STUDIA UNIV. BABEŞ-BOLYAI, INFORMATICA, VOLUME XLII, NUMBER 2, 1997

THE U LANGUAGES – A SUBCLASS OF THE Ψ LANGUAGES

A. F. BOER

Abstract. In this paper we define a subclass of the class of Ψ-languages, and we prove some closure properties. The idea for this class was a stronger correlation of the derivation, which may be achieved through the unambiguous indexes.

1. Definitions, remarks

Definition 1.1. *Let $G = (N, T, F, P. S)$ be an indexed grammar [1], i.e. a grammar in which the following conditions hold: a) $N = N_1 \cup N_2$, $N_1 \cap N_2 = \varnothing$, $S \in N_1$. b) The rules of P have the form $A \to Bf$ with $A \in N_1$, $B \in N$, $f \in F$. c) The rules in the indexes have the form $A \to z$, with $A \in N$, $z \in (N_2 \cup T)^+$. We will say that the index $f \in F$ is unambiguous, if it does not contain any two rules with the same left side.*

Definition 1.2. *A U-grammar is a Ψ-grammar [3], in which each index is unambiguous.*

Remark 1.3. *In a U-grammar after the application of the first index the generated (produced) word is uniquely defined.*

Example 1.4. $G = (N, T, F, P, S)$, where $N = \{S, A, B_a, B_b, C, C_a, C_b\}$; the application of indexes begins always by the nonterminal symbol C; $T = \{a, b\}$; the set F contains three unambiguous indexes: $g = [B_a \to a, C_a \to a, B_b \to b, C_b \to b]$, $f = [B_a \to C_a B_a, C_a \to a, B_b \to C_b B_b, C_b \to b]$, $h = [C \to B_a B_b]$, P contains three rules: $S \to Ag$, $A \to Af$, $A \to Ch$. Each derivation in G has the following form: $S \to Ag \to^* Af^n g \to Ch f^n g \to (B_a B_b) f^n g = B_a f^n g B_b f^n g$; forth, the derivation from B_a is made such that:

$$B_a f^n g \to C_a f^{n-1} g B_a f^{n-1} g \to a B_a f^{n-1} g \to aC_a f^{n-2} g B_a f^{n-2} g \to^*$$
$$a^2 B_a f^{n-2} g \to^* \ldots \to^* a^{n-2} C_a g B_a g \to^* a^n.$$

And the derivation from B_b is made similarly changing a with b and obtaining b^n with the same n; finally the language $L = \{ a^n b^n \mid n \in N^* \}$ is obtained.

Remark 1.5. *A U-grammar for the language $\{ a_1^n a_2^n \ldots a_1^n \mid n \in N\}$ for each $t \in N^*$ may be built similarly.*

Received by the editors: January 28, 1998.
1991 Mathematics Subject Classifications: 68Q45, 68Q50.
CR Categories and Subject Descriptors: F.4.2 [**Mathematical Logic and Formal Languages**]: Grammars and Other Rewriting Systems - *grammar types*; F.4.3. [**Mathematical Logic and Formal Languages**]: Formal Languages - *operations in languages, classes defined by grammars*.

Proposition 1.6. *Each finite language can be generated by a U-grammar.*

Proof. Let $L = \{x_1, ..., x_n\} \subset T^*$. The grammar $G = (\{S, C\}, T, F, P, S\}$ where F contains the indexes $f_j = [c \to w_j]$ for each w_j from L, and P contains the rules $S \to Cf_i$. It is obviously that G is a **U**-grammar and $L = L(G)$.

Proposition 1.7. *$L_3 \subset$ **L** (U), i.e. each regular language can be generated by a U-grammar, but not inversely, and the inclusion is strong.*

Proof. Let $G = (N, T, P, S)$ be a 3-type grammar, thus all the rules in P have the form $A \to a$ or $A \to aB$ where A, B are nonterminal symbols, a is terminal symbol. We construct the **U**-grammar $G' = (N', T, F, P', S')$ so: $N' = N \cup \{S', C\}$ where S' and C are new nonterminals; F contains an index f for each rule $A \to a$ and $A \to aB$ from P and an index more: $h = [C \to S]$; P contains the rules $S' \to S'f$ for each f from F and the rule $S' \to Ch$; the application of the rules from indexes-properly speaking of the rules from P- begins with the symbol C. It is obviously that $L(G') = L(G)$ and that G' is a **U** grammar. That the inclusion is strong results e.g. from the fact that the language $\{a^n b^n \mid n \geq 1\}$ is not a regular, but it is a **U**-language as is shown in the Example 1.4.

Proposition 1.8. L(U) $\subseteq$ L (Ψ) $\subseteq$ L (Ind).

The proof results from the definition of the corresponding grammars.

2. Closure properties of the family L(U)

Lemma 2.1. *For each **U** grammar $G = (N, T, F, P, S)$ a U-grammar $G' = (N', T, F', P', S'')$ can be effectively constructed, in which the application of index rules begins always from the same nonterminal symbol.*

Proof. We denote by N_C the set of such nonterminals $A \in N$ from which it can begin the application of the index rules. (This set can be effectively found: it contains from such symbols $a \in N$ for which exists a $B \in N$ and $f \in F$ so that $B \to Af$ is in P and f contains a rule with the left side A, i.e. $A \to z$, $z \in (N_2 \cup T)^+$.) For each symbol A from N_C we introduce a new, „dual“ symbol A' and we denote the set of these „dual“ symbols with N'_C. Let $C \notin N$ a new nonterminal symbol. We define $N' = (N_1 - N_C) \cup N'_C \cup N_2 \cup \{C\}$. F' contains all indexes from F and for each $A \in N_C$ contains a new index $c_A = [C \to A]$ where C is the new nonterminal.

P' contains the rules from P, changing all symbols $A \in N_C$ with the dual symbols A', and we add to P' the rules $A' \to Cc_A$ for each A' from N'_C.

The new initial symbol S'' is S, if S is not in N_C, otherwise is the dual S' of S.

The new Grammar G' is equivalent to G, the application of index rules begins always from the same nonterminal symbol C, and all indexes are unambiguous (so were the indexes in G).

Lemma 2.2. *The family **L(U)** is effectively closed under the union.*

Proof. Let $G' = (N', T', F', P', S')$ and $G'' = (N'', T'', F'', P'', S'')$ two **U**-grammars. We can consider that the sets N' and N'', and F' and F'' are distinct (otherwise we can rename the symbols). Let S be a new nonterminal symbol and h a new index. The grammar $G = (N, T, F, P, S)$ with $N = N' \cup N'' \cup \{S\}$, $T = T' \cup T''$, $F = F' \cup F'' \cup \{h = [A \to A$ for all A from $N' \cup N''] \}$, $P = P' \cup P'' \cup \{S \to S'h, S \to S''h\}$ is a **U**-grammar which generates the union, i.e. $L(G) = L(G') \cup L(G'')$.

Lemma 2.3. *The family* **L(U)** *is not closed under the intersection.*

Proof. We consider two **U**-languages L_1 and L_2 and show that their intersection has greater density [2] than linear. So the intersection can not be a Ψ-language [3] and then from the proposition 3 follows that it is not a **U**-language.

The two languages and their grammars are: $L_1 = \{a^p \mid p = 2^n, n = 1, 2, \ldots \}$ and $L_2 = \{a^p \mid p = n2^n, n = 1, 2, \ldots \}$, $L_1 = L(G)$, $L_2 = L(G')$,

$G = \{N, T, F, P, S\}$, where $N = \{S, T, A\}$, $T = \{a\}$, $F = \{f, h\}$, with $f = [T \to AA, A \to AA]$, $h = [A \to a]$, and P contains the rules $S \to Th$, $T \to Tf$, and $G' = (N', T, F', P', S)$, where $N = \{S, T, A, B\}$, $F' = \{h', f'\}$ with $h' = [a \to a, B \to a]$, $f' = [A \to AA, B \to aB, T \to AAB]$, and P' contains the rules $S \to TH$, $T \to Tf$, as in G too.

To find the intersection $L_1 \cap L_2 = \{a^p|p = 2^n = m2^m)\}$ we must solve the diophantine equation $2^n = m2^m$. It is obviously that for $m \geq 1$ we have $m \leq n$. Dividing the equation by 2^m we obtain $2^{n-m} = m$, with $n\text{-}m \geq 0$, so m has the form 2^k. Replacing m with 2^k we have $2^{n-(2\wedge k)} = 2^k$, i.e. $n\text{-}2^k = k$, and from this we have $n = 2^k + k$. So the general form of the solutions is $(n, m) = (2^k + k, 2^k)$ and in conclusion the intersection is $L_1 \cap L_2 = \{a^p \mid p = 2^k \cdot 2^{(2\wedge k)}, k = 1, 2, \ldots \}$.

Now we will show that this language has greater density as all linear functions. (it has the cn^2). For this it is sufficient to show this fact for the set of lengths: $M = \{2^{(2\wedge k)} \mid k > 0\}$; we suppose that the density of this set is linear, that is that there exists a natural constant n_0 and a natural constant c so that for all $n \geq n_0$ there is an m in M for which $n \leq m < cn$ [2]. I.e. for each $n \geq n_0$ there is a natural number k which satisfy the relations $n \leq 2^{(2\wedge k)} < cn$. From this: $\log_2 n \leq 2^k < \log_2 c + \log_2 n$, and so $\log_2(\log_2 n) \leq k < \log_2(\log_2 n + \log_2 c)$ for all k (beginning at a value). But this is not possible because for an n sufficient great the difference between the two margins – which generally are not integers – becomes less than the unity. We have $\log_2(\log_2 n + \log_2 c) - \log_2(\log_2 n) < 1$ if and only if $\log_2 (\log_2 n + \log_2 c) / \log_2 n < 1$, and so if $\log_2 (1 + \log_2 c / \log_2 n) < 1$, i.e. if $\log_2 c / \log_2 n < 1$, what is true for $n < c$. So, if $\log_2(\log_2 n)$ is not an integer and $n > c$ then we have no element of the set M between n and $n + cn$, and so the density cannot be linear.

Lemma 2.4. *The family is effectively closed under the concatenation.*

Proof Let $L' = L(G')$, $L'' = L(G'')$, where G' and G'' are **U**-grammars. We will show that from the grammars G' and G'', the grammar G can be construct which generates the concatenation $L'L''$. We may suppose that all the symbols in the two grammars are different (because when it is not so, we can rename them), and according to the Lemma 2.1 we can consider that in the two grammars $G' = (N', T', F', P', S')$ and $G'' = (N'', T'' F'', P'', S'')$ the application of the index rules begins from the same nonterminals C' and C'', respectively. We will denote the grammar which generates the concatenation by $G = (N, T, F, P, S')$, where $T = T' \cup T''$, $N. = N' \cup N'' \cup \{ Z, D', D''\}$, Z, D', D'' are new nonterminals ($Z \notin N' \cup N''$, which will not appear in the left side of any rule from P or from indexes, from D will begin the generation of the words according to the grammar G', form D''-according to G''). P is built in the following way: it contains all the rules from P', but each rule of the form $A \to C'$ f is changed in the rule $A \to D'f$; it contains all the rules from P'', but each rule of the form $A \to C''f$ is changed in the rule $A \to D''f$; we add two new rules: $D' \to S''h'$, where $h' = [\, C' \to C', \to \,]$ is a new index, and $D'' \to Dh''$ where $h'' = [\, D \to C'C''\,]$ is another new index, which really make the concatenation.

The set of indexes of G is formed in the following way: to each index f'' from F'' we add the rule $C' \to C'$; through this process the unambiguously of the indexes is preserved, because we supposed that all the symbols from the two grammars G' and G'' are different; maintaining the notation of the indexes from F'' with the new rules, we take for the new index set $F = F' \cup F'' \cup \{ h', h''\}$; we remark that the new indexes h' and h'' are unambiguously too.

Each derivation in the grammar G begins with: $S \to^* D'z'$, where $z' \in F'$; forth – when in the grammar G' the application of the index rules begins – only the rule $D' \to S''h'$ can be applied, because the new nonterminal symbol D' does not appear in the left side of any index rule from F' and in any other rule from, thus it obtains $D'z' \to S''h'z'$; forth it can apply only the rules from P'', while the application of the index rules from F'', and for a derivation from G'' of the form $S'' \to C''z''$ it is obtained $S''h'z' \to D''z''h'z'$; similarly to the first part of the derivation (which models the derivation in G'), now too, when it begins the application of the index rules, the only possibility is the application of the new rule $D'' \to Dh''$ from P, after that no rule from P can be applied any rule from P, only the rule $D \to C'C''$ from h'', and so we obtain $S \to^* D'z' \to S''h'z' \to^* D''z''h'z' \to Dh''z''h'z' \to (C'C'')z''h'z' = C'z''h'z'C''z''h'z'$. Forth for the first part, the only possibility is $C'z''h'z' \to^* C'k'z' \to C'z'$; from this place it continues exactly like in G': if the derivation is not terminal in G', then it will not be terminal in G too, and if we had in G' the derivation $S' \to^* C'z' \to^* x' \in T'^*$, then we have here the same derivation $C'z' \to^* x' \in T'^*$. For the second part $C''z''h'z'$ cu $C''z''$ we think similarly for the string $C'z'$: if the derivation in C'' was not terminal, it will be not terminal in G too and it stops at last by the index h' which doesn't contain any rule with the nonterminals from N'', and if we had in G'' a derivation of the form $C''z'' \to {}^* x'' \in T''^*$, then we will have the same derivation in G.

So, for each two terminal derivations $S' \to^* x' \in T'^*$ in the grammar G' and $S'' \to^* x'' \in T''^*$ in the grammar G'' we have the terminal derivation $S' \to^* x'x'' \in T'^*T''^* \subseteq T^*$, and any other terminal derivation isn't possible.

We remark that the proof is constructive: it gives effectively a method for building of the grammar G from the two given grammars G' and G''.

Lemma 2.5. *The family* **L(U)** *is effectively closed under the Kleene-closure.*

The proof may be similarly to the proof of previous lemma. It is easy to see that the Kleene closure is formed from words of the form $x_1x_2 \ldots x_n$, where $x_j \in L$, and from the empty word e. Let $G = (N, T, F, P, S)$ a **U**-grammar in which the application of the index rules begins always with the same nonterminal symbol C (see the Lemma 2.1). We construct (effectively) a new **U**-grammar $G' = (N', T, F', P', S')$ which will generate the language L^* and in which the application of the index rules begins always with the nonterminal C'. The terminal alphabet is, obviously, the same, T. For each nonterminal A from N we introduce a „dual“ A_d and we denote the set of these „duals“ with N_d. Let $N' = N \cup N_d \cup \{ C', M\}$ (M is a new nonterminal symbol). P' will contain all rules of P, all „dualised“ rules, i.e. if $A \to Bf$ was in P then $A_d \to B_d f$ is in P' too, and the new rules: $C_d \to C'h$, $C_d \to S h'$, $C \to S h$, $C \to C' h$, where h and h' are new indexes: $h = [C' \to C'M, M \to C'M]$, $h' = [M \to C', C' \to C']$. In each index from F we change the symbol C in C' and we add the rule $M \to \underset{.}{M}$. F' will contain all this changed indexes from F and the two new indexes h and h'.

Each derivation in G' begins with: $S' \to^* C_d z_1$, if $S \to^* Cz_1$ $(z1 \in F'^*)$ was a derivation in G, exactly as in the grammar G, but changing all nonterminal symbol A with A_d. If in G it follows the application of the index rules, here the rule $C_d \to C'h$ or the rules $C_d \to Sh'$ from P P' can be applied; in the first case follows the application of the rule $C' \to C'$ from h (doesn't exist any other possibility) and continues exactly as in the grammar G, and in the second case a new sequence of indexes $S \to^* C z_2$ may be generated from S and so in the grammar G' we obtain the derivation $S' \to^* C_d z_1 \to Sh' z_1 \to^* Cz_2 h' z_1$. From C we can obtain Sh or $C'h$; in the first case it continues similarly again, in the second case it follows the application of the index rules. Anyway, after a finite number of steps we obtain: $S' \to^* C_d z_1 \to S h' z_1 \to^* C z_2 h' z_1 \to^* \ldots \to^* C z_n h z_{n-1} h \ldots h z_2 h' z_1$ (for $j = 1, \ldots, n$ we have $z_j \in F'^*$). If it applies now the rule $C \to C'h$, then follows the application of index rules, because C' don't appears in the right part of any rule from P'. From the index h it applies the rule $C' \to C'M$, and it obtain the sequence $C' z_n h z_{n-1} h \ldots h z_2 h' z_1 M z_n h z_{n-1} h \ldots h z_2 h' z_1$; from $C'z_n$ it obtains the same-terminal or nonterminal-word as in the grammar G from Cz_n, and from $M z_n h z_{n-1} h \ldots h z_2 h' z_1$ it obtains $M h z_{n-1} h \ldots h z_2 h' z_1$. From M applying again the index h, it obtains $C'M$ again, and all repeats until it goes to $Mh'z_1$. It applies the rule $M \to C'$, and forth it proceeds with $C'z_1$ as previously. If a derivation $C'z_k \to x_k$ does not result a terminal word, the all the derivation stops at a word that isn't in T^*; if all derivations $C'z_k \to x_k$ give words x_k in T^*, then it obtains finally the word $x_nx_{n-1} \ldots x_2x_1$ in T^*. **L(U)** is closed (effectively) under the union and it contains all the finite languages, so we can add the empty word e to L', if it is necessary. The constructions was effective. So the lemma is proved.

Lemma 2.6. *The family* **L(U)** *is effectively closed under the e-free homomorphisms.*

Proof. Let the language $L = L\ (G)$, where $G = (N, T, F, P, S)$ is a grammar of the type **U**, and $h\colon T \to T'^{+}$ an e-free homomorphism. We construct a new **U**-grammar $G' = (N', T', F', P', S')$ which will generate the language $L' = h(L)$. For each terminal symbol a in T we introduce a new nonterminal symbol N_a and we denote their set with N_T : $N_T = \{N_a | a \in T\}$. Let $N' = N \cup N_T \cup \{ S' \}$, where S' is a new nonterminal, the new initial symbol. P' will contain all rules from P and a new rule: $S' \to Sh$, $h = [N_a \to h(a)$ for all a from $T]$ is a new index; we observe that h is unambiguous. In each index from F we change each nonterminal symbol a which appears (in the right part) of the rules with its „dual“ N_a and we add the rules $N_a \to N_a$ for each a in T. F' will contain all such modified indexes and the new index h.

Each derivation in G' begins with cu $S' \to Sh$, after that follows a derivation as in G, and we obtain a word in which each terminal symbol a is changed in its „dual“ N_a; if the word obtained in G was a terminal word x, then through the application of the index h we obtain $h(x)$-the homomorphic image of the word x, and if the word obtained in G was not in T^*, then it contains nonterminal symbols for which the index h cannot be apply and so in the grammar G' we do not obtain a terminal word too. Since G' is a **U**-grammar, it is obviously that $L\ (G') = h\ (\ L)$ and the construction was effective, so the lemma is proved.

Lemma 2.7. *The family* **L(U)** *is not closed under its complement.*

The *proof* results from the relation $A \cap B = \mathbf{C}(\ \mathbf{C}\ (A) \cup \mathbf{C}\ (B)\ )$ and from the Lemmas 2.2 and 2.3.

Remark 2.8. To exemplify the utility of these grammars we consider the mathematical formulas and expressions. These – e.g. from the algebra or mathematical analysis – can be generated through a context-free grammar. But if we have a function with one or more variables and we want to generate the replacement of the variables with given values, then the context-free grammars are not sufficient. This problem may be easy solved by the **U**-grammars. To generate polynomials of any degree, ordered (ascending or descending) on the X, the context-free grammars are not sufficient again, but the **U**-grammars can be used successful in this case too.

REFERENCES

1. Aho, A.V., *Indexed grammars-an extension of context-free grammars*, Journal of the ACM, 1968, vol. 13, No. 4, 647-671.
2. Boer, A. F., *The density – a numerical characteristic for languages*, Studia Universitatis „Babes-Bolyai“, Informatica, 1996, 97-105.
3. Boer, A. F.: *The Ψ languages – a subclass of the indexed languages*, Studia Universitatis „Babes-Bolyai“, Informatica, 1997, 85-90.
4. Chomsky, N., *On certain formal properties of grammars*, Information and Control, No. 2, 1959, 137-167.

Institutul de Informatica "Gabor Denes", Oradea
E-mail address: boer@oradea.iiruc.ro