Upper bound for the number of closed and privileged words

Josef Rukavicka[*]

November 25, 2019
Mathematics Subject Classification: 68R15

Abstract

A non-empty word w is a *border* of the word u if $|w| < |u|$ and w is both a prefix and a suffix of u. A word u with the border w is *closed* if u has exactly two occurrences of w. A word u is *privileged* if $|u| \leq 1$ or if u contains a privileged border w that appears exactly twice in u.

Peltomäki (2016) presented the following open problem: "Give a nontrivial upper bound for $B(n)$", where $B(n)$ denotes the number of privileged words of length n. Let $\mathrm{D}(n)$ denote the number of closed words of length n. Let $q > 1$ be the size of the alphabet. We show that there is a positive real constant c such that

$$\mathrm{D}(n) \leq c \ln n \frac{q^n}{\sqrt{n}}, \text{ where } n > 1.$$

Privileged words are a subset of closed words, hence we show also an upper bound for the number of privileged words.

1 Introduction

A non-empty word w is a *border* of the word u if $|w| < |u|$ and w is both a prefix and a suffix of u. A border w of the word u is the *maximal border* of u if for every border $\bar{w}$ of u we have that $|\bar{w}| \leq |w|$. A word u with the border

[*]Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CZECH TECHNICAL UNIVERSITY IN PRAGUE (josef.rukavicka@seznam.cz).

w is *closed* if u has exactly two occurrences of w. It follows that w occurs only as a prefix and as a suffix of u. A word u is *privileged* if $|u| \leq 1$ or if u contains a privileged border w that appears exactly twice in u. Obviously privileged words are a subset of closed words.

The properties of closed and privileged words have been studied in recent years [2], [5], [6]. One of the questions that has been investigated is the enumeration of privileged words. In [3], it was proved that there are constants c and n_0 such that for all $n > n_0$, there are at least $\frac{cq^n}{n(\log_q n)^2}$ privileged words of length n. This improves the lower bound for the number of privileged words from [1]. Since every privileged word is a closed word, the result from [3] forms also a lower bound for the number of closed words.

Concerning an upper bound for the number of privileged words we have found only the following open problem [4]: "Give a nontrivial upper bound for $B(n)$", where $B(n)$ denotes the number of privileged words of length n. Also in [4], the author presents an idea how to improve the lower bound from [3]. On the other hand, in [4], there is no explicit suggestion how to approach the problem of determining the upper bound.

In the current article we construct an upper bound for the number of closed words of length n. Since the privileged words are a subset of closed words, we present also a response to the open problem from [4].

We explain in outline our proof. Let A be an alphabet with $q > 1$ letters, let A^m denote the set of all words of length m, and let $\mathrm{A}^* = \bigcup_{m\geq 0} \mathrm{A}^m$. It is known that $|\,\mathrm{A}^m\,| = q^m$. Let $\mathrm{A}_w(n)$ denote the number of words of length n that do not contain the factor $w \in \mathrm{A}^*$. Let $\mu(n,m)$ be the maximal value of $\mathrm{A}_w(n)$ for all w of length m; formally

$$\mu(n,m) = \max\{\mathrm{A}_w(n) \mid w \in \mathrm{A}^m\}.$$

Let $\hat{\mathrm{D}}(n)$ denote the set of all closed words of length n and let $\hat{\mathrm{D}}(n,m)$ denote the set of all closed words of length n having a maximal border of length m. Let $\mathrm{D}(n) = |\hat{\mathrm{D}}(n)|$ and $\mathrm{D}(n,m) = |\hat{\mathrm{D}}(n,m)|$.

Obviously $\hat{\mathrm{D}}(n) = \bigcup_{m=1}^{n-1} \hat{\mathrm{D}}(n,m)$ and $\hat{\mathrm{D}}(n,m) \cap \hat{\mathrm{D}}(n,\bar{m}) = \emptyset$, where $m \neq \bar{m}$. We show that if $2m > n$ then $\mathrm{D}(n,m) \leq q^{\lceil \frac{n}{2} \rceil}$ and if $2m \leq n$ then $\mathrm{D}(n,m) \leq q^m \mu(n-2m,m)$; see Lemma 2.5. It follows that

$$\mathrm{D}(n) = \sum_{m=1}^{n-1} \mathrm{D}(n,m) \leq \sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} q^m \mu(n-2m,m) + \sum_{m=\lfloor \frac{n}{2} \rfloor + 1}^{n-1} q^{\lceil \frac{n}{2} \rceil}. \qquad (1)$$

Let $\mathbb{N}$ denote the set of positive integers. Let $\omega(n) = \frac{1}{\ln q}(\ln n - \ln\ln n)$. Let Π denote the set of all functions $\pi(n) : \mathbb{N} \to \mathbb{N}$ such that $\pi(n) \in \Pi$ if and only if $1 \leq \pi(n) \leq \max\{1, \omega(n))\}$ and $\pi(n) \leq \pi(n+1)$ for all $n \in \mathbb{N}$. We apply the function max, because $\omega(n) < 1$ for some small n.

The key observation in our article is that the number of words of length n that do not contain some “short" factor of length $\pi(n) \in \Pi$ has the same growth rate as the number of words of length $n - \lfloor \frac{\ln n}{\ln q} \rfloor$. Formally said, for each $\pi(n) \in \Pi$ there is a positive real constant c such that $\mu(n, \pi(n)) \leq cq^{n-\frac{\ln n}{\ln q}}$; see Theorem 2.3. This observation allows us to show that there are real positive constants c_1, c_2 such that

$$\sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} q^m \mu(n-2m, m) \leq c_1 \ln n \sum_{m=\lfloor c_2 \ln n \rfloor}^{\lfloor \frac{n}{2} \rfloor} q^m \mu(n-2m, m). \qquad (2)$$

In consequence we may count only closed words having a maximal border longer than $c_2 \ln n$ in order to find an upper bound for $\mathrm{D}(n)$. Applying that $\mu(n-2m, m) \leq q^{n-2m}$ for $n \geq 2m$, we derive from (1) and (2) our result for the number of closed words.

2 Upper bound for the number of closed words

We present an upper bound for the number of words of length n that avoid some factor of length m; it means an upper bound for $\mu(n, m)$.

Lemma 2.1. *If $n, m \in \mathbb{N}$ then*

$$\mu(n, m) \leq q^n \left(1 - \frac{1}{q^m}\right)^{\lfloor \frac{n}{m} \rfloor}.$$

Proof. Given $w \in \mathrm{A}^m$, let $U_{n,w}$ be a set of words $u = u_1 u_2 \dots u_{k-1} u_k \in \mathrm{A}^*$, where $|u| = n$, $|u_i| = m$, $w \neq u_i$ for all $1 \leq i < k$, and $|u_k| = n \bmod m$. It follows that $|u_k| < m = |w|$ and thus $u_k \neq w$. Obviously

$$|U_{n,w}| = (q^m - 1)^{\lfloor \frac{n}{m} \rfloor} q^{n \bmod m} = q^n \left(1 - \frac{1}{q^m}\right)^{\lfloor \frac{n}{m} \rfloor}.$$

Note that $|\,\mathrm{A}^m \setminus \{w\}| = q^m - 1$. It is clear that the set of words of length n not containing the factor w is a subset of $U_{n,w}$. The lemma follows. □

For the proof of Theorem 2.3 we need the following limit.

Proposition 2.2. *We have that*

$$\lim_{n\to\infty} n\left(1-\frac{\ln n}{n}\right)^n = \mathrm{e}\,.$$

Proof. Let

$$y = \lim_{n\to\infty} n\left(1-\frac{\ln n}{n}\right)^n . \tag{3}$$

From (3) we have that

$$\ln y = \lim_{n\to\infty} \ln\left[n\left(1-\frac{\ln n}{n}\right)^n\right] = \lim_{n\to\infty}\left[\ln n + n\ln\left(1-\frac{\ln n}{n}\right)\right] . \tag{4}$$

Let us consider the second term on the right side of (4):

$$\begin{gathered}\lim_{n\to\infty} n\ln\left(1-\frac{\ln n}{n}\right) = \lim_{n\to\infty}\frac{\ln\left(1-\frac{\ln n}{n}\right)'}{\left(\frac{1}{n}\right)'} = \\ \lim_{n\to\infty}\frac{\frac{(-1)\left(\frac{1-\ln n}{n^2}\right)}{\left(1-\frac{\ln n}{n}\right)}}{-\frac{1}{n^2}} = \lim_{n\to\infty}\frac{n(1-\ln n)}{n-\ln n}.\end{gathered} \tag{5}$$

Since $\lim_{n\to\infty}\frac{n}{n-\ln n} = 1$, it follows from (4) and (5) that

$$\ln y = \lim_{n\to\infty}\left[\ln n + \frac{n(1-\ln n)}{n-\ln n}\right] = \lim_{n\to\infty}\left[\ln n + 1 - \ln n\right] = 1.$$

It follows that $y = \mathrm{e}$. This completes the proof. □

Let $\mathbb{R}^+$ denote the set of positive real numbers.

Let $\beta = \frac{1}{\ln q} \in \mathbb{R}^+$. The following theorem states that the number of words of length n avoiding some given "short" factor (of length shorter than $\pi(n) \in \Pi$) has the same growth rate as the number of all words of length $n - \beta\ln n$.

Theorem 2.3. *If $\pi(n) \in \Pi$ then there is a constant $c \in \mathbb{R}^+$ such that for all $n \in \mathbb{N}$ we have that*

$$\frac{\mu(n,\pi(n))}{q^{n-\beta\ln n}} \le c.$$

Proof. From Lemma 2.1 we have that

$$\frac{\mu(n,\pi(n))}{q^{n-\beta\ln n}} = \frac{q^n\left(1-\frac{1}{q^{\pi(n)}}\right)^{\lfloor\frac{n}{\pi(n)}\rfloor}}{q^{n-\beta\ln n}} = n\left(1-\frac{1}{q^{\pi(n)}}\right)^{\lfloor\frac{n}{\pi(n)}\rfloor}. \tag{6}$$

Realize that $q^{\beta\ln n} = n$.

Obviously there is $n_0 \in \mathbb{N}$ such that $q^{\pi(n)} \leq \frac{n}{\ln n}$ for all $n > n_0$; recall that $\pi(n) \leq \omega(n) = \frac{1}{\ln q}(\ln n - \ln\ln n)$ as n tends to infinity. Consequently for all $n > n_0$ we have that

$$n\left(1-\frac{1}{q^{\pi(n)}}\right)^n \leq n\left(1-\frac{\ln n}{n}\right)^n. \tag{7}$$

Proposition 2.2 and (7) imply that

$$\lim_{n\to\infty} n\left(1-\frac{1}{q^{\pi(n)}}\right)^n \leq \mathrm{e}\,. \tag{8}$$

Clearly $\lim_{n\to\infty}(f(n))^{\frac{1}{\pi(n)}} \leq \mathrm{e}$ for each function $f(n)$ such that $f(n) \geq 0$ and $\lim_{n\to\infty} f(n) \leq \mathrm{e}$; recall that $\pi(n) \geq 1$. Then the theorem follows from (6) and (8). This completes the proof. □

Let $h(n) = \lfloor \beta\ln n\rfloor$. We present Theorem 2.3 in a slightly different manner that will be more useful for us in the following.

Corollary 2.4. *If $\pi(n), \bar{\pi}(n) \in \Pi$, and $\bar{\pi}(n) \leq \pi(n)$ then there is a constant $c \in \mathbb{R}^+$ such that for all $n \in \mathbb{N}$ we have that*

$$\frac{\mu(n-2\bar{\pi}(n),\bar{\pi}(n))}{q^{n-h(n)}} \leq c.$$

Proof. It is easy to verify that $\mu(n-2\bar{\pi}(n),\bar{\pi}(n)) \leq \mu(n,\pi(n))$, since the number of words of length n avoiding some factor of length $\pi(n)$ is bigger or equal to the number of words of length $n-2\bar{\pi}(n)$ avoiding some factor of length $\bar{\pi}(n) \leq \pi(n)$.

Obviously $h(n) = \lfloor\frac{\ln n}{\ln q}\rfloor \leq \frac{\ln n}{\ln q} = \beta\ln n$. In consequence we have that $q^{n-h(n)} \geq q^{n-\beta\ln n}$.

The corollary follows from Theorem 2.3. This completes the proof. □

We show an upper bound for $\mathrm{D}(n,m)$ for the cases where $2m > n$ and $2m \leq n$.

Lemma 2.5. *Suppose $n, m \in \mathbb{N}$.*

- *If $2m > n$ then $\mathrm{D}(n,m) \leq q^{\lceil \frac{n}{2} \rceil}$.*
- *If $2m \leq n$ then $\mathrm{D}(n,m) \leq q^m \mu(n-2m, m)$.*

Proof. If $2m > n$, $w \in \mathrm{A}^*$, and $|w| = m$ then there is obviously at most one word u with $|u| = n$ having a prefix and a suffix w; the prefix w and the suffix w would overlap with each other. If such u exists then the first half of u uniquely determines the second half of u. If follows that $\mathrm{D}(n,m) \leq q^{\lceil \frac{n}{2} \rceil}$.

Let $\mathrm{F}(w)$ denote the set of all factors of $w \in \mathrm{A}^*$. If $n \geq 2m$ then let

$$Z(n,m) = \{wuw \mid u \in \mathrm{A}^{n-2m} \text{ and } w \in \mathrm{A}^m \text{ and } w \notin \mathrm{F}(u)\}.$$

If $n \geq 2m$ then $\mathrm{D}(n,m) \subseteq Z(n,m)$. It is easy to see that

$$|Z(n,m)| \leq |\,\mathrm{A}^m\,| \mu(n-2m, m).$$

This completes the proof. □

Let $\kappa > 1$ be a real constant and $\bar{h}(n) = \max\{1, \lfloor \frac{1}{\kappa}\omega(n) \rfloor\}$. Again we use the function max to guarantee that $\bar{h}(n) \geq 1$ for small n.

Remark 2.6. The function $\bar{h}(n)$ defines the maximal length of a "short" border of a closed word. In the proof of Theorem 2.9 the closed words from $\hat{\mathrm{D}}(n,m)$ will be enumerated differently for $m < \bar{h}(n)$ and for $m \geq \bar{h}(n)$.

The next auxiliary lemma shows an upper bound for $q^{-h(n)+\bar{h}(n)}$, that we will use in the proof of Proposition 2.8.

Lemma 2.7. *There is a constant $c_1 \in \mathbb{R}^+$ such that for all $n \in \mathbb{N}$ we have that*

$$q^{-h(n)+\bar{h}(n)} \leq c_1 q^{\frac{1}{\ln q}\left(\frac{1}{\kappa}-1\right)\ln n}$$

Proof. Let

$$y = \lim_{n\to\infty}(-h(n) + \bar{h}(n) - \frac{1}{\ln q}\left(\frac{1}{\kappa} - 1\right)\ln n).$$

We have that

$$\begin{aligned} y &= \lim_{n\to\infty}\left(-\lfloor\frac{1}{\ln q}\ln n\rfloor + \lfloor\frac{1}{\kappa\ln q}(\ln n - \ln\ln n)\rfloor - \frac{1}{\ln q}\left(\frac{1}{\kappa}-1\right)\ln n\right) \\ &= \lim_{n\to\infty}\left(\frac{\ln n}{\ln q}\left(-1+\frac{1}{\kappa}\right) - \frac{1}{\ln q}\left(\frac{1}{\kappa}-1\right)\ln n\right) \qquad (9) \\ &= 0. \end{aligned}$$

This implies that

$$\lim_{n\to\infty} \frac{q^{-h(n)+\bar{h}(n)}}{q^{\frac{1}{\ln q}\left(\frac{1}{\kappa}-1\right)\ln n}} = 1.$$

The lemma follows. □

The next proposition shows an upper bound for the number of closed words of length n having a maximal border of length $\leq \lceil \frac{n}{2} \rceil$.

Proposition 2.8. *There is a constant $c \in \mathbb{R}^+$ such that*

$$\sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m, m) \leq c \ln n \frac{q^n}{\sqrt{n}}, \textit{ where } n > 1.$$

Proof. Since $\mu(n-2m, m) \leq q^{n-2m}$ we have that

$$\sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m, m) \leq \sum_{m=1}^{\bar{h}(n)-1} q^m \mu(n-2m, m) + \sum_{m=\bar{h}(n)}^{\lceil \frac{n}{2} \rceil} q^m q^{n-2m}. \tag{10}$$

Corollary 2.4 implies that $\mu(n-2m, m) \leq cq^{n-h(n)}$ for some constant $c \in \mathbb{R}^+$. It follows that

$$\begin{aligned} \sum_{m=1}^{\bar{h}(n)-1} q^m \mu(n-2m, m) &\leq \sum_{m=1}^{\bar{h}(n)} q^m c q^{n-h(n)} \\ &\leq \bar{h}(n) q^{\bar{h}(n)} c q^{n-h(n)}. \end{aligned} \tag{11}$$

Lemma 2.7 and (11) imply that

$$\sum_{m=1}^{\bar{h}(n)-1} q^m \mu(n-2m, m) \leq c_1 \bar{h}(n) c q^{n-\frac{\ln n}{\ln q}(1-\frac{1}{\kappa})}, \tag{12}$$

where c_1 is some real positive constant.

It is easy to verify that

$$q^{-\bar{h}(n)} \leq q^{-\frac{1}{\kappa \ln q}(\ln n - \ln \ln n)+1} = q(\ln n)^{\frac{1}{\kappa}} q^{-\frac{1}{\kappa \ln q}\ln n}. \tag{13}$$

Thus using (13)

$$\sum_{m=\bar{h}(n)}^{\lceil \frac{n}{2} \rceil} q^m q^{n-2m} \leq q^n \sum_{m=\bar{h}(n)}^{\lceil \frac{n}{2} \rceil} q^{-m} \leq \frac{q^{n-\bar{h}(n)}}{1-q^{-1}} \leq \frac{q(\ln n)^{\frac{1}{\kappa}} q^{n-\frac{1}{\kappa \ln q}\ln n}}{1-q^{-1}}. \tag{14}$$

Obviously $\bar{h}(n) \leq \frac{\ln n}{\kappa \ln q}$. Hence taking $\kappa = 2$, we get from (10), (12), and (14) that

$$\begin{aligned}\sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m,m) &\leq c_1 \bar{h}(n) c q^{n - \frac{1}{2\ln q}\ln n} + \frac{q(\ln n)^{\frac{1}{2}} q^{n-\frac{1}{2\ln q}\ln n}}{1-q^{-1}} \\ &\leq q^{n-\frac{1}{2\ln q}\ln n}\left(c_1 c \frac{\ln n}{2\ln q} + \frac{q(\ln n)^{\frac{1}{2}}}{1-q^{-1}}\right) \\ &\leq q^{n-\frac{1}{2\ln q}\ln n}(c_2 \ln n + c_3 (\ln n)^{\frac{1}{2}}),\end{aligned} \tag{15}$$

for some constants $c_2, c_3 \in \mathbb{R}^+$. Since $\sqrt{n} = q^{\frac{1}{2\ln q}\ln n}$ the proposition follows from (15). □

We show an upper bound for $\mathrm{D}(n)$.

Theorem 2.9. *There is a constant $c \in \mathbb{R}^+$ such that*

$$\mathrm{D}(n) \leq c \ln n \frac{q^n}{\sqrt{n}}, \textit{ where } n > 1.$$

Proof. We have that

$$\mathrm{D}(n) = \sum_{m=1}^{n-1} \mathrm{D}(n,m) = \sum_{m=1}^{\lceil \frac{n}{2} \rceil} \mathrm{D}(n,m) + \sum_{m=\lceil \frac{n}{2} \rceil+1}^{n-1} \mathrm{D}(n,m). \tag{16}$$

From Lemma 2.5 and (16) we get that

$$\mathrm{D}(n) \leq \sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m,m) + \sum_{m=\lceil \frac{n}{2} \rceil+1}^{n-1} q^{\lceil \frac{n}{2} \rceil}. \tag{17}$$

Realize that

$$\sum_{m=\lceil \frac{n}{2} \rceil+1}^{n-1} q^{\lceil \frac{n}{2} \rceil} \leq \frac{n}{2} q^{\lceil \frac{n}{2} \rceil}$$

and

$$\lim_{n\to\infty} \frac{nq^{\frac{n}{2}}}{\frac{\ln n q^n}{\sqrt{n}}} = 0.$$

Then it follows that from (17), and Proposition 2.8 that there are constants $c_2, c_3 \in \mathbb{R}^+$ such that

$$c_2 \sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m,m) \geq \sum_{m=\lceil \frac{n}{2} \rceil + 1}^{n-1} q^{\lceil \frac{n}{2} \rceil} \text{ and}$$

$$\mathrm{D}(n) \leq c_3 \sum_{m=1}^{\lceil \frac{n}{2} \rceil} q^m \mu(n-2m,m). \tag{18}$$

The theorem follows from (18), and Proposition 2.8 □

Remark 2.10. Note that the some of the constants c, c_1, c_2, c_3, that we used in our results and in particular in Theorem 2.9, depend on q.

Acknowledgments

The author acknowledges support by the Czech Science Foundation grant GAČR 13-03538S and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/205/OHK4/3T/14.

References

[1] M. Forsyth, A. Jayakumar, J. Peltomäki, and J. Shallit, *Remarks on privileged words*, International Journal of Foundations of Computer Science, Vol. (27), No. 04, available at https://doi.org/10.1142/S0129054116500088, (2016), pp. 431–442.

[2] J.Kellendonk, D.Lenz, and J.Savinien, *A characterization of subshifts with bounded powers*, Discrete Mathematics, Volume 313, Issue 24, available at https://doi.org/10.1016/j.disc.2013.08.026, (2013), pp. 2881–2894.

[3] J. Nicholson and N. Rampersad, *Improved estimates for the number of privileged words*, Journal of Integer Sequences, 21 (2018).

[4] J. Peltomäki, *Privileged words and sturmian words*, Turku Centre for Computer Science, TUCS Dissertations No 214, August 2016, available at http://urn.fi/URN:ISBN:978-952-12-3422-4.

[5] ——, *Introducing privileged words: Privileged complexity of sturmian words*, Theoretical Computer Science, Volume 500, available at https://doi.org/10.1016/j.tcs.2013.05.028, (2013), pp. 57–67.

[6] L. Schaeffer and J. Shallit, *Closed, palindromic, rich, privileged, trapezoidal, and balanced words in automatic sequences*, Electr. J. Comb., 23 (2016), p. P1.25.