STUDIA UNIV. BABEŞ-BOLYAI, INFORMATICA, VOLUME II, NUMBER 1, 1997

THE Ψ-LANGUAGES – A SUBCLASS OF THE INDEXED LANGUAGES

A. F. BOER

Abstract. In this paper we define a subclass of the class of indexed languages [1], which is equivalent to a subclass of conditional languages, and prove that the density [2] of these languages is linear, which gives examples of non-Ψ-languages.

1. Introduction

The main idea for the definition of Ψ-grammar was a kind of „correlation“ which realises the indexes by the application of index-rules. This is more obviously if at first appear all indexes and afterwards all index rules are applied. The Ψ-grammar realises such a derivation. So the study of properties is easier too.

Remark 1.1. We will denote the length of the word x by $|x|$, the cardinal of a set M by $|M|$ or by $n(M)$, the empty word by e, the reflexive and transitive closure of the relation $\rightarrow$ by $\rightarrow^*$, the class of indexed grammars by **Ind**, the class of Ψ-grammars by Ψ; if G is a class of grammars then the corresponding class of languages we denote by $L(G)$.

Definition 1.2. A Ψ-grammar is an indexed grammar {1] $G=(N, T, F, P, S)$, in which the following conditions are satisfied:

1. $N = N_1 \cup N_2$, $N_1 \cap N_2 = \varnothing$, $S \in N_1$.
2. The rules of P have the form $A \rightarrow Bf$ with $A \in N_1$, $B \in N$, $f \in F$.
3. The rules in the indexes have the form $A \rightarrow z$, with $A \in N$, $z \in (N_2 \cup T)^+$.

We will denote the class of Ψ-grammars by Ψ, and the corresponding family of languages by **L(Ψ)**.

Received by the editors: January 12, 1998.
1991 Mathematics Subject Classification. 68Q45, 68Q50
1991 CR Categories and Descriptors. F4.2 [**Mathematical Logic and Formal Languages**]: Grammars and Other Rewriting Systems - *grammar types*; F4.3 [**Mathematical Logic and Formal Languages**]: Formal Languages - *Operations in languages, Classes defined by grammars.*

Example 1.3. Let $G = (\{ S, Z, A \}, \{ a \}, \{f, g\}, P, S)$, where P contains the rules $S \to Zg$, $Z \to Zf$, and the indexes are $f = \{Y \to A, A \to A^k\}$, $g = \{A \to a\}$, with k a natural number.

Each derivation in G which gives a terminal word (i.e. a word from T^*) have the form $S \to Zg \to Zfg \to Zffg = Zf^2g \to^* Zf^{n+1}g \to Af^ng \to^* (Af^{n-1}g)^k \to^* (Af^{n-2}g)^{k\wedge 2} \to^* (Ag)^{k\wedge n} \to^* a^{k\wedge n}$. $(n \geq 0)$. We observe that $L(G) = \{ a^{k\wedge n} \mid n = 0, 1, 2, \dots \}$.

Remark 1.4. (1) From the definition results that in a Ψ-grammar each derivation has two parts: first only the rules of P are applied, and is obtained $S \to^* Az$, $A \in N$, $z \in F^*$, after them are applied only the rules from the indexes, because after the application of any rule from any index appears a nonterminal symbol from N_2 or a terminal symbol on which cannot apply the rules from P.

(2) With the rules of P a regular language (i.e. from L_3) is obtained in the alphabet $N \cup F^*$; we will denote it with L'.

Definition 1.5. We call (nonterminal) *composition set* of the word $x \in I^*$ the set $V(x) = K_N(x) = \{ A \in N \mid A \text{ is in } x\}$. (i.e. the set of the nonterminal symbols which appears in x).

Notation 1.6. If all index sequences after all nonterminal symbols of the word $x \in I^*$ ends with the same index sequence $z \in F^*$, then we will write $(x)z$ (i.e. if $x = x_1A_1z_1z \dots x_kA_kz_kzx_{k+1}$, then we will write $(x_1A_1z_1 \dots x_kA_kz_kx_{k+1})z$, and conversely).

Definition 1.7. The number $s(f) = \max \{ |x| \mid A \to x \in f \}$ is the *degree of the index f*

Definition 1.8. The number $s_N = \max \{ |pr_N(x)| \mid A \to x \in f \}$ (where $pr_N(x)$ is the projection of the word x on the neterminal alphabet N) is the *nonterminal degree of the index f.*

Definition 1.9. *The nonterminal degree of the grammar G* is the number $S_N = \max \{ s_N(f) \mid f \in F \}$.

Definition 1.10. *The degree of the grammar G* is the number $s = \max \{s(f) \mid f \in F\}$. (If the grammar is not obviously, we can write $s_{G,N}$, or s_G).

Remark 1.11. We have $s \geq s_N \geq 1$ for each nonterminal, and for each f in F $s(f) \geq s_N(f) \geq 1$, $s \geq s(f)$ and $s_N \geq s_N(f)$. If the language is infinite, then $s > 1$ (but $s_N = 1$ is possible for some infinite languages too, e.g. for the grammars with the rules in indexes of the form $A \to Ba$).

Notations 1.12. We will denote: $|N_1| = n(N_1) = n$, $|N_2| = n(N_2) = nn$, $|F| =$ $n(F) = q$ (the number of indexes in F), $|P| = n(P) = p$ (the number of rules in P), $|T|$ $= n(T) = t$ (the number of terminal symbols); the number of terminal symbols is $n + nn$.

If $x \in I^*$, $z \in (N_2 \cup T)^*$ °ï $x \to^* z$, and the derivation is made by application of all indexes from x, then we will write $R(x) = z$.

Remark 1.13. Using this notation, each terminal derivation in a Ψ-grammar (i.e. derivation which end in a terminal word from T^*) may be written in the form:

$$S \to^* Af_rf_{r-1} \dots f_1, \text{ and } y_1 = A,\ y_j = R(y_{j-1}f_{r-j+2}) \text{ for } j = 2, 3, \dots r+1$$

where $A \in N$, $f_j \in F$, for $j = 1, 2, \dots r$, $y_{r+1} \in T^*$, and in the first part of the derivation only the rules from P are applied (and in the second part, obviously, only the index rules).

2. The linear density of the Ψ-languages

Theorem 2.1. Each infinite Ψ-language has no more than linear density.

The proof of the theorem can be made using the following lemmas.

Remark 2.2. If a language is finite, then it is regular, and may be generated by grammar of type 3.

Lemma 2.3. In each infinite Ψ-language there exists a word x which derivation begins with: $S \to^* Az_1fqfz_2$, where only the rules from P are applied, with $a \in N$, z_1, q, $z_2 \in F^*$, $f \in F$, and such that the following conditions are satisfied:

a) the two appearances of the index f are obtained through the same rule from P;

b) $|R(Az_1f)| < |R(Az_1fqf)|$ (i.e. through the application of the index string qf the length of the word doesn't increase);

c) $V(R(Az_1f)) = V(R(Az_1fqf)) \neq \varnothing$, and we denote by V_0 (i.e. the composition set of $R(Az_1f)$ doesn't change through the application of the index string qf).

Proof. We number the rules of P such that: v: $A \to Bf$ $(1 \leq v \leq p)$. The number of a rule determines the index which appears through its application.

The composition sets which appear after the application of the first index are subsets of N_2, we number these subsets from 1 to 2^{nn}, and the number of the subset $H \subseteq N_2$ we note by $u(H)$.

We remark that if $|z| = a$ then $|R(Az)| \leq s^a$ $(z \in F^*, A \in N)$; we denote $|R(Az)|$ with b and then we have $a \geq \log_s b$. (For an infinite language we have $s > 1$.)

If $L(G)$ is an infinite language then for each natural number n there exists a word x from $L(G)$ with the length greater than n.

We will see the derivations by „steps“, where each step means the application of an index on all the nonterminal symbols to which this refers (i.e. each step has the form $(x)f \to^* R(x(f))$). To each step it corresponds the pair (u_j, v_j) formed from the number of the composition set of the word obtained through the application of the index of rang j, and from the number of the rule from P, through which application was this index of rang j obtained.

The total number of such pairs (u, v) is $p2^{nn}$; now, if we take a word $x \in L(G)$ with $|x| > s^{p(2\wedge nn)}$, then at least a pair (u, v) repeats. By each step the length of the word increases at most s times, so for to obtain a word of length a we need at least $\log_s a$ steps. Furthermore, if $a > s^{p(2\wedge nn)}$, then we have at least a repetition of a pair (u, v) so, that between the two appearances of the pair (u, v) the length of the word (in the alphabet $V = N \cup T$) increases. So, with the followings, the lemma is proved.

Indeed, the derivation of a word x with $|x| > n_0 = s^{p(2\wedge nn)}$ shows so:

$$S \to^*G_{,p}\, Af_1 \dots f_l = y_0 f_1 \dots f_l \to^*G\, (y_1)f_2 \dots f_l \to^*$$

$$_G\, (y_{i-1})f_i \dots f_l \to^* (y_{j-1})f_j \dots f_l \to^* (y_j)f_{j+1} \dots f_l \to^* y_l = x \in T^*,$$

where $(u_i, v_i) = (u_j, v_j)$, consequently $V(y_j) = V(y_i)$ and $f_j = f_i$, where the two indexes are obtained through the same rule from P and between y_i and y_j we have a growth of the length: $|y_j| > |y_i|$, and so $V(y_i) \neq \emptyset$. From y_j, by the help of the index string $f_{j+1} \dots f_l$ it obtains the terminal word $x \in T^*$, and so $f_{j+1} \dots f_l \neq e$. If we note $f_i = f_j = f$, $f_1 \dots f_{i-1} = z_1$, $f_{i+1} \dots f_{j-1} = q$, $f_{j+1} \dots f_l\, 0\, z_2$, then we have $y_i = R(Az_1f)$, $y_j = R(Az_1fqf)$.

Lemma 2.4. In each Ψ-grammar $|R(y)z| \le |y|s^{|z|}$, for $y \in V^*$, $z \in F^*$.

The proof can be made by induction on $|z|$, using the fact that, from the definition of s, $|R(Af)| \le s$, and $|R(af)| = |a| = 1$, where $A \in V, f \in F, a \in T$.

Lemma 2.5. In each infinite language L, which is generated through an indexed grammar $G \in \Psi$, it is a sequence of words $x_0, x_1, \dots, x_m, \dots, x_m \in L(G)$, so that $|x_m| < |x_{m+1}|$ and it is a natural number c so that we have $|x_{m+1}| \le c|x_m|$ for all $m = 0, 1, 2, \dots$

Proof. From Lemma 2.3 and from the fact that the rules of P have the form of grammars of the type 3 grammars follows that $S \to^* Az_1f(qf)^m z_2$ for $m = 0, 1, 2, \dots$, $Az_1f(qf)^m z_2 \to^* x_m \in T^*$ and $|x_m| < |x_{m+1}|$.

The derivation is made as follows: according to Lemma 2.3, $S \to^* Az_1fqfz_2$, i.e. $S \to^* B_1z_2 \to^* B_2fz_2$, $B_2 \to^* B_2fq$, $B_2 \to^* Az_1$; obviously, the derivation $B_2 \to^* B_2fq$ may be repeat however often, obtaining $Az_1f(qf)^m z_2$; further, from Lemma 2.3 too, we have:

$$Az_1fqf \rightarrow^* (u_0)qfz_2 \rightarrow^* (u_1)z_2 \rightarrow^* x \in T^*,$$

and $V(u_0) = v(u_1) = \{A_1, \ldots A_k\}$, $k \geq 1$. $(u_1)z_2 \rightarrow^* x \in T^*$ means that each A_j from $V(u_1)$ gives a non empty terminal word (because aren't rules of form $A \rightarrow e$): $A_jz_2 \rightarrow^* w_j \in T^*$ for $j = 1, \ldots, k$.

We consider $Az_1f(qf)^m z_2 \rightarrow^* (u_i)(qf)^{m-1}z_1 \rightarrow^* (u_m)z_2$, and having $V(u_0) = V(u_1)$ follows $V(u_i) = V(u_0) = \{A_1, \ldots A_k\}$, so for each $m = 1, 2, \ldots$ we have $(u_m)z_2 \rightarrow^* x_m$.

From Lemma 2.3 follows that $|u_1| > |u_0|$, what is possible only if there exists at least an $A_j \in V(u_0)$ for which $|R(A_jqg)| > |u_j| = 1$; this A_j appears in each u_i, so the length increase from u_i to u_{i+1}: $|u_i| < |u_{i+1}|$, $i = 0, 1, 2, \ldots, m-1$. From this follows that $|R((u_i)z_2)| < |R((u_{i+})z_2)|$, and so $|x_i| < |x_{i+1}|$. In this way we obtain the sequence of words $x_0, x_1, \ldots, x_m, \ldots$, with $x_m \in L(G)$ and $|x_m| < |x_{m+1}|$ for $m = 0, 1, 2,$.

We have still to prove that there is a constant c for which $|x_{m+1}| < c|x_m|$. Let $|qf| = d$; then $|R((u_m)qf)| \leq |u_m|s^d$; but $u_{m+1} = R((u_m)qf)$, and so $|u_{m+1}| \leq |u_m|s^d$.

From the equality $x_m = R((u_m)z_2)$, and noting $|z_2|$ with d' we obtain:

$$| R((u_{m+1})z_2)| \leq = |u_{m+1}|s^{d'} \leq s^{d'}s^d|u_m| \leq s^{d'+d}|x_m|.$$

since $|u_m| \leq |x_m|$.

We take $c = s^{d'+d}$, and so we obtain the needed equality.

Proof of the theorem. The density is linear when there exists a constant k such that for any natural number $n \geq n_0$ there exists a word $x \in L(G)$ such that $n < |x| < n+kn = (1+k)n = cn$ (see [2]).

We consider the sequence $x_0, x_1, \ldots, x_m, \ldots$, built in the Lemma 2.5. From the properties of this sequence follows that for each $n \geq |x_0|$ there exists an x_m such that: $|x_m| \leq n < |x_{m+1}|$. Then from the Lemma 2.5 we can write: $n < |x_{m+1}| \leq c\,|x_m| \leq cn$, $c = s^{d'+d}$. With $n_0 = |x_0|$ the theorem is proved.

Corollary 2.6. From this theorem follows that if an (infinite) language has a density greater than linear, then it cannot be generated by a Ψ-grammar. E.g. the language $L = \{a^{2\wedge(2\wedge n)} | n \geq 0\}$ is not in **L(Ψ)**.

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 „Babeš-Bolyai", Informatica, 1997.

E-mail address: `anti@lego.rdsor.ro`