ADDITIVE FUNCTIONS WITH RESPECT TO NUMERATION SYSTEMS ON REGULAR LANGUAGES

PETER J. GRABNER† AND MICHEL RIGO‡

Abstract. Asymptotic formulæ for the summatory function of additive arithmetic functions related to numeration systems given by regular languages are derived.

1. Introduction

Additive numeration systems and the corresponding additive arithmetic functions have been studied from various points of view since the seminal papers of H. Delange [4, 5], where such functions were investigated for the usual q-adic numeration system. Later more exotic systems of numeration, such as general linear numeration systems [13, 14], especially such systems defined by linear recurring sequences were considered. Furthermore, digital representations with respect to substitutions over a finite alphabet were studied (cf. [8, 9]). All these numeration systems have in common that the number of integers represented by words of length n satisfies a pure exponential law $\sim C^n$ for some constant $C > 1$. Different aspects of such representations of the integers were studied: dynamics of corresponding adding machine ("odometer") [15], topological dynamics of the odometer [1], asymptotic properties of summatory functions of additive functions such as the "sum-of-digits" function [8, 9, 12, 16, 17], local and global versions of central limit theorems for the values of additive functions [6, 7, 10], existence of distribution functions of additive functions [2].

In the present paper we take the opposite approach compared to the existing literature on the subject. We start with a regular language $\mathcal{L}$, order its words by the genealogical ordering induced by an ordered alphabet, and assign the number n the $(n+1)$-st word in the language. The above mentioned expansions, which come from finite linear recurrences are special cases of this setting. For the question of recognizability of the language generated by an increasing sequence of integers we refer to [13, 20, 21].

The paper is organized as follows. In Section 2 we introduce the basic notation of numeration systems related to regular languages; for this purpose we summarize the contents of the paper [18]. In Section 3 we state the main theorem, which is the appropriate analogue of the summation formula for the sum-of-digits function discovered by H. Delange

Date: October 5, 2007.

2000 *Mathematics Subject Classification.* Primary: 11A67; Secondary: 68Q45, 11B85, 11K55.

Key words and phrases. regular language; additive function; numeration system.

† This author is supported by the START project Y96-MAT of the Austrian Science Fund.

‡ This author warmly thanks P. Grabner for his kind invitation to Graz and his hospitality.

[5]. Furthermore, we state two corollaries which simplify the asymptotic expression given in the theorem for special cases. Section 4 is devoted to the proofs of the theorem and its corollaries. The final Section 5 presents a number of examples which show that the results given in Section 3 cannot be improved in general.

2. Preliminaries

Let Σ be a finite alphabet. The free monoid generated by Σ with identity ε is Σ^*. If w is a word over Σ, $|w|$ denotes its length. We assume that the reader is familiar with classical notions of formal languages theory like (minimal) automaton, regular language or transducer (see for instance [11]).

For a totally ordered alphabet $(\Sigma, <)$, a word w is *genealogically less* than w' if $|w| < |w'|$ or if $|w| = |w'|$ and there exist letters $\sigma < \tau$ such that $w = x\sigma y$ and $w' = x\tau y'$, $x, y, y' \in \Sigma^*$. In the literature, the terminology *radix order* is also used.

Describing an infinite regular language $\mathcal{L}$ over a totally ordered alphabet $(\Sigma, <)$ with respect to the genealogical ordering gives a one-to-one and onto increasing mapping between $\mathbb{N}$ and $\mathcal{L}$. If w is the n-th word of the genealogically ordered language $\mathcal{L}$, $n \in \mathbb{N} \setminus \{0\}$, then we denote by $\mathrm{val} : \mathcal{L} \to \mathbb{N}$ the application mapping w onto $n - 1$. The integer $\mathrm{val}(w)$ is said to be the *numerical value* of w. So each non-negative integer n is represented by a unique word $\mathrm{val}^{-1}(n) \in \mathcal{L}$ and this leads to the notion of *numeration system on a regular language*. These systems have been introduced in [18] and generalize classical numeration systems like the q-adic systems, the Fibonacci system and the linear numeration systems whose characteristic polynomial is the minimal polynomial of a Pisot number (for the properties of these latter systems we refer to [3]). As an example, consider the alphabet $\Sigma = \{0, 1\}$ where we assume that $0 < 1$. If we consider the regular language $\mathcal{L} = \varepsilon \cup 1\Sigma^* \setminus \Sigma^*11\Sigma^* = \varepsilon \cup 1 \cup 10\{10, 0\}^*$ then the numeration system on $\mathcal{L}$ is exactly the Fibonacci numeration system (cf. [23]).

In this paper, $\mathcal{L}$ always refers to an infinite regular language having $\mathcal{M}_{\mathcal{L}} = (Q, \Sigma, s, \delta, F)$ as trimmed minimal automaton where Q is the set of states, $s \in Q$ is the initial state, $\delta : Q \times \Sigma \to Q$ is the transition function and $F \subset Q$ is the set of terminal states (to obtain unambiguous constructions, we only consider minimal automata; in order to relate the size of the language with the eigenvalues of the incidence matrix, we assume the automaton to be trimmed) and we represent integers using the numeration system built upon $\mathcal{L}$ for a given ordering of the alphabet. We extend δ to $Q \times \Sigma^*$ by $\delta(q, \sigma_1 \cdots \sigma_k) = \delta(\delta(q, \sigma_1), \sigma_2 \cdots \sigma_k)$ and write $q.w$ as a shorthand for $\delta(q, w)$, $q \in Q$, $w \in \Sigma^*$.

For each state $q \in Q$, we define the language

$$\mathcal{L}_q = \{w \in \Sigma^* \mid \delta(q, w) \in F\}$$

of the words accepted by $\mathcal{M}_{\mathcal{L}}$ with initial state q. In particular, $\mathcal{L} = \mathcal{L}_s$. For each state $q \in Q$, we define two functions $u_q(n)$ and $v_q(n)$ counting the number of words in $\mathcal{L}_q$ respectively of length n and of length less or equal to n,

$$u_q(n) = \#(\mathcal{L}_q \cap \Sigma^n) \quad \text{and} \quad v_q(n) = \#(\mathcal{L}_q \cap \Sigma^{\leq n}).$$

If $Q = \{q_1 = s, q_2, \ldots, q_k\}$ then we denote by $\mathbf{u}(n)$ the k-tuple $(u_{q_1}(n), \ldots, u_{q_k}(n))$. Let us recall that for a regular language, $u_s(n)$ is either $\Theta(n^k)$ for some integer $k \geq 0$ or an exponential function of the order $2^{\Omega(n)}$. For a characterization of polynomial languages we refer to [22].

The incidence matrix of the minimal automaton defining the language $\mathcal{L}$ is given by $A_{p,q} = \#\{\sigma \in \Sigma \mid p.\sigma = q\}$ for $p, q \in Q$.

From now on we will assume that $\mathcal{L}$ is an exponential language, whose trimmed minimal automaton has one dominating eigenvalue $\lambda > 1$. Thus we can write for each $q \in Q$: $u_q(n) = P_q(n)\lambda^n + o(\lambda^n)$ for (possibly zero) polynomials P_q, and assume that P_s (which is non-zero) has degree $d \geq 0$. For each $q \in Q$, there exists a word w of length bounded by $\#Q$ such that $s.w = q$. So $u_s(n + l) \geq u_q(n)$ for all $n \geq 0$ and consequently the degree of P_s is the highest degree among the $\deg P_q$, $q \in Q$. We define the sets

$$\begin{aligned} Q_1 &= \{q \in Q \mid u_q(n) = P_q(n)\lambda^n + o(\lambda^n), \quad \deg P_q = \deg P_s\} \\ Q_2 &= \{q \in Q \mid u_q(n) = P_q(n)\lambda^n + o(\lambda^n), \quad \deg P_q = \deg P_s - 1\} \\ Q_3 &= Q \setminus (Q_1 \cup Q_2). \end{aligned}$$

Since $\mathcal{L}_q$ is also regular, it can be genealogically ordered and we obtain a new numeration system. The function mapping the n-th word of $\mathcal{L}_q$ onto $n-1$ is denoted $\mathrm{val}_q : \mathcal{L}_q \to \mathbb{N}$ (in particular, $\mathrm{val} = \mathrm{val}_s$). If $\mathcal{L}_q$ is finite then the domain of val_q is also finite and its image is restricted to $\{0, \ldots, \#\mathcal{L}_q - 1\}$.

Using the definition of the genealogical ordering, a formula for computing numerical values was derived in [18]. Let $w = \sigma_1 \cdots \sigma_n \in \mathcal{L}$, then we have

$$\mathrm{val}(w) = \sum_{q \in Q} \sum_{i=1}^{|w|} \beta_{q,i}(w) u_q(|w| - i) \tag{2.1}$$

where

$$\beta_{q,i}(w) = \#\{\sigma < \sigma_i \mid s.\sigma_1 \cdots \sigma_{i-1}\sigma = q\} + \delta_{q,s} \quad \textit{for } i = 1, \ldots, |w| \tag{2.2}$$

Observe that these coefficients are bounded :

$$0 \leq \sum_{q \in Q} \beta_{q,i}(w) \leq \#\Sigma.$$

Formula (2.1) can be proved by observing that the summand for (q, i) for $q \neq s$ is the number of words v of length $|w|$ which have prefix $\sigma_1 \ldots \sigma_{i-1}\sigma$ with $\sigma < \sigma_i$ (which means that $v < w$), the state q is reached after reading the first i letters of v, and the postfix $v_{i+1} \ldots$ is accepted by the automaton with initial state q. For $q = s$ the summand for (s, i) equals the number with the same descriptions as above plus the number of words of length $|w| - i$ which are accepted by the automaton starting from s. Summing over all possible pairs (q, i) first gives the number of words $v < w$ with $|v| = |w|$, the extra summand for $q = s$ equals the number of words v with $|v| < |w|$. Altogether this equals $\mathrm{val}(w)$.

Remark 1. For a given $q \in Q$, the coefficients $\beta_{q,i}(w)$ can be computed by a transducer built upon $\mathcal{M}_{\mathcal{L}}$. For all $p, r \in Q$ and $\sigma \in \Sigma$ such that $p.\sigma = r$, if we replace in $\mathcal{M}_{\mathcal{L}}$ the edge $p \overset{\sigma}{\to} r$ with $p \overset{\sigma|\beta}{\to} r$ where

$$\beta = \#\{\tau < \sigma \mid p.\tau = q\} + \delta_{q,s}$$

then we obtain a transducer that reading w produces the output $\beta_{q,1}(w) \cdots \beta_{q,n}(w)$. Consequently, if two words x and y in $\mathcal{L}$ have the same prefix of length $k \geq 1$ then for all states $q \in Q$

$$\beta_{q,i}(x) = \beta_{q,i}(y) \quad \text{for } i = 1, \ldots, k.$$

In [19] numeration systems on regular languages are extended to the representation of real numbers. (These numbers are therefore represented by right-infinite words instead of finite words). With those considerations, a particular set $\mathcal{L}_\infty$ of infinite words is defined, which is used for the representation of real numbers in the interval $[\frac{1}{\lambda}, 1]$. A word $\omega \in \Sigma^\omega$ belongs to $\mathcal{L}_\infty$ if there exists a sequence of words in $\mathcal{L}$ converging to ω. If we denote by $x[\ell]$ the prefix of length ℓ of the (finite or infinite) word x, then we have by definition of the set $\mathcal{L}_\infty$ that a word ω belongs to $\mathcal{L}_\infty$ if and only if

$$\exists (x_n)_{n\in\mathbb{N}} \in \mathcal{L}^{\mathbb{N}}, \forall \ell > 0, \exists N_\ell \in \mathbb{N} : \forall n > N_\ell, x_n[\ell] = \omega[\ell].$$

The set $\mathcal{L}_\infty$ is equipped with the topology induced by the infinite product topology on $\Sigma^{\mathbb{N}}$. As a consequence of Remark 1, to any word $\omega \in \mathcal{L}_\infty$ there corresponds a unique infinite sequence of coefficients $(\beta_{q,n}(\omega))_{n\in\mathbb{N}}$ for each $q \in Q$ because ω has arbitrarily long common prefixes with words in $\mathcal{L}$. Actually, the sequence $(\beta_{q,n}(\omega))_{n\in\mathbb{N}}$ is obtained by feeding the transducer built on $\mathcal{M}_{\mathcal{L}}$ with the infinite word ω. In [19], it is shown that for any sequence of words V_k converging to a word $\omega \in \mathcal{L}_\infty$, the limit

$$\lim_{k\to\infty} \frac{\operatorname{val}(V_k)}{v_s(|V_k|)}$$

only depends on ω and we denote its value by $\operatorname{val}_\infty(\omega)$.

For one of our theorems we will need the following technical hypothesis.

Hypothesis 1. Let V and $\tilde{V}$ be two prefixes of words in $\mathcal{L}$. Then there exist two words η and $\tilde{\eta}$ such that $V\eta, \tilde{V}\tilde{\eta} \in \mathcal{L}$ and $|V\eta| = |\tilde{V}\tilde{\eta}|$.

This is equivalent to the following more technical hypothesis, which we will actually use.

Hypothesis 2. There exist constants $C, D \geq 0$ such that for any two prefixes V and $\tilde{V}$ of words in $\mathcal{L}$ with $||V| - |\tilde{V}|| \leq C$ there are words η and $\tilde{\eta}$ such that $|\eta|, |\tilde{\eta}| \leq D$, $V\eta, \tilde{V}\tilde{\eta} \in \mathcal{L}$, and $|V\eta| = |\tilde{V}\tilde{\eta}|$.

Proof of equivalence of Hypotheses 1 and 2. Assume that $\mathcal{L}$ satisfies Hypothesis 1 and consider two words $V, \tilde{V} \in \operatorname{pref}(\mathcal{L})$. By Hypothesis 1 there are two words $\eta, \tilde{\eta}$ with $V\eta, \tilde{V}\tilde{\eta} \in \mathcal{L}$ and $|V\eta| = |\tilde{V}\tilde{\eta}|$. Let us define the automaton $\mathcal{B} = (Q \times Q, \Sigma \times \Sigma, (s,s), \delta_{\mathcal{B}}, F \times F)$ where the transition function $\delta_{\mathcal{B}}$ is defined by

$$\delta_{\mathcal{B}}((q,q'),(\sigma,\sigma')) = (\delta(q,\sigma), \delta(q',\sigma'))$$

where δ is the transition function of $\mathcal{M}_{\mathcal{L}}$. Feeding $(V\eta, \tilde{V}\tilde{\eta})$ into the automaton $\mathcal{B}$, after $\max(|V|, |\tilde{V}|)$ steps we reach some state. From this state we can reach a terminal state after at most $(\#Q)^2$ steps. Thus we can choose a constant $D = C + (\#Q)^2$.

Now assume that $\mathcal{L}$ satisfies Hypothesis 2. We use induction on k with $(k-1)C < |\tilde{V}| - |V| \leq kC$. Let $\tilde{W}$ be the prefix of $\tilde{V}$ of length $|V| + (k-1)C$. Then an application of the induction hypothesis to $\tilde{W}$ and V gives the existence of words η and η' such that $|\tilde{W}\eta'| = |V\eta|$. If $|\tilde{W}\eta'| \geq |\tilde{V}|$ we apply Hypothesis 2 to the prefix of $V\eta$ of length $|\tilde{V}|$ and $\tilde{V}$; if $|\tilde{W}\eta'| < |\tilde{V}|$ we apply Hypothesis 2 to $V\eta$ and $\tilde{V}$. $\square$

3. An asymptotic formula for the summatory function of additive functions

Let us consider a regular language $\mathcal{L}$ of exponential growth and its corresponding sequence $v_s(n)$. We define the strictly increasing continuous function $g : \mathbb{R}^+ \to \mathbb{R}^+$ by

$$g(n+x) = v_s(n)^{1-x} v_s(n+1)^x \quad \text{for } 0 \leq x \leq 1 \quad \text{and } n \in \mathbb{N}. \tag{3.1}$$

This function has the property $g(n) = v_s(n)$ for all $n \in \mathbb{N}$. We define the function $h : \mathbb{R}^+ \to \mathbb{R}^+$ as the inverse function of g. By (3.1) we also have

$$h(y) = n + \frac{\log y - \log v_s(n)}{\log v_s(n+1) - \log v_s(n)} \quad \text{for } v_s(n) \leq y \leq v_s(n+1), \quad n \in \mathbb{N}. \tag{3.2}$$

Then we have $|W| = \lfloor h(\mathrm{val}(W)) \rfloor + 1$, because a word $w \in \mathcal{L}$ has length n, if and only if $v_s(n-1) \leq \mathrm{val}(w) < v_s(n)$.

Theorem 1. *Let $\mathcal{L}$ be a regular language of exponential growth on the alphabet Σ and $f : \Sigma \to \mathbb{R}$ a function. We assume further that the incidence matrix of the automaton $\mathcal{M}_{\mathcal{L}}$ has a unique dominating eigenvalue. For a word $w = \sigma_1 \cdots \sigma_k \in \mathcal{L}$ we define $f(w) = \sum_{\ell=1}^{k} f(\sigma_\ell)$. There exist bounded functions $F, G : \mathcal{L} \to \mathbb{R}$ such that*

$$\sum_{\substack{w<W \\ w\in\mathcal{L}}} f(w) = \mathrm{val}(W)|W|F(W) + \mathrm{val}(W)G(W) + \mathcal{O}\left(\frac{\mathrm{val}(W)}{|W|}\right) \tag{3.3}$$

holds for $W \in \mathcal{L}$. Moreover, if W tends to a limit $\omega \in \mathcal{L}_\infty$ then $F(W)$ and $G(W)$ also tend to a limit. The functions $\omega \mapsto \lim_{W\to\omega} F(W)$ and $\omega \mapsto \lim_{W\to\omega} G(W)$ are continuous on $\mathcal{L}_\infty$.

Corollary 2. *Under the hypothesis of Theorem 1 and under Hypothesis 1 the following asymptotic formula holds*

$$\sum_{\substack{w<W \\ w\in\mathcal{L}}} f(w) = N \cdot h(N)\mathcal{F}(h(N)) + \mathcal{O}(N), \tag{3.4}$$

where $N = \mathrm{val}(W)$ and $\mathcal{F}$ is a Lipschitz-continuous periodic function of period 1.

Corollary 3. *Under the hypothesis of Theorem 1 and if the limit*

$$\lim_{n\to\infty}\frac{1}{nu_q(n)}\sum_{\substack{|w|=n\\ w\in\mathcal{L}_q}} f(w)=C_f \quad \textit{for all states } q\in Q_1$$

then the following asymptotic formula holds

$$\sum_{\substack{w<W\\ w\in\mathcal{L}}} f(w)=C_fN\cdot h(N)+\mathcal{O}(N), \tag{3.5}$$

where $N=\mathrm{val}(W)$.

4. Proof of the Results

Proof of Theorem 1. We introduce the notation

$$F_q(n)=\sum_{\substack{|w|=n\\ w\in\mathcal{L}_q}} f(w)$$

and consider the vector $\mathbf{F}(n)=(F_q(n))_{q\in Q}$. We have the recursion formula

$$F_q(n)=\sum_{\sigma\in\Sigma}\sum_{\substack{|w|=n-1\\ w\in\mathcal{L}_{q.\sigma}}} f(\sigma w)=\sum_{\sigma\in\Sigma}\left(F_{q.\sigma}(n-1)+u_{q.\sigma}(n-1)f(\sigma)\right). \tag{4.1}$$

The incidence matrix of the minimal automaton defining the language $\mathcal{L}$ is denoted by A. Furthermore, we introduce the matrix B given by

$$B_{p,q}=\sum_{\substack{\sigma\in\Sigma\\ p.\sigma=q}} f(\sigma).$$

Then we can rewrite (4.1) as

$$\mathbf{F}(n)=A\mathbf{F}(n-1)+B\mathbf{u}(n-1),\quad \text{with } \mathbf{F}(0)=\mathbf{0},$$

which has the solution

$$\mathbf{F}(n)=\left(A^{n-1}B+A^{n-2}BA+\cdots+ABA^{n-2}+BA^{n-1}\right)\mathbf{u}(0). \tag{4.2}$$

Here we have used that $\mathbf{u}(n)=A^n\mathbf{u}(0)$.

Lemma 1. *Let A be a non-negative square matrix with one dominating eigenvalue $\lambda>0$ and assume that the matrix B satisfies $A_{p,q}=0\Rightarrow B_{p,q}=0$ for all indices p and q. Then the solutions of the equation*

$$\begin{aligned}\mathbf{F}(n)&=A\mathbf{F}(n-1)+B\mathbf{u}(n-1)\quad \mathbf{F}(0)=\mathbf{0}\\ \mathbf{u}(n)&=A\mathbf{u}(n-1)\end{aligned}$$

satisfies

$$|F_q(n)|\le Cnu_q(n) \tag{4.3}$$

for some positive constant C and all indices q. Furthermore, the asymptotic relation

$$F_q(n) = C_q n u_q(n) + D_q u_q(n) + \mathcal{O}(u_q(n) n^{-1}) \quad \textit{for real constants } C_q, D_q \tag{4.4}$$

holds for those indices q for which $\lim_{n\to\infty} \lambda^{-n} u_q(n) \neq 0$.

Proof of Lemma 1. We can assume that the matrix B is also non-negative. Then there exists a constant C such that $C \cdot A \geq B$ (we interpret inequalities applied to matrices or vectors component-wise). By the positivity of A we have

$$\mathbf{F}(n) = \left(A^{n-1}B + A^{n-2}BA + \cdots + ABA^{n-2} + BA^{n-1}\right)\mathbf{u}(0) \leq CnA^n\mathbf{u}(0) = Cn\mathbf{u}(n).$$

This is the first inequality.

In order to prove (4.4) we assume that $\lambda = \lambda_1, \lambda_2, \ldots, \lambda_t$ are the eigenvalues of A; then by our assumptions we have $\lambda > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_t|$. From the general theory of matrix recurrences we know that every entry of the matrix A^n can be written as

$$(A^n)_{p,q} = \sum_{\ell=1}^{t} P_{p,q}^{(\ell)}(n)\lambda_\ell^n$$

for polynomials $P_{p,q}^{(\ell)}$. Inserting this into (4.2) and summing up the expressions for the entries we obtain

$$F_q(n) = \sum_{\ell=1}^{t} R_q^{(\ell)}(n)\lambda_\ell^n$$

for polynomials $R_q^{(\ell)}$. Now $u_q(n) = P_q(n)\lambda^n + o(\lambda^n)$ holds for a non-zero polynomial P_q by our hypothesis. The inequality $|F_q(n)| \leq Cnu_q(n)$ implies that $\deg R_q^{(1)} \leq \deg P_q + 1$. This gives the desired asymptotic relation. □

Up to now we have only studied the summatory function of f for blocks of a given length. We now turn to the evaluation of the sum (3.5) for general $W \in \mathcal{L}$. We write

$W=\sigma_1\sigma_2\cdots\sigma_\ell$. Then we have

$$\sum_{\substack{w<W\\ w\in\mathcal{L}}} f(w)=\sum_{\substack{|w|<|W|\\ w\in\mathcal{L}}} f(w)+\sum_{\substack{|w|=|W|\\ w<W\\ w\in\mathcal{L}}} f(w)=$$

$$\sum_{k=1}^{|W|-1} F_s(k)+\sum_{k=1}^{|W|}\sum_{\sigma<\sigma_k}\sum_{\substack{w\in\mathcal{L}_{s.\sigma_1\cdots\sigma_{k-1}\sigma}\\ |w|=|W|-k}} f(\sigma_1\cdots\sigma_{k-1}\sigma w)=$$

$$\sum_{k=1}^{|W|-1} F_s(k)+\sum_{k=1}^{|W|}\sum_{\sigma<\sigma_k}\left(F_{s.\sigma_1\cdots\sigma_{k-1}\sigma}(|W|-k)+u_{s.\sigma_1\cdots\sigma_{k-1}\sigma}(|W|-k)f(\sigma_1\cdots\sigma_{k-1}\sigma)\right)=$$

$$\sum_{q\in Q}\sum_{k=1}^{|W|}\beta_{q,k}(W)F_q(|W|-k)+\sum_{q\in Q}\sum_{k=1}^{|W|}\left(\sum_{i=1}^{k-1} f(\sigma_i)\right)(\beta_{q,k}(W)-\delta_{q,s})u_q(|W|-k)+$$

$$\sum_{k=1}^{|W|}\sum_{\sigma<\sigma_k} f(\sigma)u_{s.\sigma_1\cdots\sigma_{k-1}\sigma}(|W|-k)$$

where we have used the additivity of f in the second last line. In order to rewrite the sum in the last line we introduce

$$\gamma_{q,i}(W)=\sum_{\substack{\sigma<\sigma_i\\ s.\sigma_1\cdots\sigma_{i-1}\sigma=q}} f(\sigma).$$

Then again using the same reasoning as above we obtain

$$\sum_{\substack{w<W\\ w\in\mathcal{L}}} f(w)=\sum_{q\in Q}\sum_{k=1}^{|W|}\Bigg[\beta_{q,k}(W)F_q(|W|-k)+ \left(\gamma_{q,k}(W)+(\beta_{q,k}(W)-\delta_{q,s})\sum_{i=1}^{k-1} f(\sigma_i)\right)u_q(|W|-k)\Bigg]=:\sum_{q\in Q} S_q \tag{4.5}$$

We now insert the asymptotic information from Lemma 1 into (4.5).

By our assumption the number $v_s(n)$ of strings of length less or equal to n can be written as $v_s(n)=T(n)\lambda^n+o(\lambda^n)$ for a polynomial T of degree d. Furthermore, we have

$$\frac{g(n+x)}{g(n)}=\lambda^x\left(1+\frac{xd}{n}+\mathcal{O}\left(\frac{1}{n^2}\right)\right)\qquad\text{for } x\in\mathbb{R}. \tag{4.6}$$

and $\lim_{x\to\infty}\frac{g(x)}{T(x)\lambda^x}=1$.

We will make use of the following asymptotic expansions frequently

$$\frac{u_q(N-n)}{g(N)} = \begin{cases} \lambda^{-n}\left(a_q - da_q\frac{n}{N} + \frac{c_q}{N} + \mathcal{O}\left(\frac{n^2}{N^2}\right)\right) & \text{for } q \in Q_1 \quad \text{and } n = o(\sqrt{N}) \\ \lambda^{-n}\left(\frac{a_q}{N} + \mathcal{O}\left(\frac{n}{N^2}\right)\right) & \text{for } q \in Q_2 \quad \text{and } n = o(N) \\ \mathcal{O}\left(\frac{1}{N^2\lambda^n}\right) & \text{for } q \in Q_3 \quad \text{and } n \le N \end{cases} \tag{4.7}$$

and

$$\frac{F_q(N-n)}{g(N)} = \begin{cases} \lambda^{-n}\left(\mu_q N - (d+1)\mu_q n + \xi_q + \mathcal{O}\left(\frac{n^2}{N}\right)\right) & \text{for } q \in Q_1 \quad \text{and } n = o(\sqrt{N}) \\ \lambda^{-n}\left(\mu_q + \mathcal{O}\left(\frac{n}{N}\right)\right) & \text{for } q \in Q_2 \quad \text{and } n = o(N) \\ \mathcal{O}\left(\frac{1}{N\lambda^n}\right) & \text{for } q \in Q_3 \quad \text{and } n \le N. \end{cases} \tag{4.8}$$

We note that a_q, c_q and μ_q, ξ_q can be computed from the two leading coefficients of the polynomials T, P_q and $R_q^{(1)}$.

In the following we will use the asymptotic formulas (4.7) and (4.8). Technically, we would have to split summation at indices of order $o(\sqrt{|W|})$ and $o(|W|)$, but we will consequently omit this, since the contribution to the error term is negligible compared to the other error terms. Assume now that $q \in Q_1$. Then we can rewrite the corresponding part of (4.5):

$$\begin{aligned} S_q = g(|W|)\sum_{k=1}^{|W|}&\left(a_q\left(\gamma_{q,k}(W) + (\beta_{q,k}(W) - \delta_{q,s})\sum_{i=1}^{k-1} f(\sigma_i)\right) - (d+1)\mu_q k\beta_{q,k}(W)\right)\lambda^{-k} + \\ &\mu_q|W|g(|W|)\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k} + \mathcal{O}\left(\frac{g(|W|)}{|W|}\right). \end{aligned} \tag{4.9}$$

Similarly, we get the following expansion for $q \in Q_2$

$$S_q = \mu_q g(|W|)\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k} + \mathcal{O}\left(\frac{g(|W|)}{|W|}\right) \tag{4.10}$$

and $q \in Q_3$

$$S_q = \mathcal{O}\left(\frac{g(|W|)}{|W|}\right). \tag{4.11}$$

Inserting (4.9), (4.10), and (4.11) into (4.5) yields

$$\sum_{\substack{w<W \\ w\in\mathcal{L}}} f(w) = |W|g(|W|)\Phi(W) + g(|W|)\Psi(W) + \mathcal{O}\left(\frac{g(|W|)}{|W|}\right) \tag{4.12}$$

with

$$\Phi(W) = \sum_{q\in Q_1}\mu_q\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k} \tag{4.13}$$

and
(4.14)

$$\Psi(W)=\sum_{q\in Q_1}\sum_{k=1}^{|W|}\left(a_q\gamma_{q,k}(W)+a_q(\beta_{q,k}(W)-\delta_{q,s})\sum_{i=1}^{k-1}f(\sigma_i)-(d+1)\mu_q k\beta_{q,k}(W)\right)\lambda^{-k}$$
$$+\sum_{q\in Q_2}\mu_q\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k}.$$

We note that Φ and Ψ extend to continuous functions on $\mathcal{L}_\infty$ by $\Phi(\omega)=\lim_{W\to\omega}\Phi(W)$ and $\Psi(\omega)=\lim_{W\to\omega}\Psi(W)$.

We finish the proof by noting that we have from (2.1) and (4.7)

$$\frac{\mathrm{val}(W)}{g(|W|)}=Y(W)+\frac{1}{|W|}Z(W)+\mathcal{O}(|W|^{-2})$$

with

$$Y(W)=\sum_{q\in Q_1}a_q\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k}$$
$$Z(W)=-\sum_{q\in Q_1}da_q\sum_{k=1}^{|W|}\beta_{q,k}(W)k\lambda^{-k}+\sum_{q\in Q_2}a_q\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k}$$
$$+\sum_{q\in Q_1}c_q\sum_{k=1}^{|W|}\beta_{q,k}(W)\lambda^{-k}.$$

Again $Y(W)$ and $Z(W)$ extend to continuous functions on $\mathcal{L}_\infty$. For $Y(W)$ this fact was used in [19] to define an expansion for real numbers. Inserting

$$g(|W|)=\mathrm{val}(W)\left(\frac{1}{Y(W)}-\frac{1}{|W|}\frac{Z(W)}{Y(W)^2}+\mathcal{O}(|W|^{-2})\right)$$

into (4.12) we obtain the desired result with

$$F(W)=\frac{\Phi(W)}{Y(W)}$$
$$G(W)=\frac{\Psi(W)}{Y(W)}-\frac{\Phi(W)Z(W)}{Y(W)^2}.$$

□

Proof of Corollary 2. From Theorem 1 we know that

$$\sum_{\substack{w<W\\ w\in\mathcal{L}}}f(w)=\mathrm{val}(W)|W|F(W)+\mathcal{O}(\mathrm{val}(W))$$

and that the limit $\lim F(W)$ exists, if $W \to \omega \in \mathcal{L}_\infty$. Thus we only have to prove that $F(W)$ can be written as a function of $\{h(N)\}$ (recall that $N = \mathrm{val}(W)$). We split the proof of this fact into two parts: first we prove that $\lim F(W)$ for $W \to \omega$ only depends on $\mathrm{val}_\infty(\omega)$, and secondly we show that $\mathrm{val}_\infty(\omega)$ can be expressed in terms of $\{h(N)\}$.

To prove the first part we consider two sequences of words $(V_k)_{k\in\mathbb{N}}$ and $(\tilde{V}_k)_{k\in\mathbb{N}}$ converging to ω and $\tilde{\omega}$ with

$$\mathrm{val}_\infty(\omega) = \lim_{k\to\infty} \frac{\mathrm{val}(V_k)}{g(|V_k|)} = \lim_{k\to\infty} \frac{\mathrm{val}(\tilde{V}_k)}{g(|\tilde{V}_k|)} = \mathrm{val}_\infty(\tilde{\omega}). \tag{4.15}$$

Without loss of generality we may assume that V_k has the prefix $\omega_1 \cdots \omega_k$ and $\tilde{V}_k$ has the prefix $\tilde{\omega}_1 \cdots \tilde{\omega}_k$. Moreover, we may consider V_k and $\tilde{V}_k$ having the same length. Indeed using Hypothesis 2, we can choose words η_k and $\tilde{\eta}_k$ of bounded length such that the words

$$\omega_1\omega_2 \cdots \omega_k\eta_k \quad \text{and } \tilde{\omega}_1\tilde{\omega}_2 \cdots \tilde{\omega}_k\tilde{\eta}_k$$

have the same length and belong to $\mathcal{L}$. So we may choose V_k (resp. $\tilde{V}_k$) to be $\omega_1\omega_2 \cdots \omega_k\eta_k$ (resp. $\tilde{\omega}_1\tilde{\omega}_2 \cdots \tilde{\omega}_k\tilde{\eta}_k$). Then by (4.15) and $|f(w)| \le M|w|$ (for some positive constant M) we have

$$\begin{aligned} &\left|\mathrm{val}(\tilde{V}_k)F(\tilde{V}_k) - \mathrm{val}(V_k)F(V_k)\right| (k + |\eta_k|) + \mathcal{O}(\mathrm{val}(\tilde{V}_k) + \mathrm{val}(V_k)) \\ &\qquad = \left| \sum_{V_k \le w < \tilde{V}_k} f(w) \right| \le \left|\mathrm{val}(\tilde{V}_k) - \mathrm{val}(V_k)\right| M(k + |\eta_k|). \end{aligned} \tag{4.16}$$

Dividing this by $g(k+|\eta_k|)(k+|\eta_k|)$ and letting $k \to \infty$ we obtain that $F(\tilde{\omega}) = F(\omega)$. Thus we can define $\mathcal{F}(\log_\lambda \mathrm{val}_\infty(\omega)) := F(\omega)$. The function $\mathcal{F}$ is then a continuous function, since F is continuous and the map val_∞ is continuous. Inserting (4.7) into (4.15) yields $\{h(\mathrm{val}(V_k))\} = \{\log_\lambda \mathrm{val}_\infty(\omega)\} + \mathcal{O}(\frac{1}{k})$ and therefore we have $\mathcal{F}(\{h(\mathrm{val}(V_k))\}) = F(\omega) + \mathcal{O}(\mu_{\mathcal{F}}(\frac{1}{k}))$, where $\mu_{\mathcal{F}}$ denotes the modulus of continuity of $\mathcal{F}$. Thus $F(\omega)$ can be written as a function of $\{h(N)\}$ and it remains to prove that the modulus of continuity satisfies $\mu_{\mathcal{F}}(\delta) = \mathcal{O}(\delta)$.

In order to prove Lipschitz-continuity of F we use the same reasoning once again for two sequences U_k and V_k with

$$\lim_{k\to\infty} \frac{\mathrm{val}(U_k)}{g(|U_k|)} = \lambda^x, \quad \lim_{k\to\infty} \frac{\mathrm{val}(V_k)}{g(|V_k|)} = \lambda^y$$

and $|U_k| = |V_k|$. We insert U_k and V_k into (4.16) to obtain

$$|\lambda^x\mathcal{F}(x) - \lambda^y\mathcal{F}(y)| \le M\,|\lambda^x - \lambda^y| \le C|x - y|.$$

Thus the function $\lambda^x\mathcal{F}(x)$ is Lipschitz continuous on $[-1, 0]$ and by the Lipschitz continuity of the function $x \mapsto \lambda^{-x}$ also $\mathcal{F}$ is Lipschitz continuous on $[-1, 0]$.

It remains to prove that $\mathcal{F}$ has a continuous periodic extension to $\mathbb{R}$. For this purpose it is enough to show that $\mathcal{F}(-1) = \lim_{x\to 0-} \mathcal{F}(x)$. For this purpose we observe that given a minimal word V_{k+1} of length $k+1$ and a maximal word W_k of length k, we have

$\mathrm{val}(W_k)+1=\mathrm{val}(V_{k+1})$. Furthermore, $f(V_{k+1})=\mathcal{O}(k)$ and thus it follows immediately that

$$\mathcal{F}(-1)=\lim_{k\to\infty}F(V_{k+1})=\lim_{k\to\infty}F(W_k)=\lim_{x\to 0-}\mathcal{F}(x).$$

This finishes the proof of Corollary 2. □

Proof of Corollary 3. If the limit

$$\lim_{n\to\infty}\frac{1}{nu_q(n)}\sum_{\substack{|w|=n\\ w\in\mathcal{L}_q}}f(w)=C_f$$

for all states $q\in Q_1$ we have $\mu_q=C_f a_q$ for $q\in Q_1$ and we can write

$$C_f\,\mathrm{val}(W)=\Phi(W)g(|W|)+\mathcal{O}\left(\frac{g(|W|)}{|W|}\right)$$

and therefore we can use $|W|=\lfloor h(N)\rfloor+1$ to obtain

$$\sum_{\substack{w<W\\ w\in\mathcal{L}}}f(w)=C_f h(N)N+\mathcal{O}(N),$$

where $N=\mathrm{val}(W)$. □

Remark 2. The only properties of the function g that we used were the strict monotonicity and (4.6). Thus any function g with these properties could be used instead; the results would be the same.

5. Examples

We present a number of examples which show that in general the results cannot be improved. Especially, Example 1 will show that the function G given in (3.3) does not necessarily extend to a continuous function of $\{h(N)\}$.

Example 1. Let $\mathcal{L}=\{a,b\}^*c\{a,b\}^*\cup\{d,e\}^*$ with $a<b<c<d<e$ and consider the function $f(a)=f(b)=f(c)=f(d)=0$, $f(e)=1$. In this case Corollary 3 applies with $C_f=0$, since the language $\{d,e\}^*$ has only 2^n words of length n, whereas $\mathcal{L}$ has $u_s(n)=n2^{n-1}+2^n$ and $v_s(n)=(n+1)2^n$. It is a simple exercise to compute

$$\sum_{w<d^{(n)}}f(w)=n2^{n-1}-2^n+1\quad\text{and}\quad\sum_{w<e^{(n)}}f(w)=(n+1)2^n-2^{n+1}-n+1.$$

Then

$$\lim_{n\to\infty}\frac{1}{v_s(n)}\#\{w\in\mathcal{L}\mid d^{(n)}\le w<e^{(n)}\}=0$$

but

$$\lim_{n\to\infty}\frac{1}{v_s(n)}\sum_{d^{(n)}\le w<e^{(n)}}f(w)=\frac{1}{2},$$

which implies that the two sequences $d^{(\omega)}$ and $e^{(\omega)}$ represent the same real number, but the values of G at these points differ.

Example 2. Let $\mathcal{L} = (\{a,b\}^*c\{a,b\}^*) \cup (\{d,e\}^*c\{d,e\}^*)$ with $a < b < c < d < e$ and consider the function $f(a) = f(b) = f(c) = f(d) = 0$, $f(e) = 1$. In this case we have $u_s(n) = n2^n$ for $n > 1$ and $v_s(n) = (n-1)2^{n+1} + 1$ for $n > 0$. Furthermore, we have

$$\sum_{w < cd^{(n-1)}} f(w) = (n^2 - 5n + 8)2^{n-2} - 2 \quad \text{and} \quad \sum_{w < e^{(n-1)}c} f(w) = (n^2 - 3n + 4)2^{n-1} - n - 1.$$

The numerical values of the words are given by $\mathrm{val}(cd^{(n-1)}) = (3n-4)2^{n-1} + 1$ and $\mathrm{val}(e^{(n-1)}c) = (n-1)2^{n+1}$. Therefore we have

$$\lim_{n\to\infty} \frac{1}{n\,\mathrm{val}(cd^{(n-1)})} \sum_{w < cd^{(n-1)}} f(w) = \frac{1}{6} \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n\,\mathrm{val}(e^{(n-1)}c)} \sum_{w < e^{(n-1)}c} f(w) = \frac{1}{4},$$

which shows that the function $\mathcal{F}$ in Corollary 2 is not constant.

Although we did not study polynomial languages in this paper, we will give an example for such a language which shows that similar phenomena can be expected.

Example 3. Let $\mathcal{L} = a^*b^*$ with $a < b$ and consider the function $f(a) = 1$, $f(b) = 0$. In this case we have $u_s(n) = n + 1$ and $v_s(n) = \frac{(n+1)(n+2)}{2}$. Furthermore, we have

$$\sum_{w < a^{(\ell)}b^{(n-\ell)}} f(w) = \frac{n^3 - n}{6} + \frac{(n-\ell)(n+\ell+1)}{2}$$

and $\mathrm{val}(a^{(\ell)}b^{(n-\ell)}) = \frac{n(n+1)}{2} + n - \ell$. Setting $N = \mathrm{val}(a^{(\ell)}b^{(n-\ell)})$ we can compute $n = \lfloor\sqrt{2N + \frac{9}{4}} - \frac{3}{2}\rfloor$ and $\ell = N - \frac{n(n+3)}{2}$. Inserting this into the above formula we obtain

$$\sum_{w < a^{(\ell)}b^{(n-\ell)}} f(w) = \frac{\sqrt{2}}{3}N^{\frac{3}{2}} + N\left(\left\{\sqrt{2N + \frac{9}{4}} - \frac{3}{2}\right\}\left(1 - \left\{\sqrt{2N + \frac{9}{4}} - \frac{3}{2}\right\}\right) - \frac{1}{2}\right) + \mathcal{O}(\sqrt{N}).$$

This shows that we have a periodic function of $h(N) = \sqrt{2N + \frac{9}{4}} - \frac{3}{2}$ in the second term of the asymptotic formula.

Acknowledgment. The authors are indebted to Jean-Paul Allouche for suggesting the subject of this paper to them. The authors thank two anonymous referees for their valuable comments.

References

1. G. Barat, T. Downarowicz, A. Iwanik, and P. Liardet, *Propriétés topologiques et combinatoires des échelles de numération*, Colloq. Math. **84/85** (2000), 285–306.
2. G. Barat and P. J. Grabner, *Distribution properties of G-additive functions*, J. Number Theory **60** (1996), 103–123.
3. V. Bruyère and G. Hansel, *Bertrand numeration systems and recognizability*, Theor. Comput. Sci. **181** (1997), 17–43, Latin American Theoretical Informatics (Valparaíso, 1995).
4. H. Delange, *Sur les fonctions q-additive ou q-multiplicatives*, Acta Arith. **21** (1972), 285–298.
5. ———, *Sur la fonction sommatoire de la fonction "Somme des Chiffres"*, Enseign. Math. II. Sér. **21** (1975), 31–47.

6. M. Drmota, *q-Additive functions and well distribution*, Demonstr. Math. **30** (1997), 883–896.
7. M. Drmota and J. Gajdosik, *The distribution of the sum-of-digits function*, J. Théor. Nombres Bordx. **10** (1998), 17–32.
8. J.-M. Dumont and A. Thomas, *Digital sum problems and substitutions on a finite alphabet*, J. Number Theory **39** (1991), 351–366.
9. ______, *Digital sum moments and substitutions*, Acta Arith. **64** (1993), 205–225.
10. ______, *Gaussian asymptotic properties of the sum-of-digits function*, J. Number Theory **62** (1997), 19–38.
11. S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.
12. P. Flajolet, P. J. Grabner, P. Kirschenhofer, H. Prodinger, and R. F. Tichy, *Mellin transforms and asymptotics: digital sums*, Theor. Comput. Sci. **123** (1994), 291–314.
13. C. Frougny, *Representation of numbers and finite automata*, Math. Syst. Theory **25** (1992), 37–60.
14. ______, *Numeration Systems*, Algebraic Combinatorics on Words (M. Lothaire, ed.), Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, 2002.
15. P. J. Grabner, P. Liardet, and R. F. Tichy, *Odometers and systems of numeration*, Acta Arith. **70** (1995), 103–123.
16. P. J. Grabner and R. F. Tichy, *Contributions to digit expansions with respect to linear recurrences*, J. Number Theory **36** (1990), 160–169.
17. ______, *α-expansions, linear recurrences and the sum-of-digits function*, Manuscr. Math. **70** (1991), 311–324.
18. P. Lecomte and M. Rigo, *Numeration systems on a regular language*, Theory Comput. Syst. **34** (2001), 27–44.
19. ______, *On the representation of real numbers using regular languages*, Theory Comput. Syst. **35** (2002), 13–38.
20. N. Loraud, *β-shift, systèmes de numération et automates*, J. Théor. Nombres Bordeaux **7** (1995), no. 2, 473–498.
21. J. O. Shallit, *Numeration systems, linear recurrences, and regular sets*, Inform. and Comput. **113** (1994), 331–347.
22. A. Szilard, S. Yu, K. Zhang, and J. Shallit, *Characterizing regular languages with polynomial densities*, Mathematical foundations of computer science 1992 (Prague, 1992), Springer, Berlin, 1992, pp. 494–503.
23. E. Zeckendorf, *A generalized Fibonacci numeration*, Fibonacci Quart. **10** (1972), no. 4, 365–372.

(P. G.)
Institut für Mathematik A
Technische Universität Graz
Steyrergasse 30
8010 Graz
Austria
E-mail address: peter.grabner@tugraz.at

(M. R.)
Institut de Mathématique
Université de Liège
Grande Traverse 12 (B 37)
4000 Liège
Belgium
E-mail address: M.Rigo@ulg.ac.be