RAIRO. INFORMATIQUE THÉORIQUE

SÁNDOR HORVÁTH

The family of languages satisfying Bar-Hillel's lemma

RAIRO. Informatique théorique, tome 12, nº 3 (1978), p. 193-199
<http://www.numdam.org/item?id=ITA_1978__12_3_193_0>

R.A.I.R.O. Informatique théorique/Theoretical Computer Science
(vol. 12, n° 3, 1978, p. 193 à 199)

THE FAMILY OF LANGUAGES SATISFYING BAR-HILLEL'S LEMMA (*) (1)

by Sándor HORVÁTH (2)

Communicated by M. NIVAT

Abstract. — *It is shown that there exist properly context-sensitive, recursive recursively enumerable, and non-recursively enumerable, languages that do satisfy the classical pumping lemma for context-free languages (resp. for regular sets). The family of these languages is briefly studied.*

INTRODUCTION

In our terminology and notation we mainly follow Hopcroft and Ullman [3]. Let Σ be a countably infinite "base alphabet", $\mathscr{L}$ the class of "languages" i. e. sets L for which there is a finite $\Sigma_1 \subset \Sigma$ with $L \subset \Sigma_1^*$. The subclasses $\mathscr{RE}$, $\mathscr{CS}$, $\mathscr{CF}$, $\mathscr{RG}$ are then the Chomsky classes (the classes of recursively enumerable, context-sensitive, context-free and regular languages respectively), and let $\mathscr{R}$ be the class of recursive languages. As is wellknown (see e. g. [3]), the following chain of proper inclusions hold:

$$\mathscr{RG} \underset{\neq}{\subset} \mathscr{CF} \underset{\neq}{\subset} \mathscr{CS} \underset{\neq}{\subset} \mathscr{R} \underset{\neq}{\subset} \mathscr{RE} \underset{\neq}{\subset} \mathscr{L}$$

(in this paper, an inclusion denoted by " $\subset$ " is not necessarily proper).

A classical result on the class $\mathscr{CF}$, known as "Bar-Hillel's lemma" (in short "*BH* lemma") or the "*uvwxy* theorem" or "$p-q$ theorem" (which was first formulated in [1] and appeared and was used later, among many others, in [2-5]), is the following.

BAR-HILLEL'S LEMMA : *For every context-free language L there exist constants p and q such that any* $z \in L$ *with* $|z| > p$ *can be written as* $z = uvwxy$ *where* $|vwx| \leqq q$ *and* $|vx| > 0$ *so that* $\{uv^i wx^i y \mid i \geqq 0\} \subset L$.

(*) Received December 1977, revised March 1978.
(1) This paper is a slightly modified version of the author's earlier paper [8].
(2) Eötvös Lorand University, Dept. Computer Math., Budapest 8.

We say briefly that every context-free language is "BH". We remark that if we are given a context-free grammar for L then we can effectively calculate suitable p and q from it, and so we can decide, by means of the BH lemma, whether L is infinite or not. Another typical application of the BH lemma is its use in proofs, that some languages are not context-free.

Here we formulated the BH lemma in its "full", "modern" form i. e. $i=0$ may stand too in $uv^i wx^i y$. Let us denote the family of "full BH" languages (as a subclass of $\mathscr{L}$) by $\mathscr{B}_0$. In the original, "weak" form of the lemma (in [1, 2]) $i \geqq 1$, and let us denote the corresponding "weaker" family by $\mathscr{B}_1$. Another restriction is the "regular case" where $|vw|=0$, and we denote the corresponding two "regular BH" families (analogously to $\mathscr{B}_0$ and $\mathscr{B}_1$) by $\mathscr{B}\mathscr{R}_0$ and $\mathscr{B}\mathscr{R}_1$. In the following proposition we relate these four "BH families" to each other, in terms of set-theoretic inclusion.

PROPOSITION 1: *Between the families $\mathscr{B}_0$, $\mathscr{B}_1$, $\mathscr{B}\mathscr{R}_0$ and $\mathscr{B}\mathscr{R}_1$ the following relations hold:*

$$\mathscr{B}_0 \underset{\neq}{\subset} \mathscr{B}_1, \qquad \mathscr{B}\mathscr{R}_1 \underset{\neq}{\subset} \mathscr{B}_1, \qquad \mathscr{B}_0 - \mathscr{B}\mathscr{R}_1 \neq \emptyset,$$

$$\mathscr{B}\mathscr{R}_1 - \mathscr{B}_0 \neq \emptyset, \qquad \textit{and} \qquad \mathscr{B}\mathscr{R}_0 \underset{\neq}{\subset} \mathscr{B}_0 \cap \mathscr{B}\mathscr{R}_1.$$

Proof: Let

$$L_1 := \{ a^m b^n a^n \mid 0 \leqq m \leqq n \}, \qquad L_2 := \{ a^m b^m \mid m \geqq 0 \},$$

$$L_3 := \{ a^{m^2} b^n \mid m \geqq 0,\ n \geqq 1 \}, \qquad \text{and} \qquad L_4 := \{ a^m b^m a^n \mid m \geqq 0,\ n \geqq 1 \}.$$

Then we have

$$L_1 \in \mathscr{B}_1 - \mathscr{B}_0, \qquad L_1 \in \mathscr{B}_1 - \mathscr{B}\mathscr{R}_1, \qquad L_2 \in \mathscr{B}_0 - \mathscr{B}\mathscr{R}_1,$$

$$L_3 \in \mathscr{B}\mathscr{R}_1 - \mathscr{B}_0 \qquad \text{and} \qquad L_4 \in (\mathscr{B}_0 \cap \mathscr{B}\mathscr{R}_1) - \mathscr{B}\mathscr{R}_0$$

($\mathscr{B}\mathscr{R}_0 \subset \mathscr{B}_0 \cap \mathscr{B}\mathscr{R}_1$ is evident).

Q.E.D.

It can be conjectured that the full BH property is only a necessary condition for a language to be context-free, and this is even stated, though without proof, e. g. in [4, 5]. The aim of the present paper is to give such a proof, together with some further (algebraic and set-theoretic) characterization of the above four BH families.

ALGEBRAIC PROPERTIES OF THE *BH* FAMILIES AND THEIR RELATION TO THE CHOMSKY CLASSES

The four *BH* families are "almost" AFL's (*see* [6]), namely we have the following.

PROPOSITION 2. – *The families $\mathscr{B}_0$, $\mathscr{B}_1$, $\mathscr{B}\mathscr{R}_0$ and $\mathscr{B}\mathscr{R}_1$ satisfy all and only those "AFL axioms" different from closedness under inverse homomorphism and intersection with regular sets.*

Proof: We prove only the two non-closedness statements (the rest is a simple checking). In view of Proposition 1 above, it suffices to prove that the application of these two kinds of operations to elements of $\mathscr{B}\mathscr{R}_0$ may result in languages even outside $\mathscr{B}_1$. To show this, let

$$L_5 := L_3 \cup a^* \quad (\textit{see}\ \text{above}),$$

$$h: a \mapsto a,\ b \mapsto ab \ \text{be a homomorphism},$$

$$L_6 := a^* b \quad (\in \mathscr{R}\mathscr{G}).$$

Then we have $L_5 \in \mathscr{B}\mathscr{R}_0$ while

$$h^{-1}(L_5) = \{ a^{m^2-1} b \mid m \geqq 1 \} \cup a^* \notin \mathscr{B}_1$$

and

$$L_5 \cap L_6 = \{ a^{m^2} b \mid m \geqq 0 \} \notin \mathscr{B}_1.$$

(For $\mathscr{B}_0$ and $\mathscr{B}_1$ only, a more complex construction is the following:

$$L_5 := \{ a^{k^2} b^m c d^m e^{n^2} \mid k,\ n \geqq 0;\ m \geqq 1 \} \cup a^* c e^*$$

$$h: a \mapsto a,\ b \mapsto ab,\ c \mapsto c,\ d \mapsto de,\ e \mapsto e,$$

$$L_6 := a^* bcde^*.)$$

Q.E.D.

In the rest of this section we relate the four *BH* families to the Chomsky classes, but for the sake of simplicity we shall speak only about $\mathscr{B}_0$, though all results will be valid verbatim for the other *BH* families too.

THEOREM 1: $\mathscr{B}_0 \cap (\mathscr{C}\mathscr{S} - \mathscr{C}\mathscr{F}) \neq \emptyset$.

First proof: We construct an element L of $\mathscr{B}_0 \cap (\mathscr{C}\mathscr{S} - \mathscr{C}\mathscr{F})$. Let L consist of exactly those words v on $\{ a, b, c \}$ obtainable by substituting in any element w of $L' := \{ r^j s^k t^m \mid j, m \geqq k \geqq 0 \}$, an arbitrary element of $a^+ b^+$ for each of the letters r and t, and an arbitrary element of $a^+ c^+$ for each s. We call the substituted words the r-, s- or t-subwords of any v according to what letter of w they substitute. Clearly $L \in \mathscr{B}_0$ (e. g. with $p=0$, $q=2$). A context-sensitive grammar for L can easily be obtained by suitably modifying such a grammar of L', it is left to the reader. We have to prove that L is not context-free. Assuming the contrary, let L be generated by some context-free grammar in whose rules the maximal length of the right sides in d. (Unlike the usual proofs of the BH lemma, this grammar is context-free in the most general sense, it need not be "normed" in any manner.) Let $z_1, z_2, \ldots,$ be an infinite sequence of elements of L such that the number k_i of the s-subwords of $z_i, \to \infty$ if $i \to \infty$. For each i let T_i be a derivation tree of z_i and T_i' be the least subtree of T_i such that its terminal string contains all the s-subwords of z_i. Among the immediate subtrees of T_i' there is one, say with root A_i, the terminal string of which contains at least $(k_i+1-d)/d$ s-subwords, and of course does not contain both an r-subword and a t-subword at a time. Then again there is a variable D, occurring in the sequence (A_i) infinitely often. If A_{i_1} and A_{i_2} are two occurrences of D such that $i_2 - i_1$ is sufficiently large, then by substituting the A_{i_2}-subtree of T_{i_2} for the A_{i_1}-subtree in T_{i_1}, we get an element of L in which the number of s-subwords arbitrarily exceeds the number of either the r-subwords or the t-subwords, contradicting the definition of L.

Q.E.D.

REMARKS: 1. In the above first proof of Theorem 1 the language L seems at first sight to be unnecessarily complicated, but the case of L_1 in the proof of Proposition 1 (of which $L_1 \in \mathscr{CS} - \mathscr{CF}$ is wellknown, this can be proved e. g. in a way similar to the above proof, or just by the BH lemma, since L_1 is not in $\mathscr{B}_0$, only in $\mathscr{B}_1$) shows that the main difficulty in constructing non-context-free elements of $\mathscr{B}_0$ is to cover $i=0$ too.

2. Hereby we have proved the nonemptiness itself too of $\mathscr{CS} - \mathscr{CF}$, and in a similar way it can be proved, without the BH lemma and any "normal form transformation", that no language of the form $\{ a^{f(i)} b^{g(i)} a^{h(i)} \mid i \geqq 0 \}$ can be context-free if the functions $f, g, h \to \infty$.

3. In this proof we used only the (quite general) notion of a context-free grammar and that of a derivation tree. The following proof uses already the fact that $\mathscr{CS} - \mathscr{CF} \neq \emptyset$, and that all and only the context-free languages are pushdown-automaton recognizable.

Second proof of Theorem 1: Let $a, b, c \in \Sigma_1$, $H \subset \Sigma_1^*$, $H \in \mathscr{CS} - \mathscr{CF}$, and

$$L := (\{ a^n bc^n \mid n \geqq 1 \} H) \cup (b \Sigma_1^*) \quad (\in \mathscr{CS}).$$

Clearly $L \in \mathscr{B}_0$ (e. g. with $p=0$, $q=3$). Suppose $L \in \mathscr{CF}$, then it is accepted by some pushdown automaton (pda) M. It is easy to see that we can construct from M another pda M_1 such that any word $w \in \Sigma_1^*$ is accepted by M_1 iff $abcw$ is accepted by M, i. e. H is accepted by the pda M_1, contradiction.

Q.E.D.

The following results concern the existence of elements of $\mathscr{B}_0$ in $\mathscr{R}-\mathscr{C}\mathscr{S}$, $\mathscr{R}\mathscr{E}-\mathscr{R}$ and $\mathscr{L}-\mathscr{R}\mathscr{E}$, and the cardinality of $\mathscr{B}_0$.

THEOREM 2: $\mathscr{B}_0 \cap (\mathscr{R}-\mathscr{C}\mathscr{S}) \neq \emptyset$.

First proof: Take an element H of $\mathscr{R}-\mathscr{C}\mathscr{S}$ (the existence of H is proved e. g. in [3]), and define L exactly as in the second proof of Theorem 1. It remains to prove only that L is not context-sensitive. Indirectly, let L be accepted by a linear bounded automaton (lba) M, then another lba M_1 which first prefixes the string abc to its input word w and then does the same as M would do with the word $abcw$ as input, accepts H, contradiction.

Q.E.D.

Second proof: It is known that the context-sensitive languages (if their words are regarded as "r-adic numbers" for suitable r) are primitive recursive sets (this is proved e. g. in [7]), on the other hand there exist recursive but not primitive recursive sets (languages). (Besides, this provides another proof of the existence of non-context-sensitive recursive languages.) If in the above definition of L, H is recursive but not primitive recursive, then the primitive recursiveness of L would imply that of H too (since prefixing abc clearly corresponds to a primitive recursive function), contradiction.

Q.E.D.

THEOREM 3: *$\mathscr{B}_0 \cap (\mathscr{R}\mathscr{E}-\mathscr{R}) \neq \emptyset$ and $\mathscr{B}_0 \cap (\mathscr{L}-\mathscr{R}\mathscr{E}) \neq \emptyset$.*

Proof: The same argument as in the first proof of Theorem 2, except that now $H \in \mathscr{R}\mathscr{E}-\mathscr{R}$ or $H \in \mathscr{L}-\mathscr{R}\mathscr{E}$ respectively, and M, M_1 are Turing machines instead of lba's.

Q.E.D.

COROLLARY: *The cardinality of $\mathscr{B}_0 \cap (\mathscr{L}-\mathscr{R}\mathscr{E})$, and consequently that of $\mathscr{B}_0$ too, is C (continuum).*

Proof: The assertion easily follows from the preceding proof and the fact that the cardinality of $\mathscr{L}-\mathscr{R}\mathscr{E}$ is C.

Q.E.D.

We remark that of course the cardinality of $\mathscr{L}-\mathscr{B}_0$ is C as well, since

$$\{L \mid L \text{ is an infinite subset of } \{a^{i^2} \mid i \geqq 1\}\} \subset \mathscr{L}-\mathscr{B}_0.$$

PROBLEMS: 1. Are the sets of grammars corresponding to $\mathscr{B}_0 \cap (\mathscr{C}\mathscr{S}-\mathscr{C}\mathscr{F})$ and $\mathscr{B}_0 \cap (\mathscr{R}\mathscr{E}-\mathscr{C}\mathscr{S})$ recursive or at least recursively enumerable?

2. For what grammars generating BH elements of $\mathscr{RE} - \mathscr{CF}$ can we compute directly from the rules the corresponding p, q constants?

3. Which of our results are valid for "Ogden's lemma" (*see* [13, 14]) too in place of (the variants of) the BH lemma? (Ogden's lemma is stronger than the BH lemma.)

CONCLUDING REMARKS AND ACKNOWLEDGEMENT

I should like to thank my colleague, Dr. L. Hunyadvári, a talk on the algebraic properties of $\mathscr{B}_0$, and that he discovered for me, though after the finishing of this research, the papers [9-11]. (So these papers together, and ours, are mutually independent.) Only our Theorem 1 and the second part of our Theorem 3 appear in them, but the attached proofs are valid only for the "weak" BH cases ($i \geqq 1$). Yet later (after the 2nd Hung. Comp. Sci. Conf., Budapest, 1977, where the first version of this paper [8] was presented), the author discovered a further independent article, [12], in which the second part of our Theorem 3 appears, with a similar proof. Our proof of non-closedness under inverse homomorphism bears the influence of an analogous proof in [12], but ours is simpler. I thank also Prof. G. Păun (Bucharest) for pointing out that $\mathscr{B}_0$ is not closed under intersection with regular sets.

REFERENCES

1. Y. Bar-Hillel, M. Perles and E. Shamir, *On Formal Properties of Simple Phrase Structure Grammars*, Zeitschr. Phonetik, Sprachwiss., Kommunikationsforsch., Vol. 14, 1961, p. 143-172.
2. S. Ginsburg, *The Mathematical Theory of Context-free Languages*, McGraw-Hill, New York, 1966.
3. J. E. Hopcroft and J. D. Ullman, *Formal Languages and their Relation to Automata*, Addison-Wesley, Reading, Mass., 1969.
4. D. F. Martin, *Formal Languages and their Related Automata*, in *Computer Science*, A. F. Cardenas, L. Presser and M. Marin, eds., Wiley-Interscience, New York, London, 1972, p. 409-460.
5. A. Salomaa, *Formal Languages*, Academic Press, New York, London, 1973.
6. S. Ginsburg and S. Greibach, *Abstract Families of Languages*, Mem. Amer. Math. Soc., Vol. 87, 1969, p. 1-32.
7. W. S. Brainerd and L. H. Landweber, *Theory of Computation*, Wiley-Interscience, New York, London, 1974.
8. S. Horváth, *BHFL: the Family of Languages Satisfying Bar-Hillel's Lemma*, 2nd Hungarian Computer Science Conf., Budapest, June 27-July 2, preprints, Vol. I, p. 479-483.

9. C. CĪSLARU and G. PĂUN, *Classes of Languages with the Bar-Hillel, Perles and Shamir's Property*, Bull. Math. Soc. Sc. Math. R. S. Roum., Bucharest, Vol. 18, No. 3-4, 1974 (received: July, 1975; appeared: 1976), p. 273-278.

10. V. COARDOS, *O clasă de limbaje neidependente de context care verifică conditia lui Bar-Hillel*, Stud. cerc. mat., Bucharest, Vol. 27, No. 4, 1975, p. 407-411.

11. G. PĂUN, *Asupra proprietătii lui Bar-Hillel, Perles si Shamir*, Stud. cerc. mat., Bucharest, Vol. 28, No. 3, 1976, p. 303-309.

12. T. KLØVE, *Pumping languages*, Internat. J. Comp. Math., R. RUSTIN, ed., Gordon and Breach Sc. Publishers, London, New York, Paris; Vol. 6, No. 2, 1977, p. 115-125.

13. W. OGDEN, *A Helpful Result for Proving Inherent Ambiguity*, Math. Syst. Theory, Vol. 2, No. 3, 1968, p. 191-194.

14. A. V. AHO and J. D. ULLMAN, *The Theory of Parsing, Translation, and Compiling*, Vol. I, "Parsing", Prentice-Hall, 1971, 2nd printing: 1972, section 2.6.