

Sets Represented as the Length- n Factors of a Word

Shuo Tan and Jeffrey Shallit

School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
{s22tan, shallit}@uwaterloo.ca

Abstract. In this paper we consider the following problems: how many different subsets of Σ^n can occur as set of all length- n factors of a finite word? If a subset is representable, how long a word do we need to represent it? How many such subsets are represented by words of length t ? For the first problem, we give upper and lower bounds of the form α^{2^n} in the binary case. For the second problem, we give a weak upper bound and some experimental data. For the third problem, we give a closed-form formula in the case where $n \leq t < 2n$. Algorithmic variants of these problems have previously been studied under the name “shortest common superstring”.

1 Introduction

Let w, x, y, z be finite words. If $w = xyz$, we say that y is a *factor* of w . De Bruijn proved [1] the existence of a set of binary words $(B_n)_{n \geq 1}$ with the property that every binary word of length n appears as a factor of B_n (and, in fact, appears exactly once in B_n). Here we are thinking of B_n interpreted as a circular word. For example, consider the case where $n = 2$, where we can take $B_2 = 0011$. Interpreted circularly, the factors of length 2 of B_2 are 00, 01, 11, 10, and these factors comprise all the binary words of length 2.

However, not every subset of $\{0, 1\}^n$ can be represented as the factors of some finite word. For example, the set $\{00, 11\}$ cannot equal the set of all factors of any word w — interpreted in the ordinary sense or circularly — because the set of factors of any w containing both letters must contain either 01 or 10.

This raises the natural question, how many different non-empty subsets S of $\{0, 1\}^n$ can be represented as the factors of some word w ? (Note that, unlike [7], we do *not* insist that each element of S appear exactly once in w .) We give upper and lower bounds for this quantity for circular words, both of the form α^{2^n} . Our upper bound has $\alpha = \sqrt[4]{10} \doteq 1.78$ while our lower bound has $\alpha = \sqrt{2} \doteq 1.41$.

If the set of length- n factors of a word w (considered circularly) equals S , we say that w *witnesses* S . We study the length of the shortest witness for subsets of $\{0, 1\}^n$, and give an upper bound.

Restriction on the length of a witness leads us to another interesting problem. Let $T(t, n)$ denote the number of subsets of $\{0, 1\}^n$ witnessed by some word of

length $t \geq n$. Is there any characterization of $T(t, n)$? We focus on ordinary (non-circular) words for this question and derive a closed-form formula for $T(t, n)$ in the case where $n \leq t < 2n$.

Algorithmic versions of related problems have been widely studied in the literature under the name “shortest common superstring”. For example, Gallant, Maier, and Storer [4] proved that the following decision problem is NP-complete:

Instance: A set S of words and an integer K .

Question: Is there a word w of length $\leq K$ containing each word in S (and possibly others) as a factor?

However, the combinatorial problems that we study in this paper seem to be new.

2 Preliminaries

Let $\Sigma = \{0, 1\}$ denote the alphabet. Let $F_n(w)$ denote the set of length- n factors of an ordinary (non-circular) word w , and let $C_n(w)$ denote the set of length- n factors of w where w is interpreted circularly. For example, if $w = 001$, then $F_2(w) = \{00, 01\}$, while if $w = 001$ is interpreted circularly, then $C_2(w) = \{00, 01, 10\}$.

We say that a word w *witnesses* (resp., *circularly witnesses*) a subset S of Σ^n if $F_n(w) = S$ (resp., $C_n(w) = S$). A subset S of Σ^n is *representable* (resp., *circularly representable*) if there exists a non-empty word (resp., circular word) that witnesses S . Let R_n denote the set of all non-empty representable subsets of Σ^n , and let \mathring{R}_n denote the set of all non-empty circularly representable subsets of Σ^n .

Let $\text{sw}(S)$ (resp., $\text{scw}(S)$) denote the length of the shortest non-circular witness (resp., circular witness) for S . Let μ_n (resp., ν_n) denote the maximum length of the shortest non-circular (resp., circular) witness over all representable subsets of Σ^n .

A *de Bruijn word* B_n of order n over the alphabet Σ is a shortest circular witness for the set Σ^n . It is known [1] that the length of a de Bruijn word of order n over Σ is 2^n .

For convenience, we let $w[i]$ denote the i 'th letter of w and $w[i..j]$ denote the factor of w with length $j - i + 1$ that starts with the i 'th letter of w . Thus $w = w[1..n]$ where $n = |w|$.

3 Bounds on the size of \mathring{R}_n

In this section, we give lower and upper bounds on the size of \mathring{R}_n , both of which are of the form α^{2^n} . Our lower bound has $\alpha = \sqrt{2}$ while our upper bound has $\alpha = \sqrt[4]{10}$. Note that our lower bound also works for the size of R_n , since every circularly representable subset is also representable.

3.1 Lower bound

Our argument for the lower bound derives from constructing a set of circularly representable subsets.

Proposition 1. *Let b_n be any de Bruijn word of order n . Then $|C_{n+1}(b_n)| = 2^n$.*

Proof. Every de Bruijn word of order n is of length 2^n ; thus there are 2^n length- $(n+1)$ factors of b_n (considered circularly). These length- $(n+1)$ factors are pairwise distinct, for if $w \in \Sigma^{n+1}$ appears more than once as a factor of b_n , then $w[1..n]$ appears more than once as a factor of b_n . However, every length- n factor appears only once in b_n , a contradiction. Hence $|C_{n+1}(b_n)| = 2^n$. \square

Lemma 2. *Given a de Bruijn word b_n , let Y denote the set $\Sigma^{n+1} \setminus C_{n+1}(b_n)$. For any $y \in Y$, the set $\{y\} \cup C_{n+1}(b_n)$ is circularly witnessed by a word w for which both the length- 2^n prefix and the length- 2^n suffix equal b_n .*

Proof. We construct such a witness for $\{y\} \cup C_{n+1}(b_n)$.

Let $t = b_n b_n b_n b_n$. Let $y_1 = y[1..n]$ and $y_2 = y[2..n+1]$. Let i_1 denote the index of the first occurrence of y_1 in t ; namely, the index i_1 is the minimal integer such that $y_1 = t[i_1..i_1+n-1]$. Let i_2 denote the index of the last occurrence of y_2 in t ; namely, the index i_2 is the maximal integer such that $y_2 = t[i_2..i_2+n-1]$.

We argue that the first occurrence of y_1 does not overlap the last occurrence of y_2 . We have $i_1 \leq 2^n$, since every possible factor of length n appears in the circular word b_n . Similarly, we obtain $i_2 > 3 \cdot 2^n - n$. Thus we have

$$i_1 + n - 1 - i_2 < -2 \cdot 2^n + 2n - 1 < 0,$$

and hence the first occurrence of y_1 does not overlap the last occurrence of y_2 .

Now consider the circular word

$$t_y = b_n b_n t[1..i_1 - 1] t[i_1..i_1 + n - 1] t[i_2 + n - 1] t[i_2 + n..2^{n+2}] b_n b_n.$$

We argue that t_y is a witness for $\{y\} \cup C_{n+1}(b_n)$. For one direction, every element of $\{y\} \cup C_{n+1}(b_n)$ appears as a length- $(n+1)$ factor of t_y . This is a consequence of the following two facts:

1. $b_n b_n$ witnesses $C_{n+1}(b_n)$.
2. $t[i_1..i_1 + n - 1] t[i_2 + n - 1] = y[1..n] y[n + 1] = y$.

For the other direction, we can see that all factors of length $n+1$ in t_y are elements of $\{y\} \cup C_{n+1}(b_n)$ by inspection. Note that the length- 2^n prefix and the length- 2^n suffix of t_y both equal b_n . Hence we conclude that there exists a word for which the prefix and the suffix equal b_n and this circular word circularly witnesses $\{y\} \cup C_{n+1}(b_n)$. \square

Example 3. Let $n = 2$. One of the de Bruijn words of order 2 is $b_2 = 0011$. We have $C_3(b_2) = \{001, 011, 110, 100\}$. Thus $Y = \{000, 010, 101, 111\}$. Let $y =$

010. The following circular word demonstrates that the set $\{y\} \cup C_{n+1}(b_n)$ is representable:

$$t_{010} = \underbrace{(00110011)}_{b_2 b_2} \underbrace{(\quad 0 \quad)}_{t[1..i_1-1]} \underbrace{(\quad 01 \quad)}_{t[i_1..i_1+n-1]=y_1} \underbrace{(\quad 0 \quad)}_{t[i_2+n-1]} \underbrace{(\quad 011 \quad)}_{t[i_2+n..2^{n+2}]} \underbrace{(00110011)}_{b_2 b_2}.$$

Proposition 4. *Given a de Bruijn word b_n , let Y denote the set $\Sigma^{n+1} \setminus C_{n+1}(b_n)$. For any subset $S \subseteq Y$, the set $S \cup C_{n+1}(b_n)$ is a circularly representable subset of Σ^{n+1} .*

Proof. We have proved this proposition for the case where $|S| = 1$ by Lemma 2. Now we turn to the general case. Let $S = \{s_1, s_2, \dots, s_m\}$. By Lemma 2, for each $1 \leq i \leq m$, there exists a circular word t_i that witnesses $\{s_i\} \cup C_{n+1}(b_n)$ and both the prefix and the suffix of t_i equal b_n . We argue that the circular word $t_S = t_1 t_2 \cdots t_m$ witnesses $S \cup C_{n+1}(b_n)$.

First, for any $1 \leq i \leq m$, s_i appears in t_i and thus in t_S . Moreover, every element of $C_{n+1}(b_n)$ appears in the prefix of t_S : $b_n b_n$. Thus, it suffices to show that every length- $(n+1)$ factor of t_S is a member of $S \cup C_{n+1}(b_n)$. This is shown by the fact that for any $1 \leq i < m$, both the suffix of t_i and the prefix of t_{i+1} equal b_n , which implies that the concatenation of t_i and t_{i+1} does not produce any new factor of length $n+1$ in t_S .

Thus, we conclude that for any subset S of Y , there exists a witness for the set $S \cup C_{n+1}(b_n)$. \square

Corollary 5. *A lower bound for the size of \mathring{R}_{n+1} is $2^{2^n} = \sqrt{2}^{2^{n+1}}$.*

3.2 Upper bound

An obvious upper bound for $|\mathring{R}_n|$ is 2^{2^n} , since $\mathring{R}_n \subseteq 2^{\Sigma^n}$, where $|2^{\Sigma^n}| = 2^{2^n}$. In this section, we will show that a tighter upper bound is α^{2^n} , where $\alpha = \sqrt[4]{10}$.

Definition 6. *Let $S \subseteq \Sigma^{n+1}$ and $T \subseteq \Sigma^n$. We say that S is incident on T if there exists a circular word w such that w witnesses both S and T .*

Example 7. For example, we fix $n = 4$. Let $w = 0110$. Then w is a witness for the set $S = \{0110, 1100, 1001, 0011\} \in \mathring{R}_4$ and $T = \{011, 110, 100, 001\} \in \mathring{R}_3$. It follows that S is incident on T . Note that $w' = 01100110$ is also a witness for S , and a witness for T as well.

In fact we can argue that if S is incident on T , then every word that witnesses S also witnesses T .

Proposition 8. *Every set $S \in \mathring{R}_{n+1}$ is incident on exactly one set in \mathring{R}_n .*

Proof. Let $T = \{t \in \Sigma^n : \exists w \in S \text{ such that } t \text{ is a length-}n \text{ prefix or suffix of } w\}$. Then a word w which witnesses S also witnesses T . Thus S is incident on T . Moreover, if S is incident on T and T' , then every witness of S must also witness T and T' . Thus we have $T = T'$. So we conclude that every set $S \in \mathring{R}_{n+1}$ is incident on exactly one set in \mathring{R}_n . \square

Now we give a partition of \mathring{R}_{n+1} . Let

$$\mathring{R}_{n+1}[T] = \{S \in \mathring{R}_{n+1} : S \text{ is incident on } T\}.$$

Proposition 8 implies that $\{\mathring{R}_{n+1}[T]\}_{T \in \Sigma^n}$ is a pairwise disjoint partition of the set \mathring{R}_{n+1} . Namely, (1) for every $T_1 \neq T_2$, we have $\mathring{R}_{n+1}[T_1] \cap \mathring{R}_{n+1}[T_2] = \emptyset$ and (2) $\bigcup_{T \in \mathring{R}_n} \mathring{R}_{n+1}[T] = \mathring{R}_{n+1}$.

Thus we have $|\mathring{R}_{n+1}| = \sum_{T \in \Sigma^n} |\mathring{R}_{n+1}[T]|$. So to give an upper bound for $|\mathring{R}_{n+1}|$, it suffices to give an upper bound for the size of $\mathring{R}_{n+1}[T]$.

Definition 9. Let x be a word of length n . We say that $P_x = \{0x, 1x\}$ is a pair of order n w.r.t x , that $S_x = \{0x, 1x, x0, x1\}$ is a skeleton of order n w.r.t. x , and $N_x = \{0x0, 0x1, 1x0, 1x1\}$ is a net of order n w.r.t x . We also say that a set S contains P_x (resp., S_x and N_x) if $P_x \subseteq S$ (resp., $S_x \subseteq S$ and $N_x \subseteq S$).

For any $T \subseteq \Sigma^n$, let $\sigma(T)$ denote the number of skeletons of order $n - 1$ in T and let $\rho(T)$ denote the number of pairs of order $n - 1$ in T . We have the following proposition:

Proposition 10. For any $T \subseteq \Sigma^n$, we have $|\mathring{R}_{n+1}[T]| \leq 7^{\sigma(T)}$.

Before giving the proof for Proposition 10, we introduce another definition.

Definition 11. A set R is feasible for a set $T \subseteq \Sigma^n$ if there exists $S \in \mathring{R}_{n+1}[T]$ such that $R \subseteq S$.

We observe that $\Sigma^{n+1} = \bigcup_{x \in \Sigma^{n-1}} N_x$ and thus any subset $S \in \Sigma^{n+1}$ is a disjoint union of subsets of nets of order $n - 1$. Formally, for any subset $S \in \Sigma^{n+1}$, we have $S = \bigcup_{x \in \Sigma^{n-1}} R_x$, where $R_x \subseteq N_x$.

Proof (of Proposition 10). Let F_x denote the set of feasible subsets (for T) of the net N_x . If $S \in \mathring{R}_{n+1}[T]$, then S is a disjoint union of feasible subsets (for T) of nets. Thus we have $|\mathring{R}_{n+1}[T]| \leq \prod_{x \in \Sigma^n} |F_x|$. In order to prove this proposition, it now suffices to show that for any $x \in \Sigma^{n-1}$, the following condition holds.

- if $S_x \subseteq T$, then $|F_x| \leq 7$;
- otherwise $|F_x| \leq 1$.

For any $x \in \Sigma^{n-1}$, we consider all the possible feasible subsets of N_x . Let F denote any feasible subset of N_x .

- For the first case where $S_x \subseteq T$, we have the following properties:
 1. Either $0x0 \in F$ or $0x1 \in F$ since $0x \in T$;
 2. Either $1x0 \in F$ or $1x1 \in F$ since $1x \in T$;
 3. Either $0x0 \in F$ or $1x0 \in F$ since $x0 \in T$;
 4. Either $0x1 \in F$ or $1x1 \in F$ since $x1 \in T$.

Hence we have at most 7 possible feasible subsets of N_x which are listed as follows: $\{0x0, 1x1\}$, $\{0x0, 0x1, 1x1\}$, $\{0x0, 1x0, 1x1\}$, $\{0x0, 0x1, 1x0, 1x1\}$, $\{0x0, 0x1, 1x0\}$, $\{0x1, 1x0\}$, $\{0x1, 1x0, 1x1\}$. Thus $|F_x| \leq 7$.

- For the second case where $S_x \not\subseteq T$, we argue that $|F_x| \leq 1$. Without loss of generality, suppose $0x \notin T$. It follows that:
 1. $0x0$ and $0x1$ cannot occur in F since $0x \notin T$;
 2. $1x0 \in F$ if and only if $x0 \in T$;
 3. $1x1 \in F$ if and only if $x1 \in T$;
 Hence, F is fixed. It follows that $|F_x| \leq 1$.

By finishing the argument on the above two cases, we conclude that $|\mathring{R}_{n+1}[T]| \leq 7^{\sigma(T)}$. \square

Now, we are close to the core part. Instead of computing the number of skeletons, which is quite complex, we consider the number of pairs.

Proposition 12. *The size of the set $|\mathring{R}_{n+1}|$ is bounded by $10^{2^{n-1}}$.*

Proof. Let $L_{k,i}$ denote the number of subsets $T \in \mathring{R}_n$, such that $|T| = k$ and $\rho(T) = i$. There are in total 2^{n-1} pairs in Σ^n , and we first choose i 's pairs from them. Then, we choose the other $k - 2i$ elements which do not form any pair from the remaining $2^{n-1} - i$ elements. Thus, we have

$$L_{k,i} = \binom{2^{n-1}}{i} \binom{2^{n-1} - i}{k - 2i} 2^{k-2i}.$$

Note that $k \geq 2i$ since a set of k elements can contain at most $\frac{k}{2}$ pairs and the term $L_{k,i}$ vanishes when $k - 2i > 2^{n-1} - i$. Thus we have

$$|\mathring{R}_{n+1}| = \sum_{T \in \Sigma^n} |\mathring{R}_{n+1}[T]| \leq \sum_{k=0}^{2^n} \sum_{i=0}^{\frac{k}{2}} L_{k,i} 7^i.$$

The inequality holds since we count the number of pairs instead of the number of skeletons and the number of pairs is always greater than or equal to the number of skeletons. Then we can see that

$$|\mathring{R}_{n+1}| \leq \sum_{k=0}^{2^n} \sum_{i=0}^{\frac{k}{2}} \binom{2^{n-1}}{i} \binom{2^{n-1} - i}{k - 2i} 2^{k-2i} 7^i \leq \sum_{i=0}^{2^{n-1}} \binom{2^{n-1}}{i} 7^i \sum_{k=2i}^{2^n} \binom{2^{n-1} - i}{k - 2i} 2^{k-2i}$$

by writing $L_{k,i}$ in closed form. Note that

$$\sum_{k=2i}^{2^n} \binom{2^{n-1} - i}{k - 2i} 2^{k-2i} = \sum_{k=0}^{2^n - 2i} \binom{2^{n-1} - i}{k} 2^k = \sum_{k=0}^{2^{n-1} - i} \binom{2^{n-1} - i}{k} 2^k = 3^{2^{n-1} - i}.$$

So we have

$$|\mathring{R}_{n+1}| \leq \sum_{i=0}^{2^{n-1}} \binom{2^{n-1}}{i} 7^i 3^{2^{n-1} - i} = 10^{2^{n-1}}.$$

\square

Proposition 12 directly implies the upper bound we claimed in the beginning of this section.

4 Shortest witness

Recall that μ_n (resp., ν_n) is the maximum length of the shortest non-circular witness (resp., circular witness) over all subsets of Σ^n . The quantities of μ_n and ν_n are of interest since we can enumerate all sequences of length less than or equal to μ_n (resp., ν_n) in order to list all the representable (resp., circularly representable) subsets of Σ^n . In this section we obtain an upper bound on μ_n and ν_n .

We need the following result of Hamidoune [6, Prop. 2.1]. Since the result is little-known and has apparently not appeared in English, we give the proof here. By a *Hamiltonian walk* we mean a closed walk, possibly repeating vertices and edges, that visits every vertex of G .

Proposition 13. *Let $G = (V, E)$ be a directed graph on n vertices. If G is strongly connected (that is, if there is a directed path from every vertex to every vertex), then there is a Hamiltonian walk of length at most $\lfloor (n+1)^2/4 \rfloor$. Furthermore, this bound is best possible.*

Proof. Let L be a longest simple path in G . (A simple path does not repeat edges or vertices.) Let $V - L = \{v_i : 1 \leq i \leq k\}$. Let v_0 be the last vertex in L and v_{k+1} be the first vertex in L . Let L_i be a simple path from v_i to v_{i+1} . Then a Hamiltonian walk W is obtained by following the edges in L_0, L_1, \dots, L_k , and then those in L . So the number of edges in W is at most $(k+2)|L| = |L|(n+1-|L|)$. But it is easy to see that $r(n+1-r)$ is maximized when $r = \lceil n/2 \rceil$, so $r(n+1-r) = \lfloor (n+1)^2/4 \rfloor$, as claimed.

To see that this bound is best possible, consider a graph where there is a directed chain of $\lfloor n/2 \rfloor$ vertices, where the last vertex has a directed edge to $\lceil n/2 \rceil$ other vertices, and each of those vertices have a single directed edge back to the start of the chain. The shortest walk covering all the vertices traverses the chain, then an edge to one of the other vertices, then a single edge back, and repeats this $\lceil n/2 \rceil$ times. The total length is then $(\lfloor n/2 \rfloor + 1)\lceil n/2 \rceil = \lfloor (n+1)^2/4 \rfloor$. So the bound is tight. \square

From this we immediately get

Proposition 14. *An upper bound for μ_n and ν_n is $2^{2n-2} + 2^{n-1}$.*

5 Numerical results

It is not feasible to enumerate every single word to verify whether a subset is circularly representable (or non-circularly representable). For this reason, we exploit ideas from graph theory.

Formally, we define $G_n = (V_n, E_n)$, where

$$V_n = \{(S, u, v) : S \subseteq \Sigma^n \text{ and } u, v \in \Sigma^n\} \text{ and}$$

$$E_n = \{((S, u, v), (S \cup \{x\}, u, x)) : S \subseteq \Sigma^n, u, v, x \in \Sigma^n, \text{ and } v[2..n] = x[1..n-1]\}.$$

We say that a node (S, u, v) is *valid* if S is witnessed by a non-circular word w for which the length- n prefix is u and the length- n suffix is v .

We use a breadth-first search strategy to compute all the possible valid nodes in G_n . Let I denote a subset of nodes $\{(\{u\}, u, u) : a \in \Sigma^n\}$ in G_n . Nodes in G_n that are connected to any node in I can be proven valid by induction. Thus, a breadth-first search begins with the subset I and enumerates all nodes that are connected to nodes in I .

The relation between valid nodes in G_n and non-empty representable subsets of order n is that any subset $S \subseteq \Sigma^n$ is representable if and only if there exist $u, v \in \Sigma^n$ such that (S, u, v) is valid. This relation can be proved by induction. Similarly, any subset $S \subseteq \Sigma^n$ is circularly representable if and only if there exists $u \in \Sigma^n$ such that (S, u, u) is valid and the minimum distance d between (S, u, u) and nodes in I satisfies the inequality $d \geq n - 1$.

With the above properties, we can enumerate all the possible non-empty representable (or circularly representable) subsets of order n . Our results are shown in the following table. The last two columns give words w of length ν_n (resp., μ_n) for which no shorter word witnesses $C_n(w)$ (resp., $F_n(w)$).

n	$ R_n $	$ R_n $	ν_n	μ_n	longest circ. witness	longest witness
1	3	3	2	2	01	01
2	6	14	4	5	0011	00110
3	27	121	9	10	000100111	0001011100
4	973	5921	24	24	000010001011100011101111	000010010101100101101111
5	2466131	20020315	82	77	—	—

6 Fixed-length witnesses

We now turn to a related question. We fix a length n and we ask, how many different subsets of Σ^n can we obtain by taking the (ordinary, non-circular factors) of a word of length t ? We call this quantity $T(t, n)$. As we will see, for $t < 2n$, there is a relatively simple answer to this question.

In order to compute $T(t, n)$, we consider the number of words that witness the same subset of Σ^n . Suppose $S \subseteq \Sigma^n$. Let $C_t(S)$ denote the number of words of length t that witness S . Then we have

$$T(t, n) = 2^t - \sum_{\substack{S \subseteq \Sigma^n \\ C_t(S) > 1}} (C_t(S) - 1).$$

It suffices to characterize what subsets S satisfy $C_t(S) > 1$ and to determine $C_t(S)$.

For $t < 2n$, we have such a characterization by Theorem 15 below. Before stating the proposition, we first introduce some notation.

Let w be a word. Let $\text{Pref}(w)$ denote the set of prefixes of w . A *period* p of w is a positive integer such that w can be factorized as

$$w = s^k s', \text{ with } |s| = p, s' \in \text{Pref}(s), \text{ and } k \geq 1.$$

Let $\pi(w)$ denote the minimal period of w .

The *root* of a word w is the prefix of w with length $\pi(w)$. Let $r(w)$ denote the root of w . Two words w and w' are *conjugate* if there exist $u, v \in \Sigma^*$ such that $w = uv$ and $w' = vu$; w and w' are *root-conjugate* if their roots $r(w)$ and $r(w')$ are conjugate.

The following theorem is crucial for our work and of independent interest.

Theorem 15. *Let t, n, k be such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. Let w and w' be distinct words of length t over an arbitrary alphabet. Then $F_n(w) = F_n(w')$ iff $\pi(w) = \pi(w') \leq k + 1$ and w, w' are root-conjugate.*

One direction is easy: if w and w' are root-conjugate with period $p \leq k + 1$, then there are p places to begin, and considering consecutive factors of length $n + p - 1$ gives exactly p distinct length- n factors.

For the other direction, we need three lemmas.

Lemma 16. *(Fine-Wilf theorem [3, Theorem 1]) Let w_1, w_2 be two words. If w_1 and w_2 have a common prefix of length $\pi(w_1) + \pi(w_2) - 1$, then $r(w_1) = r(w_2)$.*

Lemma 17. *For any $w \in \Sigma^+$, if there exists a factorization $w = xyz$ such that $xy = yz$ and $x, y, z \in \Sigma^+$, then w is periodic with $\pi(w) \leq |x|$.*

Proof. By the Lyndon-Schützenberger theorem [5, Lemma 2], there exist $u \in \Sigma^+, v \in \Sigma^*$ and an integer $e \geq 0$ such that $x = uv, y = (uv)^e u, z = vu$. Thus $w = (uv)^{e+2}u$. Thus w is periodic with $\pi(w) \leq |x|$. \square

Lemma 18. *Let t, n, k be integers such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. Let w be a word of length t with $\pi(w) \leq k + 1$. If w' is any word such that $F_n(w) = F_n(w')$, then w and w' are root-conjugate.*

Carpi and de Luca proved a stronger proposition [2, Proposition 6.2] which directly implies this lemma. We first introduce some relevant notation from that paper.

A factor s of a word w is said to be *right-special* in w if there exist two distinct symbols a and b such that sa and sb are factors of w . Let R_w denote the minimal length m such that there exists no factor of length m that is right-special.

A factor s of a word w is said to be *right-extendable* (resp., *left-extendable*) in w if there exists a symbol a such that sa is a factor of w (resp., as is a factor of w). Let K_w and H_w denote the length of the shortest factor which is not right-extendable (resp., left-extendable).

A word is *semiperiodic* if $R_w < H_w$.

Proof (of Lemma 18). Carpi proved [2, Lemma 3.2] that $\pi(w) > R_w$. Also, we have $H_w \geq \pi(w)$ since the length- $(\pi(w) - 1)$ prefix of w is left-extendable. Thus w is semiperiodic. Moreover we have $F_n(w) = F_n(w')$ where $n \geq k + 1 \geq \pi(w) \geq 1 + R_w$. Then we can apply [2, Proposition 6.2] to prove this lemma. \square

Proof (of Theorem 15). We give a proof for Theorem 15 by induction on k .

The base case is when $k = 0$. In this case $t = n$ and thus $F_n(w) = \{w\}$ and $F_n(w') = \{w'\}$. Thus $w = w'$.

Now we deal with the induction step. We assume the result holds for $k - 1$ and we prove it for k . For convenience, we let $p_i(w)$ denote the length- i prefix of the word w ; let $s_i(w)$ denote the length- i suffix of the word w .

We first consider the case where $H_w < n$. We have $p_n(w) \in F_n(w) = F_n(w')$. If $p_n(w) \neq p_n(w')$, then there exists $a \in \Sigma$ such that $ap_{n-1}(w) \in F_n(w')$. Thus we have $ap_{n-1}(w) \in F_n(w)$ which leads to the contradiction that $H_w \geq |ap_{n-1}(w)| = t$. Hence $p_n(w) = p_n(w')$.

Now let $s = w[2..t]$ and $s' = w'[2..t]$. Clearly $|s| = |s'| = t - 1$. The prefix $p_n(w)$ appears only once as a factor of w , otherwise $p_{n-1}(w)$ is left-extendable in w which contradicts the fact that $H_w < n$. Thus we have $F_n(s) = F_n(w) \setminus \{p_n(w)\}$. Similarly we have $F_n(s') = F_n(w') \setminus \{p_n(w)\}$. Thus $F_n(s) = F_n(s')$. Let $k' = k - 1$. We have $t - 1 = n + k - 1 = n + k'$ and $p \geq k + 1 > k' + 1$. By induction, we have either

Case 1: $s = s'$; or

Case 2: s and s' are root-conjugate and $\pi(s) = \pi(s') = \rho$, where $\rho \leq k' + 1 = k$.

In Case 1, it follows that $w = w'$, contradicting the fact that w, w' are distinct. In Case 2, we prove that $s = s'$ by showing that their roots are identical. Suppose s and s' have a common prefix of length d . We have $d \geq n - 1$, since w and w' have a common prefix of length at least n . If $d \geq \rho$, then the root of s is identical to the root of s' . Otherwise, we have the chain of inequalities $k \geq \rho \geq d + 1 \geq n \geq k + 1$, which is trivially a contradiction. Thus neither Case 1 nor Case 2 can occur and we are done with the case where $H_w < n$.

Similarly we can prove the induction step when $K_w < n$. Thus it suffices to consider the case where $H_w \geq n$ and $K_w \geq n$. We first claim $\pi(w) \leq k + 1$. There are several cases to settle:

- The first case is when $p_{n-1}(w) = s_{n-1}(w)$ and the occurrence of $p_{n-1}(w)$ and $s_{n-1}(w)$ do not overlap; namely we have $w = p_{n-1}(w)Lp_{n-1}(w)$, where $L \in \Sigma^*$. We have the inequality $n + k = t = |w| = 2|p_{n-1}(w)| + |L| = 2(n - 1) + |L|$. Thus $|L| = k + 2 - n$. Hence $\pi(w) \leq |p_{n-1}(w)L| = n - 1 + k + 2 - n = k + 1$.
- The second case is when $p_{n-1}(w) = s_{n-1}(w)$ and these occurrences overlap. Formally we put it as follows: there exist $x, y, z \in \Sigma^+$, such that $p_{n-1}(w) = xy = yz$ and $w = xyz$. It follows that $\pi(w) \leq |x| \leq k + 1$ by Lemma 17.
- The last case is when $p_{n-1}(w) \neq s_{n-1}(w)$. Let i_p denote the index of the last occurrence of $p_{n-1}(w)$; namely $i_p = \sup\{i \geq 0 : p_{n-1}(w) = w[i..i + n - 2]\}$. Note that $i_p > 0$ since $p_{n-1}(w)$ is left-extendable and $i_p \leq t - n + 2$ since $p_{n-1}(w) \neq s_{n-1}(w)$. Thus, the first occurrence of $p_{n-1}(w)$ (the prefix of w) overlaps the last occurrence of $p_{n-1}(w)$. By Lemma 17, we get that $w_1 = w[1..i_p + n - 2]$ is periodic with $\pi(w_1) \leq i_p - 1$. Similarly we let i_q denote the index of the first occurrence of $s_{n-1}(w)$ and $w_2 = w[i_q..t]$. We have $0 < i_q \leq t - n + 2$ and $\pi(w_2) \leq t - n + 2 - i_q$. The factors w_1 and w_2 overlap

for at least $|w_1| + |w_2| - t \geq \pi(w_1) + \pi(w_2) - 1$ symbols. Let D denote the overlap of w_1 and w_2 . We have $|D| \geq \pi(w_1) + \pi(w_2) - 1$. Also $\pi(w_1)$ is a period of D since $|D| \geq \pi(w_1)$ and D can be factorized as

$$D = d^l d', \text{ where } d \text{ is conjugate to the root of } w_1, d' \in \text{Pref}(d), \text{ and } l \geq 1.$$

By Lemma 16, the overlap D has the same root as w_2 . Since root-conjugacy is an equivalence relation, we have w_1 and w_2 are root-conjugate. It follows that w is periodic with $\pi(w) = \pi(w_1) \leq k + 1$.

Finally by Lemma 18, we get that w and w' are root-conjugate and their periods $\pi(w) = \pi(w') \leq k + 1$. By all cases, we finish the induction and complete the proof of Theorem 15. \square

The following corollary gives $T(t, n)$ when $t < 2n$.

Corollary 19. *For $n \leq t < 2n$, we have $T(t, n) = 2^t - \sum_{k=1}^{t-n+1} \frac{k-1}{k} \sum_{d|k} \mu\left(\frac{k}{d}\right) 2^d$, where $\mu(\cdot)$ is the Möbius function.*

Proof. Let $k = t - n$. We have $n \geq t - n + 1 = k + 1$. By Theorem 15, we know that for any set $S \subseteq \Sigma^n$, $C_t(S) > 1$ if and only if there exists a word w that witnesses S with $\pi(w) \leq k + 1$. In this case we have $C_t(S) = \pi(w)$; that is, the set of words that witness S is the same as the set of the words that are root-conjugate to w . Thus each S such that $C_t(S) > 1$ corresponds to a set of root-conjugate words, which can be represented by their lexicographically least roots (the Lyndon words).

Thus we have

$$\begin{aligned} T(t, n) &= 2^t - \sum_{\substack{S \subseteq \Sigma^n \\ C_t(S) > 1}} (C_t(S) - 1) = 2^t - \sum_{\substack{w \text{ is a Lyndon word} \\ \pi(w) \leq k+1}} (\pi(w) - 1) \\ &= 2^t - \sum_{i=1}^{k+1} (i - 1) \cdot L(i), \end{aligned}$$

where $k = t - n$ and $L(i) = \frac{1}{i} \sum_{d|i} \mu\left(\frac{i}{d}\right) 2^d$ is the number of Lyndon words of length i . \square

Example 20. To finish this section, we give a table listing some numerical results for $T(t, n)$. The numbers in bold follow from Corollary 19.

$n \backslash t$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
2		4	7	11	12	12	12	12	12	12	12	12	12	12	12	12
3			8	15	27	48	72	94	100	103	101	103	101	103	101	103
4				16	31	59	114	216	391	677	1087	1621	2246	2928	3595	4235
5					32	63	123	242	474	933	1795	3421	6399	11682	20704	35914
6						64	127	251	498	986	1965	3899	7709	15171	29710	57726
7							128	255	507	1010	2010	4013	8001	15969	31789	63256
8								256	511	1019	2034	4058	8109	16193	32367	64671

7 Open Problems and Future Work

- In Section 3, we gave lower and upper bounds on $|\mathring{R}_n|$, both of the form α^{2^n} . Does the limit $\lim_{n \rightarrow \infty} |\mathring{R}_n|^{\frac{1}{2^n}}$ exist?
- Find better bounds for μ_n and ν_n . For example, is $\mu_n \leq (n-1)2^n$ for $n \geq 2$?
- It is easy to see that Theorem 15 fails for $t < k + 1$. Indeed, it is possible to have $F_n(x) = F_n(y)$ in this case, and yet $\pi(x) \neq \pi(y)$. For example, take $n = k - 1$ so that $t = 2k - 1$, and consider $x = 0^k 10^{k-2}$ and $y = 0^{k-1} 10^{k-1}$. Then $F_n(x) = F_n(y)$ but $\pi(x) = k + 1$ and $\pi(y) = k$.
The remaining case is $n = k$, so that $t = 2k$. We conjecture that if x and y are distinct binary words of length $2n$ with $F_n(x) = F_n(y)$ then $\pi(x) = \pi(y)$ and furthermore x and y are root-conjugate. However, it is possible in this case that $\pi(x) > n + 1$. Furthermore it seems that if $\pi(x) > n + 1$, then $x = uv01v^R u$ and $y = uv10v^R u$ (or vice versa) for some nonempty words u, v where u is a palindrome and $\pi(x) = n + |u|$.
As an example, consider $x = 010110$, $y = 011010$. Then $F_3(x) = F_3(y) = \{010, 011, 101, 110\}$ but $\pi(x) = \pi(y) = 5$. Here $u = 0$, $v = 1$.

References

- N. G. de Bruijn. A combinatorial problem. *Nederl. Akad. Wetensch., Proc.* **49** (1946), 758–764. (= *Indagationes Math.* **8** (1946), 461–467.)
- A. Carpi and A. de Luca. Semiperiodic words and root-conjugacy. *Theoret. Comput. Sci.* **292** (2003), 111–130.
- N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965), 109–114.
- J. Gallant, D. Maier, and J. A. Storer. On finding minimal length superstrings. *J. Comput. System Sci.* **20** (1980), 50–58.
- R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.
- Y. O. Hamidoune. Sur les sommets de demi-degré h d’un graphe fortement h -connexe minimal. *C. R. Acad. Sci. Paris Sér. A-B* **286** (1978), A863–A865.
- E. Moreno. De Bruijn sequences and De Bruijn graphs for a general language. *Info. Proc. Letters* **96** (2005), 214–219.