

MAXIMAL STATE COMPLEXITY AND GENERALIZED DE BRUIJN WORDS

DANIEL GABRIC, ŠTĚPÁN HOLUB, AND JEFFREY SHALLIT

ABSTRACT. We compute the exact maximum state complexity for the language consisting of m words of length N , and characterize languages achieving the maximum. We also consider a special case, namely languages $C(w)$ consisting of the conjugates of a single word w . The words for which the maximum state complexity of $C(w)$ is achieved turn out to be a natural generalization of de Bruijn words. We show that generalized de Bruijn words exist for each length and consider the number of them.

1. INTRODUCTION

Let x, y be words. We say x and y are *conjugates* if one is a cyclic shift of the other; equivalently if there exist words u, v such that $x = uv$ and $y = vu$. For example, the English words **listen** and **enlist** are conjugates.

The set of all conjugates of a word w is denoted by $C(w)$. Thus, for example, $C(\text{eat}) = \{\text{eat}, \text{tea}, \text{ate}\}$. We also write $C(L)$ for the set of all conjugates of elements of the language L .

For a regular language L let $\text{sc}(L)$ denote the *state complexity* of L : the number of states in the smallest complete DFA accepting L . State complexity is sometimes also called *quotient complexity* [1]. The state complexity of the cyclic shift operation $L \rightarrow C(L)$ for arbitrary regular languages L was studied in Maslov's pioneering 1970 paper [2]. More recently, Jirásková and Okhotin [3] improved Maslov's bound, and Jirásek and Jirásková studied the state complexity of the conjugates of prefix-free languages [4].

In this note we investigate the state complexity of *uniform-length* languages, that is, of languages $L \subseteq \Sigma^N$. The language $C(w)$ is a special case of a uniform-length language. Clearly $\text{sc}(C(x))$ achieves its minimum — namely, $N + 2$ — at words of the form a^N , where a is a single letter. By considering random words, it seems likely that $\text{sc}(C(x)) = \Theta(N^2)$ in the worst case.

In Theorem 7, we prove an exact bound for the state complexity of (almost all) uniform-length languages and characterize languages that attain the bound. In particular, this means that we determine the state complexity of cyclic shift on languages consisting of a single word. Moreover, the characterization of words w for which $C(w)$ achieves the maximum turns out to be a natural generalization of de Bruijn words to words of arbitrary length. Therefore, in Section 2, we introduce

(A1,A3) SCHOOL OF COMPUTER SCIENCE, UNIVERSITY OF WATERLOO, WATERLOO, ONTARIO N2L 3G1, CANADA

(A2) DEPARTMENT OF ALGEBRA, FACULTY OF MATHEMATICS AND PHYSICS, CHARLES UNIVERSITY, PRAGUE, CZECH REPUBLIC

E-mail addresses: dgabric@uwaterloo.ca, holub@karlin.mff.cuni.cz, shallit@uwaterloo.ca.

the concept of *generalized de Bruijn word* and show that such words exist for each length.

This paper is the journal version of the conference paper [5]. It differs in several respects from that paper: we have reworked the discussion of the necessary concepts from graph theory (in Section 2), providing more details; we have characterized uniform-length languages achieving maximum state complexity in Theorem 7 which includes a corrected statement of Theorem 3 of the conference paper; and we have provided additional enumeration details in Tables 1 and 4.

2. GENERALIZED DE BRUIJN WORDS

De Bruijn words (also called de Bruijn sequences) have a long history [6, 7, 8, 9, 10], and have been extremely well studied [11, 12]. Let Σ_k denote the k -letter alphabet $\{0, 1, \dots, k-1\}$. Traditionally, there are two distinct ways of thinking about these words: for integers $k \geq 2$, $n \geq 1$ they are

- (a) the words w having each word of length n over Σ_k exactly once as a factor; or
- (b) the words w having each word of length n over Σ_k exactly once as a factor, when w is considered as a “circular word”, or “necklace”, where the word “wraps around” at the end back to the beginning.

For example, for $k = 2$ and $n = 4$, the word

0000111101100101000

is an example of the first interpretation and

0000111101100101

is an example of the second.

In this paper, we are concerned with the second (circular) interpretation of de Bruijn words. Obviously, such words exist only for lengths of the form k^n . Is there a sensible way to generalize this class of words so that one could speak fruitfully of (generalized) de Bruijn words of every length?

One natural way to do so is to use the notion of *subword complexity* (also called *factor complexity* or just *complexity*). For $0 \leq i \leq N$ let $\gamma_i(w)$ denote the number of distinct length- i factors of the word $w \in \Sigma_k^N$ (considered circularly). For all words w , there is a natural upper bound on $\gamma_i(w)$ for $0 \leq i \leq N$, as follows:

$$\gamma_i(w) \leq \min(k^i, N). \tag{1}$$

This is immediate, since there are at most k^i words of length i over Σ_k , and there are at most N positions where a word could begin in w (considered circularly).

Ordinary de Bruijn words are then precisely those words w of length k^n for which $\gamma_n(w) = k^n$. But even more is true: a de Bruijn word w also attains the upper bound in (1) for *all* $i \leq k^n$. To see this, note that if $i \leq n$, then every word of length i occurs as a prefix of some word of length n , and every word of length n is guaranteed to appear in w . On the other hand, all k^n (circular) factors of each length $i \geq n$ are distinct, because their length- n prefixes are all distinct.

This motivates the following definition:

Definition 1. A word w of length N over a k -letter alphabet is said to be a *generalized de Bruijn word* if $\gamma_i(w) = \min(k^i, N)$ for all $0 \leq i \leq N$.

Table 1 gives the lexicographically least de Bruijn words for a two-letter alphabet, for lengths 1 to 31, and the number of such words (counted up to cyclic shift). This forms sequence [A317586](#) in the *On-Line Encyclopedia of Integer Sequences* (OEIS) [13].

We point out an alternative characterization of our generalized de Bruijn words.

Proposition 1. *A word $w \in \Sigma_k^N$ is a generalized de Bruijn word iff both of the following hold:*

- (1) $\gamma_r(w) = k^r$; and
- (2) $\gamma_{r+1}(w) = N$,

where $r = \lfloor \log_k N \rfloor$.

Proof. A generalized de Bruijn word trivially has these properties. An argument similar to the discussion before Definition 1 shows that the two properties imply the equality $\gamma_i(w) = \min(k^i, N)$ for all $0 \leq i \leq N$. \square

We now show that generalized de Bruijn words exist. Since one of the most powerful tools for studying de Bruijn words are de Bruijn graphs, we shall need some results from (directed) graph theory. Let us first set the terminology. A closed sequence of edges $(v_0, v_1), (v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_0)$ is called a *cycle* (of length n) if all vertices v_0, v_1, \dots, v_{n-1} are distinct. If all edges in the sequence are distinct (vertices may repeat), the sequence is called a *circuit* (of length n). A cycle that visits all vertices of a graph is called a *Hamiltonian cycle*. A circuit traversing all edges is an *Eulerian circuit*. A directed graph is an *Eulerian graph* if, for each vertex v , the number of edges incoming to v is the same as the number of edges outgoing from v . It is well known that each connected component of an Eulerian graph admits an Eulerian circuit. If, for all vertices, the number of incoming edges, as well as the number of outgoing edges is k , then the graph is said to be *regular of degree $2k$* . The *degree* of a vertex is the total number of its incoming and outgoing edges.

A *factor* (more precisely a 2-factor) of a graph is the set of vertex-disjoint cycles that together cover all vertices. Note, for example, that a Hamiltonian cycle is a special case of a factor. One of the first published results in graph theory is the following fact, proved in [14, Claim 9, p. 200]. (For a more contemporary proof, see, for example, [15, Theorem 3.3.9, p. 140].)

Lemma 2 (Petersen). *Let G be a regular graph of degree $2k$. Then the edges of G can be partitioned into k distinct factors.*

The k -ary de Bruijn graph of order n , denoted G_n^k , is a directed graph where the vertices are the k -ary words of length n , and edges join a word x to a word y if $x = at$ and $y = tb$ for some letters a, b and a word t . An ordinary de Bruijn word $a_0 a_1 \dots a_{k^n - 1}$ of length k^n can be represented by the cycle $(v_0, v_1), (v_1, v_2), (v_2, v_3), \dots, (v_{k^n - 1}, v_0)$ where $v_i = a_i a_{i+1} \dots a_{i+n-1}$, indices taken modulo k^n . This establishes a one-to-one correspondence between Hamiltonian cycles of G_n^k and de Bruijn words of length k^n . Similarly, there is a one-to-one correspondence between such words and Eulerian circuits in G_{n-1}^k of the form $(v'_0, v'_1), (v'_1, v'_2), (v'_2, v'_3), \dots, (v'_{k^n - 1}, v'_0)$ where $v'_i = a_i a_{i+1} \dots a_{i+n-2}$, indices again taken modulo k^n . More generally, edges in G_{n-1}^k are in one-to-one correspondence with vertices of G_n^k , where the edge (at, tb) corresponds to the vertex atb . Circuits in G_{n-1}^k then correspond to cycles in G_n^k .

Every vertex of G_n^k has k incoming edges, and k outgoing edges, and therefore G_n^k is a regular graph of degree $2k$. The fact that such a graph is Eulerian yields the existence of ordinary de Bruijn words. By Proposition 1, it also becomes clear that building a generalized de Bruijn word of length $N = k^n + j$, where $0 \leq j \leq (k-1)k^n$, over a k -letter alphabet amounts to constructing a circuit of length N in G_n^k that visits every vertex.

The existence of generalized de Bruijn words of any length is almost proved in a paper by Lempel [16]. Lempel proved that for all $k \geq 2$, $n \geq 1$, $N \leq k^{n+1}$, there exists a circular word $w = w(k, n, N)$ of length N for which the factors of size n are distinct. (Also see [17, 18].) In other words, Lempel shows the existence of a connected Eulerian graph with N edges in G_n^k . However, his proof does not explicitly state that the circuit visits all vertices if $k^n \leq N$. The resulting word therefore satisfies condition (2) of Proposition 1, but not necessarily condition (1). For example, the binary word 10011110000 of length 11 contains 11 distinct circular factors of length 4, but only 7 factors of length 3: the factor 101 is missing (see Figure 1).

A further analysis of Lempel's construction nevertheless reveals that this additional required property is satisfied. For sake of completeness, we reconstruct the proof below. In fact, our proof more closely follows the proof by Yoeli [19] for the binary case, which, in turn, was followed by Lempel. (A similar analysis of Yoeli's proof in the binary setting can be found in [20].)

The core of the proof are the following two facts about de Bruijn graphs.

Lemma 3. *Let $k \geq 2$ and $n \geq 1$. Then every cycle in G_n^k can be completed to a factor.*

Proof. For $n = 1$, the graph G_1^k contains a loop, i.e., the edge (a, a) , for each vertex a where a is a letter. A given cycle C can be therefore completed with loops in vertices that are not contained in C .

Let $n \geq 2$ and let C be a cycle in G_n^k . Consider the complement H of the connected Eulerian graph corresponding to C in G_{n-1}^k . The graph H is Eulerian, and the cycles in G_n^k corresponding to Eulerian circuits of connected components of H together with C form a factor of G_n^k . \square

Lemma 4. *Let H' be an Eulerian subgraph of G_n^k in which each vertex of G_n^k has degree at least two. Then there exists a connected Eulerian subgraph H of G_n^k in which each vertex has the same degree as in H' . In particular, the number of edges in H is the same as in H' .*

Proof. Suppose that H' is not connected and proceed by induction on the number of its connected components. There exist vertices at and tb in G_n^k , where a and b are letters, such that $at \in C_1$ and $tb \in C_2$, where C_1 and C_2 are two distinct connected components of H' . Let (at, tc) be an edge in C_1 and (dt, tb) be an edge in C_2 . Define H'_1 by replacing edges (at, tc) and (dt, tb) in H' with edges (at, tb) and (dt, tc) . The graph H'_1 satisfies the hypothesis of the lemma and has a strictly smaller number of connected components. Moreover, the degrees of all vertices are not affected by the exchange of edges. This completes the proof. \square

We can now reprove [16, Theorem 1] (see also [19, Theorem A and Theorem B]) in the form suitable for our purposes.

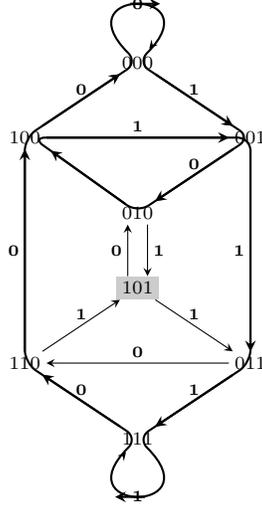


FIGURE 1. The circuit representing the word 10011110000 in G_3^2

Theorem 5. *Let $k \geq 2$ and $n \geq 1$. Then for every N , $0 < N \leq k^{n+1}$, the graph G_n^k contains a connected Eulerian graph H , with N edges and $\min\{k^n, N\}$ vertices. In other words, H is a cycle if $N \leq k^n$, and H contains all vertices of G_n^k otherwise.*

Proof. We proceed by induction on n . Let first $0 < N \leq k^n$. If $n = 1$, then G_1^k contains a cycle of length N , since G_1^k is the clique on k vertices (with loops). If $n > 1$, then, by the induction hypothesis, the graph G_{n-1}^k contains a circuit of length N , which corresponds to a cycle of length N in G_n^k .

Let now $N = jk^n + N'$ where $1 \leq j \leq k - 1$ and $0 < N' \leq k^n$. Let C be a cycle in G_n^k of length N' obtained in the previous paragraph, and let $F_1 = \{C, C_1, \dots, C_m\}$ be a factor of G_n^k obtained by Lemma 3. The complement of F_1 is a regular graph of degree $2k - 2$, whose edges can be partitioned into $k - 1$ factors F_2, F_3, \dots, F_k by Lemma 2. The edges of $C, F_2, F_3, \dots, F_{j+1}$ together yield an Eulerian graph H' with N edges. Each vertex of G_n^k has degree at least two in H' . We obtain H from H' using Lemma 4. \square

We therefore have proved the desired result.

Corollary 1. *For all integers $k \geq 2$ and $N \geq 1$ there exists a generalized de Bruijn word of length N over a k -letter alphabet.*

Remark. We have not been able to find this precise notion of generalized de Bruijn word in the literature anywhere, although there are some papers that come very close. For example, Iványi [21] considered the analogue of the upper bound (1) for ordinary (non-circular) words. He called a word w *supercomplex* if the bound is attained not only for w , but also for all prefixes of w . However, binary supercomplex words do not exist past length 9. The third author also considered the analogue of the bound (1) for ordinary words [20]. However, Lemma 3 of that paper actually implies the existence of our generalized (circular) de Bruijn words of every length over a binary alphabet, although this was not stated explicitly. Anisiu, Blázsik, and

Kása [22] discussed a related concept: namely, those length- N words w for which $\max_{1 \leq i \leq N} \rho_i(w) = \max_{x \in \Sigma_k^N} \max_{1 \leq i \leq N} \rho_i(x)$ where $\rho_i(w)$ denotes the number of distinct length- i factors of w (here considered in the ordinary sense, not circularly). Also see [23].

3. STATE COMPLEXITY

In this section we show that a generalized de Bruijn word can be characterized as a word w with the maximum state complexity of $C(w)$. To this end, we first consider a more general setting of languages $L \subseteq \Sigma^N$. In other words, L is a language containing some words of length N only.

N	lexicographically least generalized binary de Bruijn word of length N	number of such words
1	0	2
2	01	1
3	001	2
4	0011	1
5	00011	2
6	000111	3
7	0001011	4
8	00010111	2
9	000010111	4
10	0000101111	3
11	00001011101	6
12	000010100111	13
13	0000100110111	12
14	00001001101111	20
15	000010011010111	32
16	0000100110101111	16
17	00000100110101111	32
18	000001001101011111	36
19	0000010100110101111	68
20	00000100101100111101	141
21	000001000110100101111	242
22	0000010001101001011111	407
23	00000100011001110101111	600
24	000001000110010101101111	898
25	0000010001100101011011111	1440
26	00000100011001010011101111	1812
27	000001000110010100111011111	2000
28	0000010001100101001110101111	2480
29	00000100011001010011101011111	2176
30	000001000110010110100111011111	2816
31	0000010001100101001110101101111	4096

TABLE 1. Generalized de Bruijn words

The following theorem determines the maximum state complexity of such a language for sufficiently large N , and characterize languages that achieve the maximum. Let $\pi_i(L)$ (resp., $\sigma_i(L)$) denote the number of prefixes (resp., suffixes) of length i of the language L .

Theorem 6. *Let Σ be an alphabet of cardinality $k \geq 2$, let $N \geq 1$ be an integer, and let $L \subseteq \Sigma^N$. Define $m = |L|$ and $r = \lfloor \log_k |L| \rfloor$ and $v = 1 + k + k^2 + \dots + k^r$. If $N \geq 3r + 1$, then*

$$\text{sc}(L) \leq 2v + m \cdot (N - 2r - 1) + 1. \quad (2)$$

If $N > 3r + 1$, then equality holds in (2) if and only if both of the following two conditions are satisfied:

- (a) $\sigma_r(L) = \pi_r(L) = k^r$
- (b) $\sigma_{r+1}(L) = \pi_{r+1}(L) = m$.

Proof. We use the standard construction of the minimal automaton \mathcal{A} accepting L as follows. The states $S_{\mathcal{A}}$ of \mathcal{A} are left quotients \bar{p} of the language L , where

$$\bar{p} = \{s \mid ps \in L\}.$$

Note that all elements in the state \bar{p} have the same length $N - |p|$. We divide the states of \mathcal{A} into subsets according to the length of words they contain, as follows:

$$S_{\mathcal{A}} = A \cup M \cup \bigcup_{\ell=1}^r T_{\ell} \cup \{f\} \cup \{\emptyset\}$$

where

- $A = \{\bar{p} \mid |p| \leq r\}$,
- $M = \{\bar{p} \mid r < |p| < N - r\}$,
- $T_{\ell} = \{\bar{p} \mid |p| = N - \ell\}$,
- $f = \{\varepsilon\}$.

The state f is the accepting state, and \emptyset is the “dead” state. For the size of A we have a bound $v = 1 + k + k^2 + \dots + k^r$, since v is the number of words p that can define a state \bar{p} .

Let $d = m - \pi_{N-r-1}$. For each length $r < \ell < N - r$, there are at most π_{N-r-1} words p of length ℓ such that \bar{p} is nonempty—namely, the prefixes of L of length ℓ . Therefore the size of M is at most $(m - d) \cdot (N - 2r - 1)$.

For T_{ℓ} , $1 \leq \ell \leq r$, we need a more detailed analysis, which exhibits a trade-off between the size of T_{ℓ} and the size of M . More precisely, we shall show that large T_{ℓ} implies large d . Consider the set T_{ℓ} for some fixed $1 \leq \ell \leq r$. Every state $\bar{p} \in T_{\ell}$ is a set of words of length ℓ . “Expected” elements of T_{ℓ} are singletons $\{s\}$, with $|s| = \ell$, which yields an “expected” size k^{ℓ} of T_{ℓ} . Assume that T_{ℓ} contains a state \bar{p} with cardinality d_p larger than one, say $\bar{p} = \{s_1, s_2, \dots, s_{d_p}\}$. Then L contains words $ps_1, ps_2, \dots, ps_{d_p}$, all having the same prefix of length $N - r - 1$. This implies that d is at least $d_p - 1$. Moreover, the contribution to d is cumulative. Indeed, assume that $\bar{p}' = \{s'_1, s'_2, \dots, s'_{d_{p'}}\}$ with $d_{p'} > 1$ for some $\bar{p} \neq \bar{p}' \in T_{\ell}$. Then $p's'_1, p's'_2, \dots, p's'_{d_{p'}}$ are pairwise distinct words in L with the same prefix of length $N - r - 1$, and they are also all distinct from any $ps \in L$. Altogether we have (still with a fixed ℓ)

$$d \geq \sum_{\bar{p} \in T_{\ell}} (d_p - 1),$$

and the size of T_ℓ is at most $k^\ell + d$. Therefore, the set $T = \bigcup_{\ell=1}^r T_\ell$ has size at most $k + k^2 + \dots + k^r + dr$.

We have shown that

$$(3) \quad \begin{aligned} \text{sc}(L) &\leq v + (m - d)(N - 2r - 1) + (v - 1 + dr) + 2 = \\ &= 2v + m(N - 2r - 1) + 1 - d(N - 3r - 1), \end{aligned}$$

which proves the bound, due to the assumption $N \geq 3r + 1$.

To show the second half of the theorem, note that (3) and $N > 3r + 1$ imply $d = 0$ if the equality holds in (2). Therefore states in T are all singletons, and all bounds in the above description have to be achieved. Then the automaton has the topology depicted in Figure 2 and the two conditions are satisfied.

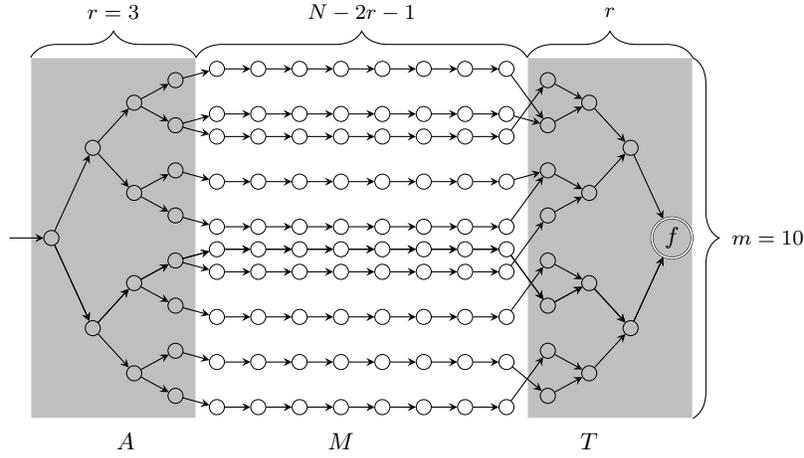


FIGURE 2. Example of the maximum automaton topology

Now assume that conditions (a) and (b) of the theorem are satisfied. Let p be a prefix of a word in L with $|p| > r$, and assume that $s_1, s_2 \in \bar{p}$ for two distinct words s_1 and s_2 . Then $ps_1, ps_2 \in L$ have the same prefix of length $r + 1$, a contradiction with $\pi_{r+1} = m$. Therefore all \bar{p} in $M \cup T$ are singletons. From $\sigma_r = k^r$ we now deduce that $T_\ell = \{\{s\} \mid \ell = |s|\}$ for each $\ell = 1, 2, \dots, r$, and T has size $k + k^2 + \dots + k^r$.

Let p_1 and p_2 be two distinct prefixes in L of length at most $N - r - 1$ such that some s is in both \bar{p}_1 and \bar{p}_2 , which are states in $A \cup M$. Then p_1s and p_2s are two distinct words in L with the same suffix of length $r + 1$, a contradiction with $\sigma_{r+1} = m$. Therefore states in $A \cup M$ are pairwise disjoint. From $\pi_{r+1} = m$ we deduce that L has m distinct prefixes for each size $r < \ell < N - r$, hence the size of M is $m \cdot (N - 2r - 1)$. Finally, from $\pi_r = k^r$ we obtain that A contains v distinct states. The “dead” state \emptyset completes the bound. \square

In the conference version of our paper we mistakenly claimed that Theorem 6 holds for $N \geq 2r + 1$ instead of $N \geq 3r + 1$. The following example shows that this claim was incorrect, and that the bound $N \geq 3r + 1$ is optimal. Consider the language

$$L = \{000000, 000001, 010000, 100010, 110101, 111011\}.$$

We have $m = 6$, $r = 2$ and $N = 3r = 6$. The state complexity of L is 22 while $2v + m(N - 2r - 1) + 1 = 21$. The minimal automaton for L is shown in Figure 3. Compared to the topology of Figure 2, there is one state missing in part M ($d = 1$) which allows two non-singleton states in T_2 and T_1 (the “dead” state is not shown).

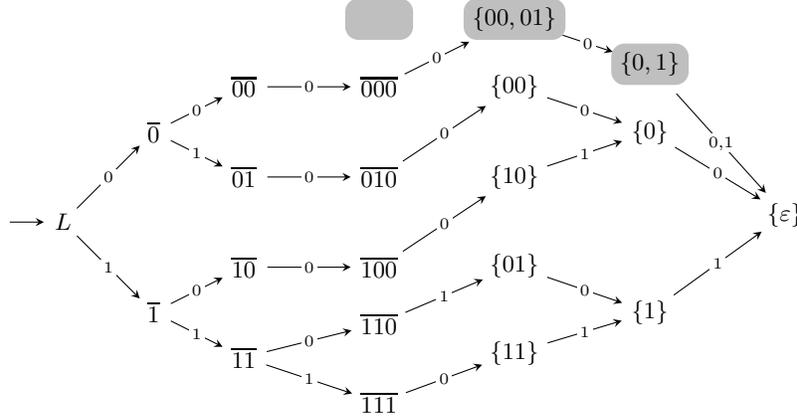


FIGURE 3. A counter-example for $N = 3r$

The slightly modified language

$$L' = \{0000000, 0000001, 0100000, 1000010, 1100101, 1110011\}$$

also shows that for $N = 3r + 1$, the maximum can be achieved with a different topology, namely with $\pi_{r+1} = \sigma_{r+1} = m - 1$, see Figure 4.

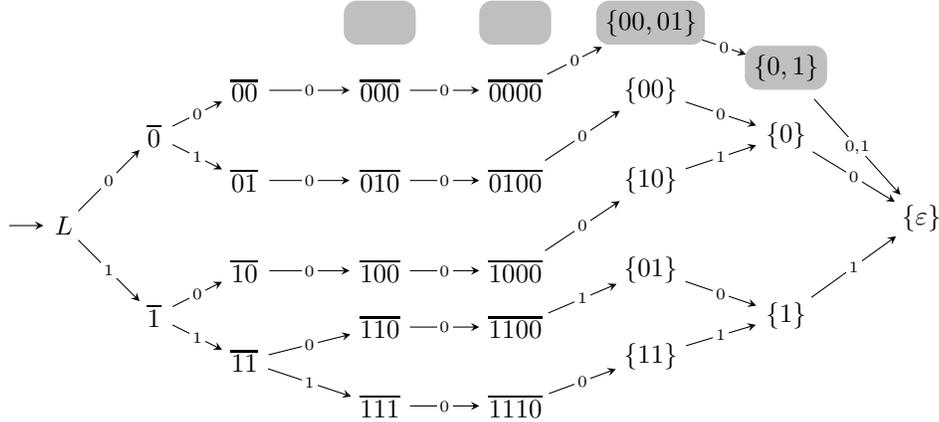


FIGURE 4. A counter-example for $N = 3r + 1$

We can now formulate our result on the state complexity of generalized de Bruijn words.

Theorem 7. *If w is a word of length N over a k -letter alphabet, with $k \geq 2$, then*

$$(4) \quad \text{sc}(C(w)) \leq 2v + N(N - 2r - 1) + 1,$$

where $r = \lfloor \log_k N \rfloor$ and $v = 1 + k + k^2 + \dots + k^r$.

Moreover, equality holds in (4) iff w is a generalized de Bruijn word.

Proof. Let w be a word of length N , and let $L = C(w)$. Note that, for each $1 \leq i \leq N$, we have $\pi_i(L) = \sigma_i(L) = \gamma_i(w)$. Therefore, the theorem follows from Theorem 6 if $N > 3r + 1$.

For $N \leq 3r + 1$, the claim has to be checked separately. This concerns the following cases:

- $N = 1$ for any $k \geq 2$;
- $1 < N \leq 10$ for $k = 2$;
- $N = 3$ and $N = 4$ for $k = 3$; and
- $N = 4$ for $k = 4$.

If $|w| = 1$, then $r = 0$, $v = 1$, and the minimal accepting automaton has three states: $\{w\}$, $\{\varepsilon\}$ and \emptyset . Moreover, w is a generalized de Bruijn word, since $\gamma_0(w) = \gamma_1(w) = k^0 = N = 1$. Therefore the theorem holds in this case.

Table 2 lists all generalized de Bruijn words (up to the conjugation and the exchange of letters) for the remaining cases not covered by Theorem 6. We verified by an exhaustive computer search that they are exactly the words for which equality holds in (4), and that no other word has a larger complexity. \square

k	N	maximum words
2	2	01
2	3	001
2	4	0011
2	5	00011
2	6	000111, 001011
2	7	0001011, 0001101
2	8	00010111
2	9	000010111, 000011101
2	10	0000101111, 0001011101
3	3	012
3	4	0012,0102
4	4	0123

TABLE 2. Maximum words not covered by Theorem 6

For $k = 2$ the maximum state complexity of $C(x)$ over length- N words x is given in Table 3 for $1 \leq N \leq 10$. It is sequence [A316936](#) in the OEIS [13].

N	$\max_{x \in \Sigma_2^N} \text{sc}(C(x))$
1	3
2	5
3	7
4	11
5	15
6	21
7	29
8	39
9	49
10	61

TABLE 3. Maximum state complexity of conjugates of binary words of length N

4. COUNTING GENERALIZED DE BRUIJN WORDS

We first count the total number of factors of a generalized de Bruijn word. This is a generalization of Theorem 2 of [20] to all $k \geq 2$, adapted for the case of circular words.

Proposition 8. *If $w \in \Sigma_k^N$ is a generalized de Bruijn word, then*

$$\sum_{0 \leq i \leq N} \gamma_i(w) = \frac{k^{r+1} - 1}{k - 1} + N(N - r),$$

where $r = \lfloor \log_k N \rfloor$.

Proof. We have

$$\begin{aligned} \sum_{0 \leq i \leq N} \gamma_i(w) &= \sum_{0 \leq i \leq N} \min(k^i, N) \\ &= \sum_{0 \leq i \leq r} k^i + \sum_{r < i \leq N} N \\ &= \frac{k^{r+1} - 1}{k - 1} + N(N - r). \end{aligned}$$

□

Counting the exact number of generalized de Bruijn words of length N appears to be a difficult task. Figures for small N can be obtained by a computer search, as in Table 1. The second author has computed these numbers up to $N = 64$ (see Table 4 for a possible independent verification).

length	number	length	number	length	number
32	2 048	43	940 878	54	36 137 280
33	4 096	44	1 457 197	55	38 730 752
34	3 840	45	2 234 864	56	41 246 208
35	7 040	46	3 302 752	57	50 774 016
36	13 744	47	4 975 168	58	60 764 160
37	28 272	48	7 459 376	59	62 619 648
38	54 196	49	10 347 648	60	70 057 984
39	88 608	50	13 841 408	61	59 768 832
40	160 082	51	17 696 256	62	88 080 384
41	295 624	52	23 404 848	63	134 217 728
42	553 395	53	30 918 336	64	268 435 456

TABLE 4. Numbers of longer binary generalized de Bruijn words

Except in a few simple cases, we do not even know an exact asymptotic expression. For example, if $N = k^n$, then it follows from known results [24] that this number is $(k!)^{k^{n-1}}/k^n$, counted up to cyclic shift. Some loose bounds could be obtained from [25], keeping in mind, however, that we are interested in circuits visiting all vertices, not just arbitrary circuits. Precise numbers seem to be relatively easily computable for $N = k^n \pm 1$, and possibly also for $N = k^n \pm 2$. In particular, the number of binary generalized de Bruijn words of length $N = 2^n \pm 1$ is twice the number of such words of length 2^n ; see the discussion in [11, p. 202]. The considerations, however, quickly become involved. It can be verified by a computer search, for example, that the formula for cycles of length $2^n - 2$ given in [11, p. 203] is wrong. Similarly, the number of cycles of length $k^n \pm 1$ we gave in the final comments of our conference paper is also wrong for $k > 2$. For example, computer search shows that the number of ternary generalized de Bruijn words of lengths 8 and 10 are 36 and 108, respectively, while the number of ternary (generalized) de Bruijn words of length 9 is 24. We therefore leave this question open for further research.

ACKNOWLEDGMENTS

We thank the anonymous referees for helpful comments and suggestions.

REFERENCES

- [1] J. Brzozowski, Quotient complexity of regular languages, *J. Automata, Languages, and Combinatorics* 15 (2010) 71–89.
- [2] A. N. Maslov, Estimates of the number of states of finite automata, *Dokl. Akad. Nauk SSSR* 194 (6) (1970) 1266–1268, in Russian. English translation in *Soviet Math. Dokl.* 11 (5) (1970), 1373–1375.
- [3] G. Jirásková, A. Okhotin, State complexity of cyclic shift, *RAIRO Inform. Théor. App.* 42 (2008) 335–360.
- [4] J. Jirásek, G. Jirásková, Cyclic shift on prefix-free languages, in: A. A. Bulatov, A. M. Shur (Eds.), *CSR 2013*, Vol. 7913 of *Lecture Notes in Computer Science*, Springer-Verlag, 2013, pp. 246–257.
- [5] D. Gabric, S. Holub, J. Shallit, Generalized de Bruijn words and the state complexity of conjugate sets, in: M. Hospodár, et al. (Eds.), *DCFS 2019*, Vol. 11612 of *Lecture Notes in Computer Science*, Springer-Verlag, 2019, pp. 137–146.
- [6] C. Flye Sainte-Marie, Question 48, *L’Intermédiaire Math.* 1 (1894) 107–110.

- [7] M. H. Martin, A problem in arrangements, *Bull. Amer. Math. Soc.* 40 (1934) 859–864.
- [8] I. J. Good, Normal recurring decimals, *J. London Math. Soc.* 21 (1946) 167–169.
- [9] N. G. de Bruijn, A combinatorial problem, *Proc. Konin. Neder. Akad. Wet.* 49 (1946) 758–764.
- [10] N. G. de Bruijn, Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n -letter word exactly once, Tech. Rep. 75-WSK-06, Department of Mathematics and Computing Science, Eindhoven University of Technology, The Netherlands (June 1975).
- [11] H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, *SIAM Review* 24 (1982) 195–221.
- [12] A. Ralston, De Bruijn sequences — a model example of the interaction of discrete mathematics and computer science, *Math. Mag.* 55 (1982) 131–143.
- [13] N. J. A. Sloane, et al., The on-line encyclopedia of integer sequences, available online at <https://oeis.org> (2019).
- [14] J. Petersen, *Die Theorie der regulären Graphs*, *Acta Math.* 15 (1891) 193–220. doi:10.1007/BF02392606. URL <https://doi.org/10.1007/BF02392606>
- [15] D. B. West, *Introduction to Graph Theory*, 2nd Edition, Prentice Hall, 2000.
- [16] A. Lempel, m -ary closed sequences, *J. Combin. Theory* 10 (1971) 253–258.
- [17] F. Hemmati, D. J. Costello, Jr., An algebraic construction for q -ary shift register sequences, *IEEE Trans. Comput.* 27 (1978) 1192–1195.
- [18] T. Etzion, An algorithm for generating shift-register cycles, *Theoret. Comput. Sci.* 44 (1986) 209–224.
- [19] M. Yoeli, Binary ring sequences, *Amer. Math. Monthly* 69 (1962) 852–855.
- [20] J. Shallit, On the maximum number of distinct factors of a binary string, *Graphs and Combinatorics* 9 (1993) 197–200.
- [21] A. Iványi, On the d -complexity of words, *Ann. Univ. Sci. Budapest. Sect. Comput.* 8 (1987) 69–90.
- [22] M.-C. Anisiu, Z. Blázsik, Z. Kása, Maximal complexity of finite words, *Pure Math. Appl.* 13 (2002) 39–48.
- [23] A. Flaxman, A. W. Harrow, G. B. Sorkin, Strings with maximally many distinct subsequences and substrings, *Electronic J. Combinatorics* 11 (1) (2004) #R8.
- [24] T. van Aardenne-Ehrenfest, N. G. de Bruijn, Circuits and trees in oriented linear graphs, *Simon Stevin* 28 (1951) 203–217.
- [25] U. M. Maurer, *Asymptotically tight bounds on the number of cycles in generalized de Bruijn-Good graphs*, *Discrete Appl. Math.* 37/38 (1992) 421–436. doi:10.1016/0166-218X(92)90149-5. URL [https://doi.org/10.1016/0166-218X\(92\)90149-5](https://doi.org/10.1016/0166-218X(92)90149-5)