Theoretical Computer Science 49 (1987) 283–309
North-Holland

ANALYTIC MODELS AND AMBIGUITY OF CONTEXT-FREE LANGUAGES*

Philippe FLAJOLET
INRIA, Rocquencourt, 78150 Le Chesnay Cedex, France

Abstract. We establish that several classical context-free languages are inherently ambiguous by proving that their counting generating functions, when considered as analytic functions, exhibit some characteristic form of transcendental behaviour. To that purpose, we survey some general results on elementary analytic properties and enumerative uses of algebraic functions in relation to formal languages. In particular, the paper contains a general density theorem for unambiguous context-free languages.

1. Introduction

We propose here to study an analytic method for approaching the problem of determining whether a *context-free* language is *inherently ambiguous.* This method (which cannot be universal since the problem is highly undecidable) is applied to several context-free languages that had resisted previous attacks by purely combinatorial arguments. In particular, we solve here a conjecture of Autebert, Beauquier, Boasson and Nivat [1] by establishing that the 'Goldstine language' is inherently ambiguous. Our technique is also applied to a number of context-free languages of rather diverse structural types.

There are relatively few types of languages that have been proved to be inherently ambiguous. This situation owes mostly to the fact that classical proofs of inherent ambiguity have to be based on a combinatorial argument of some sort considering *all possible grammars* for the language. Such proofs are therefore scarce and relatively lengthy.

At an abstract level, our methodology is related to a more general principle, namely the construction of *analytic models for combinatorial problems.* Informally the idea is as follows:

> To determine if a problem P belongs to a class $\boldsymbol{C}$, associate to elements ω of $\boldsymbol{C}$ adequately chosen analytic objects $\vartheta(\omega)$ so that a (possibly partial) characterisation of $\vartheta(\boldsymbol{C})$ can be obtained. If $\vartheta(P) \notin \vartheta(\boldsymbol{C})$, then P does not belong to $\boldsymbol{C}$.

* A preliminary version of some of the results in this paper has been presented under the title "Ambiguity and transcendence" at the ICALP85 Conference (Lecture Notes in Computer Science **194** (Springer, Berlin, 1986) 179–188).

At such a level of generality, this principle is of course of little use. However it has been successfully applied in the past in the derivation of nontrivial lower bounds in complexity theory, as the two following examples demonstrate.

(1) Shamos and Yuval [39] have obtained interesting lower bounds for the complexity of computing the mean distance of points in a Euclidean space by considering *the Riemann surface associated to the complex multivalued function* (especially its *branch points*) that continues the function defined by the original problem. They obtain in this way an $\Omega(n^2)$ lower bound on the complexity of the problem. The fact that the proof of this particular result was subsequently made algebraic by Pippenger [33] does not limit the interest of their approach.

(2) More recently Ben-Or [4] has obtained a number of lower bounds for membership problems, including for instance the distinctness problem, set equality and inclusion.... His method consists in considering *the topological structure of the real algebraic variety* (the number of connected components) associated to a particular problem and relate it to the inherent complexity of that problem.

Our approach here is to examine properties of *generating functions of context-free languages* especially when these functions are considered as *analytic functions* instead of plain formal power series. The situation in this case is greatly helped by the fact that, from an old theorem of Chomsky and Schutzenberger [9], the ordinary generating function of an *unambiguous* context-free language is *algebraic* as a *series*, and thus also as an analytic *function*. Therefore, we can simply prove that a context-free language is inherently ambiguous provided we establish that its generating function is a *transcendental function*. Thus, in the previously described framework, ϑ is the mapping that associates to a formal language its counting generating function, and with $\boldsymbol{C}$ the class of unambiguous context-free languages, $\vartheta(\boldsymbol{C})$ is a subset of the set of $\boldsymbol{Q}[z]$-algebraic functions.

Proofs of transcendence for analytic functions appear to be fortunately appreciably simpler than for real numbers. A method of choice consists in establishing the transcendence of a function by investigating its *singularities*, in particular, showing that it has a *non-algebraic singularity* (the way algebraic functions may become singular is well characterised), or infinitely many singularities or even a *natural boundary*. Alternatively, one may study Taylor coefficients of functions since many of their properties, especially their asymptotic growth, are reflections of functions singularities.

In the sequel, we shall state some useful transcendence criteria for establishing inherent ambiguity of context-free languages (Section 3), and then present a number of applications to specific languages (Sections 4–8).

Note about our presentation: It should be clear in what follows that we have made no attempt at deriving the simplest or most elementary proofs of inherent ambiguities of languages. We have instead tried to demonstrate the variety of techniques that may be employed here as they should prove useful in future applications. It should also be clear that a very large number of languages are amenable to these techniques

and some random sampling has been exercised to keep this paper within reasonable size limits.

2. Some inherently ambiguous languages

A context-free grammar G is *ambiguous* iff there exists at least one word in the language generated by G that can be parsed according to G in two different ways. A context-free language L is *inherently ambiguous* iff *any* grammar that generates L is ambiguous.

A prototype of an inherently ambiguous language is

$$L = \{a^m b^n c^p \mid n = m \text{ or } n = p\} \tag{1}$$

and the proof of its inherent ambiguity proceeds by showing, by means of some iteration theorem, that *any* grammar for L needs to generate words of the form $a^n b^n c^n$ at least twice for large enough n. (See, e.g., Harrison's book [21] for similar classical proofs.)

In this paper, we propose to prove the inherent ambiguity of a number of languages of various types that are structurally more complex than the above example.

Theorem 1 (Languages with constraints on the number of occurrences of letters). *The languages O_3, O_4, Ω_3 are inherently ambiguous, where*

$$O_3 = \{w \in \{a, b, c\}^* \mid |w|_a = |w|_b \text{ or } |w|_a = |w|_c\},$$
$$O_4 = \{w \in \{x, \bar{x}, y, \bar{y}\}^* \mid |w|_x = |w|_{\bar{x}} \text{ or } |w|_y = |w|_{\bar{y}}\},$$
$$\Omega_3 = \{w \in \{a, b, c\}^* \mid |w|_a \neq |w|_b \text{ or } |w|_a \neq |w|_c\}.$$

Theorem 2 (Crestin's language formed with products of palindromes). *The language C is inherently ambiguous, where*

$$C = \{w_1 w_2 \mid w_1, w_2 \in \{a, b\}^*;\ w_1 = w_1^t,\ w_2 = w_2^t\}$$

with w^t denoting the mirror image of w.

Theorem 3 (A simple linear language). *The language S is inherently ambiguous, where*

$$S = \{a^n b v_1 a^n v_2 \mid n \geq 1;\ v_1, v_2 \in \{a, b\}^*\}.$$

Theorem 4 (Languages with a comb-like structure). *The languages P_1, P_2 are inherently ambiguous, where*

$$P_1 = \{\underline{n}_1 \underline{n}_2 \ldots \underline{n}_{2k} \mid (\text{for all } j,\ n_{2j} = n_{2j-1}) \text{ or } (\text{for all } j,\ n_{2j} = n_{2j+1},\ n_{2k} = n_1)\},$$
$$P_2 = \{\underline{n}_1 \underline{n}_2 \ldots \underline{n}_k \mid (n_1 = 1, \text{for all } j,\ n_{2j} = 2n_{2j-1}) \text{ or } (\text{for all } j,\ n_{2j} = 2n_{2j+1})\}$$

and, for an integer $n \geq 0$, $\underline{n}$ denotes the unary representation of n in the form of $a^n b$.

Remark. In these definitions, k runs over the integers $\geqslant 1$ and the n_j over integers $\geqslant 0$.

Theorem 5 (Languages deriving from the Goldstine language). *The languages* $G_{\neq}$, $G_{<}$, $G_{>}$, $H_{\neq}$ *are inherently ambiguous, where*

$$G_{\neq} = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_p \,|\, \text{for some } j,\ n_j \neq j\},$$

$$G_{<} = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_p \,|\, \text{for some } j,\ n_j < j\},$$

$$G_{>} = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_p \,|\, \text{for some } j,\ n_j > j\},$$

$$G_{=} = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_p \,|\, \text{for some } j,\ n_j = j\},$$

$$H_{\neq} = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_p \,|\, \text{for some } j,\ n_j \neq p\}.$$

Remark. Variable p runs over all integers $\geqslant 1$ and the n_j over integers $\geqslant 0$.

Theorem 6 (Languages obeying local constraints). *The languages* K_1, K_2 *are inherently ambiguous, where*

$$K_1 = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_k \,|\, \text{for some } j,\ n_{j+1} \neq n_j\},$$

$$K_2 = \{\underline{n}_1\underline{n}_2 \ldots \underline{n}_k \,|\, \text{for some } j,\ n_{j+1} \neq 2n_j\}.$$

Remark. Variable k runs over the integers $\geqslant 2$ and the n_j over integers $\geqslant 0$.

Theorem 7 (A language based on binary representations of integers). *The language* B *is inherently ambiguous, where*

$$B = \{\tilde{n}_1\tilde{n}_2 \ldots \tilde{n}_k \,|\, n_1 \neq 1 \text{ or, for some } j,\ n_{j+1} \neq n_j + 1\}$$

in which $\tilde{n} \in \{0, 1, c\}^*$ *denotes the standard binary representation (starting with a "1") of integer* n *followed by marker "c".*

Some of these results are actually known but have been included here for the sake of illustrating the power of the methods we employ. The case of languages like O_2, O_3 is easily reduced to the ambiguity of languages like L defined in (1). The ambiguity of language Ω_3 is included here because it is related to a stronger conjecture of Autebert et al., namely that the language

$$L' = \{a^m b^n c^p \,|\, n \neq m \text{ or } n \neq p\} \tag{2}$$

is inherently ambiguous.

The language C has been studied combinatorially by Crestin [12] who proved that it is of *inherent unbounded ambiguity.* We establish here the transcendence of its generating function, which settles a conjecture of Kemp. The result concerning language S is akin to a result due to Shamir [38] by which the 'more general' language

$$\{ucv_1u^tv_2 \,|\, u, v_1, v_2 \in \{a, b\}^*\}$$

is infinitely ambiguous. The language P_2 has been studied by Kemp [26] who proved that the asymptotic density of a closely related language is a transcendental number, thereby establishing its ambiguity. Finally, the case of the language $G_{\neq}$, which is exactly the Goldstine language, solves the conjecture of Autebert et al.

Although it seems quite plausible at first sight that such languages must be inherently ambiguous, the difficulty owes to the fact that when attempting to apply iteration theorems (like Ogden's lemma), some of them (most notably the Goldstine language) behave 'almost' like regular languages.

3. An overview of transcendence criteria used for establishing inherent ambiguity

To any infinite language $L \subset A^*$ (A a finite alphabet) we associate its *enumeration sequence* (also called counting sequence) defined by

$$l_n = \text{card}\{w \in L \mid |w| = n\}.$$

This sequence is characterised by its generating function, called the *generating function* of language L:

$$l(z) = \sum_{n \geq 0} l_n z^n.$$

This function is an *analytic* function in a neighbourhood of the origin, and its radius of convergence ρ satisfies

$$\frac{1}{\text{card } A} \leq \rho \leq 1$$

since $l_n \leq (\text{card } A)^n$.

Consideration of analytical properties of the function $l(z)$ or, in an often equivalent manner, of asymptotic properties of the sequence $\{l_n\}$ permits in a number of cases to establish inherent ambiguity of the context-free language L by means of the following classical theorem of Chomsky and Schutzenberger [9].

Theorem. *Let $l(z)$ be the generating function of a context-free language L. If L is unambiguous, then $l(z)$ is an algebraic series (function) over $\boldsymbol{Q}$.*

We recall that a series $l(z)$ is algebraic over a field $\boldsymbol{K}$ (or over $\boldsymbol{K}[z]$ if one prefers) if it satisfies an algebraic equation in the ring $\boldsymbol{K}[[z]]$ of formal power series in one variable, of the form $P(z, l(z)) = 0$ for some bivariate polynomial $P(z, y) \in \boldsymbol{K}[z, y]$. It is also known that an algebraic series (over $\boldsymbol{Q}$ or $\boldsymbol{C}$) represents (a branch of) an algebraic function in a neighbourhood of the origin. Last, from classical elimination theory follows that a component of a solution to a finite system of algebraic equations is also an algebraic function in the above sense. In other words, sets of equations can be reduced rationally to a single equation (see, e.g., [30]).

The classical Chomsky-Schutzenberger theorem is established in a constructive manner by transforming an unambiguous grammatical specification of the language into a set of polynomial equations. Since we later need repeatedly to transform specifications of context-free languages into equations for corresponding generating functions, we shall illustrate the use of this theorem by means of an example.

Example. Consider the grammar with axiom A (assignment), nonterminals B (boolean), E (expression) and V (variables) over the terminal alphabet

$$\{:=, \text{not}, \text{or}, \leqslant, *, \log, \text{if}, \text{then}, \text{else}, \text{fi}, x, y, w\}$$

and production rules

$$A \to V := E,$$

$$B \to \text{not } B + \text{or } BB + \leqslant EE,$$

$$E \to *EE + \log E + \text{if } B \text{ then } E \text{ else } E \text{ fi} + V,$$

$$V \to x + y + w.$$

This grammar describes simple assignment statements in a rudimentary programming language. Letting $a(z)$, $b(z)$, $e(z)$, $v(z)$ be the generating functions of the languages associated to the nonterminals A, B, E, V respectively, we have

$$a(z) = zv(z)e(z),$$

$$b(z) = zb(z) + zb(z)^2 + ze(z)^2,$$

$$v(z) = ze(z)^2 + ze(z) + z^4 b(z)e(z)^2 + v(z),$$

$$v(z) = 3z.$$

By elimination, we find that $a(z)$ is an algebraic function of degree 10:

$$a(z)^8 - 27(z^3 - z^2)a(z)^5 + \cdots + 59049z^{10} = 0.$$

Thus the theorem simply expresses the fact that disjoint unions and (unambiguous) catenation products correspond to sums and ordinary products of generating functions (also, equations correspond to equations ...).

We shall need a related principle also to be found in [9]: if two languages are such that $L = M^*$, then the corresponding generating functions satisfies

$$l(z) = \frac{1}{1 - m(z)}$$

provided the 'star' operation on M defines L unambiguously.

The theorem will be used in the sequel under the following trivially equivalent form.

Corollary. *If the generating function $l(z)$ of a context-free language L is transcendental over $\mathbf{Q}$, then L is inherently ambiguous.*

The above corollary (see [35] for general information on languages and formal power series) therefore permits to conclude as to the inherent ambiguity of a language provided the following two conditions are met:

(i) (counting condition): one has at one's disposal a combinatorial decomposition of the language, in a way that gives access to the sequence l_n and permits to 'express' $l(z)$;

(ii) (transcendence condition): a transcendence criterion is available to establish the non-algebraic character of $l(z)$.

We now proceed with the statement of a few simple transcendence criteria of which applications will be given in the following sections.

3.1. Transcendence of values

This method constitutes in principle the most straightforward transcendence criterion for functions, although it is almost invariably the most difficult to apply. (Transcendence results are usually much easier for functions than for numbers.)

Theorem A. *Let $l(z)$ be an algebraic series over $\mathbf{Q}$ and ω an algebraic number. Then $l(\omega)$ is algebraic.*

The proof simply follows from eliminating ω from the set of two equations:

$$\begin{cases} R(\omega) = 0, \\ P(\omega, l(\omega)) = 0. \end{cases}$$

Hence, we can formulate the following criterion.

Criterion A (Transcendence of values at an algebraic point). *If $l(z)$ is a (convergent) series of $\mathbf{Q}[[z]]$ and if $l(\omega)$ is transcendental for some algebraic ω, then $l(z)$ is transcendental.*

3.2. Nature of singularities

The next criterion is based on the fact that an algebraic function has a finite number of singularities[1] that can be explicitly determined.

Theorem B. *An algebraic function $l(z)$ over $\mathbf{Q}$ defined by an equation $P(z, l(z)) = 0$ has a finite number of singularities that are algebraic numbers z satisfying one of the*

[1] Singularities are meant here in the sense of analytic functions (not in the sense of algebraic curves): for us, $\sqrt{1-z}$ is singular at $z = 1$.

equations:

$$\text{(i)} \quad \text{Resultant}_y\left[P(z, y), \frac{\partial P(z, y)}{\partial y}\right] = 0,$$

$$\text{(ii)} \quad p_d(z) = 0,$$

where $p_d(z)$ *is the coefficient of the term of* $P(z, y)$ *of highest degree in* y.

This result is of course a very classical one (see, for instance, [27, 37]).

If in equation $P(z, y) = 0$ the coefficient of the highest degree term in y vanishes, then some of the points of the algebraic curve $y(z)$ are rejected to infinity and one has a pole (at least for some branch of the analytic function).

Otherwise, around a point (z_0, y_0) satisfying $P(z_0, y_0) = 0$, one has a locally linear relation:

$$(z - z_0)\frac{\partial P}{\partial z}(z_0, y_0) + (y - y_0)\frac{\partial P}{\partial y}(z_0, y_0) \sim 0. \tag{3}$$

If $\partial P/\partial y$ is not zero, then relation (3) locally defines y as an analytic function of z by the implicit function theorem. Else, one has a branch point.

Example. Let ε denote the empty word. The grammar

$$D \to aDbD + \varepsilon$$

defines the usual parenthesis language. The corresponding generating function satisfies the equation in y:

$$z^2y^2 - y + 1 = 0. \tag{4}$$

From Theorem B, singularities are to be found amongst:
- the roots of $p_2(z) \equiv z^2 = 0$, that is $z = 0$;
- the roots in z of the system:

$$z^2y^2 - y + 1 = 0, \qquad 2z^2y - 1 = 0;$$

that is to say, $z = \pm\frac{1}{2}$.

This can be checked here by solving (4) directly. The two solutions of (4) are

$$y_1 = \frac{1 - \sqrt{1 - 4z^2}}{2z^2}; \qquad y_2 = \frac{1 + \sqrt{1 - 4z^2}}{2z^2}.$$

The branch y_2 has a pole at the origin and hence, cannot represent the generating function of language D. The branch y_1 (which represents the generating function of D) admits $z = \pm\frac{1}{2}$ as singularities.

Criterion B. *A function having infinitely many singularities (for instance, a natural boundary) is transcendental.*

In the sequel, this result is used to establish the ambiguity of Crestin's language C taking advantage of Kemp's determination of its generating function which appears to have infinitely many singularities. Other applications stem from the existence of natural boundaries for *lacunary series* (also called gap series [34]) as an application of theorems of Hadamard, Borel and Fabry.

Theorem (Lacunary series theorem). *A series of the form*

$$y(z) = \sum_{i=0}^{\infty} a_i x^{c_i}$$

such that the c_i's are integers satisfying the 'lacunary' condition:

$$\sup_i (c_{i+1} - c_i) = +\infty$$

admits its circle of convergence as a natural boundary.

Thus, such a series cannot represent the expansion of an algebraic function around the origin. Examples of such series related to some of our future applications are:

$$\sum_{n\geq 0} z^{n(n+1)/2}; \qquad \sum_{n\geq 0} z^{2^n}; \qquad \sum_{n\geq 0} z^{[\log n!]}.$$

These functions all have the unit circle as a natural boundary and thus fail to be algebraic.

3.3. Algebraicity and transcendence of local expansions

A more refined way of establishing the transcendence of a series consists in observing the appearance of transcendental elements in *local expansions* around a singularity. Indeed, for an algebraic function, one has the following theorem.

Theorem C. *If $l(z)$ is algebraic over $\mathbf{Q}$, it admits, in the vicinity of a singularity, a fractional power series expansion of the type*

$$l(z) = \sum_{k\geq -m} a_k \left(1 - \frac{z}{\alpha}\right)^{kr}, \quad m \in \mathbf{N}, r \in \mathbf{Q}^+,$$

where the coefficients a_k are algebraic.

The above expansion is nothing but the familiar Puiseux expansion of an algebraic function. The exponents may be determined explicitly by Newton's polygon rule [14].

Example. Consider the grammar

$$S \to fSSS + x$$

defining the language S of functional schemes (terms) with x as a nullary symbol and f as a ternary symbol. By Theorem B, the singularities of the generating function $s(z)$ of language S are found to be

$$\rho, \quad \rho\, e^{2i\pi/3}, \quad \rho\, e^{-2i\pi/3} \qquad \text{with } \rho = \frac{2^{2/3}}{3}.$$

At $z=\rho$, one has $s(\rho)=2^{-1/3}$. Setting $Z=2^{1/3}(z-\rho)$ and $Y=2^{1/3}(s(z)-s(\rho))$, one gets the relation

$$-9Z-6Y^2-9ZY-2Y^3-9ZY^2-3ZY^3=0$$

so that when $Z\sim 0$, we have $Y\sim \pm i\sqrt{\frac{3}{2}Z}$ and a full expansion of Y in powers of $\sqrt{Z}$ can be obtained. (Note: one has there a branch point of order 1 and the generating function $s(z)$ corresponds to the minus sign.)

The case of equations with the particular form $y=z\varphi(y)$ is discussed by Meir and Moon [32], and it corresponds to so-called simple families of trees (or, equivalently, to terms formed with a fixed set of functional symbols).

Criterion C. *If $l(z)$ has, in the vicinity of a singularity, an asymptotic equivalent that is not of the form*

$$\omega\left(1-\frac{z}{\alpha}\right)^r$$

with ω and α algebraic and r rational, then $l(z)$ is transcendental.

In particular, the occurrence of logarithmic terms in a local expansion of a function will immediately reveal that the function is transcendental.

3.4. Density results for coefficients

It is also well known that the local behaviour of a function in the vicinity of its singularities is closely reflected by the asymptotic behaviour of its Taylor coefficients. Corresponding 'transfer' lemmas rely on contour integration techniques. From Cauchy's formula

$$l_n = \frac{1}{2i\pi}\int_{O^+} l(z)\frac{dz}{z^{n+1}} \tag{5}$$

using an adequate contour of integration, one can relate the local behaviour of function $l(z)$ to the asymptotic form of the coefficients l_n. For that purpose, one may use either the classical Darboux method [22; 11, p. 277] (i.e., integration on the circle of convergence) or the type of contour of [16] (i.e., integration on a contour extending outside the circle of convergence). Basically, these methods guarantee that the coefficients of a function $l(z)$ satisfying the expansion of Theorem

C can be obtained asymptotically by extracting the coefficients of the expansion, noticing that[2]:

$$[z^n](1-z)^{-d} = \binom{n+d-1}{d-1} \sim \frac{n^{d-1}}{\Gamma(d)}.$$

Finally, the contributions from all dominant singularities have to be added. Thus using these classical results, one finds the following theorem.

Theorem D (General density theorem for unambiguous context-free languages). *If $l(z)$ is an algebraic function over $\mathbf{Q}$ that is analytic at the origin, then its n-th Taylor coefficient l_n has an asymptotic equivalent of the form*

$$(\Delta): \quad l_n = \frac{\beta^n n^s}{\Gamma(s+1)} \sum_{i=0}^{m} C_i \omega_i^n + O(\beta^n n^t),$$

where $s \in \mathbf{Q}/\{-1, -2, -3, \ldots\}$, $t < s$; β is a positive algebraic number and the C_i and ω_i are algebraic with $|\omega_i| = 1$.

Criterion D. *Let $l(z)$ be a function analytic at the origin; if its Taylor coefficients l_n do not satisfy an asymptotic expansion of type (Δ), then $l(z)$ is transcendental.*

In passing, Criterion D generalises a result of Berstel [5] who observed that if there exists an integer β such that the limit

$$\lambda = \lim \frac{l_n}{\beta^n}$$

exists and λ is a transcendental number, then $l(z)$ is a transcendental function, so that L cannot be an unambiguous context-free language. Theorem D does provide a *generalised density* characterisation for unambiguous context-free languages that extends Berstel's results.

Examples. A particularly useful set of applications of Theorem D is for coefficients with asymptotic equivalents of the form

$$l_n \sim \gamma \beta^n n^r.$$

If either r is irrational, β transcendental or $\gamma\Gamma(r+1)$ is transcendental, then $l(z)$ is a transcendental function. Therefore, the following asymptotic behaviours are characteristic of transcendental functions:

$$O(e^n n^r); \quad O(\beta^n n^{\sqrt{2}}); \quad O\left(\frac{\beta^n}{n}\right); \quad O\left(\frac{\beta^n}{n^2}\right); \quad \pi^{1/2} 4^n n^{-3/2}; \ldots.$$

[2] We let, as usual, $[z^n]f(z)$ denote the coefficient of z^n in the Taylor expansion of $f(z)$.

The third example corresponds to a logarithmic singularity. For the fifth example, it suffices to notice that we can write it as

$$-\frac{\pi}{2}\frac{4^n n^{-3/2}}{\Gamma(-\frac{1}{2})}$$

(since $\Gamma(\frac{1}{2})=\sqrt{\pi}$) and use the fact that π is transcendental.

In contrast, $\pi^{-1/2}4^n n^{-3/2}$ *does* occur in the expansion of algebraic functions as the classical example of the Catalan numbers that count words in the Dyck language (the language D defined above by well-parenthesised expressions) demonstrate:

$$D_{2n} \equiv [z^{2n}]\frac{1-\sqrt{1-z^2}}{2z^2} = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^3}}.$$

Similarly, by the Lagrange inversion theorem [22], the language S of ternary functional terms satisfies

$$S_{3n} = \frac{1}{2n+1}\binom{3n}{n} \sim \left(\frac{27}{4}\right)^n \sqrt{\frac{3}{4\pi n^3}}.$$

On the other hand, it is easy to see, using again Stirling's formula, that amongst the sequences

$$a_n^{(k)} = \binom{2n}{n}^k$$

for integral k, it is only for $k=1$ that the $a_n^{(k)}$ are coefficients of an algebraic function (see [40] for related questions). For instance, $a_n^{(2)} \sim 16^n/\pi n$ and the factor n^{-1} corresponds to a logarithmic singularity.

3.5. Polynomial recurrences

The last batch of methods is based on a theorem by Comtet [10] to the effect that any algebraic function satisfies a linear differential equation with polynomial coefficients, a fact itself reflected on its Taylor coefficients by the following theorem.

Theorem E. *If l_n is the sequence of coefficients of an algebraic function, there exist a set of polynomials $q_0(u), \ldots, q_m(u)$ such that, for all $n \geq n_0$,*

$$\sum_{j=0}^{m} q_j(n) l_{n-j} = 0.$$

Criterion E. *Let $l(z)$ be an analytic function. If there does not exist a finite sequence of polynomials $q_0, q_1, \ldots, q_m$ such that, for n large enough,*

$$\sum_{j=0}^{m} q_j(n) l_{n-j} = 0,$$

then $l(z)$ is transcendental.

The reader is referred to Stanley's paper [40] for additional information regarding sequences satisfying polynomial-linear recurrences that have been named *P*-recursive sequences. Notice also, in passing that Comtet's result makes it possible to determine in linear time the number of words of given length in an unambiguous context-free language L. It can thus be used to generate 'at random' words of given length in an unambiguous context-free language efficiently (thereby improving some of the complexity bounds of [23]).

Note on the application of the transcendence criteria: In some cases, the above transcendence criteria can be used directly on the generating functions l of context-free languages. In some cases however, one has to proceed indirectly as follows: From $l(z)$, build a new function $\varphi(z)$ by means of an adequately chosen 'algebraic functional': $\varphi(z) = \Omega(z, l(z))$. (An algebraic functional is defined here as a functional transforming algebraic functions into algebraic functions.) Then use one of the above criteria to prove $\varphi(z)$—whence $l(z)$—transcendental.

Also, we make an occasional use of an extension of the basic Chomsky-Schutzenberger theorem under the following form: *Let $l_{n_1 \ldots n_k}$ denote the number of words in language L with n_1 occurrences of letter $a_1, \ldots, n_k$ occurrences of letter a_k. Then the multivariate generating function*

$$l(z_1, \ldots, z_k) = \sum l_{n_1 \ldots n_k} z_1^{n_1} \cdot \ldots \cdot z_k^{n_k}$$

is an algebraic function over $\boldsymbol{Q}[z_1, \ldots, z_k]$.

The two methods may be combined. Thus, in the case of a binary alphabet, if e.g. the function

$$\varphi(z_2) = \frac{\partial}{\partial z_1} l(z_1 z_2^2, z_2 z_1^2)\big|_{z_1 = 1}$$

is transcendental, then the language L is inherently ambiguous.

4. Transcendence of values of generating functions

This method is in principle the most direct. However, in practice, it turns out to be rather hard to apply because of the relative scarcity of transcendence results for real numbers. (Actually, it is even the case that many arithmetic transcendence results are established by function-theoretic techniques). That method can be applied to the following languages:

- the language O_4 defined by occurrences constraints (Theorem 1);
- the 'comb-like' language P_2 (Theorem 4).

The reader is referred to either [18, 36] for an exposition of classical transcendental number theory.

Language O_4. This language is the union of two unambiguous (actually deterministic) context-free languages. In the sense of multisets, one can write the equation

$O_4 = L_1 + L_2 - I$ where $I = L_1 \cap L_2$ and L_1, L_2 are defined by:

$$L_1 = \{w \in \{x, \bar{x}, y, \bar{y}\}^* \,|\, |w|_x = |w|_{\bar{x}}\}, \qquad L_2 = \{w \in \{x, \bar{x}, y, \bar{y}\}^* \,|\, |w|_y = |w|_{\bar{y}}\}.$$

Corresponding generating functions[3] $l_1(z)$, $l_2(z)$ are algebraic, so that $O_4(z)$ has the same transcendence status as the generating function $I(z)$ of I. Note in passing that language I encodes 2-dimensional walks on a square lattice starting and ending at the origin.

A direct computation shows that:

$$I_{2n} \equiv [z^{2n}]I(z) = \sum_{k=0} \binom{2n}{2k}\binom{2k}{k}\binom{2n-2k}{n-k} \tag{6}$$

$$= \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k}^2 \tag{7}$$

$$= \binom{2n}{n}^2. \tag{8}$$

Here, (6) is the basic counting of I as a shuffle of two languages whose enumerating sequence is the central binomial sequence; (7) comes from simplification of factorials and (8) relies on Vandermonde convolution.

From there, we find that $I(z)$ is a hypergeometric function [41, p. 499] but also an elliptic integral, as can be checked by direct expansion using Wallis' integrals:

$$I(z) = \frac{2}{\pi} \int_0^{\pi/2} (1 - 16z^2 \sin^2 \vartheta)^{-1/2} \, \mathrm{d}\vartheta.$$

One can then use a classical result in transcendence theory [36] concerning values of such integrals at algebraic points to deduce the transcendence of $I(z)$ and hence of the generating function of O_4 which is thus inherently ambiguous.

Language P_2. This language also presents itself as the union of two deterministic context-free languages. One can write $P_2 = L_1 + L_2 - I$ with now:

$$L_1 = \{\underline{n}_1 \underline{n}_2 \ldots \underline{n}_k \,|\, [n_1 = 1, \text{for all } j,\ n_{2j} = 2n_{2j-1}]\},$$

$$L_2 = \{\underline{n}_1 \underline{n}_2 \ldots \underline{n}_k \,|\, [\text{for all } j,\ n_{2j} = 2n_{2j+1}]\},$$

and $I = L_1 \cap L_2$. Languages L_1, L_2 are again deterministic, whence unambiguous, and with algebraic bivariate generating functions. Since we have

$$I = \{aba^2ba^{2^2}ba^{2^3}b \ldots a^{2^p}b \,|\, p \geqslant 0\},$$

we find that the bivariate generating function of I is

$$I(a, b) = \sum_{p \geqslant 1} b^p a^{2^p - 1}.$$

[3] We adhere from now on to the implicit convention of denoting a language L, its generating function $l(z)$ (or $L(z)$) and its counting sequence l_n (or L_n) by the same group of letters.

The function $x+xI(x,1)$ is exactly the *Fredholm series*: $F(x)=\sum_{n\geq 0} x^{2^n}$ and the approximation theorem of Thue-Siegel-Roth shows the value of the series to be transcendental at any point $x=1/q$ for integral $q>1$. Thus, $I(x,1)$ and $I(a,b)$ are transcendental functions so that P_2 is inherently ambiguous.

The last example of language P_2 has been inspired by the construction due to Kemp [26] of a context-free language with a transcendental density.

Let us recall that a language L over an alphabet A of cardinality α has *asymptotic density* δ iff

$$\delta = \lim_{n\to\infty} \frac{l_n}{\alpha^n}.$$

In general, there is a relation between *values* of generating functions at particular rational points and densities of languages: let B be a proper superset of A with $b \in B/A$. Then the language

$$M = LbB^* \tag{9}$$

has a generating function that satisfies

$$m(z) = l(z)\frac{z}{1-\beta z} \tag{10}$$

where $\beta = \text{card } B$. Thus, $m(z)$ has a simple pole at $z=1/\beta$, and a direct residue calculation with Cauchy's integral formula (5) shows that

$$m_n \sim \frac{1}{\beta} l\left(\frac{1}{\beta}\right)\beta^n$$

so that $\beta^{-1}l(\beta^{-1})$ is the asymptotic density of M (see also [6, p. 23]).

Therefore, taking an alphabet B with at least five symbols and $L \equiv O_4$ or an alphabet B with at least three symbols and $L \equiv P_2$, construction (9) furnishes two examples of (ambiguous) context-free languages with a transcendental density. The second example (built from P_2) is exactly Kemp's construction.

5. Functions with infinitely many singularities

Criterion B expresses that any function with infinitely many singularities is transcendental. Such a property may either be apparent from the very expression of the function or it may result from the theorem on *lacunary series* cited in Section 3. That method is applied here to the following examples:

- the simple linear language S (Theorem 3);
- Crestin's palindrome-related language (Theorem 2);
- the Goldstine language $G_{\neq}$ and the related languages $H_{\neq}$ and $G_{<}$ (Theorem 5).

Language S. We can decompose S unambiguously, recording the first occurrence of a group a^n as follows:

$$S = \sum_{n \geq 1} a^n b R_n a^n \{a, b\}^*, \tag{11}$$

where R_n is the regular language formed with all strings over $\{a, b\}$ ending with a b—or empty—that do not have n consecutive occurrences of letter a:

$$R_n = ((\varepsilon + a + a^2 + \cdots a^{n-1})b)^*. \tag{12}$$

In terms of generating functions, decompositions (11) and (12) lead to

$$S(z) = \frac{z}{1-2z} \sum_{n \geq 1} z^{2n} R_n(z), \tag{13}$$

$$R_n(z) = \frac{1}{1 - z(1 + z + z^2 + \cdots + z^{n-1})} = \frac{1-z}{1-2z+z^{n+1}} \tag{14}$$

so that, finally,

$$S(z) = \frac{z(1-z)}{1-2z} \sum_{n \geq 1} \frac{z^{2n}}{1-2z+z^{n+1}}. \tag{15}$$

The terms in the sum of (15) are defined except at the roots of their denominator. Let $P_n(z)$ denote $1-2z+z^{n+1}$, and consider the P_n's for $|z|<1$. Each P_n for $n \geq 2$ has a unique real zero ρ_n between $\frac{1}{2}$ and 1. Using the principle of the argument, it is easy to check [28] that this is the unique zero satisfying $|z| \leq \frac{3}{4}$. Furthermore, as n increases, these zeros tend to $\frac{1}{2}$ and are clearly all distinct.

Therefore, for any complex z, $|z|<\frac{3}{4}$, that is not equal to one of the ρ_n's or to $\frac{1}{2}$, the sum in (15) converges and defines an analytic function (observe the presence of the 'convergence factor' z^{2n}). On the other hand, each of the ρ_n's is a pole of $S(z)$.

We have thus shown that $S(z)$ is analytic in $|z|<\frac{3}{4}$ except for infinitely many poles ρ_n and their accumulation point $\frac{1}{2}$. Thus, $S(z)$ is transcendental and S is ambiguous.

We may mention here that, not too surprisingly, functions $S(z)$ and $R_n(z)$ are related to classical statistics on *runs* [15] and, accordingly, R_n occurs in an analysis by Knuth of carry propagation in some binary adders [28].

Language C. The language C has been introduced by Crestin [12] and Kemp [25] has shown that its generating function is

$$C(z) = 1 + 2 \sum_{m \geq 1} \psi(m) \frac{z^m(1+z^m)(1+2z^m)}{(1-2z^{2m})^2},$$

where $\psi(m) = \prod (1-p)$, the product being extended to all prime divisors of m. From that expression follows, as in the previous argument, that, for $|z| \leq 1$, $C(z)$ has isolated singularities (double poles) at points

$$z_{m,j} = 2^{-1/(2m)} e^{ij\pi/m}$$

that cancel the denominator of one of the terms composing $C(z)$. (Note: this observation answers a question of Kemp regarding the transcendence of function C.)

Language $G_{\neq}$. The Goldstine language can be characterised via its complement w.r.t. $\{a, b\}^*$ which consists of two types of words:

(A) words in $\{a, b\}^* a$ since they fail to have the formal $\underline{n}_1 \underline{n}_2 \ldots \underline{n}_p$;

(B) words of the form:

$$\varepsilon; \quad ab; \quad aba^2b; \quad aba^2ba^3b; \quad \ldots.$$

Thus the generating function $G(z)$ of $G_{\neq}$ is

$$G(z) = \frac{1}{1-2z} - A(z) - B(z)$$

with $A(z)$ and $B(z)$ being the generating functions for words of type (A) and (B):

$$A(z) = \frac{z}{1-2z}; \qquad B(z) = \sum_{n \geq 1} z^{n(n+1)/2-1}$$

so that

$$G(z) = \frac{1-z}{1-2z} - \frac{1}{z} \sum_{n \geq 1} z^{n(n+1)/2}.$$

From this last equation results that $G(z)$ has the same transcendence status as the series

$$\Theta(z) = \sum_{n \geq 1} z^{n(n+1)/2}.$$

Function Θ is an elliptic *theta* function; it is a lacunary series and, as such, admits the unit circle as a natural boundary. Thus it cannot be algebraic and the Goldstine language is inherently ambiguous.

Language $H_{\neq}$. The argument is almost the same as for the Goldstine language. Only, for words of type B, substitute the set

$$\varepsilon; \quad ab; \quad (a^2b)^2; \quad (a^3b)^3; \ldots; \quad (a^nb)^n; \ldots$$

with generating function

$$B(z) = \sum_{n \geq 0} z^{n(n+1)}$$

again a lacunary series.

Language $G_{<}$. As in the previous two examples, consider the language

$$B = (\{a, b\}^* / G_{<}) \cap \{a, b\}^* b.$$

This language admits the decomposition:

$$B = a^2a^*b + a^2a^*ba^3a^*b + a^2a^*ba^3a^*ba^4a^*b + \cdots$$

so that

$$B(a,b)=\sum_{k\geq 1} b^k \frac{a^2}{1-a}\frac{a^3}{1-a}\cdots\frac{a^{k+1}}{1-a}=\frac{1}{a}\sum_{k\geq 1}\left(\frac{b}{1-a}\right)^k a^{(k+1)(k+2)/2}$$

which is rationally expressible in terms of the Θ function and is again transcendental.

The case of the Goldstine language is the one that initially motivated our study because of the previously mentioned conjecture of [1] regarding its inherent ambiguity. The reader may consult [3] for several related enumeration issues. We observe that a similar argument based on lacunary series could have been used to treat the Fredholm series and hence the generating function of language P_2. Also, since the Fredholm series satisfies the functional equation $F(z)=z+F(z^2)$ and $F(1^-)=+\infty$, a direct argument might have been employed to establish that F has the unit circle as a natural boundary.

6. Local behaviour around singularities

Studying the local behaviour around singularities is certainly the most comfortable method to apply. The mere appearance of logarithmic terms in the local expansion of a function around a singularity is sufficient to establish its transcendence. Such local analyses may often be treated by Mellin transform techniques, a not too surprising fact considering the arithmetical character of many of the languages we study. We shall apply this method here to the following languages:

- languages K_1 and K_2 (Theorem 6);
- the 'comb-like' language P_1 (Theorem 4);
- the Goldstine-like language G_-.

Language K_1. As in the case of the Goldstine language, we enumerate K_1 by considering its complement. Define the language

$$D=(\{a,b\}^*/K_1)\cap\{a,b\}^*b.$$

It suffices to establish that the generating function of D is transcendental. But D has the simple form

$$D=\sum_{\substack{m\geq 1\\ n\geq 0}}(a^nb)^m$$

so that its generating function is

$$D(z)=\sum_{m,n\geq 1} z^{mn}=\sum_{p\geq 1} d(p)z^p,$$

where $d(p)$ is the *divisor* function counting the number of divisors of integer p.

We propose here to establish the transcendence of $D(z)$ by showing that, as $z\to 1^-$,

$$D(z)\sim(1-z)^{-1}\log(1-z)^{-1}, \tag{16}$$

a typically transcendental behaviour. To do so, one can consider the function $\Delta(t) = D(e^{-t})$ and determine its asymptotic behaviour as $t \to 0^+$.

The Mellin transform (see, e.g., [8, 13] and, for uses in analysis of algorithms, [17]) of the function Δ is, by definition, the function Δ^* given by

$$\Delta^*(s) = \int_0^\infty \Delta(t) t^{s-1} \, dt, \tag{17}$$

which, for $\mathrm{Re}(s) > 1$, is equal to:

$$\Delta^*(s) = \sum_{n \geq 1} \frac{d(n)}{n^2} \Gamma(s) = \zeta^2(s) \Gamma(s). \tag{18}$$

($\zeta(s)$ is the Riemann *zeta* function and $\Gamma(s)$ is the Euler *gamma* function.) From the general inversion theorem for Mellin transforms

$$\Delta(t) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} \Delta^*(s) t^{-s} \, ds \tag{19}$$

which can here be taken with $c = 2$, calculating the residue of the integrand of (19) at $s = 1$ and shifting the line of integration to $\mathrm{Re}(s) = \frac{1}{2}$, one finds

$$\Delta(t) = +\frac{1}{t} \log \frac{1}{t} + O\left(\frac{1}{\sqrt{t}}\right). \tag{20}$$

This last equation entails (16). Thus, $D(z)$ is a transcendental function and language K_1 is ambiguous.

Language K_2. The argument is quite similar. Consider now the language E formed with the complement of K_2:

$$E = (\{a, b\}^* / K_2) \cap (a^+ b)^2 (a^+ b)^*.$$

Its bivariate generating function is

$$\begin{aligned} E(a, b) &= \sum_{m,k \geq 1} a^m b a^{2m} b a^{2^2 m} b \ldots a^{2^k m} b \\ &= \sum_{k,m \geq 1} a^{m(2^{k+1}-1)} b^{k+1}. \end{aligned} \tag{21}$$

Function $E(z, 1)$ thus has the expression

$$E(z, 1) = \sum_{n \geq 1} d_2(n) z^n,$$

where $d_2(n)$ is the number of divisors of n of the form $2^k - 1$ with $k \geq 2$. To prove that $E(z, 1)$ is transcendental, one may again determine its asymptotic behaviour at $z = 1^-$. The Mellin transform of $E(e^{-t}, 1)$ is

$$\zeta(s) \Gamma(s) \sum_{k \geq 2} \frac{1}{(2^k - 1)^s}.$$

To the left of the line $\mathrm{Re}(s)=-1$ the Dirichlet series $\omega(s)\equiv\sum_{k\geq 2}(2^k-1)^{-s}$ has simple poles at all points of the form $(2ik\pi)/\log 2$ with $k\in\mathbf{Z}$. Thus, around $t=0$, $E(\mathrm{e}^{-t},1)$ has an expansion of the form

$$E(\mathrm{e}^{-t},1)=\frac{1}{t}\sum_{k\geq 2}\frac{1}{2^k-1}+\lambda\log t+\sum_{k\in\mathbf{Z}}c_k t^{2ik\pi/\log 2}+\mathrm{O}(\sqrt{t}), \tag{22}$$

a typically transcendental expansion due to both the logarithmic terms and the imaginary exponents. Thus K_2 is also ambiguous.

Language P_1. A combinatorial decomposition like that used for language P_2 reduces the problem to proving that the generating function of language I, here defined by

$$I=\{(a^m b)^{2k}\,|\,k\geq 1,\ m\geq 0\},$$

is transcendental. But this function is directly related to the divisor function since

$$I(z,z)=\sum_{k,m\geq 1}z^{2km}=\sum_{n\geq 1}d(n)z^{2n},$$

so that $I(z,z)$ is transcendental by the argument given for language K_1.

Language $G_=$. We shall prove this language to be ambiguous by showing essentially that around a singularity a derived function behaves like

$$\frac{1}{\mathrm{e}}\frac{1}{1-z}$$

with e the transcendental number $\mathrm{e}=2.71828\ldots$. A somewhat related reduction (though concluding with a density argument instead) will be used in the next section when dealing with language $G_>$.

Words of the format $\underline{n}_1\ldots\underline{n}_p$ $(p\geq 0)$ that are *not* in $G_=$ are described by

$$B=\sum_{k\geq 1}(a^*/a)b(a^*/a^2)b(a^*/a^3)b\ldots(a^*/a^k)b$$

so that their bivariate generating function reads

$$B(a,b)=\sum_{k\geq 0}\frac{b^k}{(1-a)^k}(1-a(1-a))(1-a^2(1-a))\times(1-a^3(1-a))\cdot\ldots\cdot(1-a^k(1-a))$$

and

$$B(a,z(1-a))=\sum_{k\geq 0}z^k(1-a(1-a))(1-a^2(1-a))\times(1-a^3(1-a))\cdot\ldots\cdot(1-a^k(1-a)).$$

Thus, as $z\to 1^-$ for fixed a, $|a|<1$,

$$B(a,z(1-a))\sim\frac{Q(a)}{(1-z)},$$

where

$$Q(a) = \prod_{j \geqslant 1} (1 - a^j(1-a)).$$

Now, if $B(a, b)$ were algebraic, $Q(a)$ would be an algebraic function. But, by a classical identity of Euler (see [11, p. 103]),

$$\prod_{j \geqslant 1} \frac{1}{1-ua^j} = 1 + \sum_{m \geqslant 1} \frac{u^m a^m}{(1-a)(1-a^2)(1-a^3)\cdot \ldots \cdot (1-a^m)}.$$

Therefore, function $Q(a)$ has the alternative form

$$\frac{1}{Q(a)} = 1 + \sum_{m \geqslant 1} \frac{a^m}{(1+a)(1+a+a^2)\cdot \ldots \cdot (1+a+a^2+\cdots+a^{m-1})}$$

so that

$$\lim_{a \to 1^-} Q(a) = \sum_{m \geqslant 0} \frac{1}{m!} \equiv \frac{1}{e}.$$

Thus $Q(a)$ is transcendental and so is $B(a, b)$. Language $G_=$ is ambiguous.

We observe that we could alternatively have used the Lambert series expansions

$$\sum_{n \geqslant 1} d(n) z^n = \sum_{m \geqslant 1} \frac{z^m}{1-z^m}; \qquad \sum_{n \geqslant 1} d_2(n) z^n = \sum_{m \geqslant 2} \frac{z^{2^m-1}}{1-z^{2^m-1}}$$

to establish that these functions have the unit circle as a natural boundary.

Also, Mellin transform techniques when applied to the Fredholm series reveal the presence of a logarithmic term together with periodic fluctuations similar to those of Eq. (22).

7. Generalised asymptotic densities

The argument here is based on the existence of generalised densities for coefficients of algebraic functions given by expansion (Δ) of Theorem D. Here it is applied to:
- languages O_3, Ω_3 defined by occurrence constraints (Theorem 1);
- the Goldstine-like language $G_>$ (Theorem 5).

Languages Ω_3 and O_3. Language Ω_3 has a complement which is

$$I = \{a, b, c\}^* / \Omega_3 = \{w \,|\, |w|_a = |w|_b = |w|_c\}.$$

Thus, the number of words in I of length $3n$ is given by the multinomial coefficient

$$\binom{3n}{n, n, n} = \frac{(3n)!}{(n!)^3},$$

so that

$$\Omega_3(z)=\frac{1}{1-3z}-I(z)=\frac{1}{1-3z}-\sum_{n\geq 0}\frac{(3n)!}{(n!)^3}z^n.$$

Function $I(z)$ is transcendental since, by Stirling's formula,

$$I_{3n}\sim 3^{3n}\frac{\sqrt{3}}{2\pi n}$$

and, because of the n^{-1} factor, this expansion fails to be of type (Δ).

Similarly, O_3 is the union of two deterministic languages (see the treatment of O_4), whose intersection is exactly the language I defined above. Thus, O_3 is transcendental.

Language $G_>$. Once more, we prove it to be ambiguous by showing that its complement has a transcendental generating function. Therefore, we consider the language $B=\{a, b\}^*/\{a, b\}^*b$. It is formed with words of the type

$$a^0b;\quad a^0b(a^0+a^1)b;\quad a^0b(a^0+a^1)b(a^0+a^1+a^2)b;\ldots$$

so that its bivariate generating function is

$$B(a, b)=\sum_{k\geq 1} b^k\frac{1-a}{1-a}\frac{1-a^2}{1-a}\frac{1-a^3}{1-a}\cdots\frac{1-a^k}{1-a}$$

from which we get

$$B(a, z(1-a))=\sum_{k\geq 1} z^k(1-a)(1-a^2)\cdot\ldots\cdot(1-a^k).$$

That function is a *basic hypergeometric function.* For $|a|<1$, $B(a, z(1-a))$ has a simple pole at $z=1$ and one has

$$B(a, z(1-a))\sim\frac{1}{1-z}\prod_{k\geq 1}(1-a^k). \tag{23}$$

Assume a contrario that $B(a, z(1-a))$ were an algebraic function; then, so would be $(1-z)B(a, z(1-a))$ together with its value at $z=1$, namely

$$Q(a)=\prod_{k\geq 1}(1-a^k). \tag{24}$$

We may now resort to a density argument. Indeed, by a celebrated theorem of Hardy and Ramanujan [20] concerning the number p_n of partitions of integer n, we have

$$p_n\equiv[x^n]\frac{1}{Q(x)}\sim\frac{e^{\pi\sqrt{2n}/3}}{4n\sqrt{3}}. \tag{25}$$

Thus, $Q(x)^{-1}$, hence also $B(a, b)$, is transcendental, and $G_>$ is inherently ambiguous.

Observations: For $G_>$, many routes are conceivable to establish the (clear) transcendence of $Q(z)$ defined by (23). One may directly observe from the infinite product expansion that $Q(z)$ has the unit circle as a natural boundary. Also, by the Euler Pentagonal Number Theorem, $Q(z)$ is a lacunary series since

$$\prod_{k\geq 1}(1-z^k)=\sum_{k\in Z}(-1)^k z^{k(3k+1)/2}.$$

Conversely, density arguments could have been used for other languages. For instance, for language O_4 studied in Section 4 (see Eq. (8)) we have

$$I_{2n}\sim\frac{16^n}{\pi n}.$$

Similarly for languages K_1 and P_1, the *mean order* results (cf. [8, 17])

$$\frac{1}{n}\sum_{k=1}^{n} d(k)\sim\log n; \qquad \frac{1}{n}\sum_{k=1}^{n} d_2(k)\sim\log_2 n$$

are evidence of the transcendental character of the generating functions of $d(n)$ and $d_2(n)$.

8. Polynomial-linear recurrences

Recall that Criterion E based on Comtet's theorem states that if no linear recurrence with polynomial coefficients exists between terms of a sequence l_n, then that sequence cannot be the sequence of coefficients of an algebraic function. It comes as a useful complement (or as an alternative) to transcendence proofs based on lacunary series mentioned in relation to Criterion B. We shall apply it here to

- language B based on binary representations of integers (Theorem 7).

Language B. The language $C=\{0, 1, c\}^*/B$ is formed with words that are the prefixes ending with a letter c of the *infinite word*:

$$\boldsymbol{b}=1c\ 10c\ 11c\ 100c\ 101c\ 110c\ 111c\ 1000\ c\ldots.$$

Let $\lambda(k)$ denote the rank of the kth c in $\boldsymbol{b}$. We have

$$\lambda(1)=2;\quad \lambda(2)=5;\quad \lambda(3)=8;\quad \lambda(4)=12;\ \ldots$$

and, in general,

$$\lambda(k)=k+1+\sum_{j=1}^{k}[\log_2 k]$$

with $[x]$ representing the ceiling function of real x: $[x]-1<x\leq[x]$.

Therefore, $C_n=1$ if n is of the form $\lambda(k)$ for some k and $C_n=0$ otherwise. Assume a contrario the existence of a polynomial linear recurrence:

$$C_N=\sum_{j=1}^{d} p_j(n)C_{N-j}. \tag{26}$$

One has, for all k, $\lambda(p)-\lambda(p-1)=[\log_2 p]$; thus, taking in (26) $N=\lambda(p)$ (with for instance $p=2^{d+1}$) one reaches a contradiction since all terms on the r.h.s. of (26) are 0 while $C_N=1$.

Of course, all other languages where lacunary series intervene could have been treated by means of Criterion E which, for our purposes, is in principle more powerful than the lacunary series theorem. Conversely, language B could have been dealt with by using that theorem since the generating function of language C is a lacunary series. At present, we do not have examples of applications of Comtet's theorem that are not also lacunary series.

9. Conclusions and open problems

(1) The first conclusion of this work is that analytic methods are well suited to proving inherent ambiguity of a variety of context-free languages since a transcendental element in a generating function can be almost invariably recognised 'at sight' using the classical arsenal of complex analysis.

Our methods seem well-suited to languages of intermediate structural complexity in the following sense: the languages have to be simple enough so that we can solve their counting problems with the available technologies of combinatorial analysis; they have to be not too simple since otherwise their generating functions could become algebraic or even rational and the method then ceases to be applicable.

At the lower end of the spectrum, we find the languages L and L' defined in Equations (1), (2), which have rational generating functions:

$$L(z)=\frac{2}{(1-z)(1-z^2)}-\frac{1}{1-z^3}; \qquad L'(z)=\frac{1}{1-z}-\frac{1}{1-z^3}.$$

At the upper end of the spectrum, there probably lie languages like the 'hardest context-free language' of Greibach or even Shamir's language (compare with language S in our Theorem 3):

$$S'=\{ucv_1uv_2 \mid u, v_1, v_2 \in \{a, b\}^*\}$$

whose counting problem is equivalent to the general enumeration of occurrences of patterns in strings [19].

Since many of the languages considered here appear to be of *unbounded ambiguity* [12, 38] a natural question is whether our methods can be extended to cover the following situation.

Question 1. Are there sufficient conditions on generating functions to ensure that a language is *infinitely* inherently ambiguous?

We believe the answer to this question is yes.

(2) The second conclusion is that there is a fairly rich analytic structure amongst generating functions of ambiguous context-free languages. Many of these functions are related to classical *special functions*, a fact perhaps not too surprising since the language definitions are often closely related to integer partitions and compositions. Thus we have the following problem.

Question 2. In which class of transcendental functions do generating functions of (general) context-free languages lie? (For instance, in our work we came nowhere close to expressions involving the exponential function.)

A closely related problem is the following.

Question 3. Are there general results on densities of (ambiguous) context-free languages? For instance, can the number of words of size n in a context-free language grow like $\exp(c\sqrt{n})$?

In relation to Question 2, it has been proved by Bertoni and Sabadini [7] that it is undecidable whether a context-free language has an algebraic generating function. In another direction, Kuich and Shyamasundar [31] have obtained characterisations of generating functions associated to (usually non-context-free) languages produced by some Lindenmeyer systems.

In relation to Question 3, Baron and Kuich [2] as well as Ibarra and Ravikumar [24] have shown that it is decidable whether a context-free language is 'sparse', meaning that its enumeration sequence grows no faster than a polynomial. Recently, Kornai [29] has employed analytic techniques to study a related notion of density for some special context-free languages. Counting results for particular languages are also given by Beauquier and Thimonier [3].

(3) Finally, it would be very interesting to have ways of establishing the inherent ambiguity of languages like (1):

$$\{a^m b^n c^p \mid n = m \text{ or } n = p\}$$

using analytic methods. This would probably require the construction of quite different analytic models that should be of interest since they would better capture inherently noncommutative properties of formal languages.

Acknowledgment

This work was started after stimulating discussions with Lois Thimonier and J. Beauquier. Thanks also to J. Gabarro for posing the conjecture relative to language *B*.

References

[1] J.-M. Autebert, J. Beauquier, L. Boasson and M. Nivat, Quelques problèmes ouverts en théorie des langages algébriques, *RAIRO Inform. Théor.* **13** (1979) 363–379.

[2] G. Baron and W. Kuich, The characterization of nonexpansive grammars by rational power series, *Inform. and Control* **48** (1981) 109–118.

[3] J. Beauquier and L. Thimonier, Formal languages and Bernoulli processes, LITP Rept. 83-30, Univ. Paris VII (1983).

[4] M. Ben-Or, Lower bounds for algebraic computation trees, *Proc. 15th ACM Symp. on Theory of Computing* (1983) 80–86.

[5] J. Berstel, Sur la densité asymptotique des langages formels, *Proc. 1st ICALP Colloquium* (North-Holland, Amsterdam, 1972) 345–368.

[6] J. Berstel, Contribution à l'étude des propriétés arithmétiques des langages formels, Thèse, Univ. Paris VII (1972).

[7] A. Bertoni and N. Sabadini, Algebricity of the generating function for context-free languages, Manuscript (1985).

[8] K. Chandrasekharan, *Arithmetical Functions* (Springer, Berlin, 1970).

[9] N. Chomsky and M.-P. Schutzenberger, The algebraic theory of context-free languages, *Computer Programming and Formal Systems* (North-Holland, Amsterdam, 1963) 118–161.

[10] L. Comtet, Calcul pratique des coefficients de Taylor d'une fonction algébrique, *Enseignement Math.* **10** (1964) 267–270.

[11] L. Comtet, *Advanced Combinatorics* (Reidel, Dordrecht, 1974).

[12] J.P. Crestin, Un langage non ambigu dont le carré est d'ambiguité inherente bornée, *Proc. 1st ICALP Colloquium* (North-Holland, Amsterdam, 1972) 377–390.

[13] B. Davies, *Integral Transforms and Their Applications* (Springer, New York, 1978).

[14] J. Dieudonné, *Calcul Infinitésimal* (Hermann, Paris, 1968).

[15] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1950).

[16] P. Flajolet and A. Odlyzko, The expected height of binary trees and other simple trees, *J. Comput. System Sci.* **25** (1982) 171–213.

[17] P. Flajolet, M. Regnier and R. Sedgewick, Some uses of the Mellin integral transform in the analysis of algorithms, in: *Combinatorics on Words* (Springer, Berlin, 1985) 241–254.

[18] A.O. Gelfond, *Transcendental and Algebraic Numbers* (Dover, New York, 1960).

[19] L. Guibas and A. Odlyzko, Strings overlap, pattern-matching and non-transitive games, *J. Comb. Theory Ser. A* **30** (1980) 183–208.

[20] G.H. Hardy, *Ramanujan, Twelve Lectures Suggested by his Life and Work* (Cambridge, University Press, London, 1940).

[21] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).

[22] P. Henrici, *Applied and Computational Complex Analysis* (Wiley, New York, 1977).

[23] T. Hickey and J. Cohen, Uniform random generation of strings in a context-free language, *SIAM J. Comput.* **12** (1983) 645–655.

[24] O. Ibarra and B. Ravikumar, On sparseness, ambiguity and other decision problems for acceptors and transducers, in: *Proc. STACS'86*, Lecture Notes in Computer Science **210** (Springer, Berlin, 1986). 171–179.

[25] R. Kemp, On the number of words in the language $\{w \in \Sigma^* \mid w = w^R\}^2$, *Discrete Math.* **40** (1980) 225–234.

[26] R. Kemp, A note on the density of inherently ambiguous context-free languages, *Acta Inform.* **14** (1980) 295–298.

[27] K. Kendig, *Elementary Algebraic Geometry* (Springer, New York, 1977).

[28] D.E. Knuth, The average time for carry propagation, *Nederl. Akad. Wetensch. Indag. Math.* **40** (1978) 238–242.

[29] A. Kornai, Quantitative comparison of formal languages, Manuscript, submitted (1985).

[30] W. Kuich and A. Salomaa, *Semirings, Automata, Languages* (Springer, New York, 1986).

[31] W. Kuich and R.K. Shyamasundar, The structure generating function of some families of languages, *Inform. and Control* **32** (1976) 85–92.

[32] A. Meir and J. Moon, On the altitude of nodes in random trees, *Canad. J. Math.* **30** (1978) 997–1015.
[33] N. Pippenger, Computational complexity in algebraic function fields, *Proc. 20th IEEE Symp. FOCS* (1979) 61–65.
[34] W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1974).
[35] A. Salomaa and M. Soittola, *Automata Theoretic Aspects of Formal Power Series* (Springer, New York, 1978).
[36] Th. Schneider, *Einfuehrung in die Transzendenten Zahlen* (Springer, Berlin, 1957).
[37] A. Seidenberg, *Elements of the Theory of Algebraic Curves* (Addison-Wesley, Reading, MA, 1965).
[38] E. Shamir, Some inherently ambiguous context-free languages, *Inform. and Control* **18** (1971) 355–363.
[39] M.I. Shamos and G. Yuval, Lower bounds from complex function theory, *Proc. 17th IEEE Symp. FOCS* (1976) 268–273.
[40] R. Stanley, Differentiably finite power series, *European J. Combin.* **1** (1980) 175–188.
[41] E.T. Whittaker and G.N. Watson, *A Course in Modern Analysis* (Cambridge University Press, London, 1927; 4th edition).