On the expressive power of the shuffle product (Extended Abstract)

Antonio Restivo*

1 Introduction

A very general problem in the theory of formal languages is, given a "basis" of languages and a set of operations, to characterize the family of languages expressible from the "basis" by using the operations. In practice, a basis of languages will consists of a set of very simple languages, such as the languages of the form $\{a\}$, where a is a letter of the alphabet. In the theory of *regular* languages, the operations taken into account are usually the Boolean operations, the concatenation and the (Kleene) star operation.

In this setting, two families of languages play a fundamental role: the family *REG* of regular languages, and the family *SF* of star-free languages. *REG* is defined as the smallest family of languages containing the languages of the form $\{a\}$, where a is a letter, and $\{\epsilon\}$, where ϵ is the empty word, and closed under union, concatenation and star. It is well known that the family *REG* is closed also under all Boolean operations. The family *SF* of star-free languages is the smallest family of languages containing the languages of the form $\{a\}$ and $\{\epsilon\}$, and closed under Boolean operations and concatenation.

Another operation that plays an important role in the theory of formal languages is the *shuffle* operation. Recall that the *shuffle product* (or simply *shuffle*) of two languages L_1, L_2 is the language

$$L_1 \sqcup\!\sqcup L_2 = \{u_1v_1...u_nv_n | n \geq 0, u_1...u_n \in L_1, v_1...v_n \in L_2\}.$$

It is well known (cf. [5]) that the family *REG* of regular languages is closed under shuffle. The study of subfamilies of regular languages closed under shuffle is a difficult problem, partly motivated by its applications to the modeling of process algebras [1] and to program verification.

In particular, we here consider the smallest family of languages containing the languages of the form $\{a\}$ and $\{\epsilon\}$, and closed under Boolean operations, concatenation and shuffle. Let us call *intermixed* the languages in this family, which is denoted by *INT*. It is perhaps

*Dipartimento di Matematica e Informatica, Università di Palermo, Palermo Italy

surprising that the following important problem in the theory of regular languages is still open, and to a large extent unexplored.

Problem 1 *Give a (decidable) characterization of the family INT.*

In this talk we discuss this problem: we present some partial results and we introduce new special problems as possible steps in the characterization of the family *INT*. Such partial results and special problems show the deep connections of Problem 1 with other relevant aspects of formal languages theory and combinatorics on words. The results here presented are essentially based on the papers [2] and [3].

2 Star-Free and Intermixed Languages

In [2] it is proved the following theorem showing that the family *INT* of intermixed languages is strictly included in the family *REG* of regular languages and strictly contains the family *SF* of star-free languages.

Theorem 2.1 $SF \subsetneq INT \subsetneq REG$

Moreover, in [2] it is shown that the family *INT* is closed under quotients, but it is not closed under inverse morphism. Therefore, the family *INT* is not a variety of languages (cf. [11]), and so it cannot be characterized in terms of syntactic monoids.

Let us recall (cf. [9])that a language $L \subseteq \Sigma^*$ is said to be *aperiodic*, or *non-counting*, if there exists an integer $n > 0$ such that for all $x, y, z \in \Sigma^*$ one has

$$xy^n z \in L \Leftrightarrow xy^{n+1} z \in L.$$

A fundamental theorem of Schutzenberger states that *a regular language is star-free if and only if it is aperiodic.*

The strict inclusion between the families *SF* and *INT* implies that the shuffle of two star-free languages in general is not star-free. This means, roughly speaking, that *the shuffle creates periodicities.*

In order to enlighten on the difficult Problem 1, in this talk we consider the following

Problem 2 *Determine conditions under which the shuffle of two star-free languages is star-free too.*

A first condition is obtained in [3] by introducing a weaker version of the shuffle product, called *bounded shuffle.*

Let k be a positive integer. The *k-shuffle* of two languages $L_1, L_2 \subseteq \Sigma^*$ is defined as follows:

$$L_1 \sqcup\!\sqcup_k L_2 = \{u_1 v_1 ... u_m v_m | m \leq k, u_1 ... u_m \in L_1, v_1 ... v_m \in L_2\}.$$

Any k-shuffle is called *bounded shuffle*. It is not difficult to show that the family REG of regular languages is closed under bounded shuffle. In [3] it is proved the following theorem.

Theorem 2.2 *SF is closed under bounded shuffle, i.e. if $L_1, L_2 \in SF$ then $L_1 \sqcup\!\sqcup_k L_2 \in SF$, for any $k \geq 1$.*

One can derive the following corollary.

Corollary 2.3 *The shuffle of a star-free language and a finite language is star-free.*

3 Partial Commutations

The family SF is closed under concatenation and it is not closed under shuffle. What is the difference between concatenation and shuffle?

In this section we introduce an operation between languages, that generalizes at the same time concatenation and shuffle, and we investigate the closure of SF with respect to this operation. The new operation is defined by introducing a partial commutation between the letters of the alphabet, and its appropriate setting is the theory of *traces* (cf [4]).

Let Γ be a finite alphabet and let $\theta \subseteq \Gamma \times \Gamma$ be a symmetric and irreflexive relation called the *(partial) commutation relation*. We consider the congruence $\sim_\theta$ of Γ^* generated by the set of pairs (ab, ba) with $(a, b) \in \theta$. If $L \subseteq \Gamma^*$ is a language, $[L]_\theta$ denoted the closure of L by $\sim_\theta$, and L is *closed by* $\sim_\theta$ if $L = [L]_\theta$. The closed subsets of Γ^* are called *trace languages*.

Let now L_1 and L_2 be two languages over the alphabet Σ

Let us consider two disjoint copies Σ_1 and Σ_2 of the alphabet Σ, i.e. such that $\Sigma_1 \cap \Sigma_2 = \emptyset$, and the isomorphism σ_1 from Σ_1^* to Σ^* and σ_2 from Σ_2^* to Σ^*.

Let L_1' (L_2' resp.) be the subset of Σ_1^* (Σ_2^* resp.) corresponding to L_1 (L_2 resp.) under the isomorphism σ_1 (σ_2 resp.). Let us consider the morphism $\sigma : (\Sigma_1 \cup \Sigma_2)^* \to \Sigma^*$ defined as follows:

$$\sigma(a) = \begin{cases} \sigma_1(a), & \text{if } a \in \Sigma_1^*; \\ \sigma_2(a), & \text{if } a \in \Sigma_2^*. \end{cases}$$

Let θ be of the form $\theta \subseteq \Sigma_1 \times \Sigma_2$. The $\theta - product$ (denoted by $\sqcup\!\sqcup_\theta$) of the languages $L_1, L_2 \subseteq \Sigma^*$ is defined as follows:

$$L_1 \sqcup\!\sqcup_\theta L_2 = \sigma([L_1' L_2']_\theta).$$

Remark that the product (concatenation) and the shuffle correspond to two special (extremal) cases of the $\theta-product$. Indeed, if $\theta = \emptyset$ then $L_1 \sqcup\!\sqcup_\theta L_2 = L_1 L_2$, and, if $\theta = \Sigma_1 \times \Sigma_2$, then $L_1 \sqcup\!\sqcup_\theta L_2 = L_1 \sqcup\!\sqcup L_2$.

The partial commutation $\theta \subseteq \Sigma_1 \times \Sigma_2$ induces a partial commutation θ' on Σ defined as follows: if $(a, b) \in \theta$ the $(\sigma_1(a), \sigma_2(b)) \in \theta'$.

In [6] it is proved the following theorem.

Theorem 3.1 *Let L_1 and L_2 be two languages closed under θ', i.e., $[L_1]_{\theta'} = L_1$ and $[L_2]_{\theta'} = L_2$. If L_1 and $L_2 \in SF$, then $L_1 \sqcup\!\sqcup_\theta L_2 \in SF$.*

The theorem states, roughly speaking, that, if *internal* commutation (i.e., the commutations allowed inside each of the languages L_1 and L_2) is the "same" as the *external* commutation (i.e., the commutations between the letters in L_1 and the letters in L_2), then the $\theta - product$ preserves the star-freeness.

Special cases of the previous theorem are the well known result that the concatenation of two star-free languages is star-free, and the result of J.F. Perrot (cf. [10]) that *the shuffle of two* commutative *star-free languages is star-free.*

4 Unambiguous Star-Free languages

In this section we investigate some conditions for Problem 2, related to the unambiguity of the product of languages.

A language $L \subseteq \Sigma^*$ is a *marked product* of the languages $L_0, L_1, ..., L_n$ if

$$L = L_0 a_1 L_1 a_2 L_2 ... a_n L_n,$$

for some letters $a_1, a_2, ..., a_n$ of Σ.

It is known (cf [13]) that the family SF of star-free languages is the smallest Boolean algebra of languages of Σ^* which is closed under marked product.

A marked product $L = L_0 a_1 L_1 a_2 L_2 ... a_n L_n$ is said to be *unambiguous* if every word u of L admits a *unique* decomposition

$$u = u_0 a_1 u_1 ... a_n u_n,$$

with $u_0 \in L_0, u_1 \in L_1, ..., u_n \in L_n$. For instance, the marked product $\{a, c\}^* a \{\epsilon\} b \{b, c\}^*$ is unambiguous.

Let us define the family USF of *unambiguous star-free* languages as the smallest Boolean algebra of languages of Σ^* containing the languages of the form A^*, for $A \subseteq \Sigma$, which is closed under unambiguous marked product (cf [13].

The family *USF* is a very robust class of languages: the languages in this family admit indeed several other nice characterizations (see [15] for a survey).

It can be shown that *USF* is strictly included in *SF*, and so we have the following chain of inclusions:

$$USF \subsetneq SF \subsetneq INT \subsetneq REG.$$

The following theorem, proved in [3], shows the role of unambiguity in Problem 2.

Theorem 4.1 *If L_1 and $L_2 \in USF$, then $L_1 \sqcup\!\sqcup L_2 \in SF$.*

5 Cyclic Submonoids and Combinatorics on Words

The languages in the family *USF* can be described by regular expressions in which the star operation is restricted to subsets of the alphabet. Furthermore, Theorem 4.1 states that the shuffle of languages in this family is star-free. Hence, the critical situations, with respect to Problem 2, occur with languages corresponding to regular expressions in which the star operation is applied to concatenation of letters. So, in this section, we consider the shuffle of languages of the form u^*, where u is a word of Σ^*. Actually, such languages correspond to *cyclic submonoids* of Σ^*.

The special interest of such languages in our context is shown by the following theorem, proved in [2].

Theorem 5.1 *If the word u contains more than one letter, then the language u^* is intermixed.*

Moreover, next theorem, firstly proved in [9], shows that the combinatorial properties of the word u play a role in Problem 2. Let us first introduce a definition. A word $u \in \Sigma^*$ is *primitive* if it is not a proper power of another word of Σ^*, i.e., if the condition $u = v^n$, for some word v and integer n, implies that $u = v$ and $n = 1$.

Theorem 5.2 *The language u^* is star-free if and only if u is a primitive word.*

We now consider the shuffle $u^* \sqcup\!\sqcup v^*$ of two cyclic submonoids generated by the words u and v, respectively. If u and v are primitive words then, by the previous theorem, u^* and v^* are star-free languages. Remark that the languages u^* and v^* do not belong to *USF*, and their shuffle, in general, is not star-free. Here we study the conditions under which the language $u^* \sqcup\!\sqcup v^*$ is star-free.

Let us consider some examples. If $u = b$ and $v = ab$, the language $b^* \sqcup\!\sqcup (ab)^* = (b + ab)^*$ is star-free. Let us consider now $u = aab$ and $v = bba$, the language $(aab)^* \sqcup\!\sqcup (bba)^*$ is not star-free. Indeed the language

$$((aab)^* \sqcup\!\sqcup (bba)^*) \cap (ab)^* = ((ab)^3)^*$$

is not star-free, by the Theorem 5.2

Problem 3 : *Characterize the pairs of primitive words $u, v \in \Sigma^*$ such that $u^* \sqcup\!\sqcup v^*$ is a star-free language.*

This last problem is closely related to some relevant questions in combinatorics on words. Recall that combinatorics on words is a fundamental part of the theory of words and languages. It is deeply connected to numerous different fields of mathematic and its applications, and it emphasizes the algorithmic nature of many problems on words (cf [7]).

Some important problems in combinatorics on words pertain to the non primitive words that appear in the set u^+v^+, where u and v are primitive words.

A remarkable result in this direction is the famous Lyndon-Schutzenberger theorem (cf [8]), originally formulated for the free groups.

Theorem 5.3 *If u and v are distinct primitive words, then the word $u^n v^m$ is primitive for all $n, m \geq 2$.*

The next theorem, proved by Shyr and Yu ([14]), can be considered as a light improvement of the previous result.

Theorem 5.4 *If u and v are distinct primitive words, then there is at most one non-primitive word in the language u^+v^+.*

Problem 3 is, in a certain sense, related to those considered in the previous theorems, with the difference that we here take into account the shuffle of the two languages u^+ and v^+, instead of their concatenation. Actually, Problem 3 leads to investigate the non-primitive words that appear in the language $u^+ \sqcup\!\sqcup v^+$, where u and v are primitive words. In particular, we are interested to investigate the exponents of the powers that appear in $u^+ \sqcup\!\sqcup v^+$.

Let us introduce further notation. Let us denote by Q the set of primitive words. For $u, v, w \in Q$, let $p(u, v, w)$ be the integer k such that

$$u^* \sqcup\!\sqcup v^* \cap w^* = (w^k)^*.$$

Remark that, if $u^* \sqcup\!\sqcup v^* \cap w^* = \{\epsilon\}$, then $p(u, v, w) = 0$. For $u, v \in Q$, let us define the set of integers

$$P(u, v) = \{p(u, v, w) |\ w \in Q\}.$$

For instance, if we consider the words $u = a^{10}b$, $v = b$, then $P(u, v) = \{0, 1, 2, 5, 10\}$.

The following problem is closely related to Problem 3.

Problem 4 : *Given two primitive words u, v, characterize the set $P(u, v)$ in terms of the combinatorial properties of u and v.*

References

[1] Baeten, J., Weijland, W.: *Process Algebra*, Cambridge University Press, 1990.

[2] Berstel, J., Boasson, L., Carton, O., Pin, J.-E., Restivo, A.: The expressive power of the shuffle product, *Inf. Comput.*, **208**(11), 2010, 1258–1272.

[3] Castiglione, G., Restivo, A.: On the Shuffle of Star-Free Languages, *Fundam. Inform.*, **116**(1-4), 2012, 35–44.

[4] Diekert, V.: *The Book of Traces*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1995, ISBN 9810220588.

[5] Eilenberg, S.: *Automata, Languages, and Machines*, Academic Press, Inc., Orlando, FL, USA, 1976.

[6] Guaiana, G., Restivo, A., Salemi, S.: Star-Free Trace Languages, *Theor. Comput. Sci.*, **97**(2), 1992, 301–311.

[7] Lothaire, M.: *Algebraic Combinatorics on Words*, Cambridge University Press, 2002.

[8] Lyndon, R. C., Schützenberger, M. P.: The equation $a^M = b^N c^P$ in a free group, *Michigan Math. J.*, **9**(4), 1962, 289–298.

[9] McNaughton, R., Papert, S.: *Counter-Free Automata*, MIT Press, Cambridge, 1971.

[10] Perrot, J. F.: Varietes de Langages et Operations, *Theor. Comput. Sci.*, **7**, 1978, 197–210.

[11] Pin, J.-E.: *Varieties of formal languages*, North Oxford, LondonPlenum, New-York, 1986, (Traduction de Variétés de langages formels).

[12] Pin, J.-E.: Syntactic semigroups, in: *Handbook of formal languages* (G. Rozenberg, A. Salomaa, Eds.), vol. 1, chapter 10, Springer, 1997, 679–746.

[13] Pin, J.-E.: Theme and variations on the concatenation product, *Proceedings of the 4th international conference on Algebraic informatics*, CAI'11, Springer-Verlag, Berlin, Heidelberg, 2011, ISBN 978-3-642-21492-9, 44–64.

[14] Shyr, H. J., Yu, S. S.: Non-primitive words in the Language p^+q^+, *Soochow J. Math.*, **20**(4), 1994, 535–546.

[15] Tesson, P., Therien, D.: Diamonds Are Forever: The Variety DA, *Semigroups, Algorithms, Automata and Languages, Coimbra (Portugal) 2001*, World Scientific, 2002, 475–500.