEMMA 6.4. *If B is bounded of degree $k\geqslant 3$, then* Min(B) = Min(B'), *where B' is a union of bounded languages of degree at most $k-1$.*

Proof of the theorem. Let B be a linear bounded language. Then, B is a finite union of bounded languages $B_1, ..., B_n$ of degree at most k, for some integer k:

$$B = B_1 \cup \cdots \cup B_n.$$

Applying Lemmas 6.4 and 6.3 repetitively, for each B_i, there exists a language B'_i which is a union of languages of degree 1 and such that Min(B_i) = Min(B'_i). Now

$$\begin{aligned}\mathrm{Min}(B) &= \mathrm{Min}(\mathrm{Min}(B_1)\cup\mathrm{Min}(B_2)\cup\cdots\cup\mathrm{Min}(B_n))\\ &= \mathrm{Min}(\mathrm{Min}(B'_1)\cup\mathrm{Min}(B'_2)\cup\cdots\cup\mathrm{Min}(B'_n))\\ &= \mathrm{Min}(B'_1\cup B'_2\cup\cdots\cup B_n)\end{aligned}$$

and the result follows directly from Lemma 6.2. ∎

Proof of Lemma 6.4. Let

$$B = \alpha\langle u, C, v\rangle\ \alpha'$$

be a bounded language of degree $k\geqslant 3$ with C of degree $k-1$. We have

$$\mathrm{Min}(B) = \mathrm{Min}(\alpha\langle u, \mathrm{Min}(C), v\rangle\ \alpha').$$

If $k=3$ then, by Lemma 6.3, we get Min(C) = Min(C'), where C' is a union of bounded languages of degree 1. If $k>3$, by induction, we have Min(C) = Min(C'), where C' is an union of bounded languages of degree $k-2$. In both cases, the language $B' = \alpha\langle u, C', v\rangle\ \alpha'$ is a union of bounded languages of degree $k-1$, and

$$\begin{aligned}\mathrm{Min}(B) &= \mathrm{Min}(\alpha\langle u, \mathrm{Min}(C'), v\rangle\ \alpha') = \mathrm{Min}(\alpha\langle u, C', v\rangle\ \alpha')\\ &= \mathrm{Min}(B')\end{aligned}$$

which proves the lemma. ∎

Before proving Lemma 6.3, we establish a combinatorial result.

LEMMA 6.5. *Let u and x be two words and let $k > |u|$. If x^k is prefix of ux^k, then $xu = ux$.*

Proof. If $u=\varepsilon$ or $x=\varepsilon$, the result is immediate. Hence, we assume that neither $u\neq\varepsilon$ nor $x\neq\varepsilon$. Then, $k\geqslant 2$. We may assume also that x is a primitive word, because if $x=t^p$, then t^k is a prefix of ut^k, and the equation $tu=ut$ implies $xu=ux$.

By the above assumptions, there exists a word y such that

$$x^k y = ux^k.$$

It follows from $x^{k-1}(xy) = ux^k$ that u is a prefix of x^{k-1}. So, we have

$$x^{k-1} = uz$$

for a word z. But now, $uzxy = x^k y = ux^k$, so that $x^k = zxy$. Since the word x is primitive, the two words z and y are powers of x. Hence, u is a power of x. ∎

COROLLARY 6.6. *Let x, y, z, and u be four words and let $k > |u|$. If xy^k is a prefix of uxy^k, then, there exists a word v such that $xv = ux$ and $yv = vy$.*

The word v is obviously a conjugate of u.

Proof. As xy^k is a prefix of uxy^k, the word x is a prefix of ux, so that $xv = ux$ for a word v. Then, xy^k is a prefix of xvy^k, and this ensures that y^k is a prefix of vy^k. Now, $k > |u| = |v|$ and the lemma completes the proof. ∎

Proof of Lemma 6.3. Let

$$B = \{\alpha u^n x y^m z t^m w v^n \beta \mid n, m \geqslant 0\}$$

be a bounded language of degree 2. Clearly,

$$\mathrm{Min}(B) = \alpha\ \mathrm{Min}(B')\ \beta,$$

where $B' = \{u^n x y^m z t^m w v^n \mid n, m \geqslant 0\}$. So, we may assume $\alpha = \beta = \varepsilon$. We have $uv \neq \varepsilon$ and $yt \neq \varepsilon$. Let

$$r = \mathrm{lcm}(|uv|, |yt|), \quad L = \frac{1}{|uv|}, \quad M = \frac{r}{|yt|}.$$

Then

$$B = \bigcup_{\substack{0 \leqslant \lambda \leqslant L \\ 0 \leqslant \mu \leqslant M}} B_{\lambda, \mu}$$

with

$$B_{\lambda, \mu} = \{u^{Ln} u^{\lambda} x y^{\mu} y^{Mm} z t^{Mm} t^{\mu} w v^{\lambda} v^{Ln} \mid n, m \geqslant 0\},$$

so that we may assume that $|uv| = |yt|$. Finally, we may assume that, in B, the exponents n and m satisfy $n, m \geqslant K$ for any fixed integer K. Indeed, the words in B for which $n < K$ or $m < K$ belong to a finite union of bounded language of degree 1.

Let then

$$B = \{u^{n+K} x y^{m+K} z t^{m+K} w v^{n+K} \mid n, m \geqslant 0\},$$

where $K > |u|$, $|y|$, $|t|$, $|v|$, and $|uv| = |yt|$. We prove that $\mathrm{Min}(B)$ is formed of the set of minimal words of a finite union of linear bounded languages of degree at most 1. Define

$$b_{n, m} = u^{n+K} x y^{m+K} z t^{m+K} w v^{n+K}.$$

We have to compute, for each integer N, the word

$$b = \mathrm{Min}\{b_{n, m} \mid 0 \leqslant n, m \leqslant N, n + m = N\}$$

because the right-hand side is made of all words in B of a given length. The computation goes by analyzing several cases.

If $uxy^K \ll xy^K$, then $b = b_{N, 0}$; similarly, if $uxy^K \gg xy^K$, then $b = b_{0, N}$. Hence, we are left with the case where xy^K is a prefix of uxy^K. By the corollary, there exists a word $\bar{u}$ such that $ux = x\bar{u}$ and $y\bar{u} = \bar{u}y$. But now, $\bar{u}$ and y are powers of some word h; so,

$$\bar{u} = h^{\lambda}, \qquad y = h^{\mu}$$

for integers $\lambda, \mu \geqslant 0$, with $\lambda + \mu > 0$, and we have

$$\begin{aligned} b_{n, m} &= x\bar{u}^{n+K} y^{m+K} z t^{m+K} w v^{n+K} \\ &= x h^{\lambda n + \mu m + (\lambda + \mu)K} z t^{m+K} w v^{n+K}. \end{aligned}$$

Case $\lambda = \mu$. We have $\bar{u} = y$ and

$$b = x y^{N+2K}\ \mathrm{Min}\{t^{m+K} w v^{n+K} \mid 0 \leqslant n, m \leqslant N, n + m = N\}.$$

As above, compare the words twv^K and wv^K. If $twv^K \ll wv^K$, then $b = b_{0, N}$; if $wv^K \ll twv^K$, then $b = b_{N, 0}$. We are left with the case where wv^K is prefix of twv^K. In this case, there exists a word $\bar{t}$ such that $tw = w\bar{t}$ and $\bar{t}v = v\bar{t}$. Now t and v have the same length, so $\bar{t} = v$. It follows that $t^{m+K} w v^{n+K} = w v^{N+2K}$ and consequently $b = b_{n, m}$ for all n, m such that $n + m = N$.

The result is proved in this case. There remains the cases where $\lambda \neq \mu$.

Case $\lambda < \mu$. Let $\mu = \lambda + \gamma$ and $c = h^{\gamma}$. We have $y = c\bar{u} = \bar{u}c$, and

$$b_{n, m} = x\bar{u}^{N+2K} c^{m+K} z t^{m+K} w v^{n+K}.$$

We remark that $|ct| = |v|$ and, moreover, $|c| < K$. It is simpler to set here

$$\begin{aligned} \bar{b}_{n, m} &= c^{m+K} z t^{m+K} w v^{n+K}, \\ \bar{b} &= \mathrm{Min}\{\bar{b}_{n, m} \mid 0 \leqslant n, m \leqslant N, n + m = N\}. \end{aligned}$$

As above, compare czt^K and zt^K. If $czt^K \ll zt^K$ (resp. $czt^K \gg zt^K$), we have $\bar{b} = \bar{b}_{0, N}$ (resp. $\bar{b} = \bar{b}_{N, 0}$). We are left with the case where zt^K is a prefix of czt^K. By the corollary, there exists a conjugate $\bar{c}$ of c such that $\bar{c}$ and t are powers of same word $\bar{h}$:

$$\bar{c} = \bar{h}^{\bar{\gamma}}, \qquad t = \bar{h}^{\bar{\mu}}$$

Set $d = \bar{h}^{\bar{\gamma} + \bar{\mu}}$. Then

$$\bar{b}_{n, m} = z d^{m+K} w v^{n+K}$$

and

$$\bar{b} = z\ \mathrm{Min}\{d^{m+K} w v^{n+K} \mid 0 \leqslant n, m \leqslant N, n + m = N\}.$$

We conclude just as in the previous case.

Case $\lambda > \mu$. This cases is almost symmetric to the previous one. Set $\lambda = \mu + \theta$ and $c = h^{\theta}$. We have $\bar{u} = cy = yc$ and

$$b_{n, m} = x y^{N+2K} c^{n+K} z t^{m+K} w v^{n+K}.$$

We remark that, in this case, $|t| = |cv|$ and $|c| \leqslant |\bar{u}| < K$. Set

$$\bar{b}_{n,m} = c^{n+K} z t^{m+K} w v^{n+K},$$

$$\bar{b} = \mathrm{Min}\{\bar{b}_{n,m} \mid 0 \leqslant n, m \leqslant N, m = N\},$$

As above, compare czt^K and zt^K. If $czt^K \ll zt^K$ (resp. $czt^K \gg zt^K$), we have $\bar{b} = \bar{b}_{N,0}$ (resp. $\bar{b} = \bar{b}_{0,N}$). So, we are now left with the case where zt^K is a prefix of czt^K. By the corollary, there exists a conjugate $\bar{c}$ of c such that $\bar{c}$ and t are powers of the same word $\bar{h}$:

$$\bar{c} = \bar{h}^{\bar{\gamma}}, \qquad t = \bar{h}^{\bar{\theta}}.$$

As $|t| = |cv|$, we have $\bar{\lambda} \leqslant \bar{\theta}$. Set $d = \bar{h}^{\bar{\theta} - \bar{\gamma}}$. Then

$$\bar{b}_{n,m} = z\bar{c}^{N+2K} d^{m+K} w v^{n+K}$$

and

$$\bar{b} = z\bar{c}^{N+2K}\,\mathrm{Min}\{d^{m+K} w v^{n+K} \mid 0 \leqslant n, m \leqslant N, n+m = N\}.$$

We conclude as above. ▮

Proof of Lemma 6.2. It suffices to prove that, if B_1 and B_2 are two linear bounded languages of degree 1, then $\mathrm{Min}(B_1 \cup B_2)$ is context-free.

Set

$$B_1 = \{\alpha u^n \beta v^n \gamma \mid n \geqslant 0\}, \qquad B_2 = \{xy^n z t^n w \mid n \geqslant 0\}.$$

Let $r = \mathrm{lcm}(|uv|, |yt|)$. Replacing $u, v, y,$ and t by appropriate powers, we may express B_1 and B_2 as union of languages such that $|uv| = |yt| = r$. Now, B_1 and B_2 have words of the same lengths if and only if $|\alpha\beta\gamma| \equiv |xzw| \bmod r$. Choosing appropriate exponents allows us to assume that $|\alpha\beta\gamma| = |xzw|$. By replacing u, v, y, t by conjugates if necessary, we may even assume that $|\alpha| = |x|$, $|\beta| = |z|$, so that $|\gamma| = |w|$. Thus, we have to compare words $\alpha u^n \beta v^n \gamma$ and $xy^n z t^n w$ for $n \geqslant 0$, with

$$|\alpha| = |x|, \qquad |\beta| = |z|, \qquad |\gamma| = |w|, \qquad |uv| = |yt|.$$

If $\alpha \neq x$, then $\mathrm{Min}(B_1 \cup B_2)$ is equal to B_1 or to B_2, according to $\alpha \ll x$ or $x \ll \alpha$. Thus there remains the case where $\alpha = x$.

If $|u| = |y|$, then $\mathrm{Min}(B_1 \cup B_2)$ is equal to B_1 or to B_2, according to $u\beta v\gamma \ll xztw$ or $xztw \ll u\beta v\gamma$. If these words are equal, then $B_1 = B_2$.

By symmetry, we may assume that $|u| < |y|$. The words $u^{|y|}$ and $y^{|u|}$ have the same length. If they are different, then $\mathrm{Min}(B_1 \cup B_2)$ is equal to B_1 or to B_2, according to $u^{|y|} \ll y^{|u|}$ or $u^{|y|} \gg y^{|u|}$.

If the words $u^{|y|}$ and $y^{|u|}$ are equal, then u and y are powers of the same primitive word h. We then get $y = h^l$ for an integer l. For all n, we then have $xy^n z t^n w = \alpha u^n h^{ln} z t^n w$, and we just have to compare $\beta v^n \gamma$ and $h^{ln} z t^n w$. Let k be an exponent such that $|\beta| < |h^k|$. Then $\mathrm{Min}(B_1 \cup B_2)$ is B_1 or B_2, according to $\beta \ll h^k$ or $\beta \gg h^k$.

The case remains where β is a prefix of h^k. In this case, there exists a factorization $h = h_1 h_2$ and an exponent λ such that $\beta = h^\lambda h_1$. We then compare $v^{|h|}$ and $(h_2 h_1)^{|v|}$. The language $\mathrm{Min}(B_1 \cup B_2)$ is equal to B_1 or to B_2 according to $v^{|h|} \ll (h_2 h_1)^{|v|}$ or $v^{|h|} \gg (h_2 h_1)^{|v|}$.

We are now left with the case where these two words are equal. Since the word $h_2 h_1$ is primitive, the word v is a power of $h_2 h_1$. Set $v = (h_2 h_1)^m$. Then $\beta v^n \gamma = h^{l+mn} h_1 \gamma$ and the computation of $\mathrm{Min}(B_1 \cup B_2)$ reduces to the comparison of the words of same lengths $h^{l+mn} h_1 \gamma$ and $h^{ln} z t^n w$. We then have $m > l$, and we conclude as above. ▮

CONCLUSION

We have shown that, given a context-free language L, the language $\mathrm{Min}(L)$ is context-free. Each step of the proof is effective. Hence, given a grammar generating L, we can construct a grammar generating $\mathrm{Min}(L)$.

The construction we presented produces a description of $\mathrm{Min}(L)$ as a union of linear bounded languages of degree 1. It is very easy to show that *any* context-free language which happens to be slender (containing at most one word of each length) can be described by such an union. Consequently, it is the context-freeness of the language $\mathrm{Min}(L)$ (and not its type with respect to the hierarchy of context-free language) which constitutes the "difficult" part of the result.

ACKNOWLEDGMENTS

The authors thank the anonymous referee for pointing out some some errors in a preliminary version and for helpful comments which have improved the clarity of the paper.

REFERENCES

1. J.-M. Autebert, "Théorie des langages et des automates," Masson, Paris, 1994.
2. J. Berstel, "Transductions and Context-Free Languages," Teubner, Leipzig, 1979.
3. L. Boasson and M. Nivat, Adherences of languages, *J. Comput. System Sci.* **10** (1980), 185–309.
4. J. Dassow, G. Paun, and A. Salomaa, On thinness and slenderness of L languages, *Bull. EATCS* **49** (1993), 152–158.

5. S. Eilenberg, "Automata, Languages and Machines," Vol. A, Academic Press, New York, 1974.
6. M. Harrison, "Introduction to Formal Language Theory," Addison–Wesley, Reading, MA, 1978.
7. M. Lothaire, "Combinatorics on Words", Cambridge Univ. Press, Cambridge, 1983.
8. A. Salomaa, "Formal Languages," Academic Press, New York, 1973.