THE DENSITY – A NUMERICAL CHARACTERISTIC FOR LANGUAGES

A. F. BOER

Abstract. In this paper we present the density of languages, that was defined by Stotzky in [3], we study some properties, and establish the density for some classes of languages. With this numerical characteristic it is possible to show that some languages can't be generated by grammars from some classes.

Definition 0.1 (3). *Let M an infinite subset of the set N of the natural numbers and $d : N \to N$ a function with $d(n) \leq d(m)$ if $n < m$. We say that the set M has the density d, if there exists a number $n_0 \in N$ so that for each natural number $n \geq n_0$ there exists an element m in M with the property $n \leq m \leq n + d(n)$.*

We remark that the density of a set isn't unique, but it is possible to find a "maximal" density.

Definition 0.2 (3). *The density of the language L is the density of the set of lengths of the words from L, i.e. the density of the set $M(L) = \{n \in N | \exists x \in L |x| = n\}$.*

We remark that we can reformulate this definition so: the infinite language L has the density d (where $d : N \to N$ is a function with $d(n) \leq d(m)$ if $n < m$), there exists a natural number n_0 such that for each $n \geq n_0$ there exists a word x in L such that $n \leq |x| \leq n + d(n)$.

Definition 0.3. *If $d = c$ (where c is a constant number) then the density is constant.*

Definition 0.4. *If the density is $d(n) = cn$ (where c is a constant number), then we say that it is a linear density.*

Remark 0.5. *In this case the definition may be reformulated in the following equivalent form: $\exists n_0 \leq 0 \exists c_1 > 1$ such that for each n from N if $n \geq n_0$ then there exists an element m in M such that $n \leq m \leq c_1 n$, where $c_1 = c + 1$.*

1991 *Mathematics Subject Classification.* Completeaza aici AMS classification.
1991 *CR Categories and Descriptors.* Completeaza aici CR categories and descriptors.

Proposition 0.6. *Let M be the set of the terms of an arithmetical progression, i.e. $M = \{a + bn | n \in N\}$ where a and $b > 0$ are natural constants. Then M has constant density.*

Proof. Let $n_0 = a$; for each n, between n and $n + b$ there exists (exactly) one term of the progression, so the constant function $d(n) = b$ is the density for the set M. □

Remark 0.7. 1. *If the set $M_1 \subseteq N$ has the density $d(n)$ and $M1 \subseteq M2 \subseteq N$ then the set $M2$ has the density $d(n)$ too.*

2. *From the proposition 0.6 and remark 0.7 it results that if a set contains an arithmetical progression then it has constant density.*
3. *If the set $M \subseteq N$ has the density $d(n)$ and $d' : N \to N$ is another function such that $d'(n) \geq d(n)$, for each n, then the set M has the density $d'(n)$ too.*

Proposition 0.8. *Let M the set of the terms of a geometrical progression, that means: $M = \{ab^n | n \in N\}$ where a and b ($b > 0, b \neq 1$) are natural constants. Then the set M has linear density and has no constant density.*

Proof. Because of $b > 1$, the progression is increasing. Let $n_0 = a$; for each n, between n and bn there exists at least one term of the progression: if m is the first natural number for which $ab^m \geq n$ then we have $ab^m < ab^{m+1} = bn$, also the linear function $d(n) = bn$ is density for the set M.

Now we show that M has no constant density. Suppose that M has constant density, i.e. we have the function $d : N \to N, d(n) = c$ for each natural n and there exists n_0 such that for each $n \geq n_0$ there exists an element m in M for which $n \leq m \leq c + n$. It means that there exists a natural number p so that $n \leq ab^p \leq c+n$. i. e. $\frac{n}{a} \leq b^p \leq \frac{n+c}{a}$, and from this $\log_b\left(\frac{n}{a}\right) \leq p \leq \log_b\left(\frac{n+c}{a}\right)$. But this is impossible because for a sufficient large p the difference between $\log_b\left(\frac{n}{a}\right)$ and $\log_b\left(\frac{n+c}{a}\right)$ is less than any real number

$$\log_b\left(\frac{n+c}{a}\right) - \log_b\left(\frac{n}{a}\right) = \log_b\left(\frac{n+c}{n}\right) < \varepsilon$$

for $n > \frac{c}{b^\varepsilon - 1}$, and so there is no natural number between n and $n + c$, and the density can not be constant. □

Proposition 0.9. *Each infinite regular set has constant density.*

Proof. From the theorem 2.8. in [2] results that if $L \subseteq T*$ is a regular set, then there exists the natural numbers (constants) p and q so that for each word z in L with $|z| > p$ there exists the words x, u, y in $T*$ such that $z = xuy$, $u1e(i.e.|u| > 0)$, $|uy| \leq q$, for each $k \geq 1$ the word $xu^k y$is in the language L. We have: $|xu^k y| = |x| + k|u| + |y| = a + kb$, where $a = |x| + |y|$ and $b = |u| > 0$, also $M(L)$ contains an arithmetical progression and then according to the proposition 0.6, the set $M(L)$ has the constant density $d(n) = b = |u|$. □

Proposition 0.10. *Each context-free language has constant density.*

Proof. We can proceed as in the proof of the proposition 0.9. Using the Bar-Hillel lemma [2, 3], where we will have for each word of the form xu^kwv_ky the relations: $|xu^kwv_ky| = |x| + k|u| + |w| + k|v| + |y| = a + kb$, where $a = |x| + |w| + |y| andb = |u| + |v|$, and so $M(L)$ contains an arithmetical progression and then according to the proposition 0.6 the set $M(L)$ (and the language L) has the constant density $d(n) = b = |u| + |v|$. □

Remark 0.11. 1. *There are languages with constant density which aren't context-free languages. For example the language $L = \{a^nb^nc^n|n \geq 1\}$ in the alphabet $\{a, b, c\}$ isn't context-free language [1], but has constant density.*

2. *The language $L = \{a^{k^n}|n \geq 1, k \in N\backslash\{0, 1\}, k$ is a constant$\}$ has linear density. In fact, we can use the proposition 0.8 when we observe that the lengths of the words from L form a geometrical progression.*

Example 0.12. *The set $M = \{2^{2^k}|k > 0\}$ has higher density than linear.*

We suppose that the density is linear, also that there exists the constant n_0 and the constant such that for each $n \geq n_0$ there exists an m in M for which $n \leq m \leq cn$. That means that for each $n \geq n_0$ it is a natural number k for which we have $n \leq 2^{2^k} \leq cn$. From this follows $\log_2 n \leq 2^k \leq \log_2 c + log_2 n$, also for each k we have $\log_2(\log_2 n) \leq k \leq \log_2(\log_2 n + \log_2 c)$ (beginning from a value). But this is impossible because for n sufficient large the difference between the two margins - which generally aren't natural numbers - will be less than any real number ε; let $\varepsilon < 1$. Moreover we have $\log_2(\log_2 n + \log_2 c) - \log_2(\log_2 n) < \varepsilon$ if and only if $\log_2 \frac{\log_2 n + \log_2 c}{\log_2 n} < \varepsilon$, also if $\log_2 \left(1 + \frac{\log_2 c}{\log_2 n}\right) < \varepsilon$. That is, if we have $\frac{\log_2 c}{\log_2 n} < \varepsilon$, which holds for $n < c^{1/\varepsilon}$. Also if $\log_2(\log_2 n)$ is not a natural number and $n > c^{1/\varepsilon}$ for a given ε, then we have no element from M between n and $n + cn$, and so the density cannot be linear.

Remark 0.13. *The density of the union of two set is less or equal to the density of each set. This is obviously from the remark 0.7.*

Remark 0.14. *The density of the intersection of two sets can be greater as the density of this sets.*

For example, let $M_1 = \{2^n|n = 1, 2,\}$ and $M_2 = \{n2^n|n = 1, 2,\}$. These sets have linear density, but the intersection $M_1 \cap M_2 = \{2^n2^{2^n}|n = 1, 2, ...\}$ has no linear density. Another example is the intersection of two sets with constant density, which has no constant, but linear density: $M_3 = \{2n|n \in N\}$,

$$M_4 = \{q_n|q_n = \begin{cases} 2n, & \text{if } \exists k \in N, n = 2^k, \\ 2n+1, & \text{otherwise} \end{cases}, n \in N\}$$

Obviously $M_3 \cap M_4 = \{2^n | n \in N^*\}$, and this set has linear, but not constant density.

Conclusion. The density characterises a set of natural numbers and so a language too. According to the remark 17, the intersection of two languages can have a greater density as the two original languages, and so it is possible e.g. to prove that some classes of languages are not closed under intersection.

References

[1] S. Ginsburg, *The Mathematical Theory of Context-Free Languages*, Mc Graw Hill Book Company, New York, 1966.

[2] Gh. Paun, *Probleme actuale n teoria limbajelor formale*, Editura Stiintifica si Enciclopedica, Bucuresti, 1984.

[3] E.D. Stotzky, *Uslovnie grammatiki s rasseiannim kontekstom*, Dokl. AN SSSR, 207(1972), No. 4.