Can. J. Math., Vol. XLI, No. 4, 1989, pp. 577–591

LYNDON WORDS, FREE ALGEBRAS AND SHUFFLES

GUY MELANÇON AND CHRISTOPHE REUTENAUER

1. Introduction. A *Lyndon word* is a primitive word which is minimum in its conjugation class, for the lexicographical ordering. These words have been introduced by Lyndon in order to find bases of the quotients of the lower central series of a free group or, equivalently, bases of the free Lie algebra [**2**], [**7**]. They have also many combinatorial properties, with applications to semigroups, pi-rings and pattern-matching, see [**1**], [**10**].

We study here the Poincaré-Birkhoff-Witt basis constructed on the Lyndon basis (PBWL basis). We give an algorithm to write each word in this basis: it reads the word from right to left, and the first encountered inversion is either bracketted, or straightened, and this process is iterated: the point is to show that each bracketting is a standard one: this we show by introducing a loop invariant (property (S)) of the algorithm. This algorithm has some analogy with the collecting process of P. Hall [**5**], but was never described for the Lyndon basis, as far we know.

A striking consequence of this algorithm is that any word, when written in the PBWL basis, has coefficients in $\mathbf{N}$ (see Theorem 1). This will be proved twice in fact, and is similar to the same property for the Shirshov-Hall basis, as shown by M.P. Schützenberger [**11**].

Our next result is a precise description of the dual basis of the PBWL basis. The former is denoted (S_w), where w is any word, and we show that

$$S_w = aS_u$$

if $w = au$ is a Lyndon word beginning with the letter a, and that

$$S_w = (k_1! \dots k_n!)^{-1} S_{l_1}^{k_1} \circ \dots \circ S_{l_n}^{k_n}$$

if $w = l_1^{k_1} \dots l_n^{k_n}$ is the decomposition of w into Lyndon words, where $\circ$ is the shuffle product and S^k means shuffle exponentiation. The latter relation may also be expressed by the following formula, "à la Hopf algebra":

$$\sum_w w \otimes w = \prod_l \exp(S_l \otimes [l])$$

in the complete tensor product $\mathbf{Q} \ll A \gg \otimes \mathbf{Q} \ll A \gg$, with the shuffle algebra on the left, and the concatenation on the right, where the sum is taken over all words w, and the product over all Lyndon words l in decreasing order, and where

Received April 29, 1987.

$[l]$ means the element of the PBWL basis associated to l. This formula holds in fact in any enveloping algebra, for any basis, as we shall indicate (remark 4).

An application of the previous result is that the elements S_l form a transcendance basis of the shuffle algebra $\mathbf{Q}\langle A\rangle$. As a consequence, we prove a theorem of Radford [**9**], who shows that the shuffle algebra $\mathbf{Z}\langle A\rangle$ admits as a basis the set of Lyndon words. More precisely, for each word

$$w = l_1^{k_1} \dots l_n^{k_n}$$

decomposed into Lyndon words, the polynomial

$$(k_1! \dots k_n!)^{-1} l_1^{k_1} \circ \dots \circ l_n^{k_n}$$

has coefficients in $\mathbf{N}$ and is of the form $w +$ smaller words (Section 4). As a corollary, $\mathbf{Z}\langle A\rangle$ (with the shuffle) is isomorphic to the algebra of integral exponential polynomials over the set of Lyndon words, hence it is a free $\mathbf{Z}$-algebra with divided powers.

2. Lyndon words and the free Lie algebra. For properties on Lyndon words which are not proved here, see [**6**] Chapter 5. Let A be a totally ordered set. The elements of A are called *letters* and the elements of the free monoid A^* generated by A are called *words*.

We totally order A^* as follows: $u < v$ if and only if: (i) there exists a non-empty word w such that $uw = v$, or (ii) there exist word x, y, z and letters a, b such that $u = xay$, $v = xbz$ and $a < b$. This is the usual lexicographical order on A^*.

A word u is a *factor* of a word v if there exist words x, y such that $v = xuy$; in case $x = 1$ is the *empty word* (resp. $y = 1$) we say that u is a left (resp. right) factor of v, *proper* if $y \neq 1$ (resp. $x \neq 1$). Two words u, v are said to be *conjugate* if there exist words x, y such that $u = xy$ and $v = yx$. A *Lyndon word* may be equivalently defined to be a word w: (i) that is strictly less than any of its conjugates; or (ii) that is strictly less than any of its proper right factors.

For example $a, ab, aabab$ are Lyndon ($a < b$). From now on, let L denote the set of Lyndon words over A. For any Lyndon word $w \in L - A$ let m be its longest proper right factor in L. Then $w = lm$ with $l \in L$ and $l < lm < m$. The couple $\sigma(w) = (l, m)$ is called the *standard factorization* of w. For example, the standard factorization of *aaabab* is (*a*, *aabab*), and not (*aaab*, *ab*). We also have the following: if $l, m \in L$ and $l < m$ then $lm \in L$.

We consider sequences of the form

$$s = [u_1][u_2] \dots [u_n] \tag{1}$$

where each u_i is a Lyndon word, with the following property:

(S) u_i is either a letter, or if (x, y) is its standard factorization, then y is greater than or equal to any $u_j, j \geqq i$.

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

Note that if each u_i is a letter, then s has property (S). Moreover, if s is decreasing, that is, $u_1 \geqq u_2 \geqq \ldots \geqq u_n$, then s has also property (S). A sequence having property (S) will be called a *standard sequence*.

We present now a *rewriting system* on the set of standard sequences: for s as in (1), s non decreasing, let $[u_i][u_{i+1}]$ be the *rightmost inversion*, that is, i is the greatest index such that $u_i < u_{i+1}$. Then define

$$s' = [u_1][u_2]\ldots[u_iu_{i+1}]\ldots[u_n] \tag{2}$$

$$s'' = [u_1][u_2]\ldots[u_{i+1}][u_i]\ldots[u_n]. \tag{3}$$

As the reader may guess, the brackettings will be interpreted in the sequel as Lie brackettings. The following key lemma shows that s', s'' are standard and that the bracketting $[u_iu_{i+1}]$ is standard.

LEMMA 1. *Let s, s', s'' be defined by* (1), (2), (3). *Then, s', s'' are standard sequences. Moreover, u_iu_{i+1} is a Lyndon word, of standard factorization (u_i, u_{i+1}).*

Proof. We prove first the second assertion. If l, m are Lyndon words with $l < m$, then lm is a Lyndon word (see [**6**] Proposition 5.1.3). Hence u_iu_{i+1} is a Lyndon word. Moreover, either u_i is a letter, hence (u_i, u_{i+1}) is the standard factorization of u_iu_{i+1}; or u_i has the standard factorization (l, m); as s is a standard sequence, we have $m \geqq u_{i+1}$; this shows, by [**6**] Proposition 5.1.4, that (u_i, u_{i+1}) is the standard factorization of u_iu_{i+1}.

We show now that s', s'' are standard sequences. By assumption,

$$u_{i+1} \geqq u_{i+2} \geqq \ldots \geqq u_n;$$

moreover, $u_{i+1} > u_iu_{i+1}$, because u_iu_{i+1} is Lyndon. This shows that s' is standard, because s is already standard. For s'', it is enough to observe that if u_{i+1} is not a letter and if (x, y) is its standard factorization, then $y > u_i$, because $u_i < u_{i+1}$ by assumption and $u_{i+1} < y$, because u_{i+1} is Lyndon. Hence s'' is standard.

We define a relation $\longrightarrow$ on the set of standard sequences: if s, s', s'' are as above, we define

$$s \longrightarrow s' \quad \text{and} \quad s \longrightarrow s''$$

Furthermore, $\overset{*}{\longrightarrow}$ will denote the transitive and reflexive closure of $\longrightarrow$.

Let $\mathbf{Z}\langle A\rangle$ denote the free associative algebra generated by A over $\mathbf{Z}$. Each element of $\mathbf{Z}\langle A\rangle$ is simply a $\mathbf{Z}$-linear combination of words on A, and called a *polynomial*. Put in another way, A^* is a $\mathbf{Z}$-basis of $\mathbf{Z}\langle A\rangle$. As we shall consider also another product on $\mathbf{Z}\langle A\rangle$, we call the product of the free algebra $\mathbf{Z}\langle A\rangle$ the *concatenation product* (because it corresponds to concatenation of words).

Let $\mathcal{L}(A)$ denote be the sub-Lie-algebra of $\mathbf{Z}\langle A\rangle$ generated by the letters in A. It is known that $\mathcal{L}(A)$ is the free Lie algebra on A and that $\mathbf{Z}\langle A\rangle$ is its enveloping algebra (see [**6**] Chapter 5, for this and what follows). An element of $\mathcal{L}(A)$ is called a *Lie polynomial*, or a *Lie element* of $\mathbf{Z}\langle A\rangle$.

Define inductively Lie polynomials $[l]$, for any Lyndon word l (we use again the notation $[l]$ as above, by a slight abuse, which will not be confusing as $l \longrightarrow [l]$ is a bijection). For any letter a, let

$$[a] = a \quad (a \in A)$$

and if u is a Lyndon word of standard factorization (k, l), then

$$[u] = [[k], [l]] = [k][l] - [l][k]$$

By a theorem of Lyndon, the set

$$\{[l] | l \in L\}$$

is a basis of $\mathcal{L}(A)$ over $\mathbf{Z}$. By the Poincaré-Birkhoff-Witt theorem, the set

$$\{[l_1] \ldots [l_n] | n \geqq 0, l_i \in L, l_1 \geqq \ldots \geqq l_n\}$$

is a basis of $\mathbf{Z}\langle A\rangle$, which we call the PBWL *basis* of $\mathbf{Z}\langle A\rangle$.

Recall that each word in A^* may be considered as a standard sequence.

The following result expresses each word in the PBWL basis, using the above rewriting system.

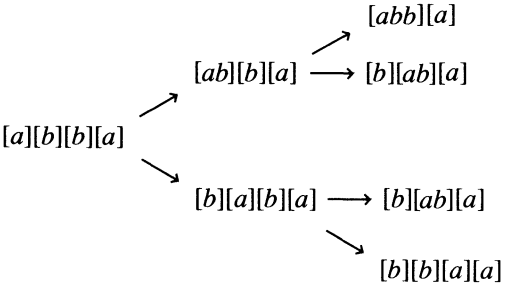
THEOREM 1. *For each word w, one has*

$$w = \sum_{\substack{w \xrightarrow{*} s \\ s \text{ decreasing}}} s$$

where each s appears with its multiplicity.

The theorem is illustrated by the two following examples

Example 1. Let $A = \{a < b\}$ and $w = abba$. One has

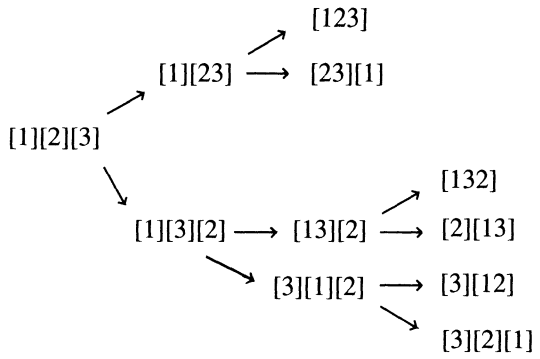


hence by the theorem

$$abba = [abb][a] + 2[b][ab][a] + [b][b][a][a].$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

Example 2. Let $A = \{1 < 2 < 3\}$ and $w = 123$. Then



hence

$$123 = [123] + [23][1] + [132] + [2][13] + [3][12] + [3][2][1].$$

More generally, it is easy to see that for $A = \{1 < 2 < \ldots < n\}$, one has

$$12 \ldots n = \sum_{w \in S_n} [w]$$

where each w in S_n is considered as a word and where $[w]$ denotes $[u_1] \ldots [u_k]$, with each u_i Lyndon and $u_1 > \ldots > u_k$. Note that in this case, $[w]$ is just the Foata transform of w (see [**3**], p. 92). As a byproduct of the rewriting system, we have obtained an algorithm to generate each permutation, in Foata form, or equivalently, in cycle decomposition form.

Proof of Theorem 1. Let $s = [u_1] \ldots [u_k]$ be a standard sequence, interpreted as an element of $\mathbf{Z}\langle A \rangle$ (that is, $[u_i]$ is interpreted as the corresponding Lie element and s as the product of these polynomials). Let s', s'' be as in (2) and (3). We have in $\mathbf{Z}\langle A \rangle$,

$$s = s' + s''$$

because

$$\begin{aligned} [u_i][u_{i+1}] &= [[u_i], [u_{i+1}]] + [u_{i+1}][u_i] \\ &= [u_i u_{i+1}] + [u_{i+1}][u_i]. \end{aligned}$$

Now, s' is shorter that s, and s'' has one fewer inversion than s. This allows to conclude, by an easy induction.

Remark 1. We have shown that in fact Theorem 1 holds for each standard sequence w.

Remark 2. The algorithm of Theorem 1 is similar to P. Hall's "collecting process" (see [**5**], or [**4**] Chapter 11; see also [**11**]; in fact, the collecting process works with commutators in the free group, but there exists a version in the free algebra). However in our algorithm, we look for the first inversion in the standard sequence; in the collecting process, one looks for the first inversion involving the smallest term of the sequence. This difference makes it impossible to find a common generalization of Hall basis and Lyndon basis, with a common generalization to both algorithms.

3. The dual basis and the shuffle product. In this section, we shall investigate the dual basis of the PBWL basis. We shall need the shuffle product. Before this, we recall the following fundamental result on Lyndon words.

THEOREM . ([**6**], Theorem 5.1.5) *Each word w in A^* may be uniquely written as*

$$w = l_1 \dots l_n \tag{4}$$

where each l_i is a Lyndon word with $l_1 \geqq \dots \geqq l_n$.

We extend the notation $[w]$ to the whole free monoid. Recall that for each Lyndon word l, the Lie polynomial $[l]$ was defined in Section 2. If w is any word, decomposed into Lyndon words as in (4), define $[w]$ to be the polynomial

$$[w] = [l_1] \dots [l_n].$$

With this notation, the PBWL basis is just the set

$$\{[u] | u \in A^*\}.$$

Note that the dual space of $\mathbf{Z}\langle A\rangle$ is naturally isomorphic to the set $\mathbf{Z} \ll A \gg$ of all formal series. Each formal series is an infinite linear combination of words. The duality

$$\begin{aligned} \mathbf{Z} \ll A \gg \times \mathbf{Z}\langle A\rangle &\longrightarrow \mathbf{Z} \\ (S, P) &\longrightarrow (S, P) \end{aligned}$$

is defined by

$$(S, P) = \sum_{w \in A^*} (S, w)(P, w)$$

where (S, w) denotes the coefficient of w in S.

The *dual basis* $S_u, u \in A^*$, is defined by

$$w = \sum_{u \in A^*} (S_u, w)[u]$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

for any word w.

Then we know by Theorem 1, that S_u is a polynomial with coefficients in $\mathbf{N}$, and which is homogeneous of degree length(u).

The *shuffle product* $\circ$ is defined inductively for any words u, v and letters a, b by (1 is the empty word):

$$1 \circ u = u \circ 1 = u$$
$$(au) \circ (bv) = a(u \circ (bv)) + b((au) \circ v)$$

It is a commutative and associative product, without zero divisors (see [**12**]).

THEOREM 2. (i) *Let* 1 *be the empty word. Then*

$$S_1 = 1.$$

(ii) *Let* bv *be a Lyndon word with first letter* b. *Then*

$$S_{bv} = bS_v.$$

(iii) *Let* w *be any word, decomposed into Lyndon words as*

$$w = l_1^{i_1} \dots l_k^{i_k} \quad (l_j \in L, l_1 > l_2 > \dots > l_k).$$

Then

$$S_w = \frac{1}{i_1! \dots i_k!} S_{l_1}^{i_1} \circ \dots \circ S_{l_k}^{i_k}$$

where exponentiation means shuffle exponentiation.

Remark 3. It is a surprising fact that exactly the same formula hold for the Hall-Shirshov basis, as shown by M. P. Schützenberger (see [**11**] IV). We have already indicated that the Lyndon basis is not a particular case of the Hall-Shirshov basis. However, this analogy is mainly surprising because of (ii) (see remark 4).

Remark 4. As the proof will show, formulas (iii) hold in any envelopping algebra. More precisely, let $\mathcal{L}$ be any Lie algebra over $\mathbf{Q}$, $\mathcal{A}$ its enveloping algebra, $(h_i)_{i \in I}$ a totally ordered basis of $\mathcal{L}$. Then the decreasing products of h_i's form a basis of $\mathcal{A}$, by the PBW theorem. Let $\mathcal{A}'$ denote the dual space of $\mathcal{A}$, and

$$c: \mathcal{A} \longrightarrow \mathcal{A} \otimes \mathcal{A}$$

be the usual coproduct of $\mathcal{A}$ defined by

$$c(h) = h \otimes 1 + 1 \otimes h$$

for h in $\mathcal{L}$. Let $\circ$ be the transpose of c: then $\mathcal{A}'$ with $\circ$ becomes a commutative and associative algebra.

Let, for $i_1, \ldots i_k \in I$, $i_1 > \ldots > i_k$, $j_1, \ldots j_k \in \mathbf{N}, j_1, \ldots, j_k \geqq 1$,

$$\varphi_{i_1,\ldots,i_k}^{j_1,\ldots,j_k} \in \mathcal{A}'$$

denote the element of the dual basis of the PBW basis corresponding to $h_{i_1}^{j_1} \ldots h_{i_k}^{j_k}$. Hence we have, for any x in $\mathcal{A}$

$$x = \sum \varphi_{i_1,\ldots,i_k}^{j_1,\ldots,j_k}(x) h_{i_1}^{j_1} \ldots h_{i_k}^{j_k}$$

For simplicity, let $\varphi_i = \varphi_i^1$. Then theorem 2 (iii) extends to

$$\varphi_{i_1,\ldots,i_k}^{j_1,\ldots,j_k} = \frac{1}{j_1! \ldots j_k!} \varphi_{i_1}^{j_1} \circ \ldots \circ \varphi_{i_k}^{j_k}$$

(with $\circ$ exponentiation).

Before proving theorem 2, we need a lemma on the rewriting system of Section 2.

LEMMA 2. (i) *Let s be the standard sequence of Eq.* (1) *and suppose that $u_1 \geqq u_2, \ldots, u_n$ and $n \geqq 2$. Then for any sequence t such that $s \xrightarrow{*} t$, t is of length at least* 2.

(ii) *Let s be the standard sequence of Eq.* (1) *with $u_2 \geqq \ldots \geqq u_n$. If $u_1 \ldots u_n$ is a Lyndon word, then $s \xrightarrow{*} [u_1 \ldots u_n]$ at multiplicity one. Otherwise, $s \xrightarrow{*} t$ implies that t is of length at least* 2.

Proof. (i) (Induction of the length of the derivations $s \xrightarrow{*} t$.) By assumption, $[u_1][u_2]$ is not an inversion. So for the rightmost inversion $[u_i][u_{i+1}]$, one has $i \geqq 2$. This implies that s' and s'' defined by (2) and (3) are of length $\geqq 2$. Moreover, they satisfy to the same condition as s: their first term is greater than the others. This is clear for s'', and for s', note that $u_i u_{i+1} < u_{i+1}$, because $u_i u_{i+1}$ is a Lyndon word. So, one concludes by induction.

(ii) (Induction on n.) If $n = 1$, there is nothing to prove. Suppose $n \geqq 2$. If $u_1 \geqq u_2$, then s is decreasing, and so there is no derivation from s; moreover, $u_1 \ldots u_n$ is not a Lyndon word (because $n \geqq 2$, and by unicity of decomposition into Lyndon words). So in this case, we are done, too.

Hence, we may suppose that $u_1 < u_2$. This will be the rightmost inversion. So

$$s \longrightarrow s' = [u_2][u_1][u_3] \ldots [u_n] \quad \text{and}$$
$$s \longrightarrow s'' = [u_1 u_2]][u_3] \ldots [u_n].$$

Note that u_2 is the greatest term of s', so by (i), $s' \xrightarrow{*} t$ implies that t is of length at least 2. Moreover, s'' is shorter than s and satisfies to the same conditions as s: so we conclude by induction on n.

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

Proof of Theorem 2(i), (ii). (i) is clear, because $[1] = 1$.

(ii) We have just to show that for any word u and any letter a, one has

$$(S_{bv}, au) = \delta_{a,b}(S_v, u).$$

We have, by definition of the S's

$$u = \sum_{v \in A^*} (S_v, u)[v].$$

This implies that

$$au = \sum_{v} (S_v, u) a[v].$$

Let $v = l_1 \dots l_n$ be decomposed into a decreasing product of Lyndon words. Then the sequence $s = [a][l_1] \dots [l_n]$ is standard, hence by remark 1

$$a[v] = \sum_{\substack{s \xrightarrow{*} t \\ t \text{ decreasing}}} t$$

holds in $\mathbf{Z}\langle A\rangle$. By Lemma 2 (ii), the only such t which is of the form $t = [l], l$ Lyndon, is $[av]$ if av is Lyndon. Thus

$$a[v] = \epsilon(av)[av] + \sum_{k \geqq 2} {}_*[u_1] \dots [u_k]$$

with $\epsilon(av) = 1$ or 0 according to $av \in L$ or $av \notin L$.

This shows that

$$\begin{aligned} au &= \sum_{v} \epsilon(av)(S_v, u)[av] + \sum_{k \geqq 2} {}_*[u_1] \dots [u_k] \\ &= \sum_{av \in L} (S_v, u)[av] + \sum_{k \geqq 2} {}_*[u_1] \dots [u_k] \\ &= \sum_{bv \in L} (S_{bv}, au)[bv] + \sum_{k \geqq 2} {}_*[u_1] \dots [u_k]. \end{aligned}$$

This proves (ii).

Before proving (iii) we need to develop a little theory. Let $p > 1$ an integer and let

$$c_p : \mathbf{Z}\langle A\rangle \longrightarrow \mathbf{Z}\langle A\rangle^{\otimes p}$$

be the homomorphism for the concatenation product defined for all $a \in$ by:

$$c_p(a) = a \otimes 1 \otimes \dots \otimes 1 + 1 \otimes a \otimes \dots \otimes 1 + \dots + 1 \otimes 1 \otimes \dots \otimes a$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

where each tensor has p factors. If $w \in A^*$ we see that

$$c_p(w) = \sum u_1 \otimes \ldots \otimes u_p$$

where the sum runs over all p-uples of words $(u_1, \ldots, u_p)$ such that $w \in u_1 \circ \ldots \circ u_p$, with multiplicity. For example,

$$c_2(abb) = abb \otimes 1 + 2ab \otimes b + bb \otimes a + a \otimes bb + 2b \otimes ab + 1 \otimes abb.$$

The following lemmas are welll-known. We give a sketch of proof for sake of completeness.

LEMMA 3. *Let $S_1, \ldots, S_p \in \mathbf{Z} \ll A \gg$ and $P \in \mathbf{Z}\langle A \rangle$; then*

$$(S_1 \circ \ldots \circ S_p, P) = (S_1 \otimes \ldots \otimes S_p, c_p(P)).$$

Proof. We only have to prove the lemma when the S_i and P are words; but in that case it follows from the remark made above.

LEMMA 4. *If P is a Lie polynomial, that is, $P \in \mathcal{L}(A)$, then*

$$c_p(P) = P \otimes 1 \ldots \otimes 1 + 1 \otimes P + \ldots + 1 \otimes \ldots \otimes 1 \otimes P.$$

Proof. The lemma may be proven by proving that the set of polynomials satisfying the lemma contains the letters, is closed under linear combinations and is closed under Lie product, hence it contains $\mathcal{L}(A)$.

LEMMA 5. *Let $T_1, \ldots, T_i$ be series in $\mathbf{Z} \ll A \gg$ with constant terms equal to zero and let $Q_1, \ldots, Q_j$ be Lie polynomials. Suppose $i > j$; then*

$$(T_1 \circ \ldots \circ T_i, Q_1 \ldots Q_j) = 0.$$

Proof. Write

$$(T_1 \circ \ldots \circ T_i, Q_1 \ldots Q_j) = (T_1 \otimes \ldots \otimes T_i, c_i(Q_1 \ldots Q_j)).$$

Then the lemma follows from Lemma 4 and by noting that each tensor in $c_i(Q_1 \ldots Q_j)$ has a factor equal to one (since $i > j$) and that each T_l has zero constant term.

LEMMA 6. *Let $T_1, \ldots, T_p$ be series in $\mathbf{Z} \ll A \gg$ with constant terms equal to zero and let $Q_1, \ldots, Q_p$ be Lie polynomials. Then:*

$$(T_1 \circ \ldots \circ T_p, Q_1 \ldots Q_p) = \sum (T_1, Q_{\sigma(1)}) \ldots (T_p, Q_{\sigma(p)})$$

where the sum runs over all permutations σ of the set $\{1, \ldots, p\}$.

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

Proof. The result follows directly from Lemma 3 and Lemma 4.

Proof of Theorem 2 (iii). We particularize the result of the preceding lemmas to the case where the series $T_i = S_u$ for a word u and the Lie polynomials are the brackets $[l](l \in L)$. Note that each $S_w(w \neq 1)$ has its constant term equal to zero. Take $u_1, \ldots, u_n \in L$ with $u_1 \geqq \ldots \geqq u_n$ and $w \in A^*$; set $u = u_1 \ldots u_n$. We have

$$(S_w, [u_1] \ldots [u_n]) = \delta_{u,w},$$

where $\delta_{u,w}$ is equal to 1 if $u = w$ and to 0 if not, since $[u] = [u_1] \ldots [u_n]$. So in case $w \in L$ and $n \geqq 2$ we have $(S_w, [u]) = 0$. Now, if $w_1, \ldots, w_i$ and $u_1, \ldots, u_j$ are Lyndon words then Lemma 5 tells us that if $i > j$ we have:

$$(S_{w_1} \circ \ldots \circ S_{w_i}, [u_1] \ldots [u_j]) = 0.$$

If on the contrary, $i < j$, the equality remains true; this follows from the fact that

$$(S_{w_1} \circ \ldots \circ S_{w_i}, [u_1] \ldots [u_j]) = (S_{w_1} \otimes \ldots \otimes S_{w_i}, c_i([u_1] \ldots [u_j]))$$

and by noting that each tensor on the right contains a factors that is a product of the form $[u_{r_1}] \ldots [u_{r_s}]$ with $u_{r_1} \geqq \ldots \geqq u_{r_s}$ and $s \geqq 2$.

So the only possibility for $(S_{w_1} \circ \ldots \circ S_{w_i}, [u_1] \ldots [u_j])$ to be non-zero is when $i = j$; and in that case, according to Lemma 6, it is equal to the sum

$$\sum (S_{w_1}, [u_{\sigma(1)}]) \ldots (S_{w_i}, [u_{\sigma(i)}])$$

taken over all permutations σ of the set $\{1, \ldots, i\}$.

Take

$$w = l_1^{i_1} \ldots l_k^{i_k}$$

as in the statement of the theorem. Set $m = \sum i_r$ and consider $u_1, \ldots, u_m \in L$ such that $u_1 \geqq \ldots \geqq u_m$. We have

$$\begin{aligned}(S_{l_1}^{i_1} \circ \ldots \circ S_{l_k}^{i_k}, [u_1] \ldots [u_m]) \\ = \sum (S_{l_1}, [u_{\sigma(1)}]) \ldots (S_{l_1}, [u_{\sigma(i_1)}]) \ldots (S_{l_k}, [u_{\sigma(m-i_k+1)}] \ldots (S_{l_k}, [u_{\sigma(m)}])\end{aligned}$$

where the sum runs over all permutations σ of the set $\{1, \ldots, m\}$. This sum is non-zero if and only if

$$l_1 = u_1 = \ldots = u_{i_1}, \ldots, l_k = u_{m-i_k+1} = \ldots = u_m.$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

In that case the permutations σ for which the product

$$(S_{l_1}, [l_{\sigma(1)}]) \dots (S_{l_1}, [l_{\sigma(i_1)}]) \dots (S_{l_k}, [l_{\sigma(m-i_k+1)}]) \dots (S_{l_k}, [l_{\sigma(m)}])$$

is non-zero are the ones permuting the i_r factors l_r among themselves. For such a permutation the product evaluates to one and since there are $i_1! \dots i_k!$ such permutations, the sum adds up to $i_1! \dots i_k!$. We conclude that the two series S_w and

$$\frac{1}{i_1! \dots i_k!} S_{l_1}^{i_1} \circ \dots \circ S_{l_k}^{i_k}$$

take the same value on the basis elements $[u]$, so they must be equal. This completes the proof of Theorem 2.

COROLLARY 1. *Let $\mathcal{T}$ denote the complete tensor product $\mathbf{Q} \ll A \gg \otimes \mathbf{Q}\langle A \rangle$, where $\mathbf{Q} \ll A \gg$ has its shuffle structure and $\mathbf{Q}\langle A \rangle$ the concatenation structure. Then*

$$\sum_{w \in A^*} w \otimes w = \prod_{l \in L} \exp(S_l \otimes [l])$$

where the product is taken in decreasing order.

Proof. The right-hand side is

$$\prod_{l \in L} \left(\sum_{i \geqq 0} \frac{1}{i!} S_l^i \otimes [l]^i \right)$$
$$= \sum_{\substack{l_1 > \dots > l_k \\ i_1, \dots, i_k \geqq 1}} \frac{1}{i_1! \dots i_k!} S_{l_1}^{i_1} \circ \dots \circ S_{l_k}^{i_k} \otimes [l_1]^{i_1} \dots [l_k]^{i_k}.$$

This is equal, by Lyndon's theorem, Theorem 2 and our notation $[w]$, to

$$\sum_u S_u \otimes [u].$$

Hence, the corollary is equivalent to

$$w = \sum_u (S_u, w)[u]$$

which is true by definition.

Remark 5. There is an equivalent formulation of Corollary 1 in terms of Hopf algebras: it says that the identity of $\mathbf{Q}\langle A \rangle$ is the product of exponentials of special projections, in the algebra $\mathrm{End}(\mathbf{Q}\langle A \rangle)$ with the product

$$(f, g) \to h$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

with $h(w) = \pi \circ (f \otimes g) \circ c_2(w)$ and $\pi(u \otimes v) = uv$ (see [**13**] p. 71 for the definition of this product).

4. Basis for the shuffle algebra. We prove a result of D. E. Radford, as a consequence of Theorem 2.

THEOREM. ([**9**])(i) *Lyndon words form a transcendance basis of the shuffle algebra* $\mathbf{Q}\langle A\rangle$.

(ii) *More precisely, for any Lyndon word w, decomposed into Lyndon words as*

$$w = l_1^{i_1} \dots l_k^{i_k} \quad (l_1 > \dots > l_k; i_1, \dots, i_k \geqq 1)$$

one has

$$\frac{1}{i_1! \dots i_k!}\, l_1^{i_1} \dots l_k^{i_k} = w + \sum_{u<w} \alpha_u u \tag{7}$$

where α_u *is some natural integer and* l^i *means shuffle exponentiation.*

Part (i) was obtained differently by Perrin, Viennot [**8**].

Proof. Note that it is enough to prove (ii); by triangularity, the polynomials

$$P_w = \frac{1}{i_1! \dots i_k!}\, l_1^{i_1} \dots l_k^{i_k}$$

will form a basis of the $\mathbf{Z}$-module $\mathbf{Z}\langle A\rangle$.

Note first that P_w has integer coefficients: indeed, in $l_1^{i_1} \dots l_k^{i_k}$, each word has a coefficient divisible by $i_1! \dots i_k!$ The point is to show that w has coefficient one. By [**6**] Lemma 5.3.2, we know that for any Lyndon word l, one has

$$[l] = l + \sum_{u>l} {}_*u.$$

Because of the properties of the lexicographical order, this implies

$$[w] = w + \sum_{u>w} {}_*u$$

for any word w. By duality, we obtain

$$S_w = w + \sum_{u<w} {}_*u. \tag{5}$$

This implies by theorem 2 (iii) that for $w = l_1^{i_1} \dots l_k^{i_k}$ (Lyndon factorization)

$$S_w = \frac{1}{i_1! \dots i_k!}\, S_{l_1}^{i_1} \circ \dots \circ S_{l_k}^{i_k} = P_w + Q \tag{6}$$

https://doi.org/10.4153/CJM-1989-025-2 Published online by Cambridge University Press

where Q has nonnegative coefficients. Comparing (5) and (6), and knowing that w occurs in P_w with coefficient $\geqq 1$, we obtain that this coefficient must be 1.

Remark 6. Another transcendance basis of the shuffle algebra is the set $S_l, l \in L$. Indeed, by duality, each word w may be written

$$w = \sum_{u \in A^*} ([u], w) S_u$$

and one concludes using Theorem 2 (iii).

Part (i) of the previous result is equivalent to the following assertion: the shuffle algebra $\mathbf{Q}\langle A\rangle$ is isomorphic with the free commutative algebra $\mathbf{Q}[L]$ generated by the set L of Lyndon words over $\mathbf{Q}$. When one is only interested in the subalgebra $\mathbf{Z}\langle A\rangle$, then one obtains the following result, where we call *algebra of integral exponential polynomials* over L, the subalgebra of $\mathbf{Q}[L]$ which is linearly generated over $\mathbf{Z}$ by the monomials

$$\frac{l_1^{i_1} \dots l_k^{i_k}}{i_1! \dots i_1!} \quad (l_j \in L). \tag{8}$$

COROLLARY. *The shuffle algebra $\mathbf{Z}\langle A\rangle$ is isomorphic with the algebra of integral exponential polynomials over the set of Lyndon words.*

Proof. Let E be the algebra of integral exponential polynomials over $\mathbf{Z}$. Define a $\mathbf{Z}$-linear homomorphism $E \longrightarrow \mathbf{Z}\langle A\rangle$ by mapping the monomial (8) onto the polynomial (7) in $\mathbf{Z}\langle A\rangle$. This mapping is well defined and onto, by triangularity of (7). Moreover, it preserves the product of both algebras, hence it is an isomorphism.

Acknowledgements. We thank A. Joyal for many stimulating discussions. He indicated to us the paper of Radford [**9**], and suggested the corollary of Section 4.

Added in proof. The first author has recently shown that Theorems 1 and 2 hold – *mutatis mutandis* – for the bases considered by Viennof (Lecture Notes Maths. *691* Springer Verlag), which generalize both Lyndon and Hall bases.

REFERENCES

1. J. P. Duval, *Factorizing words over an ordered alphabet*, J. Algorithms *4* (1983), 363–381.
2. K. T. Chen, R. H. Fox and R. C. Lyndon, *Free differential calculus IV. – The quotient groups of the lower central series*, Ann. Math. *68* (1958), 81–95.
3. D. Foata, *La série génératrice exponentielle dans les problèmes d'énumération* (Presses Univ. Montréal, 1974).
4. M. Hall Jr, *The theory of groups* (Macmillan, New York, 1964).
5. P. Hall, *A contribution to theory of groups of prime-power order*, Proc. London Math. Soc. *36* (1933), 29–95.
6. M. Lothaire, *Combinatorics on words* (Reading, Massachusetts, 1983).

7. R. C. Lyndon, *On Burnside problem I*, Trans. Amer. Math. Soc. *77* (1954), 202–215.
8. D. Perrin and G. Viennot, *A note on shuffle algebras*, unpublished manuscript (1981).
9. D. E. Radford, *A natural ring basis for the shuffle algebra and an application to group schemes*, Journal of Algebra *58* (1979), 432–453.
10. C. Reutenauer, *Mots de Lyndon et un théorème de Shirshov*, Ann. Sci. Maths. Québec *10* (1986), 237–245.
11. M. P. Schützenberger, *Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un problème de mathématiques appliquées* (Algèbre et Théorie des Nombres) Paris (1958/59).
12. R. Ree, *Lie elements and an algebra associated with shuffles*, Ann. Math *68* (1958), 210–220.
13. M. E. Sweedler, *Hopf algebras* (Benjamin, 1969).

UQAM,
Montréal, Québec;
UQAM and CNRS,
Montréal, Québec