Publ. Math. Debrecen
42 / 3–4 (1993), 315–321

Formal languages and primitive words[1]

By PÁL DÖMÖSI[2] (Debrecen), SÁNDOR HORVÁTH[2] (Budapest)
and MASAMI ITO (Kyoto)

Dedicated to Professor Lajos Tamássy on his 70th birthday

Abstract. The mathematical theory of formal languages has a very important role in theoretical computer science. In this paper we study various formal language problems related to the class of all primitive words over a fixed alphabet. Some results and problems are presented.

1. Introduction

The interest in combinatorial properties of words over a finite alphabet dates back to at least as far as THUE's 1906 and 1912 papers (see [20] and [21]). There exist a number of sistematical studies on combinatorics of words (see, for example, [6], [13], [19]). The concept of primitive words is defined and the unique existence of primitive roots is proved in [14]. Disjunctive languages are introduced in [17]. Disjunctive languages and primitive words are intensivity studied in [18] and [19]. Primitive words are considered with respect to the CHOMSKY-hierarchy in [10] and [11]. Classical works on formal languages and automata with respect to the CHOMSKY-hierarchy are, for example, [3], [4], [6], [7], [15] and [16]. In this paper we overview some results and problems on formal languages and primitive words.

---

[1]This paper was presented at the Conference "Sesiunea Anuală de Comunicări Ştiinţifice Universitatea Oradea", Oradea, Roumania, 6-8 June, 1991.

[2]The work of the first and second authors was supported in part by the Hungarian National Science Foundation "OTKA", Grants Nos. 1654/91, 1655/91 and 4295/92, and Nos. 334/88 and 4295/92, respectively.

2. Preliminaries

In this part we provide some notions and notations on formal languages. (For notions and notations not defined here see, for example, [6], [7], [15], [16], [19].) The elements of an *alphabet* X are called *letters* (X is supposed to be finite and nonempty). A *word* over an alphabet X is a finite string consisting of letters of X. The string consisting of zero letters is called the *empty word*, written λ. The *length* of a word w, in symbols $|w|$, means the number of letters in w when each letter is counted as many times as it occurs. By definition, $|\lambda| = 0$. At the same time, for any set H, $|H|$ denotes the cardinality of H. If u and v are words over an alphabet X, then their *catenation* uv is also a word over X. Catenation is an associative operation and the empty word λ is the identity with respect to catenation: $w\lambda = \lambda w = w$ for any word w. For a word w and natural number n, the notation w^n means the word obtained by catenating n copies of the word w. w^0 equals the empty word λ. w^m is called the m-th *power* of w for any nonnegative integer m. A word p is *primitive* iff it is nonempty and not of the form w^n for any word w and $n \geq 2$. Throughout this paper, the set of all primitive words over X is denoted by Q. Let X^* be the set of all words over X, moreover, let $X^+ = X^* - \{\lambda\}$. X^* and X^+ are a *free monoid* and a *free semigroup*, respectively, generated by X under catenation. Every subset L of X^* is called a (formal) *language* over X. L is said to be *dense* iff $X^*uX^* \cap L \neq \emptyset$ for any $u \in X^*$. (For $u \in X^*$ we use the shorthand u instead of $\{u\}$.) Obviously, a dense language is an infinite language. It can easily be seen that Q is a dense language, whenever $|X| \geq 2$. Throughout this paper, $\subseteq$ and $\subset$ denote (set-theoretic) inclusion and proper inclusion, respectively, and N stands for the set $\{0, 1, 2, ...\}$.

Let $L \subseteq X^*$. The congruence relation P_L on X^*, called the *principial congruence* determined by L, is defined as $u \equiv v(P_L)$ if and only if $xuy \in L \Leftrightarrow xvy \in L$ for any $x, y \in X^*$. A language $L \subseteq X^*$ is said to be *regular* iff P_L has finite index, i.e., the number of the equivalence classes of P_L is finite. In opposition to regular languages, a language $L \subseteq X^*$ is *disjunctive* iff every congruence class of P_L consists of a single element. It is clear that every disjunctive language is a dense language.

3. Chomsky classification of grammars

A generative (Chomsky-type) grammar [4] is an ordered quadruple $G = (V_N, V_T, S, P)$ where V_N and V_T are disjoint alphabets, $S \in V_N$, and P is a finite set of ordered pairs (u, v) such that v is a word over the alphabet $V = V_N \cup V_T$ and u is a word over V containing at least one letter of V_N. The elements of V_N are called *nonterminals* and those of V_T *terminals*. S is called the *start symbol*. Elements (u, v) of P are called *productions* and are written $u \to v$. A word u over V *derives directly* a word v, in symbols, $u \Rightarrow v$, iff there are words u_1, u_2, u_3, v_1 such that

$u = u_2u_1u_3,\ v = u_2v_1u_3$, and $u_1 \to v_1$ belongs to P. w *derives* z, or in symbols, $w \Rightarrow *z$ (w *really derives* z, or in symbols, $w \Rightarrow +z$) iff there is a finite sequence of words

$$w_0, w_1, ..., w_k, \quad k \geq 0 \quad (k > 0)$$

over X where $w_0 = w,\ w_k = z$ and $w_i \Rightarrow w_{i+1}$ for $0 \leq i \leq k-1$. In other words, $\Rightarrow *(\Rightarrow +)$ is the reflexive transitive closure (the transitive closure) of the binary relation $\Rightarrow$. The (formal) *language* $L(G)$ *generated by* G is defined by

$$L(G) = \{w \mid w \in V_T^*, S \Rightarrow +w\}.$$

G is *regular* (or G is of the type 3) iff each production is of one of the two forms $U \to vV$ or $U \to v$ where $U, V \in V_N$ and $v \in V_T^*$ (and then $P_{L(G)}$ has finite index).

G is *context-free* (or G is of type 2) iff each production is of the form $X \to u$ where $X \in V_N$ and $u \in (V_N \cup V_T)^*$. G is *context-sensitive* (or G is of type 1) iff each production is of the form $q_1Xq_2 \to q_1uq_2$, where $q_1, q_2 \in (V_N \cup V_T)^*,\ X \in V_N$, and $u \in (V_N \cup V_T)^+$, with the possible exception of the production $S \to \lambda$ whose occurrence in P implies, however, that S does not occur on the right side of any production in P. Finally, G is *phrase-structure* (or G is of type 0) if P has no restriction.

If there exists a generative grammar G of type $i(= 0, 1, 2, 3)$ such that $L = L(G)$ holds for a language $L \subseteq X^*$ then we also say that L is of type i. $\mathcal{L}_i$ $(i = 0, 1, 2, 3)$ denotes the class of type i languages. It is well-known that they form the *Chomsky-hierarchy* with $\emptyset \neq \mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$. It is well-known too that to each language class $\mathcal{L}_i$ there corresponds a class $\mathcal{A}_i$ $(i = 0, 1, 2, 3)$ of abstract *nondeterministic* discrete *automata* in the sense that for any $L \subseteq X^*, L \in \mathcal{L}_i$ holds iff there is an $A \in \mathcal{A}_i$ "*accepting*", from among all words of X^*, exactly those belonging to L. In the latter case we also say that A *accepts* L. Nondeterminism means here that A always freely chooses its "next move" from a finite number of actions possible at that stage of its operation. By definition, A *accepts an (input) word* w iff there is a finite sequence of consecutive possible moves of A during the "processing" of w, leading to an *accepting* or *final state of* A. *Deterministic automata* are special cases of nondeterministic automata, in which during the processing of any (input) word, at any stage at most one next move is possible. A language is called a *deterministic language* iff it is accepted by a deterministic automaton. For any type i, let $\det \mathcal{L}_i$ denote the class of deterministic languages of type i. It is known that $\det \mathcal{L}_3 = \mathcal{L}_3, \det \mathcal{L}_2 \subset \mathcal{L}_2$, and $\det \mathcal{L}_0 = \mathcal{L}_0$, but it is a famous open question, the so-called "*lba problem*", whether $\det \mathcal{L}_1 = \mathcal{L}_1$ or $\det \mathcal{L}_1 \subset \mathcal{L}_1$. Here "lba" is a shorthand for "*linear bounded automaton*", as the elements of $\mathcal{A}_1$ are termed. (For a detailed discussion of these notions and results, see, e.g., [6], [7] or [12].)

4. Some results and problems related to primitive words

In this section we suppose $|X| \geq 2$, and we consider only words, languages and language classes over X. (The results and problems discussed in this part are trivial or even untrue if X is a singleton.) We first study where Q is in the Chomsky-hierarchy.

A typical example of a disjunctive language is Q. Thus Q is not regular. To prove that Q is not deterministic context-free we use well-known results.

The following Theorem I, a classical result on the class of context-free languages, is widely known as "Bar-Hillel's lemma", or more precisely, "Bar-Hillel, Perles and Shamir's lemma" [1]. Here we formulate this lemma in its "full", "modern" form (i.e. $m = 0$ may stand too in $uv^m wx^m y$). Moreover, we note that the second author of the present paper showed in [8] that there exist properly context-sensitive, recursive, recursively enumerable, and non–recursively-enumerable languages that do satisfy this lemma. (For further combinatorial properties of context-free languages see, e.g., [2] and [9].)

Theorem I (Bar-Hillel's lemma, [1]). *For each context-free language L there exists a positive integer n with the following property: each word z in L, $|z| > n$, is of the form $uvwxy$, where $|vwx| \leq n, |vx| > 0$, and $uv^m wx^m y$ is in L for all $m \geq 0$.*

We also use the following

Theorem II (for a proof, see [5] or [7]). *L is deterministic context-free iff $X^* - L$ is deterministic context-free, i.e., $L \in \det \mathcal{L}_2$ iff $X^* - L \in \det \mathcal{L}_2$.*

Now we are ready to show the following

Proposition 1. *Q is not deterministic context-free, i.e., $Q \notin \det \mathcal{L}_2$.*

Proof. By Theorem II it is enough to prove that $X^* - Q$ does not satisfy the conditions of Bar-Hillel's lemma (Theorem I). Suppose the contrary and let $a, b \in X,\ a \neq b,\ n \geq 1$ (with n having the property described in Theorem I) such that $(a^{n+1}b^{n+1})^2$ is of the form $uvwxy$ with $|vwx| \leq n, |vx| > 0,\ uv^m wx^m y \in X^* - Q,\ m \geq 0$. Then for $m = 0$ we have

$$uwy \in \{a^i b^j a^s b^t \mid i, j, s, t \geq 1,\ (i, j) \neq (s, t)\} \subseteq Q,$$

contradicting $uwy \in X^* - Q$. $\square$

It can easily be seen that Q is accepted by a deterministic linear bounded automaton. Thus we have the following

Proposition 2. $Q \in \det \mathcal{L}_1 - \det \mathcal{L}_2$.

Conjecture. *Q is not context-free, i.e. $Q \notin \mathcal{L}_2$.*

Problem (Ito and Katsura [11]). Does L disjunctive imply $L \cap Q$ *disjunctive*?

We give a negative answer for the case $L \in \det \mathcal{L}_1 - \mathcal{L}_2$ in

Proposition 3. *There is a disjunctive language $L \in \det \mathcal{L}_1 - \mathcal{L}_2$ such that $L \cap Q$ is dense but not disjunctive (and $L \cap Q \in \mathcal{L}_2$).*

Proof sketch. Let $L = L' \cup Q^{(2)}$ where

$$L' = \{wba^{|w|} \mid w \in X^*\}, \quad Q^{(2)} = \{q^2 \mid q \in Q\}.$$

Similarly to the case of Q, it is easy to see that L too can be accepted by a deterministic linear bounded automaton, so $L \in \det \mathcal{L}_1$. On the other hand, $L \notin \mathcal{L}_2$ can be shown exactly as $X^* - Q \notin \mathcal{L}_2$ was shown in the proof of Proposition 1 above. Further, it can easily be seen that $L' \subseteq Q$ (and $L' \in \mathcal{L}_2$). So $L \cap Q = L' \in \mathcal{L}_2$ (since $Q \cap Q^{(2)} = \emptyset$).

For any $w \in X^*$ we have $wba^{|w|} \in L'$ $(a, b \in X,\ a \neq b)$. Thus L' is dense. On the other hand, $ab \equiv bb(P_{L'})$ $(a, b \in X,\ a \neq b)$. Therefore, L' is not disjunctive. Finally, by [19] we have that for the disjunctivity of L it is enough to check the case $|w_1| = |w_2|, w_1 \neq w_2$ $(w_1, w_2 \in X^*)$. Indeed, we obtain $w_1ba^{|w_1|}w_1ba^{|w_1|} \in Q^{(2)} \subseteq L$ and $w_2ba^{|w_1|}w_1ba^{|w_1|} \notin L$. □

We note that the above problem is still open for $L \in \mathcal{L}_2$. We conclude this paper with proving three further propositions.

Proposition 4. *There is a disjunctive language $L \in \mathcal{L}_2$ such that $L - Q^{(1)} \neq \emptyset,\ L \cap Q \neq \emptyset$ (where $Q^{(1)} = Q \cup \lambda$ as usual).*

Proof. Let $L = \{xyz \mid y \in X,\ x, z \in X^+,\ |x| = |z|,\ x \neq z\}$. It is easy to see that $L \in \mathcal{L}_2$. Furthermore, $(abb)^3 = abbabbabb \in L - Q^{(1)}$ $(x = abba,\ y = b,\ z = babb,\ |x| = |z|,\ x \neq z)$. On the other hand we have for any pair $w_1, w_2 \in X^*$, with $w_1 \neq w_2,\ |w_1| = |w_2|$, that

$$w_1a^{2|w_1|+1}bw_1a^{2|w_1|+1} \notin L,$$

and

$$w_2a^{2|w_1|+1}bw_1a^{2|w_1|+1} \in L \cap Q,$$

so by [19] L is disjunctive. It is clear that even both $L - Q^{(1)}$ and $L \cap Q$ are infinite. □

Proposition 5. *There are infinitely many dense languages in $\mathcal{L}_1 - \mathcal{L}_2$ and $\mathcal{L}_0 - \mathcal{L}_1$, and continuum-many outside $\mathcal{L}_0$.*

Proof. Concerning dense languages outside $\mathcal{L}_0$, the statement follows from:

1. there are continuum-many disjuctive languages (see [19]),
2. there are only denumerably many type 0 languages, and
3. disjunctivity implies density (this simply follows from the definitions).

Concerning the existence of infinitely many dense languages in $\mathcal{L}_1-\mathcal{L}_2$ and $\mathcal{L}_0-\mathcal{L}_1$, let $f : N \to N$ be a function and $L_f = \{a^{f(|w|)}bwba^{f(|w|)}|w \in X^*\}$. By suitably choosing f, L_f will be in $\mathcal{L}_1-\mathcal{L}_2$ or $\mathcal{L}_0-\mathcal{L}_1$, respectively. $\square$

Remark. From the above construction we can see that dense languages can in fact be arbitrarily "thin" in the "statistical sense".

Proposition 6. *There are infinitely many nondisjunctive languages in $\mathcal{L}_1 - \mathcal{L}_2$ and $\mathcal{L}_0 - \mathcal{L}_1$, and continuum-many outside $\mathcal{L}_0$.*

Proof. Let again $f : N \to N$ be a function and

$$L_f = \{a^{f(n)}b^{f(n)}a^{f(n)} \mid n \in N\}.$$

Clearly $(w_1, w_2 \in L_f - \{\lambda\},\ w_1 \neq w_2) \Rightarrow w_1 \equiv w_2(P_{L_f})$ and again by suitably choosing f, the statement follows. $\square$

References

[1] Y. Bar-Hillel, M. Perles and S. Shamir, On formal properties of simple phrase structure grammars, *Zeitschr. Phonetik, Sprachwiss. Kommunikationsforsch.,* **14** (1961), 143–172.

[2] L. Boasson and S. Horváth, On languages satisfying Ogden's lemma, vol. 12, *R. A. I. R. O. Informatique théorique*, 1978, pp. 201–202.

[3] N. Chomsky, Context-free grammars and pusdown storage, *M. I. T. Res. Lab. Electron. Quart. Prog. Rept.* **65** (1962).

[4] N. Chomsky, Formal properties of grammars, *Handbook of Math. Psychology* **2** (1963), 328–418.

[5] S. Ginsburg and S. A. Greibach, Deterministic context-free languages, *Inform. and Control* **9** (1966), 620–648.

[6] N. A. Harrison, Introduction to Formal Language Theory, *Addison-Wesley Publishing Company, Reading, Mass*, 1978.

[7] J. E. Hopcroft and J. D. Ullman, Introduction to Automata Theory, Languages, and Computation, *Addison-Wesley*, *Reading, Mass.*, 1979.

[8] S. Horváth, The family of languages satisfying Bar-Hillel's Lemma, *R. A. I. R. O. Informatique théorique* **12** (1978), 193–199.

[9] S. Horváth, A comparison of iteration conditions on formal languages, *Colloquia Math. Soc. János Bolyai* **42** Proc. Conf. Algebra, Combinatorics and Logic in Computer Science, Győr (Hungary), (1983), 453–463.

[10] M. Ito, M. Katsura, H. J. Shyr and S. S. Yu, Automata accepting primitive words, *Semigroup Forum* **37** (1988), 45–52.

[11] M. Ito and M. Katsura, Context-free languages consisting of non-primitive words, *Int. Journ. of Comp. Math.* **40** (1991), 157–167.

[12] S. Y. Kuroda, Classes of languages and linear-bounded automata, *Inform. and Control* **7** (1964), 207–223.

[13] M. Lothaire, Combinatorics on Words, *Addison-Wesley, Reading, Mass.* 1983, and *Cambridge Univ. Press*, 1984.

[14] R. C. Lyndon and M. P. Schützenberger, On the equation $a^M = b^N c^P$ in a free group, vol. 9, *Michigan Math. Journ*, 1962, pp. 289–298.

[15] A. Salomaa, Theory of Automata, *Pergamon Press, New York*, 1969.

[16] A. Salomaa, Formal Languages, *Academic Press, New York, London*, 1973.

[17] H. J. Shyr, Disjunctive languages on a free monoid, *Inform. and Control* **34** (1977), 123–129.
[18] H. J. Shyr, Thierrin, G., Disjunctive languages and codes, LNCS 56 (Proc. FCT' 77, ed.: M. Karpinski), *Springer-Verlag*, 1977, pp. 171–176.
[19] H. J. Shyr, Free Monoids and Languages, Lect. Notes, Dept. Math., Soochow Univ., *Taipei, Taiwan*, 1979.
[20] A. Thue, Über unendliche Zeichenreihen, *Norske Videnskabers Selskabs Skrifter Mat.-Nat. Kl. (Kristiania),* **7** (1906), 1–22.
[21] A. Thue, Über die gegenseitige Lage gleicher Theile gewisser Zeichenreihen, *Norske Videnskabers Selskabs Skrifter Mat.-Nat. Kl. (Kristiania)* **1** (1912), 1–67.

PÁL DÖMÖSI
L. KOSSUTH UNIVERSITY
H–DEBRECEN

SÁNDOR HORVÁTH
L. EÖTVÖS UNIVERSITY
H–BUDAPEST

MASAMI ITO
SANGYO UNIVERSITY
KYOTO, JAPAN

(Received December 8, 1992)