ELSEVIER

Theoretical Computer Science 273 (2002) 47–67

Theoretical Computer Science

www.elsevier.com/locate/tcs

Shuffle factorization is unique

Jean Berstel[a], Luc Boasson[b,*]

[a] *Institut Gaspard Monge (IGM), Université Marne-la-Vallée, 5, boulevard Descartes, 77454 Marne-la-Vallée Cédex 2, France*

[b] *Laboratoire d'informatique algorithmique: fondements et applications (LIAFA), Université Denis-Diderot, 2, place Jussieu, 75251 Paris Cédex 05, France*

Abstract

We prove that, given a finite set of words S, there exists at most one (normalized) multiset $\mathscr{P}$ such that S is the shuffle of the words in $\mathscr{P}$. The multiset $\mathscr{P}$ is effectively computable.

1. Introduction

Let A be an alphabet. The *shuffle* $x \sqcup\!\sqcup y$ of words x, y over A is a binary commutative and associative operation. The paper addresses the following problem: "Given a finite set S of words over A, do there exist words $x_1, x_2, \ldots, x_n$ such that

$$S = x_1 \sqcup\!\sqcup x_2 \sqcup\!\sqcup \cdots \sqcup\!\sqcup x_n \tag{1}$$

and, in the positive case, is the solution unique?"

A set S such that Eq. (1) holds, is called a *shuffle set*. Observe that if Eq. (1) holds, then all words in S have same length and even have same Parikh vector. This means that words in a solution $x_1, x_2, \ldots, x_n$ have bounded length, and thus an exhaustive search answers the question of existence, with a considerable cost. We give here a different method. This method could presumably lead to an efficient algorithm. This will be a topic of later work.

Next, if a solution $x_1, x_2, \ldots, x_n$ exists, there are easy cases where it is not unique, even up to permutation. Consider a single letter a, and any power a^n with $n \geqslant 2$. Then $a^n = a^{p_1} \sqcup\!\sqcup a^{p_2} \sqcup\!\sqcup \cdots \sqcup\!\sqcup a^{p_k}$ for every $p_1 + p_2 + \cdots + p_k = n$, showing that the solution is not unique at all. An extreme case is when a^n is replaced by n copies of a. It appears that this is the only case of ambiguity. More precisely, we show that,

* Corresponding author.

E-mail address: luc.boasson@liafa.jussieu.fr (L. Boasson).

0304-3975/02/$ - see front matter
PII: S0304-3975(00)00433-3

when powers of a single letter are decomposed into sums of letters (a process we call normalization) then the solution is unique. We call *shuffle root* of S any normalized solution.

Shuffle factorization has been studied extensively when multiplicities are considered. There are quite precise descriptions of the shuffle algebra $Q\langle A\rangle$ of polynomials with rational coefficients. A theorem of Radford [5] states that the shuffle algebra $Q\langle A\rangle$ is freely generated by the Lyndon words (see, e.g., [6, Theorem 6.1]). This means in particular that every finite set S, viewed as a polynomial with coefficients 0 and 1 is a shuffle of a finite polynomial over Lyndon words, but perhaps with negative coefficients. For instance, since $a \sqcup\!\sqcup b = ab + ba$ and ab is a Lyndon word (assuming $a < b$), one has $ba = ab - a \sqcup\!\sqcup b$.

Our approach is different. We consider only ordinary sets of words that are obtained as shuffles of (multi)sets of words, without considering multiplicities in the result, and without considering "negative" multiplicities in the words to be shuffled. The main result is the following.

Theorem 1.1. *A finite shuffle set has a unique shuffle root.*

Obviously, this root is effectively computable. In other words, given a finite set S of nonempty words, there exists at most one effectively computable normalized multiset $\mathscr{P}$ such that $S = \sqcup\!\sqcup \mathscr{P}$.

Our method is constructive. Given the set S, we try to compute some polynomial $\mathscr{P}$ that is a candidate for the equality $S = \sqcup\!\sqcup \mathscr{P}$. We show that if the computation fails, the set S is not a shuffle set. If the computation succeeds, it yields a unique polynomial $\mathscr{P}$. If $S \neq \sqcup\!\sqcup \mathscr{P}$, then S is not a shuffle set. Otherwise, $\mathscr{P}$ is the unique polynomial such that $S = \sqcup\!\sqcup \mathscr{P}$. Thus, our proof gives at most one "candidate" polynomial. Even if this candidate exists, it must be checked that it indeed produces the set S.

The construction of the candidate polynomial is in two parts. The first is a reduction step that we call separation. The second part is a divide-and-conquer method that strips off parts of the words, and parts in the words to get simpler and simpler sets. A full example of the separation process is given at the end of Section 3. A large example of the second step appears in Section 6.

The separation reduces the computation to the case where all words in S start with a fixed letter (and all end with this letter, or all end with other letters). The candidate polynomials for the separated sets are computed in the second part. Then they are composed to yield, if the composition succeeds, a candidate that is checked for being the shuffle root.

The initial letter, in the separation process, is called later the pivot. It is used as a marker for decomposing words into blocks. Each block starts with pivot letters, and contains no other pivot letters. It appears that shuffling words can be restricted, with some care, to shuffling sequences of blocks: the blocks play the role of some metaalphabet.

The second part uses two kinds of reductions, called shadow and trim reductions. Both reduce a given separated set as produced in the first step to smaller sets of the same kind. The problem is solved for the two subcases, and their solutions are combined to a solution of the initial set, provided a solution exists. Both reductions operate on blocks. A shadow reduction consists in selecting minimal initial block prefixes, and a trim reduction cuts initial blocks. Considering both reductions is necessary to guarantee that the composed solution is uniquely determined. A commented example is given after Lemma 5.2.

As it will appear, the proof of the theorem also provides an algorithm to test whether a finite set is a shuffle set. This algorithm seems to be much faster than the naive algorithm, however we have no proof for this. Algorithms exist for related problems. Thus, Ref. [7] computes the shuffle of two words without repetition. An algorithm for the computation of the shuffle of a finite set of words, that is efficient in time and in space, has been developed recently by Allauzen [1]. It is based on a clever use of suffix trees. He shows that the shuffle of n words $x_1,\dots,x_n$ can be computed in time and space the multinomial

$$\mathrm{O}\left(\binom{|x_1|+\cdots+|x_n|}{|x_1|,\dots,|x_n|}\right),$$

and that the shuffle set can be stored in linear space, i.e., in space $\mathrm{O}(\mathrm{Card}\, x_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup x_n)$. Ref. [4] gives a parallel algorithm for testing whether a word is in the shuffle of two words. See also [3]. General properties of the shuffle operation are given in [2, Chapter 6].

2. Notation

We use standard notation. The empty word is denoted by ε. We recall the definition of radix order. Let A be an ordered alphabet. The *radix order* over A^* is a total order defined by $u<v$ if $|u|<|v|$ or $|u|=|v|$ and $u<v$ in the lexicographic order on words (of same length). This order is clearly prefix preserving, i.e., $x<xy$ for every nonempty word y.

Shuffle: Let A be an alphabet. The *shuffle* of two words over A is defined for words x, y and letters a, b by

$$x \sqcup\!\sqcup \varepsilon = \varepsilon \sqcup\!\sqcup x = x,$$

$$ax \sqcup\!\sqcup by = a(x \sqcup\!\sqcup by) \cup b(ax \sqcup\!\sqcup y).$$

The shuffle is a binary commutative and associative operation. We will be concerned with the shuffle of nonempty words. The shuffle operation is extended to sets as usual by

$$X \sqcup\!\sqcup Y = \bigcup_{x\in X, y\in Y} x \sqcup\!\sqcup y.$$

We may also consider shuffling all words in a given set. Thus, for a set $X = \{x_1, x_2, \ldots, x_n\}$, we can define $\sqcup\!\sqcup(X) = x_1 \sqcup\!\sqcup x_2 \sqcup\!\sqcup \cdots \sqcup\!\sqcup x_n$. This *unary* shuffle operation is in fact defined for multisets. Indeed consider for instance

$$a \sqcup\!\sqcup ab \sqcup\!\sqcup ab = \{ababa, aba^2b, a^2bab, a^2b^2a, a^3b^2\}.$$

This is quite different from $a \sqcup\!\sqcup ab = \{aba, a^2b\}$. We call *shuffle set* any finite set S of nonempty words that is the shuffle of a multiset of words.

Multisets: A finite *multiset* is a finite set $\mathscr{P}$ with a multiplicity for each of its elements. It can also be viewed as a *polynomial* with nonnegative coefficients. We denote by $(\mathscr{P}, w)$ the multiplicity of w as an element of $\mathscr{P}$. It is convenient to write $\mathscr{P}$ as a polynomial, each word carrying its multiplicity. Thus, $\mathscr{P} = a + 2ab$ means that ab appears two times in $\mathscr{P}$ and a once.

The *underlying set* or *support* of $\mathscr{P}$ is the set of words with positive multiplicity. As an example, the support of the multiset $\mathscr{P} = a + 2ab$ is $\{a, ab\}$. We will say that a word w is an element of the multiset, or *is in* the multiset $\mathscr{P}$, if it is an element of the support of $\mathscr{P}$.

Observe that, in our problem, the result of shuffling a multiset is viewed as an ordinary set without multiplicities. One could also consider the multiset obtained by shuffling words in $\mathscr{P}$ by counting the number of times each word is obtained. In the above example, this would give $2ababa + 4aba^2b + 8a^2bab + 4a^2b^2a + 12a^3b^2$. However, in this paper, we only consider the ordinary sets produced by shuffling multisets. That is, we forget the coefficients in the result.

The difference $\mathscr{P} - \mathscr{P}'$ of two multisets $\mathscr{P}$ and $\mathscr{P}'$ is defined if $(\mathscr{P}, w) \geqslant (\mathscr{P}', w)$ for all words w. It is then given by $(\mathscr{P} - \mathscr{P}', w) = (\mathscr{P}, w) - (\mathscr{P}', w)$.

Normalization: A multiset $\mathscr{P}$ is *proper* if it does not contain the empty word, that is if $(\mathscr{P}, \varepsilon) = 0$. A proper multiset $\mathscr{P}$ is *normal for the letter* a if the only power of a that is an element of $\mathscr{P}$ is a itself. Thus, $\mathscr{P} = 37a + 2ab$ is normal for a, but $\mathscr{P}' = \varepsilon + a^2 + 7a^5 + 2ab$ is not. A *normal* multiset is a multiset that is proper and normal for all letters.

The *normalization* of a multiset $\mathscr{P}$ consists in two steps. First, the empty word is removed. Second, the powers of a letter a that are in $\mathscr{P}$ are "summed up" to a single monomial $H * a$, with $H = \sum_{n>0} (\mathscr{P}, a^n) * n$. Thus, for $\mathscr{P}' = \varepsilon + a^2 + 7a^5 + 2ab$ and for the letter a, one obtains $H = 2 + 7 * 5 = 37$, thus the normalization of $\mathscr{P}'$ yields $\mathscr{P} = 37a + 2ab$. The normalized multiset of $\mathscr{Q}$ is denoted by $\nu(\mathscr{Q})$.

The reason for doing normalization is the following. The powers of a letter a contribute a single word in the shuffle obtained by the formula given for H. Thus, $\sqcup\!\sqcup(a^2 + 7a^5) = a^{37}$, and $\sqcup\!\sqcup(a^2 + 7a^5 + \mathscr{Q}) = \sqcup\!\sqcup(37a + \mathscr{Q})$ for every polynomial $\mathscr{Q}$. Therefore, replacing powers of a letter by an appropriate multiplicity of the letter will produce the same shuffle set.

A normalized multiset $\mathscr{P}$ such that $S = \sqcup\!\sqcup \mathscr{P}$, is called a *shuffle root* of S.

3. Separation

In this section, we introduce a reduction called separation. This operation reduces a shuffle set and its shuffle root to a simpler pair, with the property that every word in the reduced shuffle root starts with a given letter (and ends with this letter or never ends with this letter).

Two kinds of separations will be defined. The first is applied to sets that are supposed to be shuffle sets, the second one to multisets that are supposed to be their shuffle root. We prove in Lemma 3.3 that the relation between a shuffle set and its shuffle root is preserved through separation. Moreover, Proposition 3.4 shows that the shuffle root of the initial set, if it exists, is determined by the shuffle roots of the separated sets.

A set of words S is called *a-simple* if $S \subset aA^* \setminus A^*a$, that is if all its words start with the letter a, and none of its words end with a. A set S is *a-symmetric* if $S \subset aA^* \cap A^*a$, that is if all its words start and end with the letter a. A set is simple (symmetric) if it is a-simple (a-symmetric) for some a. A multiset is simple (symmetric) if its support is simple (symmetric).

For instance, the set $S = \{a^2bc, acb\}$ is a-simple. Clearly, every set $S \subset A^+$ is a disjoint union of simple sets and symmetric sets.

Let B be a set of letters. For each word $w \in A^*$, we denote by $\langle B\rangle w$ the longest suffix of w starting with a letter in B, if such a suffix exists, the empty set otherwise. Equivalently, $\langle B\rangle w$ is the word that remains when the longest prefix in $(A\setminus B)^*$ is removed from w (and is the empty set when this longest prefix is w itself). For an intuitive explanation, see also the comment following Eq. (2).

For instance, if $B = \{b, c\}$, one has $\langle B\rangle a^2dba^2c = ba^2c$. This operation is extended to multisets by linearity. Symmetrically, one defines $w\langle B\rangle$.

Example. Consider $\mathscr{P} = a + 2ab$. Then $\langle a\rangle\mathscr{P} = \mathscr{P}$, $\langle b\rangle\mathscr{P} = 2b$, $\mathscr{P}\langle a\rangle = 3a$, $\mathscr{P}\langle b\rangle = 2ab$, $(\langle a\rangle\mathscr{P})\langle b\rangle = \langle a\rangle(\mathscr{P}\langle b\rangle) = 2ab$.

The following lemma will be used in the sequel for two special partitions. In the first partition, each A_i will consist of a single letter. The second partition will have two sets, one composed of a single letter, the other being the rest of the alphabet.

Lemma 3.1. *Let $A = A_1 \cup A_2 \cup \cdots \cup A_n$ be a partition of the alphabet A. Given normalized multisets $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_n$, there exists at most one, effectively computable normalized multisets $\mathscr{P}$ such that $\langle A_i\rangle\mathscr{P} = \mathscr{P}_i$ for $i = 1, \ldots, n$.*

Proof. Clearly, if $\langle A_i\rangle\mathscr{P} = \mathscr{P}_i$ then every word in $\mathscr{P}_i$ starts with a letter in A_i. Let ℓ_i be the maximal length of the words in $\mathscr{P}_i$, for $i = 1, \ldots, n$. If $\ell_i = 0$, then $\mathscr{P}_i = 0$. If $\ell_i = 1$, then $\mathscr{P}_i$ is just a linear combination of letters: $\mathscr{P}_i = \sum_{a\in A_i} (\mathscr{P}_i, a)a$. Thus, if $\ell_i \leqslant 1$ for $i = 1, \ldots, n$, then the multiset $\mathscr{P} = \mathscr{P}_1 + \cdots + \mathscr{P}_n$ is a solution, and it is the only solution. Arguing by induction on $\ell = \max(\ell_1, \ldots, \ell_n)$, consider the polynomial $\mathscr{Q} = \sum_{i=1}^{n} \sum_{|x|=\ell} (\mathscr{P}_i, x)x$. Each word x in $\mathscr{Q}$ is also in $\mathscr{P}$ and conversely, each word

of length ℓ in $\mathscr{P}$ is in $\mathscr{Q}$, because the support of $\mathscr{P}$ is contained in the union of the supports of the P_i. Thus $\mathscr{Q} = \sum_{|x|=\ell} (\mathscr{P}, x)x$. Consider $\mathscr{P}' = \mathscr{P} - \mathscr{Q}$ and $\mathscr{P}_i' = \mathscr{P}_i - \langle A_i \rangle \mathscr{Q}$. Then $\langle A_i \rangle \mathscr{P} = \mathscr{P}_i$ if and only if $\langle A_i \rangle \mathscr{P}' = \mathscr{P}_i'$, and the lemma is proved by induction. □

Example. Consider the alphabet $A = \{a, b\}$. We look for a multiset $\mathscr{P}$ such that $\langle a \rangle \mathscr{P} = 2a + ab$ and $\langle b \rangle \mathscr{P} = b + ba$. Words of length 2 in $\mathscr{P}$ are ab and ba, so $\mathscr{P} = ab + ba + \mathscr{P}'$ for some $\mathscr{P}'$ composed of letters. Next, we compute $\langle a \rangle \mathscr{P} = ab + a + \langle a \rangle \mathscr{P}'$, and since $\langle a \rangle \mathscr{P} = 2a + ab$, it follows that $\langle a \rangle \mathscr{P}' = a$. Similarly, $\langle b \rangle \mathscr{P} = b + ba + \langle b \rangle \mathscr{P}'$, whence $\langle b \rangle \mathscr{P}' = 0$. Thus, $\mathscr{P}' = a$ and $\mathscr{P} = a + ab + ba$.

As a second example, we look for a multiset $\mathscr{P}$ such that $\langle a \rangle \mathscr{P} = a$ and $\langle b \rangle \mathscr{P} = 2ba$. The word ba is the only word of length 2 in $\mathscr{P}$. Thus, $\mathscr{P} = 2ba + \mathscr{P}'$, and $a = \langle a \rangle \mathscr{P} = 2a + \langle a \rangle \mathscr{P}'$. Thus, the multiset we look for does not exist.

Let A be an ordered alphabet, and let B be a subset of A. Given a finite set S, we consider the set of words over the alphabet $A \backslash B$ that are prefixes of words in S. Among these, we choose the word p that is maximal for the radix order. We denote by $[B]S$ the set

$$[B]S = p^{-1}S \cap BA^*.$$

Symmetrically, one defines $S[B]$. The choice of p in the definition above is rather arbitrary. It is only necessary, for the propositions below, that p is a prefix of maximal length.

Example. Let $S = \{ababa, aba^2b, a^2bab, a^2b^2a, a^3b^2\}$. Then $[a]S = S$, $[b]S = (a^3)^{-1}S = \{b^2\}$, $S[a] = S(b^2)^{-1} = \{a^3\}$, $S[b] = Sa^{-1} = \{abab, a^2b^2\}$.

The next proposition shows that, for shuffle sets, this has no consequence.

Proposition 3.2. *Let S be a finite set, let B be a subset of A, and let p, q be words over the alphabet $A \backslash B$ of maximal length that are prefixes of words in S. If S is a shuffle set, then $p^{-1}S = q^{-1}S$.*

Proof. Assume that $S = x_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup x_n$, and set $x_i = p_i y_i$, where p_i is the longest prefix of x_i that is in $(A \backslash B)^*$ and $y_i \in \varepsilon \cup BA^*$. Then p and q both are in the set $p_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup p_n$, and

$$p^{-1}S = q^{-1}S = y_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup y_n. \qquad \square$$

The lemma does not hold if the set S is not a shuffle set. Indeed, the set $S = \{badc, abcd\}$ is not a shuffle set. For $B = \{c, d\}$, the prefixes ba and ab have maximal length. However, $(ba)^{-1}S = \{dc\} \neq (ab)^{-1}S = \{cd\}$.

The next lemma shows how a pair (shuffle set, shuffle root) goes through a separation.

Lemma 3.3. *Let $A = A_1 \cup A_2 \cup \cdots \cup A_n$ be a partition of the alphabet A, and let S be a finite set over A. If $S = \sqcup\!\sqcup \mathscr{P}$ for some multiset $\mathscr{P}$, then $[A_i]S = \sqcup\!\sqcup \langle A_i\rangle \mathscr{P}$ for each $i = 1, \ldots, n$.*

Proof. As in the previous lemma, we factorize any word x in $\mathscr{P}$ into $x = p_i(x)y_i(x)$, where $p_i(x)$ is the longest prefix of x that is in $(A\backslash A_i)^*$ and $y_i(x) \in \varepsilon \cup A_iA^*$. Then $[A_i]S = \sqcup\!\sqcup \sum(\mathscr{P}, x)y_i(x)$. Since $\sum(\mathscr{P}, x)y_i(x) = \langle A_i\rangle \mathscr{P}$, the lemma is proved. □

The lemma can also be stated in a more compact form as follows. If $\mathscr{P}$ is a polynomial and B is a subset of the alphabet A, then

$$\sqcup\!\sqcup \langle B\rangle \mathscr{P} = [B] \sqcup\!\sqcup \mathscr{P} \tag{2}$$

showing how the shuffle operation passes through the [ ]-operator.

As an explanation of (2), consider three words $u = pax$, $v = qay$, $w = raz$, where p, q, r do not contain the letter a. Any word in the shuffle of these words has the form tas where t does not contain the letter a. The prefix t is in the shuffle of some prefixes of u, v, w, and at least one of the words p, q, r has been "used" entirely for the construction of t. If we look for words in the shuffle with the longest possible prefix t, then t is in the shuffle $p \sqcup\!\sqcup q \sqcup\!\sqcup r$, and as is in the shuffle $ax \sqcup\!\sqcup ay \sqcup\!\sqcup az$.

The next proposition is the first special case of partition, concerning a partition into letters.

Proposition 3.4. *Let S be a finite set.*

(i) *If S is a shuffle set, then $[a]S$ is a shuffle set for each letter $a \in A$.*

(ii) *Conversely, if $[a]S$ is a shuffle set with shuffle root $\mathscr{P}_a$ for each $a \in A$, then there exists at most one, effectively computable multiset $\mathscr{P}$ such that $S = \sqcup\!\sqcup \mathscr{P}$ and $\langle a\rangle \mathscr{P} = \mathscr{P}_a$ for $a \in A$.*

Proof. The first implication is a direct consequence of Lemma 3.3. The converse implication is just Lemma 3.1. □

It is not difficult to show that, given subsets B, C of the alphabet, and a set S of words, the sets $([B]S)[C]$ and $[B](S[C])$ are equal. Similarly, for a polynomial $\mathscr{P}$, one has $(\langle B\rangle \mathscr{P})\langle C\rangle = \langle B\rangle(\mathscr{P}\langle C\rangle)$. For clarity, we introduce the notation ${}_BS_C = [B]S[C]$ when S is a set, and ${}^B\mathscr{P}^C = \langle B\rangle \mathscr{P}\langle C\rangle$ when $\mathscr{P}$ is a polynomial.

We will be interested in the sets ${}_aS_a$ and ${}_aS_{A\backslash a}$, and in the polynomials ${}^a\mathscr{P}^a$ and ${}^a\mathscr{P}^{A\backslash a}$ for $a \in A$.

Theorem 3.5. *Let S be a finite set of words.*

(i) *If S is a shuffle set, then the sets ${}_aS_a$ and ${}_aS_{A\backslash a}$ ($a \in A$) are shuffle sets.*

(ii) *Conversely, if the sets ${}_aS_a$ and ${}_aS_{A\backslash a}$ ($a \in A$) are shuffle sets, and if polynomials $\mathscr{Q}_a$ and $\mathscr{R}_a$ are given such that ${}_aS_a = \sqcup\!\sqcup \mathscr{Q}_a$ and ${}_aS_{A\backslash a} = \sqcup\!\sqcup \mathscr{R}_a$ for $a \in A$, then there exists at most one effectively computable polynomial $\mathscr{P}$ such that $S = \sqcup\!\sqcup \mathscr{P}$ and ${}^a\mathscr{P}^a = \mathscr{Q}_a$ and ${}^a\mathscr{P}^{A\backslash a} = \mathscr{R}_a$.* □

In view of this result, the computation of a shuffle root of a set can be reduced to the cases we called a-simple and a-symmetric. In other words, a given set S for which a shuffle root is looked for, is first decomposed into simple and symmetric sets ${}_aS_a$ and ${}_aS_{A\setminus a}$ $(a\in A)$. For each of these sets, a shuffle root is computed, provided it exists. There is at most one polynomial that can be constructed by composing these roots. The initial set S is a shuffle set only if the composed polynomial exists. Even then, the set S might not be a shuffle set. For a detailed example, see Section 6. For further reference, we state this fact as a corollary.

Corollary 3.6. *Let S be a finite set of nonempty words. If, for each $a\in A$, the a simple set ${}_aS_a$ and a-symmetric set ${}_aS_{A\setminus a}$ $(a\in A)$ is a shuffle set and has a unique shuffle root, then the set S has at most one effectively computable shuffle root.* □

The proof of Theorem 1.1 thus reduces to the case of simple and symmetric sets. We finish this section with an example illustrative of the separation process. See also Section 6.

Example. Let S be the set

$$aabab\ aabba\ ababa\ \ abaab\ abbaa$$
$$baaab\ baaba\ babaa.$$

Then

$$[a]S=\{aaab,aaba,abaa\},\quad [b]S=\{bab,bba\}$$

and

$$[a]S[a]=\{a^3\},\quad [a]S[b]=\{ab\},\quad [b]S[a]=\{ba\},\quad [b]S[b]=\{b^2\}.$$

Since these four sets are singletons, they are their own shuffles, and after normalization, we get

$$[a]S[a]=\sqcup\!\sqcup\, 3a,\quad [a]S[b]=\sqcup\!\sqcup\, ab,\quad [b]S[a]=\sqcup\!\sqcup\, ba,\quad [b]S[b]=\sqcup\!\sqcup\, 2b.$$

If $[a]S=\sqcup\!\sqcup\,\mathscr{P}$ for a polynomial $\mathscr{P}$, then $\mathscr{P}=\alpha ab+\beta a$ for some nonnegative integers α,β. From $\mathscr{P}\langle a\rangle=(\alpha+\beta)a=3a$ and $\mathscr{P}\langle b\rangle=\alpha ab=ab$ Content-Length: 22322.

From $\mathscr{P}\langle a\rangle=(\alpha+\beta)a=3a$ and $\mathscr{P}\langle b\rangle=\alpha ab=ab$ we get $\mathscr{P}=ab+2a$.

Thus the only possibility is $[a]S=\sqcup\!\sqcup(2a+ab)$. Similarly, one must have $[b]S=\sqcup\!\sqcup(b+ba)$. Next, if $S=\sqcup\!\sqcup\,\mathscr{Q}$, then

$$\mathscr{Q}=\alpha a+\beta ab+\gamma b+\delta ba$$

for some nonnegative integers α, β, γ, δ, and since

$$\langle a\rangle\mathscr{Q}=(\alpha+\delta)a+\beta ab=2a+ab,\quad \langle b\rangle\mathscr{Q}=(\beta+\gamma)b+\delta ba=b+ba$$

we get $\alpha=\beta=\delta=1,\gamma=0$ and $\mathscr{Q}=a+ab+ba$. A final check shows that indeed $S=a\sqcup\!\sqcup\, ab\sqcup\!\sqcup\, ba$. Observe that in the seperation process, only five of the words in S

were used. This means that if we had omitted any of the unused words *ababa*, *abaab* or *abbaa*, the candidate polynomial $\mathcal{Q}$ would have been the same. This shows that the final check is mandatory.

4. Blocks

In this section, we consider a canonical decomposition of words induced by a distinguished letter a, assuming all words start with this letter. An intricate-order relation will be defined that allows to order blocks. The main observation is that minimal words (for the radix order) in a shuffle set can be computed by considering minimal words for the block order in the shuffle root. These properties will be used in the next section.

Pivot and blocks: We fix a distinguished letter a to be the *pivot* and set $B = A\setminus\{a\}$. A *block* is any word in a^+B^*, a *regular block* is a word in a^+B^+. A block in a^+ is *irregular*. Any word w in aA^* has a unique factorization into blocks. Only the last block of this factorization may be irregular. The *initial block* of $x \in aA^*$ is the first block in its factorization. The *block degree* $\deg(x)$ of a word $x \in aA^*$ is the number of blocks in its factorization.

Order: We consider any total order relation $<$ on B^* that is compatible with the prefix ordering, i.e., such that $x < xy$ for $x \in B^*$, $y \in B^+$. The radix order is an example of such an order. We define an order on blocks called *block order*, and also denoted by $<$, by the following requirements. First, irregular blocks are ordered by their exponent: $a^p < a^n$ iff $p < n$. Next, any irregular block is greater than any regular block: $a^p u < a^n$ for any n, p and every nonempty word u over B. Finally, regular blocks are ordered first by the exponent of the pivot and, if these are equal, by the *opposite* of the order on B^*: for $u, v \in B^+$,

$$a^p u < a^n v \Leftrightarrow p < n \quad \text{or} \quad (n = p \text{ and } u > v).$$

Observe the inversion of the order on u and v.

The block order is extended on aA^* by the lexicographic extension of the order on blocks. For this, consider

$$x = \beta_1\beta_2\cdots\beta_n, \quad x' = \beta'_1\beta'_2\cdots\beta'_{n'}.$$

We say that x is a proper *block prefix* of x' if $n < n'$ and $\beta_i = \beta'_i$ for $1 \leqslant i \leqslant n$. We next define $x \sqsubset x'$ if and only if there is an integer $k \leqslant n, n'$ such that $\beta_i = \beta'_i$ for $1 \leqslant i < k$ and $\beta_k < \beta'_k$, and we set $x < x'$ if and only if $x \sqsubset x'$ or x is a proper block prefix of x'.

For blocks β and β', one has $\beta < \beta'$ iff $\beta \sqsubset \beta'$.

Example. Consider the block order induced by the radix order on $B = \{b, c\}$ with $b < c$, and consider words $x = a^2b^5c^5ba^2b^3c^5b$, $x' = a^2b^3c^3b$, $x'' = a^2b^5c^5ba$. We get block factorizations $x = \beta_1\beta_2$, $x' = \beta'_1$, $x'' = \beta_1\gamma$ with $\beta_1 = a^2b^5c^5b$, $\beta_2 = a^2b^3c^5b$, $\beta'_1 = a^2b^3c^3b$, $\gamma = a$. To show that $x < x'$, we compare β_1 and β'_1. They have the same initial power of

the letter a, namely a^2. Since $|\beta_1| > |\beta_1'|$, we get $\beta_1 < \beta_1'$ and consequently $x < x'$. Next consider x and x''. The word x'' is a prefix of x as a word of A^*. However, $\beta_2 < \gamma$ and therefore $x < x''$.

Properties of the order: The order relation over blocks helps in reducing (frequently) shuffle of words to rearrangements of blocks by inspection of special representatives. Given a set $X \subset aA^*$, we denote by $\min(X)$ the smallest element in X for the block order. We shall see that if $x = \beta_1 \cdots \beta_n$ and $y = \gamma_1 \cdots \gamma_m$, then $\min(x \sqcup\!\sqcup y)$ is the product of $\beta_1, \ldots, \beta_n, \gamma_1, \ldots, \gamma_m$ in some order.

First, given blocks β, γ, one has

$$\beta \leqslant \gamma \Rightarrow \min(\beta \sqcup\!\sqcup \gamma) = \beta\gamma.$$

Observe that $\beta\gamma$ might be a single block if β, γ are both irregular. Next, if x has the block decomposition $x = \beta_1\beta_2 \cdots \beta_n$ for some $n \geqslant 2$, then

$$\min(\beta \sqcup\!\sqcup x) = \beta_1\beta_2 \cdots \beta_i\beta\beta_{i+1} \cdots \beta_n,$$

where i is the greatest index such that $\beta_j < \beta$ for $j = 1, \ldots, i$. For instance, if $\beta = a^2b^2$ and $x = ab^3 \cdot a^2b^3 \cdot a^2b$, then $\min(\beta \sqcup\!\sqcup x) = ab^3 \cdot a^2b^3 \cdot a^2b^2 \cdot a^2b$. The relation extends to words in aA^* as follows. Let $x = \beta x'$, $y = \gamma y'$ for (regular) blocks β, γ and words $x', y' \in aA^*$. Then

$$\min(x \sqcup\!\sqcup y) = \begin{cases} \beta \min(x' \sqcup\!\sqcup y) & \text{if } \beta < \gamma, \\ \beta \min(\min(x' \sqcup\!\sqcup y), \min(x \sqcup\!\sqcup y')) & \text{if } \beta = \gamma. \end{cases}$$

Indeed, assume $\beta \leqslant \gamma$ and set $\beta = a^n u$ and $\gamma = a^p v$ $(u, v \neq \varepsilon)$. The word $\min(x \sqcup\!\sqcup y)$ starts with the smallest possible numbers of a's, that is by a^n. This group of a's is followed by the longest possible word containing no a. If $n < p$, this must be u. If $n = p$, it is also u if $\beta < \gamma$, and if $\beta = \gamma$ then it is $u = v$. Thus, in all cases, $\min(x \sqcup\!\sqcup y)$ starts with β. It is followed by the smallest possible word. This is $\min(x' \sqcup\!\sqcup y)$ if $\beta < \gamma$. If $\beta = \gamma$, the smallest possible word is either $\min(x' \sqcup\!\sqcup y)$ if the initial block β of $\min(x \sqcup\!\sqcup y)$ is the initial block of x, or $\min(x \sqcup\!\sqcup y')$ if the block β is the initial block of y. This proves the formula. The three formulas allow to compute the minimal word in the shuffle of two words in aA^*.

Next, for blocks β, γ and the word $x \in aA^*$,

$$\beta < \gamma \Rightarrow x\beta \sqsubset x\gamma$$

and for blocks β, γ and a words $x, y \in aA^*$,

$$\beta < \gamma \quad \text{and} \quad \gamma \text{ regular} \Rightarrow \beta x \sqsubset \gamma y.$$

The last relation does not hold if γ is irregular. Consider for instance $\beta = a$, $\gamma = a^2$ and take $x = a^3$ and $y = a$. Then $\beta < \gamma$ and $a^4 = \beta x \sqsupset a^3$.

Next, we show that for $x, y, y' \in aA^*$,

$$y \sqsubset y' \Rightarrow \min(x \sqcup\!\sqcup y) \sqsubset \min(x \sqcup\!\sqcup y').$$

Indeed, set $y = p\beta s$, $y' = p\gamma s'$ for blocks $\beta < \gamma$ and words $p, s, s' \in aA^* \cup \{\varepsilon\}$. Consider $w' = \min(x \sqcup\!\sqcup y')$. There is a factorization $x = x_1 x_2$ of x into products of blocks such that $w' = u\gamma v'$ for words $u \in x_1 \sqcup\!\sqcup p$, $v' \in x_2 \sqcup\!\sqcup s'$. Consider $w = u\beta v$, for $v \in x_2 \sqcup\!\sqcup s$. If γ is regular, then $w \sqsubset w'$. If γ is irregular, then s' is the empty word and $v' = x_2$. Thus, x_2 is an irregular block, because $\gamma x_2 = \gamma v'$ is the last block of w'. Since $x_2\gamma = \gamma x_2$, we may assume $x = x_1$, $x_2 = \varepsilon$ in the factorization of x. Thus, $w = u\beta$ and again $w \sqsubset w'$. Thus, $\min(x \sqcup\!\sqcup y) \leqslant w \sqsubset w' = \min(x \sqcup\!\sqcup y')$.

Lemma 4.1. *Let Y be a subset of aA^*. If no word in Y is a proper block prefix of another word in Y, then for any word x in aA^*,*

$$\min(x \sqcup\!\sqcup Y) = \min(x \sqcup\!\sqcup \min(Y)).$$

Proof. By definition, $\min(x \sqcup\!\sqcup Y) = \min_{y \in Y}(x \sqcup\!\sqcup y)$. Since $y \sqsubset y'$ implies that $\min(x \sqcup\!\sqcup y) \sqsubset \min(x \sqcup\!\sqcup y')$, the result follows by induction on the size of Y. □

Example. Consider $x = a^2bab$ and $Y = \{a^2bab, a^2ba^2b, a^2ba^3b, a^4b\}$. Then $\min(x \sqcup\!\sqcup a^2bab) = a^2baba^2bab$, $\min(x \sqcup\!\sqcup a^2ba^2b) = a^2baba^2ba^2b$, $\min(x \sqcup\!\sqcup a^2ba^3b) = a^2baba^2ba^3b$, $\min(x \sqcup\!\sqcup a^4b) = a^2baba^4b$, and the smallest of these words is the first one, which is $\min(x \sqcup\!\sqcup \min(Y))$.

The lemma implies that

$$\min(x \sqcup\!\sqcup y \sqcup\!\sqcup z) = \min(x \sqcup\!\sqcup \min(y \sqcup\!\sqcup z))$$

since $x \sqcup\!\sqcup y \sqcup\!\sqcup z = x \sqcup\!\sqcup Y$ with $Y = y \sqcup\!\sqcup z$. This formula, together with the formulas given earlier, allow to compute the minimal word in the shuffle of any polynomial. Observe that, up to irregular blocks, this minimal word is obtained as some permutation of the blocks composing the words in the support, each block appearing with the appropriate multiplicity.

Parikh vector: The *Parikh vector* $\mathrm{par}(x)$ of a word x is defined as usual. Let k be the size of the alphabet A, and for each letter $c \in A$ let $|x|_c$ be the number of c's in x. Then $\mathrm{par}(x) = (|x|_c)_{c \in A}$. As an example, $\mathrm{par}(a^2bca^3ba^2) = (7, 2, 1)$. The Parikh vector $\mathrm{par}(\mathscr{P})$ of a polynomial $\mathscr{P}$ is defined by linearity:

$$\mathrm{par}(\mathscr{P}) = \sum_{x \in P} (\mathscr{P}, x)\,\mathrm{par}(x)$$

For instance, $\mathrm{par}(2a^2bab + a^2ba^2b + a^2ba^3b + 2a^4b) = (23, 10)$.

If $S = \sqcup\!\sqcup(\mathscr{P})$, then all words in S have the same Parikh vector, and this Parikh vector is $\mathrm{par}(\mathscr{P})$.

Vectors as ordered as usual. For two vectors p, p', we set $p \leqslant p'$ iff $p_c \leqslant p'_c$ for all letters c. We will use this order for proof by induction.

5. Two reductions

We assume that the pivot is the letter a. Given a set of words S in aA^*, the *minimal initial block* of S is the smallest of the initial blocks of the words in S. It is denoted μ_S. The minimal initial block of a polynomial $\mathscr{P}$ is the minimal initial block of its support. It is denoted $\mu_{\mathscr{P}}$.

5.1. Shadow reduction

Let β be a block. Let x be a word in aA^* and let $x = \beta_1\beta_2\cdots\beta_n$ be its block decomposition. We define a word $\pi_\beta(x)$ by

$$\pi_\beta(x) = \beta_1\beta_2\cdots\beta_i,$$

where i is the smallest index such that $\beta_{i+1} > \beta$. Of course, $\pi_\beta(x) = \varepsilon$ if $\beta_1 > \beta$, and $\pi_\beta(x) = x$ if $\beta_1, \beta_2, \ldots, \beta_n \leqslant \beta$. This construction will be used in two slightly different ways.

Let $y = \gamma_1\cdots\gamma_m$ be a second word in aA^*. Set $x = \pi_\beta(x)x'$ and $y = \pi_\beta(y)y'$. Then $\min(x \sqcup\!\sqcup y)$ starts with $\min(\pi_\beta(x) \sqcup\!\sqcup \pi_\beta(y))$ and more precisely

$$\min(x \sqcup\!\sqcup y) = \min(\pi_\beta(x) \sqcup\!\sqcup \pi_\beta(y))\ \min(x' \sqcup\!\sqcup y').$$

Example. Let $x = a^2b^3 \cdot a^2b^2 \cdot a^3b$ and $y = a^2b^2 \cdot ab \cdot a^2b$. For $\beta = a^2b^2$, one gets $\pi_\beta(x) = a^2b^3 \cdot a^2b^2$ and $\pi_\beta(y) = a^2b^2 \cdot ab$. Next $\min(x \sqcup\!\sqcup y) = a^2b^3 \cdot a^2b^2 \cdot ab \cdot a^2b^2 \cdot a^2b \cdot a^3b$. Also, $\min(a^2b^3 \cdot a^2b^2 \sqcup\!\sqcup a^2b^2 \cdot ab) = a^2b^3 \cdot a^2b^2 \cdot ab \cdot a^2b^2$.

Let S be a finite subset of aA^*, and let $\mu = \mu_S$ be the minimal initial block of S. Let $s = \beta_1\beta_2\cdots\beta_n$ be the smallest word in S for the block order. Clearly $\beta_1 = \mu$. We define

$$\hat{\hat{S}} = (\pi_\mu(s))^{-1}S \cap aA^*.$$

This set of words is computed in two steps. First, only words starting with $\pi_\mu(s)$ are kept. Next, this prefix is removed from each of the remaining words, and the empty word, if any, is discarded. As a first example, consider the set

$$S = \{ab \cdot ab \cdot a, ab \cdot aab, aab \cdot ab, aabb \cdot a, aaabb\}.$$

Here $s = ababa$ and $\mu = \mu_S = ab$, $\pi_\mu(s) = abab$ and consequently $\hat{\hat{S}} = \{a\}$. Other examples will be given below.

Consider next a polynomial $\mathscr{P}$. Recall that $\mu = \mu_{\mathscr{P}}$ is the smallest of the initial blocks of the words in the support of $\mathscr{P}$. Define

$$\mathscr{P}' = \sum(\mathscr{P}, x)(\pi_\mu(x)^{-1}x) \quad \text{and} \quad \hat{\hat{\mathscr{P}}} = v(\mathscr{P}')$$

Here, for each word in the support of x, the prefix $\pi_\mu(x)$ is computed and removed. The resulting polynomial is then normalized. As a first example, consider $\mathscr{P} = a + 2ab$. Here $\mu_{\mathscr{P}} = ab$, and $\mathscr{P}' = 2\varepsilon + a$, whence $\hat{\hat{\mathscr{P}}} = a$.

The operations $S \mapsto \hat{\hat{S}}$ and $\mathscr{P} \mapsto \hat{\mathscr{P}}$ are called *shadow reductions*. The first is the shadow reduction of a set, the second the shadow reduction of a polynomial.

Example 1. Let

$$\mathscr{P} = 2a^2bab + a^2ba^2b + a^2ba^3b + 2a^4b.$$

The initial blocks are a^2b and a^4b, and the first one is smaller, thus $\mu = \mu_{\mathscr{P}} = a^2b$. We get $\pi_\mu(a^2bab) = a^2bab$, $\pi_\mu(a^2ba^2b) = a^2ba^2b$, $\pi_\mu(a^2ba^3b) = a^2b$, $\pi_\mu(a^4b) = \varepsilon$, whence $\mathscr{P}' = 2\varepsilon + \varepsilon + a^3b + 2a^4b$ and $\hat{\mathscr{P}} = a^3b + 2a^4b$. Consider the set

$$S = \sqcup\!\sqcup(\mathscr{P})$$

The smallest word s in S for the block order is $s = a^2baba^2bab(a^2b)^3a^3b(a^4b)^2$. Here $\pi_{\mu_S}(s) = a^2baba^2bab(a^2b)^3$. It is not difficult to check that $\mu_{\mathscr{P}} = \mu_S$. It appears moreover that $\hat{\hat{S}} = \sqcup\!\sqcup(\hat{\mathscr{P}})$. This is a general fact, stated in Lemma 5.1 below.

Example 2. This is a variation of the previous example. Let

$$\mathscr{Q} = 2a + 2a^2baba + a^2ba^2ba^2 + 2a^4ba.$$

The initial blocks are a, a^2b and a^4b, and the second one is the smallest. Thus $\mu_{\mathscr{Q}} = a^2b$. We get $\mathscr{Q}' = 2a + 2a + a^2 + 2a^4ba$. Normalization groups $2a + 2a + a^2$ into $6a$, whence $\hat{\mathscr{Q}} = 6a + 2a^4ba$. Consider $T = \sqcup\!\sqcup(\mathscr{Q})$. The minimal word t in T for the block order is $t = a^2b \cdot ab \cdot a^2b \cdot ab \cdot a^2b \cdot a^2b \cdot a^4b \cdot a^4b \cdot a^8$. We get $\pi_{\mu_T}(t) = a^2baba^2baba^2ba^2b$, and $\hat{\hat{T}} = \pi_{\mu_T}(t)^{-1}T = \sqcup\!\sqcup(2a + 2a + a^2 + 2a^4b) = \sqcup\!\sqcup(\hat{\mathscr{Q}})$.

Lemma 5.1. *If* $S = \sqcup\!\sqcup(\mathscr{P})$, *then* $\hat{\hat{S}} = \sqcup\!\sqcup(\hat{\mathscr{P}})$.

Proof. Let $s = \mu\beta_2 \cdots \beta_n$ be the smallest word in $S = \sqcup\!\sqcup(\mathscr{P})$. Then μ is the smallest of the initial blocks of words in the support P of $\mathscr{P}$. Thus $\mu = \mu_{\mathscr{P}} = \mu_S$.

Let Q be the set of words in P with initial block μ. Set $\mathscr{Q} = \sum_{q \in Q}(\mathscr{P}, q)q$ and define $\mathscr{R} = \mathscr{P} - \mathscr{Q}$. Each word q in Q is written as $q = \pi_\mu(q)q'$ with $\pi_\mu(q)$ nonempty. Then $\hat{\mathscr{P}} = \mathscr{R} + \mathscr{T}$, with

$$\mathscr{T} = \sum_{q \in Q, q' \neq \varepsilon} (\mathscr{P}, q)q'.$$

Next $\pi_\mu(s) = \min(\sqcup\!\sqcup(\sum_{q \in Q}(\mathscr{P}, q)\pi_\mu(q)))$. By the discussions above, this word is a product of blocks. It follows that $\hat{\hat{S}} = \sqcup\!\sqcup(\mathscr{R} + \mathscr{T})$, that is $\hat{\hat{S}} = \sqcup\!\sqcup(\hat{\mathscr{P}})$. □

Observe that in Lemma 5.1, the Parikh vector of $\hat{\mathscr{P}}$ is strictly less than the Parikh vector of $\mathscr{P}$.

5.2. Trim reduction

For any word $x = \beta_1\beta_2 \cdots \beta_n$, we define

$$\bar{x} = \beta_2 \cdots \beta_n.$$

Given a polynomial $\mathscr{P}$, we define $\bar{\mathscr{P}}$ by linearity and normalization as

$$\bar{\mathscr{P}} = v(\sum(\mathscr{P}, x)\bar{x}).$$

Observe that words composed of a single block disappear. As a first example, consider $\mathscr{P} = a + 2ab$. Here $\bar{\mathscr{P}} = v(\varepsilon + 2\varepsilon) = 0$.

Given a set $S \subset aA^*$, the *leading block* λ_S of S is one of the initial blocks of words of S. It is the initial block $a^n u$ of S with n is maximal and $|u|$ is maximal among the initial blocks with exponent n. Observe that this choice is not related to the block order we defined previously. Define

$$\bar{\bar{S}} = \lambda_S^{-1} S \cap aA^*.$$

The operations $S \mapsto \bar{\bar{S}}$ and $\mathscr{P} \mapsto \bar{\mathscr{P}}$ are called *trim reductions*. Note that these reductions strictly again decrease the corresponding Parikh vectors.

Example 1 (*continued*). Let

$$\mathscr{P} = 2a^2bab + a^2ba^2b + a^2ba^3b + 2a^4b.$$

We get

$$\bar{\mathscr{P}} = 2ab + a^2b + a^3b.$$

Next, we consider $S = \sqcup\!\sqcup(\mathscr{P})$ without computing it explicitly. The maximal number of initial a's in words in S is obtained by putting all a's of the initial blocks of $\mathscr{P}$ at the beginning of the words. This gives a^{16}. Next, we look for a longest word over $A \backslash a$ that can follow this power of a. This is obtained here by concatenating all b's occurring in all initial blocks, that is b^6. Thus, the block we look for is $\lambda_S = a^{16}b^6$. Clearly, $\bar{\bar{S}} = \lambda_S^{-1} S$ is the shuffle of what remains of $\mathscr{P}$, that is $\bar{\bar{S}} = \sqcup\!\sqcup(\bar{\mathscr{P}})$.

Example 2 (*continued*). Let

$$\mathscr{Q} = 2a + 2a^2baba + a^2ba^2ba^2 + 2a^4ba.$$

We get $\bar{\mathscr{Q}} = 2aba + a^2ba^2 + 2a$. The first term $2a$ in $\mathscr{Q}$ disappears, and the last term in $\bar{\mathscr{Q}}$ comes from $2\overline{a^4ba} = 2a$. Consider $T = \sqcup\!\sqcup(\mathscr{Q})$. The maximal number of initial a's in words in T is a^{16}, and the maximal number of b's that can follow the power if a is b^5. Thus, $\lambda_T = a^{16}b^5$ and $\lambda_T^{-1} T = \sqcup\!\sqcup(2aba + a^2ba^2 + 2a)$.

Lemma 5.2. *If* $S = \sqcup\!\sqcup(\mathscr{P})$, *then* $\bar{\bar{S}} = \sqcup\!\sqcup(\bar{\mathscr{P}})$.

Proof. Clear from the previous examples. □

Observe that in Lemma 5.2, the Parikh vector of $\bar{\mathscr{P}}$ is strictly less than the Parikh vector of $\mathscr{P}$.

Example. The aim of this example is to illustrate the two reductions introduced earlier and to give a feeling why these will lead to an effective procedure.

Consider three words $p = a^2b^2ab^3a^3b$, $q = a^2b^2ab^3a^4b^2$ and $r = a^3ba^2b^2$ and let $S = p \sqcup\!\sqcup q \sqcup\!\sqcup r$ be the shuffle set generated by these words.

The maximum number of letters a at the beginning of a word in S is clearly $2 + 2 + 3 = 7$. Among the words in S starting with a^7, those which have the maximum number of b following a^7 will clearly begin with $b^{2+2+1} = b^5$. In order to get these words in S, we have to use the three initial blocks of the words p, q, r. Hence, the set $(a^7b^5)^{-1}S$ is a simpler shuffle set. More precisely, setting $p' = ab^3a^3b$, $q' = ab^3a^4b^2$ and $r' = a^2b^2$, we have $(a^7b^5)^{-1}S = p' \sqcup\!\sqcup q' \sqcup\!\sqcup r'$. The set $(a^7b^5)^{-1}S$ is precisely what has been defined as $\bar{\bar{S}}$, and the polynomial $p' + q' + r'$ is the trimmed reduction of $p + q + r$.

Now, we look for words in S beginning with the minimum number of a. These are the words beginning with a^2. Among these words, those which have the maximum number of b following a^2 will clearly begin with a^2b^2. These are the words in S beginning with the smallest initial block. However, the set $(a^2b^2)^{-1}S$ is not a shuffle set. It is the union of two shuffle sets, obtained by removing the block a^2b^2 either from word p or from word q. In other words, one of the shuffle sets is $p' \sqcup\!\sqcup q \sqcup\!\sqcup r$ and the other is $p \sqcup\!\sqcup q' \sqcup\!\sqcup r$.

Let us repeat the process of looking for "minimal" blocks of words in S until a block appears that is greater than a^2b^2. We have to repeat the process four times, giving raise to the word $h = a^2b^2ab^3a^2b^2ab^3$ composed of four blocks. Then the quotient $h^{-1}S$ is a shuffle set: indeed, the word h has used all blocks less or equal to a^2b^2 which is the minimal initial block in the root of S; so the quotient is exactly the set obtained by shuffling the remaining suffixes of the original words, that is $h^{-1}S = a^3b \sqcup\!\sqcup a^4b^2 \sqcup\!\sqcup r$. With the previous notation, $h = \pi_\mu(s)$, where $\mu = a^2b^2$ and s is the smallest word in S. The quotient is precisely what has been defined as $\hat{\hat{S}}$.

Both of these operations lead to a simpler shuffle set. The rest of the proof follows from the fact that, given the solutions of each subproblem, there is at most one (computable) solution of the original problem.

Degree: The *block degree* $\deg(x)$ of a word $x \in aA^*$ was defined to be the number of blocks in its decomposition. For a polynomial $\mathscr{P}$, the block degree $\deg(\mathscr{P})$ is the maximum of the degrees of the words in its support. As an example, the polynomial of Example 1 has degree 2, and the polynomial of Example 2 has degree 3. It will be convenient to denote $\mathscr{P}_k$ the *homogeneous component* of degree k of $\mathscr{P}$, that is $\mathscr{P}_k = \sum_{\deg(x)=k} (\mathscr{P}, x)x$. A normalized polynomial $\mathscr{P}$ of degree d then writes as $\mathscr{P} = \mathscr{P}_1 + \mathscr{P}_2 + \cdots + \mathscr{P}_d$.

A polynomial $\mathscr{P}$ is *simple* if no word in its support contains an irregular block, it is *symmetric* if every word in its support contains (and thus ends with) an irregular block. The polynomial in Example 1 is simple, the polynomial in Example 2 is symmetric, the polynomial $a + 2ab$ is neither simple nor symmetric. The shadow polynomial and the trim polynomial of a simple (symmetric) polynomial are again simple (symmetric).

For any polynomial $\mathscr{P}$ of degree $d \geqslant 1$, the polynomial $\hat{\mathscr{P}}$ has degree d or less, and the trim polynomial $\bar{\mathscr{P}}$ has degree $d-1$.

Proposition 5.3. *Given two simple normalized polynomials $\mathscr{H}$ and $\mathscr{T}$, a vector p, and a block μ, there is at most one simple normalized effectively computable (simple) polynomial $\mathscr{P}$ such that $\hat{\mathscr{P}} = \mathscr{H}$, $\bar{\mathscr{P}} = \mathscr{T}$, $\mu_{\mathscr{P}} = \mu$ and $\mathrm{par}(\mathscr{P}) = p$.*

Proof. By induction on the degree of $\mathscr{T}$. If $\deg(\mathscr{T}) = 0$, then $\mathscr{T} = 0$ because $\mathscr{T}$ is proper. If $\mathscr{P}$ exists and $\mathscr{P} \neq 0$, then $\deg(\mathscr{P}) = 1$. Set $\mathscr{P} = h\mu + \mathscr{Q}$ with $h = (\mathscr{P}, \mu)$. Since $\hat{\mathscr{P}} = \mathscr{H}$, the support of $\mathscr{P}$ does not contain an irregular block. Since $\mu = \mu_{\mathscr{P}}$, all blocks in $\mathscr{Q}$ are greater than μ, and consequently $\hat{\mathscr{P}} = \mathscr{Q}$, hence $\mathscr{Q} = \mathscr{H}$. Thus, if $\mathscr{P}$ exists, it is of the form

$$\mathscr{P} = h * \mu + \mathscr{H}.$$

Finally, since $p = h\,\mathrm{par}(\mu) + \mathrm{par}\,\mathscr{H}$, the number h can be computed. If it is a positive integer, then $\mathscr{P}$ is uniquely determined, otherwise the polynomial $\mathscr{P}$ does not exist.

Assume next that $\deg(\mathscr{T}) = d - 1 \geqslant 1$. If a polynomial $\mathscr{P}$ exists with $\bar{\mathscr{P}} = \mathscr{T}$ and $\hat{\mathscr{P}} = \mathscr{H}$, then $\deg(\mathscr{P}) = d = 1 + \deg(\mathscr{T})$ and $\deg(\mathscr{P}) \geqslant \deg(\mathscr{H})$. Write $\mathscr{P} = \mathscr{P}' + \mathscr{P}_d$, with $\mathscr{P}' = \mathscr{P}_1 + \cdots + \mathscr{P}_{d-1}$. Set $\mathscr{P}' = \mu\mathscr{C} + \mathscr{N}$, $\mathscr{P}_d = \mu\mathscr{D} + \mathscr{M}$, such that the initial blocks of the words of the supports of $\mathscr{N}$ and $\mathscr{M}$ are greater than μ. Also, $\mathscr{C}$ has degree at most $d-2$, $\mathscr{D}$ has degree of degree $d-1$ and $\mathscr{M}$ has degree d. From $\mathscr{P} = \mu\mathscr{C} + \mathscr{N} + \mu\mathscr{D} + \mathscr{M}$, it follows that $\mathscr{T} = \bar{\mathscr{P}} = \bar{\mathscr{C}} + \bar{\mathscr{N}} + \mathscr{D} + \bar{\mathscr{M}}$ and $\mathscr{H} = \hat{\mathscr{P}} = \hat{\mathscr{C}} + \mathscr{N} + \hat{\mathscr{D}} + \mathscr{M}$. Thus $\mathscr{H}_d = \mathscr{M}$, and $\mathscr{T}_{d-1} = \mathscr{D} + \bar{\mathscr{M}}$. This determines $\mathscr{D} = \mathscr{T}_{d-1} - \overline{\mathscr{H}_d}$. If the difference has negative coefficients, the polynomial $\mathscr{P}$ does not exist. Otherwise, since $\mathscr{M}$ and $\mathscr{D}$ are determined, the homogeneous component $\mathscr{P}_d$ is uniquely determined, and we repeat the process. □

Example 1 (*continued*). Let $\mathscr{T} = 2ab + a^2b + a^3b$, $\mathscr{H} = a^3b + 2a^4b$, $p = (23, 10)$ and $\mu = a^2b$. The polynomial $\mathscr{P}$ has degree 2. Set $\mathscr{P}_2 = a^2b\mathscr{D} + \mathscr{M}$. Since $\mathscr{H}$ has degree 1, $\mathscr{M} = 0$. Next, $\mathscr{D} = \overline{\mathscr{P}_2} = \mathscr{T}$. Thus $\mathscr{P}_2 = 2a^2bab + a^2ba^2b + a^2ba^3b$. Set $\mathscr{P}_1 = h * a^2b + \mathscr{Q}$. From $\hat{\mathscr{P}} = \mathscr{Q} + a^3b = \mathscr{H}$, one gets $\mathscr{Q} = 2a^4b$. Content-Length: 12 821.

From $\hat{\mathscr{P}} = \mathscr{Q} + a^3b = \mathscr{H}$, one gets $\mathscr{Q} = 2a^4b$. Thus, $\mathscr{P} = h * a^2b + 2a^2bab + a^2ba^2b + a^2ba^3b + 2a^4b$, whence $\mathrm{par}(\mathscr{P}) = h * (2,1) + (23,10)$. Thus $h = 0$ and $\mathscr{P} = 2a^2bab + a^2ba^2b + a^2ba^3b + 2a^4b$.

We now consider the symmetric case, which is more involved. We will consider both polynomial and their reversals. More precisely, for any polynomial $\mathscr{Q}$, we denote $\tilde{\mathscr{Q}}$ the polynomial $\tilde{\mathscr{Q}} = \sum (\mathscr{Q}, x)\tilde{x}$. Observe that neither the shadow nor the trim reduction commutes with reversal. As an example, consider the polynomial

$$\mathscr{Q} = 2a + 2a^2baba + a^2ba^2ba^2 + 2a^4ba$$

of Example 2 above. We get

$$\tilde{\mathscr{Q}} = 2a + 2ababa^2 + a^2ba^2ba^2 + 2aba^4$$

from which $\mu_{\mathscr{Q}} = a^2b$, $\mu_{\tilde{\mathscr{Q}}} = ab$. Also $\bar{\mathscr{Q}} = 2aba + a^2ba^2 + a^4$, $\bar{\tilde{\mathscr{Q}}} = 2aba^2 + a^2ba^2 + a^{10}$ and $\hat{\mathscr{Q}} = a^6 + 2a^4ba$, $\hat{\tilde{\mathscr{Q}}} = a^{14} + a^2ba^2ba^2$.

Proposition 5.4. *Given symmetric normalized polynomials $\mathscr{H}, \mathscr{H}_\sim$ and $\mathscr{T}, \mathscr{T}_\sim$, a vector p, and two blocks μ, $\mu_\sim$, there is at most one symmetric normalized effectively computable polynomial $\mathscr{P}$ such that $\hat{\mathscr{P}} = \mathscr{H}$, $\hat{\tilde{\mathscr{P}}} = \mathscr{H}_\sim$, $\bar{\mathscr{P}} = \mathscr{T}$, $\bar{\tilde{\mathscr{P}}} = \mathscr{T}_\sim$, $\mu_{\mathscr{P}} = \mu$, $\mu_{\tilde{\mathscr{P}}} = \mu_\sim$ and* $\mathrm{par}(\mathscr{P}) = p$.

Proof. The proof is by induction as for Proposition 5.3. However, the initial step of the induction is different.

If $\deg(\mathscr{T}) = 0$, then $\deg(\mathscr{P}) = 1$ and $\mathscr{P} = h * a$, where h is given by the vector p.

Assume $\deg(\mathscr{T}) = 1$. Then $\deg(\mathscr{P}) = 2$, and μ has to be a regular block. Set

$$\mathscr{P} = h * a + \mu\mathscr{C} + \mathscr{N},$$

where the polynomial $\mathscr{C}$ has degree 1, the polynomial $\mathscr{N}$ is homogeneous of degree 2 and all initial blocks of words in $\mathscr{N}$ are greater than μ. One gets $\hat{\mathscr{P}} = H * a + \mathscr{N}$, where $H * a$ is obtained from normalization of the shadow reduction of $h * a + \mu\mathscr{C}$. This determines $\mathscr{N}$ to be $\mathscr{H}_2$. Similarly, write

$$\tilde{\mathscr{P}} = h * a + \mu_\sim\mathscr{D} + \mathscr{M}.$$

By the same argument, $\mathscr{M}$ is the homogeneous component of degree 2 of $\mathscr{H}_\sim$. Taking the reversal of the first relation, we get

$$\tilde{\mathscr{C}}\tilde{\mu} + \tilde{\mathscr{N}} = \mu_\sim\mathscr{D} + \mathscr{M}. \tag{3}$$

Set $\mu = a^N u$, and $\mu_\sim = a^M v$ for nonempty words u, v over $A - \{a\}$. By the definition of μ and $\mu_\sim$, every word of degree 2 in the support of $\mathscr{P}$ is of the form $a^n w a^m$ with $n \geqslant N$, $m \geqslant M$ and w a word over $A - \{a\}$. Set also $\mathscr{C} = \sum_{m \geqslant M} c_m a^m$, $\mathscr{D} = \sum_{n \geqslant N} d_n a^n$, where the coefficients c_n and d_n are nonnegative integers. Then

$$\tilde{\mathscr{C}}\tilde{\mu} = c_M a^M \tilde{u} a^N + \sum_{m > M} c_m a^m \tilde{u} a^N, \quad \mu_\sim\mathscr{D} = d_N a^M v a^N + \sum_{n > N} d_n a^M v a^n,$$

whence

$$\tilde{\mathscr{N}} + c_M a^M \tilde{u} a^N + \sum_{m > M} c_m a^m \tilde{u} a^N = d_N a^M v a^N + \sum_{n > N} d_n a^M v a^n + \mathscr{M}.$$

This entirely determines the coefficients c_m for $m > M$ and d_n for $n > N$. Moreover, if the polynomial $\mathscr{P}$ exists, then $c_M = d_N$. If $\tilde{u} \neq v$, then $c_M = d_N = 0$.

Thus, assume $\tilde{u} = v$. It remains to compute the coefficients h and c_M. From $\bar{\mathscr{P}} = \mathscr{T}$, one gets for the coefficient t:

$$t = h + \sum_{m \geqslant M} c_m * m \tag{4}$$

Finally, the a-component p_a of the Parikh vector $p = \mathrm{par}(\mathscr{P})$ has the expression

$$p_a = h + \sum_{m \geqslant M} c_m * (m + N) + \mathrm{par}(\mathscr{N})_a. \tag{5}$$

These two equations allow to compute h and c_M.

If $\deg(\mathscr{T}) > 1$, the homogeneous parts of degree at least 3 are obtained as in Proposition 5.3, and the homogeneous parts of degree at most 2 are obtained as above. □

Example 2 (*continued*). Consider the following data: $\mathscr{H} = 6a + 2a^4ba$, $\mathscr{H}_\sim = 14a + a^2ba^2ba^2$, $\mathscr{T} = 4a + 2aba + a^2ba^2$, $\mathscr{T}_\sim = 10a + 2aba^2 + a^2ba^2$, $\mu = a^2b$, $\mu_\sim = ab$, $p = (26, 8)$. Since $\mathscr{T}$ has degree 2, the polynomial $\mathscr{P}$ we look for has degree 3. We set

$$\mathscr{P} = h * a + \mu\mathscr{C} + \mathscr{N} + \mu\mathscr{E} + \mathscr{R}$$

with $\deg(\mathscr{R}) = 3$, $\deg(\mathscr{E}) = \deg(\mathscr{N}) = 2$, $\deg(\mathscr{C}) = 1$. From $\hat{\mathscr{P}} = \cdots + \mathscr{R} = \mathscr{H}$, we get $\mathscr{R} = 0$, and from $\bar{\mathscr{P}} = \cdots + \mathscr{E} = \mathscr{T}$ we get $\mathscr{E} = 2aba + a^2ba^2$, thus the component $\mathscr{P}_3 = \mu\mathscr{E} + \mathscr{R} = \mu\mathscr{E}$ is $\mathscr{P}_3 = 2a^2baba + a^2ba^2ba^2$. Consider again $\hat{\mathscr{P}} = h * a + v(\mathscr{C}) + \mathscr{N} + 4a = \mathscr{H}$. The coefficient $4a$ comes from $4a = \widehat{\mu\mathscr{E}}$. Thus $\mathscr{N} = 2a^4ba$. Set $\tilde{\mathscr{P}} = h * a + \mu_\sim \mathscr{D} + \mathscr{M} + \tilde{\mathscr{P}}_3$. Then $\hat{\tilde{\mathscr{P}}} = h * a + v(\mathscr{D}) + \mathscr{M} + 4a + a^2ba^2ba^2 = \mathscr{H}_\sim$ showing that $\mathscr{M} = 0$. It follows that $\tilde{\mathscr{N}} + \tilde{\mathscr{C}}\tilde{\mu} = \mu_\sim \mathscr{D}$. This gives the equation

$$2aba^4 + c_1aba^2 + \sum_{m>1} c_m a^m ba^2 = d_2aba^2 + \sum_{n>2} d_n aba^n$$

which implies $\mathscr{C} = c_1a$, $\mathscr{D} = d_2a^2 + 2a^4$ and $\mathscr{P} = h * a + c_1a^2ba + 2a^4ba + 2a^2baba + a^2ba^2ba^2$. The coefficient of a in $\bar{\mathscr{P}}$ is $h + c_1 + 2$, and since $(\mathscr{P}, a) = (\mathscr{T}, a) = 4$, we get $h + c_1 = 2$. Next, the a-coordinate of the Parikh vector $\mathrm{par}(\mathscr{P})$ is $h + 3c_1 + 10 + 8 + 6$ and must be 26. This gives the equation $2 = h + 3c_1$. It follows that $h = 2$ and $c_1 = 0$. Thus, our solution is precisely the polynomial $\mathscr{Q}$ of the example.

5.3. Proof of the main theorem

We are now able to give the proof of Theorem 1.1.

Let $\mathscr{P}$ and $\mathscr{Q}$ be normalized polynomials such that $\sqcup\!\sqcup(\mathscr{P}) = \sqcup\!\sqcup(\mathscr{Q})$, and set $S = \sqcup\!\sqcup(\mathscr{P}) = \sqcup\!\sqcup(\mathscr{Q})$. In view of Corollary 3.6, we may assume that S is simple or symmetric. It follows that $\mathscr{P}$ and $\mathscr{Q}$ are simple or are symmetric. Also, it is clear that $\mathrm{par}(\mathscr{P}) = \mathrm{par}(\mathscr{Q})$.

Assume first that $\mathscr{P}$ and $\mathscr{Q}$ are a-simple. The proof is by induction on this Parikh vector. By Lemma 5.1, one gets $\sqcup\!\sqcup(\hat{\mathscr{P}}) = \sqcup\!\sqcup(\hat{\mathscr{Q}})$, and by Lemma 5.2, one gets $\sqcup\!\sqcup(\bar{\mathscr{P}}) = \sqcup\!\sqcup(\bar{\mathscr{Q}})$. As observed, the Parikh vectors of $\bar{\mathscr{P}}$ and of $\hat{\mathscr{P}}$ are less than the Parikh vector of $\mathscr{P}$. By induction, $\hat{\mathscr{P}} = \hat{\mathscr{Q}}$ and $\bar{\mathscr{P}} = \bar{\mathscr{Q}}$. By Proposition 5.3, $\mathscr{P} = \mathscr{Q}$.

The same argument holds if $\mathscr{P}$ and $\mathscr{Q}$ are symmetric, using Proposition 5.4. This completes the proof. □

6. A large example

Over the alphabet $A=\{a,b\}$, consider the following set S of 58 words given in alphabetical order (for the moment, forget about the marks $*$ and $+$):

$$
\begin{array}{llllll}
+aaaaababbb & aaaaabbabb & aaaaabbbab & & & \\
+aaaabaabbb & aaaabababb & aaaababbab & *aaaababbba & aaaabbaabb & \\
aaaabbabab & *aaaabbabba & aaaabbbaab & *aaaabbbaba & & \\
+aaabaaabbb & aaabaababb & aaabaabbab & *aaabaabbba & aaababaabb & \\
aaababababb & *aaabababba & aaababbaab & *aaababbaba & aaabbaaabb & \\
aaabbaabab & *aaabbaabba & aaabbabaab & *aaabbababa & & \\
+aabaaaabbb & aabaaababb & aabaaabbab & *aabaaabbba & aabaabaabb & \\
aabaababab & *aabaababba & aabaabbaab & *aabaabbaba & aababaaabb & \\
aababaabab & *aababaabba & aababababb & *aababababa & aabbaaabab & \\
aabbaabaab & *aabbaababa & & & & \\
abaaaabbb & abaaaabbab & abaaabaabb & abaaababab & *abaaababba & \\
abaaabbaab & *abaaabbaba & abaabaaabb & abaabaabab & *abaabaabba & \\
abaababaab & *abaabababa & ababaaabab & ababaabaab & *ababaababa. &
\end{array}
$$

We start by some preliminary observations:

- All words of S contain 6 letters a and 4 letters b.
- All words of S start with the letter a.
- Some words end with a letter a, others with a b.

So, we do not need to split S according to the first letter; letter a will be the pivot. On the other hand, we have to split S into an a-simple set, say $T=S[b]$ and an a-symmetric set, say $U=S[a]$. In view of the definition in Section 3, we look at the maximal number of a's at the end of words in S. This number is 1, hence $T=Sa^{-1}$. This set is exactly the set of the 19 words marked with a $*$ in S.

Similarly, we look at the maximal number of b's at the end of words in S. This number is 3, hence $U=Sb^{-3}$. This set is exactly the set of four words marked with a $+$ in S. In view of Theorem 3.5, we consider separately these two sets. If they are shuffle sets, then we get shuffle roots for them. These will be later composed to get the unique candidate for a shuffle root for S.

6.1. Computing the shuffle root of U

The set U is

$$aaaaaba \quad aaaabaa \quad aaabaaa \quad aabaaaa.$$

We look for a polynomial $\mathscr{U}$ such that $U=\sqcup\!\sqcup(\mathscr{U})$. For this, we apply shadow and trim reduction, as described in Section 5.1. The minimal initial block of U is $\mu_U=aab$. Since there is only one word in U beginning by μ, the set $\hat{\hat{U}}$ is composed of the single word $aaaa$. Similarly, the leading block λ_U is a^5b; it gives raise to $\bar{U}=\{a\}$. Hence, the polynomial $\mathscr{U}$ associated to U will look like $\mathscr{U}=aaba^i+ja$ with $i+j=4$

(= the number of a in words of U minus the number of a in μ). In order to compute i and j, we look at the reversal of U. The minimal initial block is $\mu_{\tilde{U}} = ba$. This yields $\tilde{\tilde{U}} = \{aaaaa\}$. The leading block $\lambda_{\tilde{U}}$ of $\tilde{U}$ is $aaaab$, giving raise to $\hat{\tilde{U}} = \{aa\}$. So, we get the new description of $\mathscr{U}$ as $a^k ba + sa$ with $k + s = 5$. So, we get immediately

$$\mathscr{U} = aaba + 3a.$$

6.2. Computing the shuffle root of T

Here is the set T again:

$$\begin{aligned}
&aaaababbb\ aaaabbabb\ aaaabbbab\\
&aaabaabbb\ aaababbab\ aaababbab\ aaabbaabb\quad aaabbabab\\
&aabaaabbb\ aabaababb\ aabaabbab\ aababaabb\quad aababbabab\ aabbaabab\\
&abaaababb\ abaaabbab\ abaabaabb\ abaababab\ ababaabab\,.
\end{aligned}$$

Again, we look for a polynomial $\mathscr{T}$ such that $T = \sqcup\!\sqcup \mathscr{T}$. In order to construct $\bar{\tilde{T}}$, we look for the leading block in T; it is $\lambda_T = a^4b^3$. The left quotient by this block yields $\bar{\tilde{T}} = \{ab\}$. Hence, $\bar{\mathscr{T}} = ab$.

In order to construct $\hat{\tilde{T}}$, we look for the minimal initial block in T; it is $\mu = ab$. The prefix used then to quotient T is $abab$ yielding $\hat{\tilde{T}} = \{aabab\}$. Hence, $\hat{\mathscr{T}} = aabab$. Knowing μ, $\bar{\mathscr{T}}$ and $\hat{\mathscr{T}}$, we construct $\mathscr{T}$. It has the form $\mathscr{T} = aabab + h * ab$ for some integer h. Using the Parikh vector $\mathrm{par}(T) = (5,4)$ of T, we get $h = 2$; hence,

$$\mathscr{T} = aabab + 2ab.$$

6.3. Constructing the polynomial S

Knowing $\mathscr{U} = aaba + 3a$ and $\mathscr{T} = aabab + 2ab$, we may try now to construct $\mathscr{S}$. The longest word is $aabab$; it is in $\mathscr{S}$ and is responsible for $aaba$ in $\mathscr{U}$. So, we are left with $3a$ and $2ab$. This implies that $2ab$ is in $\mathscr{S}$ and leaves a alone in $\mathscr{U}$. Hence,

$$\mathscr{S} = aabab + 2ab + a.$$

The final step is now to compute the shuffle of $\mathscr{S}$ and to check that it is precisely S. Note that this last step *must* be performed. Indeed, if you suppress from S any unmarked word, the resulting polynomial $\mathscr{S}$ remains unchanged. Similarly, if you add to S a word such as $aaabbbaab$, then

- it has the correct Parikh vector,
- it does not end by a, so it will not appear in T,
- it does not end by bbb, so that it will not appear in U.

Hence, the computations of $\mathscr{U}$, $\mathscr{T}$ and $\mathscr{S}$ will be the same as above.

References

[1] C. Allauzen, Calcul efficace du shuffle de k mots, Tech. Report IGM2000-02, Institut Gaspard Monge, Université Marne-la-Vallée, 2000.
[2] M. Lothaire, Combinatorics on Words, Encyclopaedia of Mathematics and its Applications, vol. 17, Addison-Wesley, Reading, MA, 1983 (Reprinted in the Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997).
[3] M. Nivat, G.D. Ramkumar, C. Pandu Rangan, A. Saoudi, R. Sundaram, Efficient parallel shuffle recognition, Parallel Process. Lett. 4 (1994) 455–463.
[4] B. Pradeep, C. Murthy, A constant time string shuffle algorithm on reconfigurable meshes, Internat. J. Comput. Math. 68 (1998) 251–259.
[5] D.E. Radford, A natural ring basis for the shuffle algebra and an application to group schemes, J. Algebra 58 (1979) 432–454.
[6] C. Reutenauer, Free Lie Algebras, London Mathematical Society Monographs, Oxford Science Publications, Oxford, 1993.
[7] J.-C. Spehner, Le calcul rapide des mélanges de deux mots. (Fast computing of the shuffle of two words), Theoret. Comput. Sci. 47 (1986) 181–203.