Lyndon Words with a Fixed Standard Right Factor

Frédérique Bassino, Julien Clément and Cyril Nicaud

Institut Gaspard Monge
Université de Marne-la-Vallée
77454 Marne-la-Vallée Cedex 2 - France
{bassino,clementj,nicaud}@univ-mlv.fr

Given a totally ordered alphabet $A = \{a_1 < a_2 < \cdots < a_q\}$, a *Lyndon word* is a word that is strictly smaller, for the lexicographical order, than any of its conjugates (*i.e.*, all words obtained by a circular permutation on the letters). Lyndon words were introduced by Lyndon [6] under the name of "standard lexicographic sequences" in order to give a base for the free Lie algebra over A. The set of Lyndon words is denoted by $\mathcal{L}$. For instance, with a binary alphabet $A = \{a, b\}$, the first Lyndon words until length five are $\mathcal{L} = \{a, b, ab, aab, abb, aaab, aabb, abbb, aaaab, aaabb, aabab, aabbb, ababb, abbbb, \ldots\}$. Note that a non-empty word is a Lyndon word if and only if it is strictly smaller than any of its proper suffixes.

The *standard (suffix) factorization* of Lyndon words plays a central role in this framework (see [5], [7], [8]). For $w \in \mathcal{L} \setminus A$ a Lyndon word not reduced to a letter, the pair (u, v) of Lyndon words such that $w = uv$ and v of maximal length is called the *standard factorization*. The words u and v are called the *left factor* and *right factor* of the *standard factorization*. Equivalently, the right factor v of the standard factorization of a Lyndon word w which is not reduced to a letter can be defined as the smallest proper suffix of w. For instance we have the following standard factorizations: $aaabaab = aaab \cdot aab$, $aaababb = a \cdot aababb$ and $aabaabb = aab \cdot aabb$.

One can then associate to a Lyndon word w a binary tree $T(w)$ called its *Lyndon tree* recursively built in the following way:
– if w is a letter, then $T(w)$ is a leaf labeled by w,
– otherwise $T(w)$ is an internal node having $T(u)$ and $T(v)$ as children where $u \cdot v$ is the standard factorization of w.
This structure encodes a non-associative operation, either a commutator in the free group [3], or a Lie bracketing [5]; both constructions lead to bases of the free Lie algebra.

Our goal is first to get a better insight of the standard factorization of Lyndon words, possibly to design better algorithms for the construction of the Lyndon tree (A naive algorithm has time complexity $O(n^2)$ in the worst case). The characterization of the set of Lyndon words with a given right standard factor is also a first step towards the average case analysis of the height of Lyndon trees in order to give a more accurate view of the observed behavior of algorithms using Lyndon trees.

To state our main results, we need to recall first what are *regular languages*. A *language* L is a set of words over a fixed alphabet A. The structurally simplest (yet non trivial) languages are the *regular languages* that can be defined in a variety of ways: by regular expressions and by finite automata. Concatenation of languages is denoted by a product ($L_1 \cdot L_2 = \{w_1 w_2 \mid w_1 \in L_1, w_2 \in L_2\}$). Union of languages is the ordinary set union. The empty word is denoted by ϵ and the Kleene star operator is understood as $L^* = \epsilon + L + L \cdot L + \cdots$. A regular language over an alphabet A is built by recursively applying concatenation, union and Kleene star operator to the singleton languages $\{\epsilon\}$ and $\{\sigma\}$ ($\forall \sigma \in A$). A regular expression is the description of a regular language (most commonly using symbols "$\cdot, +, *$").

Our main result can hence be stated as follows: *the set of Lyndon words with a fixed standard right factor is a regular language whose regular expression is given in Theorem 1.* This result may seem a bit surprising. The fact that the set of Lyndon words is not even context-free [1] gives indeed an idea of the structural complexity of Lyndon words. Theorem 2 gives the corresponding enumerative generating function.

Let $A = \{a_1 < \cdots < a_q = \gamma\}$ be a totally ordered q-ary alphabet where γ denotes the last symbol of A. We denote A^* the set of finite words. We consider the lexicographical order $<$ over all non-empty words of A^* defined by the extension of the order over A. Let w be a word of $A^* \setminus \{\gamma\}^*$, the *successor* $S(w)$ of $w = v\alpha\gamma^i$, where α is a symbol of $A \setminus \{\gamma\}$ and $i \geq 0$, is defined by $S(w) = v\beta$ with β

the immediate next symbol after α in A. For any Lyndon word v, we define the set of words

$$\forall v \neq \gamma \quad \mathcal{X}_v = \{v, S(v), S^2(v), \dots, S^k(v) = \gamma\} \qquad \text{and} \qquad \mathcal{X}_\gamma = \{\gamma\}.$$

Note that $k = 1 + q \times |v| - \sum_{i=1}^{q} i \times |v|_i$ where q is the cardinality of the alphabet A, $|v|$ is the length of v and $|v|_i$ is the number of occurrences of the ith letter of the alphabet A in v.

Example.

(1) for $A = \{a, b\}$, $v = aabab$: $\mathcal{X}_{aabab} = \{aabab, aabb, ab, b\}$.
(2) for $A = \{a, b, c\}$, $v = abb$: $\mathcal{X}_{abb} = \{abb, abc, ac, b, c\}$.

By construction, v is the smallest element of $\mathcal{X}_v A^*$ for the lexicographical order.

Denote, for any letter $\alpha \in A$, the set of letters $A_{\leq\alpha} = \{a \in A \mid a \leq \alpha\}$. For a language L, we use the convenient notation $L^+ = L + L \cdot L + \cdots$. One should also know that the class of regular languages is closed under the set difference operator "\".

Theorem 1. *Let v be a Lyndon word whose first letter is α and $u \in A^*$. Then uv is a Lyndon word with $u \cdot v$ as standard factorization if and only if $u \in (A_{\leq\alpha}\mathcal{X}_v^*) \setminus \mathcal{X}_v^+$. Hence the set $\mathcal{F}_v$ of Lyndon words having v as right standard factor is a regular language.*

One of the basic properties of the set of Lyndon words is that every word is uniquely factorizable as a non increasing product of Lyndon words. We remark that when we apply this decomposition to "shifted" Lyndon words (words obtained by deleting the first symbol) the last factor is exactly the right standard factor. Another important fact is that for any two Lyndon words u and v we have $u > v$ if and only if $u \in \mathcal{X}_v^+$.

We define the generating functions $X_v(z)$ of $\mathcal{X}_v$ and $X_v^*(z)$ of $\mathcal{X}_v^*$ where $|w|$ is the length of a word w:

$$X_v(z) = \sum_{w \in \mathcal{X}_v} z^{|w|} \quad \text{and} \quad X_v^*(z) = \sum_{w \in \mathcal{X}_v^*} z^{|w|}.$$

As the set $\mathcal{X}_v$ is a code, the elements of $\mathcal{X}_v^*$ are (finite) sequences of elements of $\mathcal{X}_v$ (see [4]):

$$X_v^*(z) = \frac{1}{1 - X_v(z)}.$$

Denote by $F_v(z) = \sum_{w \in \mathcal{F}_v} z^{|w|}$ the generating function of the set

$$\mathcal{F}_v = \{uv \in \mathcal{L} \mid u \cdot v \text{ is the standard factorization}\}.$$

Theorem 2. *Let v be a Lyndon word over a q-ary alphabet. The generating function of the set $\mathcal{F}_v$ of Lyndon words having a right standard factor v can be written*

$$F_v(z) = z^{|v|}\left(1 + \frac{qz - 1}{1 - X_v(z)}\right).$$

This generating function can further be used to enumerate or get the asymptotic behavior of the numbers of Lyndon words with fixed right standard factor (see [4]).

References

[1] J. Berstel, L. Boasson, The set of Lyndon words is not context-free, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS 63 (1997) 139–140.

[2] J. Berstel, M. Pocchiola, Average cost of Duval's algorithm for generating Lyndon words, Theoret. Comput. Sci. 132 (1-2) (1994) 415–425.

[3] K. Chen, R. Fox, R. Lyndon, Free differential calculus IV: The quotient groups of the lower central series, Ann. Math. 58 (1958) 81–95.

[4] P. Flajolet, R. Sedgewick, Analytic combinatorics–symbolic combinatorics, Book in preparation, (Individual chapters are available as INRIA Research reports at http://www.algo.inria.fr/flajolet/publist.html) (2002).

[5] M. Lothaire, Combinatorics on Words, Vol. 17 of Encyclopedia of mathematics and its applications, Addison-Wesley, 1983.

[6] R. Lyndon, On Burnside problem I, Trans. American Math. Soc. 77 (1954) 202–215.

[7] C. Reutenauer, Free Lie algebras, Oxford University Press, 1993.

[8] F. Ruskey, J. Sawada, Generating Lyndon brackets: a basis for the n-th homogeneous component of the free Lie algebra, Journal of Algorithms 46 (2003) 21–26.