Prefixes of infinite words and ambiguous context-free languages

Jean-Michel Autebert, Philippe Flajolet, Joaquim Gabarro

IRIA
CENTRE DE ROCQUENCOURT

Rapports de Recherche

N° 492

PREFIXES OF INFINITE WORDS AND AMBIGUOUS CONTEXT-FREE LANGUAGES

Jean-Michel AUTEBERT
Philippe FLAJOLET
Joaquim GABARRO

Mars 1986

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. : (1) 39 63 55 11

PREFIXES OF INFINITE WORDS AND AMBIGUOUS CONTEXT-FREE LANGUAGES

Jean-Michel AUTEBERT

Philippe FLAJOLET

Joaquim GABARRO

Abstract: *Two "gap" theorems are shown for languages formed with words that fail to be prefixes of an infinite word: such languages can never be described by unambiguous context-free grammars.*

Résumé: On établit deux théorèmes "lacunaires" pour les langages formés de mots qui ne sont pas préfixes d'un mot infini: de tels langages ne peuvent jamais être décrits par une grammaire context-free non ambiguë.

PREFIXES OF INFINITE WORDS AND AMBIGUOUS CONTEXT-FREE LANGUAGES

Jean-Michel AUTEBERT

Université Paris 7, U.E.R. Mathématiques et Informatique,
2 Place Jussieu, 75251 Paris (France)

Philippe FLAJOLET

INRIA, Rocquencourt
78150 Le Chesnay (France)

Joaquim GABARRO

Universitat Politecnica, Facultat Informatica
5 Pau Gargallo, 08028 Barcelona (Spain)

ABSTRACT

Two "gap" theorems are shown for languages formed with words that fail to be prefixes of an infinite word: such languages can never be described by unambiguous context-free grammars.

1. Infinite words and their prefixes.

Let A be a fixed finite alphabet with card $A \geq 2$. Given an infinite word of A^ω:

$$\mathbf{w} = w_1 w_2 w_3 \cdots \quad ; \; w_j \in A \tag{1}$$

one defines classically its *prefix language* and its *coprefix language* by:

$$\text{Pref}(\mathbf{w}) = \{w_1 w_2 \cdots w_m \mid m \geq 0\} \text{ and } \text{Copref}(\mathbf{w}) = A^* / \text{Pref}(\mathbf{w}) . \tag{2}$$

In this paper, we propose to explore some language-theoretic properties of these sets, especially the coprefix languages, and we start with a few examples.

1. The "*simple repetitive*" word:

$$\mathbf{r} = (ab)^\omega = a\,b\,a\,b\,a\,b\,a\,b\,a\,b \cdots \tag{3}$$

is such that

$$\text{Pref}(\mathbf{r}) = (ab)^*(\varepsilon + a) \quad ; \quad \text{Copref}(\mathbf{r}) = b\{a,b\}^* + a\{a,b\}^*(aa+bb)\{a,b\}^* \tag{4}$$

and both languages are regular languages.

2. The "***all-integers***" words:

$$\mathbf{i} = ab\ a^2b\ a^3b\ a^4b\ \cdots \tag{5}$$

is such that its coprefix language satisfies:

$$\text{Copref}(\mathbf{i}) \cap \{a,b\}^*b = \{a^{n_1}ba^{n_2}b \cdots a^{n_p}b \mid \text{for some } k, n_k \neq k\}. \tag{6}$$

The context-free language given by (6) is nothing but the classical Goldstine language that was proved inherently ambiguous in [Fl85,Fl86].

3. The "***doubly exponential***" word [Gr85]:

$$\mathbf{d} = a^2b\ a^{2^2}b\ a^{2^4}b \cdots a^{2^{2^k}}b \cdots \tag{7}$$

is such that, because of fast growth properties of blocks of a's, the coprefix language associated to it is non context-free.

We propose to show here that the above situations are the only ones that may appear: in other words, there is a ***gap*** in the sense that the coprefix language of an infinite word can never be an ***unambiguous*** (non-regular) context-free language.

Theorem 1: [Main Gap Theorem] *Let* $\mathbf{w}$ *be an infinite word of* A^ω. *Then, for the language* L=Copref($\mathbf{w}$) *there are only 3 possibilities:*

a. L *is a non-context-free language;*

b. L *is an* inherently ambiguous *context-free language;*

c. L *is a regular language.*

It is of course case (b) that is of particular interest: if a coprefix language has been recognized to be context-free and non regular, then it is inherently ambiguous.

There is a particular class of infinite words (and associated coprefix languages) that appears recurrently in formal language theory and combinatorics on words [Lo83], namely words defined by *iterated morphisms*. If h is a monoid homomorphism: $h \in A \to A^*$, and $h(a)$ starts with an a, then provided that natural growth conditions are satisfied (*i.e.* $|h(a)|>1$ and the lengths of the $h^{(n)}(a)$ are unbounded), the sequence:

$$a\ ;\ h(a)\ ;\ h^2(a)\ ;\ h^3(a)\ ;\ \cdots \tag{8}$$

"converges" to an infinite word denoted by $h^\omega(a)$. The following observation is due to Berstel [Be85]:

> "*The coprefix language of an infinite word* $h^\omega(a)$ *defined by an iterated morphism is always a context-free language.*"

Furthermore, it has been proved lately by Harju-Linna [HL86] and independently by Pansiot [Pa86], that it is decidable whether an infinite word defined by iterated morphism is eventually periodic. Thus, we can deduce immediately from our main gap theorem:

Theorem 2: [Gap Theorem for Words defined by Iterated Morphisms] *Let* $\mathbf{w}$ *be a word defined by an iterated morphism. Then for the language* L=Copref($\mathbf{w}$)

there are only 2 possibilities:

a. *L is a regular language;*

b. *L is an inherently ambiguous context-free language.*

Furthermore, given morphism h, it is decidable which of cases (a) or (b) holds.

To illustrate the use of Theorem 2, let us indicate now two examples of application; ambiguity of the first example solves an earlier conjecture of Autebert and Gabarro [AG85].

1. The *Thue-Morse* word **t**. Consider the integer sequence $\{\mu_n\}_{n\geq 0}$ defined by the recurrence:

$$\mu_0 = +1 \; ; \; \mu_{2n} = \mu_n \; ; \; \mu_{2n+1} = -\mu_n \; ; \tag{9}$$

if $\nu(n)$ denotes the number of ones in the binary representation of integer n, then clearly:

$$\mu_n = (-1)^{\nu(n)} . \tag{10}$$

In words: μ_n is ± 1 depending on the parity of the number of 1-bits in the representation of n.

If one lists the pattern of signs in sequence (1), one obtains an infinite word called the Thue-Morse sequence:

$$\mathbf{t} = +\,-\,-\,+\,-\,+\,+\,-\,-\,+\,+\,-\,+\,-\,-\,+\,-\,+ \cdots \tag{11}$$

Many authors in formal language theory (see *e.g.* [Lo83]) have investigated combinatorial properties of that sequence: for instance it is cube-free, and from it one can easily construct square-free sequences. It also appears in the analysis of some probabilistic estimation algorithms [FM85], in number theory [AC85] where one has "curious identities" of Shallit, Woods and Robbins like:

$$\frac{1}{\sqrt{2}} = (\frac{1}{2})^{\mu_0}(\frac{3}{4})^{\mu_1}(\frac{5}{6})^{\mu_2} \cdots$$

and it plays a role in the theory of algebraic functions over finite fields [Fu67]: as an example, the generating function of the 0-1 sequence $(1+\mu_n)/2$ is algebraic over $GF_2[z]$ but transcendental over $\mathbb{Q}[z]$. The reader will find an amusing discussion of several related issues in [DMF82].

We shall see that our gap theorems enable us to establish the *inherent ambiguity* of the (context-free) language Copref(**t**). Observe that, rewriting "a" for "+" and "b" for "−", the Thue Morse sequence is generated by iteration of the morphism:

$$h(a) = ab \; ; \; h(b) = ba . \tag{12}$$

2. The *Fibonacci word* is obtained by iterating the morphism

$$h(a) = ab \; ; \; h(b) = a \tag{13}$$

(a is an adult rabbit and b is a baby!), which gives rise to the infinite word:

$$\mathbf{f} = a\,b\,a\,a\,b\,a\,b\,a\,a\,b\,a\,a\,b \cdots . \tag{14}$$

Again Copref(**f**) will appear to be an inherently ambiguous context-free language.

2. Proof techniques.

Our proofs rely mainly on the analytic techniques introduced by Flajolet [Fl85,Fl86], themselves based on a combination of the Chomsky-Schutzenberger Theorem:

> *"An unambiguous context-free language has an algebraic generating function."*

and classical nineteenth century complex analysis enabling us in a large number of cases to recognize that an analytic function is transcendental:

> *"If the generating function of a context-free language considered as an analytic function is a transcendental function, then the language is inherently ambiguous."*

The main contribution of this paper is to introduce in this range of questions an application of the following deep theorem of Polya and Carlson, first conjectured by Polya and established by Carlson [Ca21] (see also *e.g.* Polya's works, [Po74], pp. 175,779):

> **Theorem:** [Polya-Carlson] *If a power series with integer coefficients converges in the unit circle, then either it represents a rational function or it has the unit circle as a natural boundary.*

Let therefore $\mathbf{w}$ be any infinite word, say without loss of generality over a binary alphabet $\{a,b\}$. The (non-commutative) bivariate generating function of L=Copref($\mathbf{w}$) is:

$$l(a,b) = \frac{1}{1-a-b} - 1 - \sum_{n\geq 1} W_n(a,b) \tag{15}$$

where $W_n \equiv W_n(a,b)$ is the non-commutative monomial formed with the prefix of $\mathbf{w}$ of length n.

From the above principles, to establish Theorem 1, our task is reduced to proving that the commutative image of $l(a,b)$ given by (15) is a transcendental function. We shall do so by proving that an "algebraically related" univariate function is transcendental using the Polya-Carlson Theorem.

3. Infinite words and indicator functions.

Let κ_n be 1 if the n-th letter of word $\mathbf{w}$ is an a and 0 otherwise. Then the univariate series:

$$\kappa(z) = \sum_{n\geq 1} \kappa_n z^n \tag{16}$$

is called the *indicator* series (function) of the infinite word $\mathbf{w}$. The following simple observation is crucial:

Lemma 1: *If $\kappa(z)$ is the indicator function of $\mathbf{w}$ and $l(a,b)$ is the bivariate generating function of* Copref($\mathbf{w}$)*, then:*

$$\kappa(z) = (1-z)\frac{\partial}{\partial u}\left[\frac{z(1+u)}{1-z(1+u)} - l(zu,z)\right]\Big|_{u=1}. \qquad (17)$$

(Here, and from now on, generating functions are taken to be commutative).

To see it, notice that if we set $u=a/b$ and $z=b$, then W_n becomes:

$$W_n(zu,z) = z^n u^{\lambda_n} \qquad (18)$$

where λ_n is the number of letters "a" in W_n. Thus differentiating (18) w.r.t. u and setting $u=1$, we get the monomial $\lambda_n z^n$. Since $\kappa_n \equiv \lambda_n - \lambda_{n-1}$, we have:

$$(1-z)\frac{\partial}{\partial u}\sum_{n\geq 1} W_n(zu,z)\Big|_{u=1} = \kappa(z) \qquad (19)$$

and comparing (19) with (15) yields the statement of the lemma.

The second easy observation is:

Lemma 2: *The indicator function $\kappa(z)$ is rational iff the word $\mathbf{w}$ is eventually periodic*

Word $\mathbf{w}$ is eventually periodic iff its n-th letter is predictable from letters of rank $n-1, n-2, \cdots n-k$ for some fixed k, and this condition is equivalent to saying that the arithmetic sequence κ_n satisfies a linear recurrence with constant integer coefficients, or that $\kappa(z)$ is rational.

We can now conclude with the proof of Theorem 1, our main gap theorem. Given an infinite word $\mathbf{w}$, the following possibilities exist:

1. The coprefix language is non context-free.
2. The coprefix language is regular.
3. The coprefix language is a non-regular context-free language. Assume *a contrario* that Copref($\mathbf{w}$) is unambiguous; then its bivariate generating function is an algebraic function (by the Chomsky-Schutzenberger Theorem). By Lemma 1, the indicator function $\kappa(z)$ is an algebraic function, since it is obtained by differentiation of an algebraic function. But by the Polya-Carlson Theorem, $\kappa(z)$ can only be regular (an algebraic function has only isolated singularities), and by Lemma 2, $\mathbf{w}$ is eventually periodic so that Copref($\mathbf{w}$) is regular. Thus, we have attained a contradiction when assuming that Copref($\mathbf{w}$) is unambiguous.

Theorem 2 follows immediately from Theorem 1 and the observations made in Section 2.

4. Thue-Morse and Fibonacci sequences.

The gap theorems can be applied to the Thue-Morse word $\mathbf{t}$ and the Fibonacci word $\mathbf{f}$.

Corollary 1: *The coprefix languages associated to the Thue-Morse and the Fibonacci sequences are inherently ambiguous context-free languages.*

All that is required is checking that **t** and **f** are not eventually periodic.

For the Thue-Morse sequence, this follows directly from the fact that the sequence is cube-free.

For the Fibonacci word, we can for instance verify that the indicator function $\kappa(z)$ is not rational. The Fibonacci word **f** is the limit of the sequence of words defined by the recurrence:

$$v_{-1} = b \quad ; \quad v_0 = a \quad ; \quad v_{k+2} = v_{k+1}v_k \tag{21}$$

and the length of v_k is the $k+2$ Fibonacci number F_{k+2} ($F_0=0$, $F_1=1$). Thus for any given n, if F_k is the largest Fibonacci number strictly smaller than n, then we have the fundamental relation:

$$\kappa_n \equiv \kappa_{n-F_k} . \tag{22}$$

Assume *a contrario* that the sequence κ_n were to satisfy a minimal linear recurrence relation of order d:

$$\kappa_{n+d} = c_1\kappa_{n+d-1} + c_2\kappa_{n+d-2} + \cdots + c_d\kappa_n \tag{23}$$

with $c_d \neq 0$. By taking n to be a large enough Fibonacci number F_l, we have by (22)

$$\kappa_{n+j} = \kappa_j \tag{24}$$

for $1 \leq j \leq d$. But κ_{F_l} is 1 if l is even and 0 if l is odd. Thus in the linear recurrence (23), we should have $c_d=0$ which contradicts the minimality assumption of recurrence (23).

Thus $\kappa(z)$ is not rational, and Copref(**f**) is inherently ambiguous.

Note on the Thue-Morse sequence. In the case of the Thue-Morse sequence, a stronger algebraic structure is present since the morphism h that defines it is *uniform*: $|h(a)| = |h(b)|$. We could also have resorted in this case to a direct argument.

The sequence $\mu_n=(-1)^{\nu(n)}$ is related to κ_n by:

$$\mu_n = -1 + 2\kappa_n \tag{25}$$

so that Copref(**t**) is ambiguous iff the generating function of μ_n is transcendental. But by a classical identity that goes back to Euler:

$$\sum_{n\geq 0} u^{\nu(n)}z^n = \prod_{k\geq 0}(1+uz^{2^k}) \tag{26}$$

so that:

$$\mu(z) \equiv \sum_{n\geq 0}\mu_n z^n = \prod_{k\geq 0}(1-z^{2^k}) \tag{27}$$

and the latter equation shows directly that $\mu(z)$ admits a natural boundary.

5. Conclusions.

Coprefix languages are of interest since they represent non trivial languages with maximum density, the number of words in the language with length n being 2^n-1. A natural question is whether the inherent ambiguity result can be extended to context-free languages with such maximal densities. Consideration of language L whose complement is

$$\bar{L} = \{ab^n a^n\} \cup \{ba^n b^{n+1}\}$$

shows that this is not the case: $\bar{L}$ is deterministic hence unmabiguous and so is L.

Because of the high number of words they contain, coprefix languages also prove useful in the investigation of structural properties of context-free languages. For instance, if one considers the square-free Morse-Hedlund sequence [Lo83]

$$\mathbf{m} = abcacbabcbacabcacb \cdots \tag{28}$$

obtained by iterating the morphism:

$$g(a) = abc \ ; \ g(b) = ac \ ; \ g(c) = b \tag{29}$$

then the coprefix language M=Copref(**m**) is such that its complement is square-free. From our Theorem 2, that language is also inherently ambiguous.

Corollary 2: *The coprefix language of the Morse-Hedlund sequence, which has a square-free complement, is inherently ambiguous.*

Finally, it can be shown by similar arguments that each of the languages in the hierarchy defined in [ABBL80, p. 102] is inherently ambiguous.

Corollary 3: *For any* $n \geq 2$, *the language* G_n=Copref($\mathbf{g}_n$) *over the alphabet* $X_n=\{b_0, b_1, \cdots, b_n\}$ *where* $\mathbf{g}_n$ *is defined by iterating on* b_n *the morphism* $h \equiv h_n$*:*

$$h(b_0) = b_0 \ ; \ h(b_1) = b_1 b_0 \ ; \ \cdots \ ; \ h(b_n) = b_n b_{n-1} \cdots b_0$$

is inherently ambiguous.

These languages generate an infinite decresing hierarchy of rational cones.

Acknowledgements: This work has benefited from support by the French-Spanish cooperation program whose help is gratefully acknowledged.

References

[AC85] J-P. Allouche, H. Cohen: Dirichlet Series and Curious Infinite Products, *Mathematika* (1985), in print.

[ABBL80] J-M. Autebert, J. Beauquier, L.Boasson, M. Latteux: Very Small Families of Algebraic Non-Rational Languages, in *Formal Language Theory*, R. V. Book Ed. Academic Press (1980), pp. 89-108.

[AG85] J-M. Autebert, J. Gabarro: "Compléments des facteurs gauches de mots infinis et ambiguité: quelques exemples", Report de Recerca **RR85/15**, Facultat d'informatica de Barcelona (1985), 8 p.

[Be79] J. Berstel: "Sur les mots sans carrés définis par morphismes" in Proc. ICALP'79, *Lecture Notes in Computer Science*, **71** (1979), pp. 16-22.

[Be85] J. Berstel: Every Iterated Morphism Yields a Co-CFL, *Information Processing Letters* (1985), to appear.

[Ca21] F. Carlson: Ueber Potenzreihen mit ganzzahlingen Koeffizienten, *Math. Zeitschrift* **9** (1921), pp. 1-13.

[CS63] N. Chomsky, M-P. Schutzenberger: "The Algebraic Theory of Context-Free Languages" in *Computer Programming and Formal Systems*, Braffort and Hirschberg Ed., North Holland P.C. (1963), pp. 118-161.

[DMP82] M. Dekking, M. Mendes France, A. van der Porten: Folds!, Mathematical Intelligencer **4** (1982), pp. 131-138, 173-181, 190-195.

[Fl85] P. Flajolet: "Ambiguity and Transcendence" in Proc. ICALP'85, *lecture Notes in Computer Science*, **194** (1985), pp. 179-188.

[Fl86] P. Flajolet: "Analytic Models and Ambiguity of Context-Free Languages", INRIA Res. Rep. **483** (Jan. 1986); to appear in *Theoretical Computer Science*, ICALP'85 special issue.

[FM85] P. Flajolet, N. Martin: Probabilistic Counting Algorithms and Data Base Applications, *J. Comp. System Sciences* (1985), to appear.

[Fu67] H. Furstenberg: Algebraic Functions Over Finite Fields, *J. Algebra* **7** (1967), pp. 271-277.

[Gr85] A. Grazon: "Contribution à l'étude des petites familles de langages", Thesis, University of Paris 7 (1985).

[HL86] T. Harju, M. Linna: On the Periodicity of Morphisms in Free Monoids, *Theoretical Informatics and Applications* (1986), to appear.

[Lo83] M. Lothaire (collective pseudonym): *Combinatorics on Words* in *Encyclopedia of Mathematics and Its Applications*, **17**, Addison Wesley, Reading (1983).

[Pa86] J-J. Pansiot: Decidability of Periodicity for Infinite Words, *Theoretical Informatics and its Applications* (1986), to appear.

[Po74] G. Polya: *Collected Papers; Volume 1, Singularities of Analytic Functions*, R-P. Boas Ed., The MIT Press, Cambridge (1974).

[SS78] A. Salomaa, M. Soittola: *Automata Theoretic Aspects of Formal Power Series*, Springer Verlag, New-York (1978).

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique