Generalized de Bruijn words and the state complexity of conjugate sets

Daniel Gabric
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
dgabric@uwaterloo.ca

Štěpán Holub
Department of Algebra
Faculty of Mathematics and Physics
Charles University
Prague
Czech Republic
holub@karlin.mff.cuni.cz

Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
shallit@uwaterloo.ca

March 14, 2019

Abstract

We consider a certain natural generalization of de Bruijn words, and use it to compute the exact maximum state complexity for the language consisting of the conjugates of a single word.

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

For a regular language L let $\mathrm{sc}(L)$ 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* [3]. The state complexity of the cyclic shift operation $L \to C(L)$ for arbitrary regular languages L was studied in Maslov's pioneering 1970 paper [17]. More recently, Jirásková and Okhotin [14] improved Maslov's bound, and Jirásek and Jirásková studied the state complexity of the conjugates of prefix-free languages [13].

In this note we investigate the state complexity of the finite language $C(x)$, over all word x of length n. Clearly $\mathrm{sc}(C(x))$ achieves its minimum — namely, $n+2$ — at words of the form a^n. By considering random words, it seems likely that $\mathrm{sc}(C(x)) = O(n^2)$.

Our main result makes this precise:

Theorem 1. *Let Σ be an alphabet of cardinality $k \geq 2$, and let $n \geq 1$ be an integer. Define $i = \lfloor \log_k n \rfloor$. Then*

$$\max_{w \in \Sigma_k^n} \mathrm{sc}(C(w)) = 2r + n(n - 2i + 1) + 1,$$

where $r = (k^{i+1} - 1)/(k - 1)$.

Furthermore, we characterize those words x achieving this maximum.

Our theorem depends on a certain natural generalization of de Bruijn words, of independent interest, which is introduced in the next section.

2 Generalized de Bruijn words

De Bruijn words (also called de Bruijn sequences) have a long history [8, 16, 10, 4, 5], and have been extremely well studied [9, 18]. Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. 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 of length $k^n + n - 1$ having each word of length n over Σ_k exactly once as a factor; or

(b) the words w of length k^n 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, and we write $D(k, n)$ for the set of all such words. According to the definition, 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: $w \in D(k, n)$ also achieves 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 factors of length $\geq i$ are distinct, because the length-n prefixes are all distinct.

This motivates the following definition:

Definition 2. A word x of length N over a k-letter alphabet is said to be a *generalized de Bruijn word* if $\gamma_i(x) = \min(k^i, N)$ for $0 \leq i \leq N$.

Example 3. 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) [20]. The second author has computed these numbers up to $n = 63$.

The main result of this section is the following.

Theorem 4. *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.*

Proof. For $k = 2$ the proof can be found in [19], although strangely it is not explicitly stated anywhere in the paper. (Lemma 3 implies it.)

For $k > 2$ we can derive this result from a paper by Lempel [15]. Lempel proved that for all $k \geq 2$, $n \geq 1$, $N \leq k^n$, there exists a circular word $w = w(k, n, N)$ of length N for which the factors of size n are distinct. (Also see [11, 6].) However, as stated, this result is not strong enough for our purposes. For example, there are circular words, such as 000101 of length 6, having 6 distinct factors of length 4, but only 3 distinct factors of length 2. For our purposes, then, we need a stronger version of the result, which can nevertheless be obtained from a further analysis of Lempel's proof.

An *Euler graph* is a directed graph in which, for each vertex v, the indegree of v is equal to the outdegree of v. By a *closed chain* we mean a sequence of edges (a, v_1), (v_1, v_2), (v_2, v_3), $\ldots$, (v_{n-1}, a), where each edge is distinct, but vertices may be repeated. Each closed chain

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	000001010011010111	68
20	00000100101100111101	141
21	00000100011010010111	242
22	000001000110100111111	407
23	000001000110011101111	600
24	000001000110010101101111	898
25	000001000110010101011111	1440
26	000001000110010100111011111	1812
27	000001000110010100111011111	2000
28	000001000110010100111010111	2480
29	000001000110010100111010111	2176
30	000001000110010110100111011111	2816
31	000001000110010100111010110111	4096

Table 1: Generalized de Bruijn words

forms an Euler graph and each connected Euler graph admits a closed chain containing all its edges.

Let G_k^n be the k-ary de Bruijn graph of order n. This is a directed graph where the vertices are the 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. So every vertex of G_k^n has k incoming edges, and k outgoing edges, and therefore G_k^n is a regular graph of degree $2k$. Building a generalized de Bruijn word of length $N = k^n + i$, where $0 < i \leq (k-1)k^n$, over a k-letter alphabet then amounts to constructing a closed chain of length N in G_k^n that visits every vertex.

One of Lempel's main results ([15, Theorem 1]) states that such a closed chain exists, but does not mention explicitly whether it visits every vertex. In the proof, the chain is obtained by constructing a connected Euler graph using [15, Lemma 6]. Now, the analysis of the proof of [15, Lemma 6] shows that the constructed Euler graph is not only connected (which is the explicit concern of the lemma) but also spanning. The closed chain is eventually obtained as a complement of a graph G (denoted as T_p in [15]), where G is an Euler graph contained in G_k^n such that the degree of each vertex in G is at most $2(k-1)$. Therefore, its complement is obviously spanning. □

Remark 5. 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 [12] considered the analogue of Eq. (1) for ordinary (non-circular) words. He called a word w *supercomplex* if the analogue of the upper bound (1) 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 Eq. (1) for ordinary words [19]. 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 [2] 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 [7].

We now turn to an alternative characterization of our generalized de Bruijn words.

Proposition 6. *A word $w \in \Sigma_k^N$ is a generalized de Bruijn word iff both of the following hold:*

(a) $\gamma_r(w) = k^r$; *and*

(b) $\gamma_{r+1}(w) = N$,

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

Proof. A generalized de Bruijn word trivially has these properties, and it is easy to see that the two properties imply the bound in Eq. (1). □

We now count the total number of factors of a generalized de Bruijn word. This is a generalization of Theorem 2 of [19] to all $k \geq 2$, adapted for the case of circular words.

Proposition 7. *If $w \in \Sigma_k^N$ is a generalized de Bruijn word, then*

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

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

Proof. We have

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

□

3 State complexity

We start with a general upper bound on state complexity.

Theorem 8. *Let Σ be an alphabet of cardinality k. Suppose $L \subseteq \Sigma^t$, and suppose $|L| = m$. Define $i = \lfloor \log_k m \rfloor$ and $r = (k^{i+1}-1)/(k-1)$. If $t \ge 2i+1$ then $\mathrm{sc}(L) \le 2r + m(t-2i-1)+1$.*

Proof. A *level* is a set of nodes at a particular distance from the root. The complete k-ary tree of $i+1$ levels therefore corresponds to words of length $\le i$, and the total number of nodes in this tree is $1 + k + \cdots + k^i = \frac{k^{i+1}-1}{k-1}$.

The language L can be accepted by a DFA with the following topology: there is a complete k-ary tree of $i+1$ levels rooted at the initial state p_ϵ. At the very next level there are at most m nodes, and these nodes form the roots of at most m chains of $t-2i-1$ nodes each. These chains need not be disjoint, but will be in the worst case. At the end, there is another complete k-ary tree of $i+1$ levels culminating in a single accepting state. Finally, there is also a single non-accepting state that captures all transitions not yet defined. The total number of states is therefore $2r + m(t-2i-1)+1$.

Define

$$\begin{aligned}X &= \Sigma^{\le i} \cup \{x \ : \ i < |x| < t-i-1 \text{ and } x \text{ is a prefix of an element of } L\} \\ Y &= \{y \ : \ |y| = t-i-1 \text{ and } y \text{ is a prefix of an element of } L\}\end{aligned}$$

More formally, the states of our DFA are d, a "dead" state; p_x, for $x \in X$; and s_z, for all z with $|z| \le i$. The states p_x correspond to prefixes of words of L and the states s_z correspond to suffixes of words of L.

The initial state is p_ϵ.

The transitions are given by $\delta(p_x, a) = p_{xa}$ for $x \in X$ and $a \in \Sigma$ and $\delta(p_y, a) = s_z$, if $y \in Y$ and $yaz \in L$; $\delta(s_{av}, a) = s_v$ for $v \in \Sigma^{<i}$ and $a \in \Sigma$. All other transitions go to d.

Finally, the unique final state is s_ϵ. □

This construction is illustrated in Figure 1 for $k = 2$, $t = 12$, $m = 10$, $i = 3$, $r = 15$, $t - 2i - 1 = 5$, and

$$L = \{000010100000, 000101100010, 011110100001, 100110011111, 101011110111, \\ 110100100110, 110101010011, 110110101101, 111001100101, 111110110100\}.$$

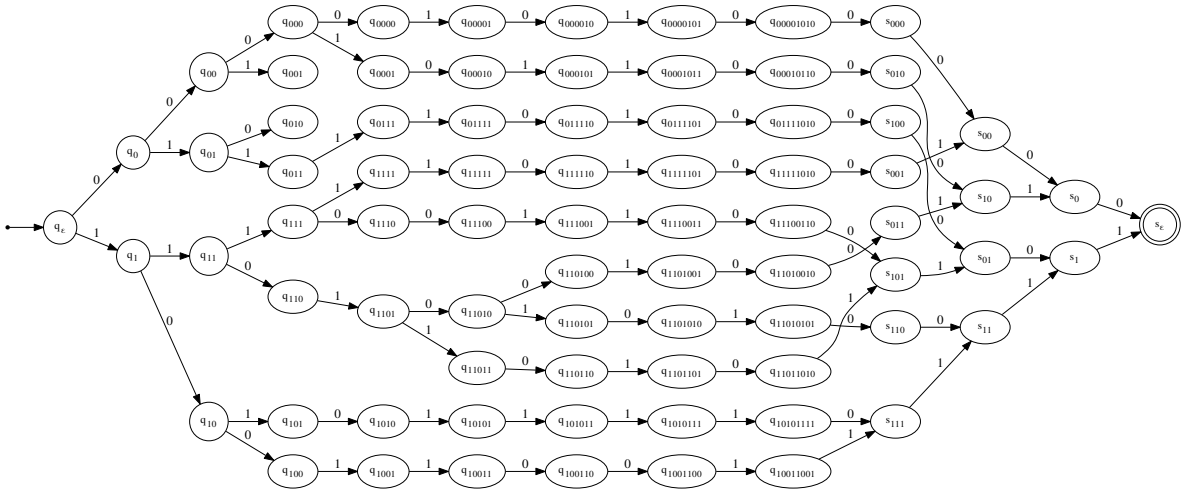

Figure 1: Example of the construction

As a corollary, we now get an upper bound on $\mathrm{sc}(C(x))$:

Corollary 9. *If x is a word of length n over a k-letter alphabet, then*

$$\mathrm{sc}(C(x)) \leq 2r + n(n - 2i - 1) + 1,$$

where $i = \lfloor \log_k n \rfloor$ and $r = (k^{i+1} - 1)/(k - 1)$.

Proof. Let x be a word of length n, and let $L = C(x)$. Then $|L| \leq n$. In Theorem 8 take $t = n$ and $m \leq n$. Set $i = \lfloor \log_k n \rfloor$ and $r = (k^{i+1} - 1)/(k - 1)$. The inequality $t \geq 2i + 1$ holds in all cases except $k = 2$ and $n = 2$; this case can be checked separately. We therefore get $\mathrm{sc}(L) \leq 2r + n(n - 2i - 1) + 1$, as desired. □

It now remains to prove that there exist words that achieve this upper bound. In fact, such words are exactly the generalized de Bruijn words defined in Section 2.

Theorem 10. *A length-n word x over a k-letter alphabet satisfies*

$$\mathrm{sc}(C(x)) = 2r + n(n - 2i - 1) + 1,$$

where $i = \lfloor \log_k n \rfloor$ and $r = (k^{i+1} - 1)/(k - 1)$ iff x is a generalized de Bruijn word.

Proof. Suppose x is a generalized de Bruijn word. We first show that there are $2r + n(n - 2i - 1) + 1$ inequivalent words for the Myhill-Nerode equivalence relation R associated with $C(x)$. This will show $\mathrm{sc}(C(x)) \geq 2r + n(n - 2i - 1) + 1$ and hence, by Corollary 9, that $\mathrm{sc}(C(x)) = 2r + n(n - 2i - 1) + 1$.

Representatives of the Myhill-Nerode classes can be classified as follows:

(a) all the words of length $\leq i$;

(b) all the factors of conjugates of x of length ℓ, for $i < \ell < n - i$;

(c) for each word w of length $\leq i$, the lexicographically least factor z of $C(x)$ of length $n - i$ for which $zw \in C(x)$.

(d) the single equivalence class corresponding to words not in $C(x)$.

There are k^i words in (a), and k^i words in (c), there are $n(n - 2i - 1)$ words in (b), and one word in (d).

We need to see that these are all inequivalent. Since all the words in $C(x)$ are of length n, no two factors of different lengths can be equivalent. It therefore suffices to examine pairs of words of identical length.

In group (a), let y, z be two distinct words of length $j \leq i$. Since x, considered circularly, contains all factors of length $i = \lfloor \log_k n \rfloor$, it contains y and z as factors. Let yy' (resp., zz') be a conjugate of x with prefix y (resp., z). Then $|y'| = |z'| = n - j \geq i + 1$. If both yz' and zz' occur in $C(x)$, we would have two separate occurrences of z' in x (considered circularly), which is impossible since x is of length n and has n distinct factors (considered circularly). So $yz' \notin C(x)$ and y, z are inequivalent under Myhill-Nerode. This gives $(k^{i+1} - 1)/(k - 1)$ equivalence classes.

In group (b), let y, z be two distinct factors of $C(x)$ (considered circularly) of length j with $i < j < n - i$. Since x is of length n and contains n distinct factors of length i, the first i symbols of y (resp., z) uniquely determines the position of y (resp., z) within x (considered as a circular word). So there is a unique y' such that $yy' \in C(x)$, and similarly, there is a unique z' such that $zz' \in C(x)$. Just as in case (a), since $|y'| = |z'| \geq i + 1$, we see that $y' \neq z'$. This gives $n(n - 2i)$ equivalence classes.

In group (c), for each word t of length $\leq i$, let x_t be the lexicographically least word of length $n - i$ such that $x_t t \in C(x)$. (We know such a word exists because each such t is a factor of x, considered circularly.) Let t, u be distinct words of length j. Then since $|x_t| \geq i + 1$, the word x_t occurs in exactly one location in x, considered circularly, and there it must be followed by t. So $x_t u \notin C(x)$, so x_t and x_u are inequivalent under Myhill-Nerode. This gives $(k^{i+1} - 1)/(k - 1)$ equivalence classes.

Now let us prove the reverse direction. Suppose x is such that $\mathrm{sc}(C(x)) = 2r + n(n - 2i - 1) + 1$. Then from the upper bound in Corollary 9 and the construction of Theorem 8 from which it is derived, we know that all the words corresponding to the states of the automaton

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

Table 2: Maximum state complexity of conjugates of binary words of length n

in Theorem 8 are pairwise inequivalent under Myhill-Nerode. But there are k^i such words of length i and n such words of length $i+1$. Hence, by Proposition 6, x is a generalized de Bruijn word. □

For $k = 2$ the maximum state complexity of $C(x)$ over length-n words x is given in Table 2. It is sequence A316936 in the OEIS [20].

4 Final comments

We do not currently know an accurate asymptotic expression for the number of generalized de Bruijn words of length n, except in few simple cases. If $n = k^i$, then it follows from known results [1] that this number is (counted up to cyclic shift) $(k!)^{k^{i-1}}/k^i$.

A generalized de Bruijn word of length k^i+1 corresponds to a closed path in the de Bruijn graph G_k^i that visits one vertex exactly twice and all others exactly once. This implies that the additional edge is a loop. Therefore, each generalized de Bruijn word of length $k^i + 1$ is obtained from an ordinary de Bruijn word of length k^i by replacing a factor a^i with a^{i+1} where a is a letter. It follows that the number of such words is $(k!)^{k^{i-1}}/k^{i-1}$. A similar argument yields the same number of generalized de Bruijn words of length $k^i - 1$.

Already for $k^i \pm 2$ these kinds of considerations become very complex. We leave this as a challenging open problem for the reader.

References

[1] T. van Aardenne-Ehrenfest and N. G. de Bruijn. Circuits and trees in oriented linear graphs. *Simon Stevin* **28** (1951), 203–217.

[2] M.-C. Anisiu, Z. Blázsik, and Z. Kása. Maximal complexity of finite words. *Pure Math. Appl.* **13** (2002), 39–48.

[3] J. Brzozowski. Quotient complexity of regular languages. *J. Automata, Languages, and Combinatorics* **15** (2010), 71–89.

[4] N. G. de Bruijn. A combinatorial problem. *Proc. Konin. Neder. Akad. Wet.* **49** (1946), 758–764.

[5] 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. Technical Report 75-WSK-06, Department of Mathematics and Computing Science, Eindhoven University of Technology, The Netherlands, June 1975.

[6] T. Etzion. An algorithm for generating shift-register cycles. *Theoret. Comput. Sci.* **44** (1986), 209–224.

[7] A. Flaxman, A. W. Harrow, and G. B. Sorkin. Strings with maximally many distinct subsequences and substrings. *Electronic J. Combinatorics* **11**(1) (2004), #R8 (electronic).

[8] C. Flye Sainte-Marie. Question 48. *L'Intermédiaire Math.* **1** (1894), 107–110.

[9] H. Fredricksen. A survey of full length nonlinear shift register cycle algorithms. *SIAM Review* **24** (1982), 195–221.

[10] I. J. Good. Normal recurring decimals. *J. London Math. Soc.* **21** (1946), 167–169.

[11] F. Hemmati and D. J. Costello, Jr. An algebraic construction for q-ary shift register sequences. *IEEE Trans. Comput.* **27** (1978), 1192–1195.

[12] A. Iványi. On the d-complexity of words. *Ann. Univ. Sci. Budapest. Sect. Comput.* **8** (1987), 69–90.

[13] J. Jirásek and G. Jirásková. Cyclic shift on prefix-free languages. In A. A. Bulatov and A. M. Shur, editors, *CSR 2013*, Vol. 7913 of *Lecture Notes in Computer Science*, pp. 246–257. Springer-Verlag, 2013.

[14] G. Jirásková and A. Okhotin. State complexity of cyclic shift. *RAIRO Inform. Théor. App.* **42** (2008), 335–360.

[15] A. Lempel. m-ary closed sequences. *J. Combin. Theory* **10** (1971), 253–258.

[16] M. H. Martin. A problem in arrangements. *Bull. Amer. Math. Soc.* **40** (1934), 859–864.

[17] 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.

[18] A. Ralston. De Bruijn sequences — a model example of the interaction of discrete mathematics and computer science. *Math. Mag.* **55** (1982), 131–143.

[19] J. Shallit. On the maximum number of distinct factors of a binary string. *Graphs and Combinatorics* **9** (1993), 197–200.

[20] N. J. A. Sloane et al. The on-line encyclopedia of integer sequences. Available online at https://oeis.org, 2019.