RAIRO. INFORMATIQUE THÉORIQUE

R. SOMMERHALDER

Classes of languages proof against regular pumping

RAIRO. Informatique théorique, tome 14, n° 2 (1980), p. 169-180

<http://www.numdam.org/item?id=ITA_1980__14_2_169_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 14, n° 2, 1980, p. 169 à 180)

CLASSES OF LANGUAGES PROOF AGAINST REGULAR PUMPING (*)

by R. SOMMERHALDER (1)

Communicated by J. BERSTEL

Abstract. — *The main purpose of this note is to collect a number of scattered properties, examples and unanswered questions concerning classes of languages which are proof against various forms of pumping of the words in the language. These forms of pumping lead to different infinite hierarchies. Requiring that it must be possible to select from a prefix the substring to be repeated independently of the suffix, results in yet another hiearchy of languages.*

Finally a semi-pumping property will be given which is a combination of the Nerode-equivalence and the classical pumping lemma.

Résumé. — *Le but principal de cette note est de rassembler des propriétés, exemples et questions ouvertes de divers types concernant des classes de langages formels qui tolèrent diverses formes d'itération (pumping) des mots de leurs langages. Ces conditions d'itération conduisent à différentes hiérarchies infinies. Si l'on exige l'existence, dans un facteur gauche, d'un facteur itérant indépendant du facteur droit restant, on obtient encore une autre hiérarchie de langages.*

Enfin, on donne une propriété de demi-iération qui combine l'équivalence de Nérode et le lemme de l'étoile classique.

INTRODUCTION

In almost every course on formal languages one or more versions of the pumping lemma for regular sets will be covered. In the classical formulation of Rabin-Scott this lemma states that from every word of the language which is long enough, one can produce other words in the language by deleting, or by repeating an arbitrary number of times, some subword of the given word. Quite a few of the common examples in formal language theory are proof against this type of pumping. To circumvent the trouble of *ad hoc* reasoning, so called stronger versions of the pumping lemma can then be used.

In this note we will study classes of languages which are proof against these sorts of pumping with the added requirement that repeating a subword is

(*) Receveid April 1979, revised August 1979.
(1) Technische Hogeschool, Onderafdeling Wiskunde, The Netherlands.

R.A.I.R.O. Informatique théorique/Theoretical Informatics, 0399-0540/1980/169/$ 5.00

allowed only, if it is done more than a minimum number of times. An other added requirement will be that the selection of the subword to be repeated must, to a certain degree, be made independently of the right (or left-) context. All of this does not result in a pumping property which characterizes the regular sets. As to a characterizing property we do not have more to offer than a combination of the Nerode-equivalence and the pumping lemma. This property however is even more easy to use than the classical pumping lemma and thence desserves to be stated.

REGULAR PUMPING

In the sequel the following notation will be used:

- N is the set of all natural numbers, including zero;
- $\mathscr{R}$ is the class of regular sets;
- $L_1 \subseteqq L_2$ means that L_1 is included in L_2;
- $L_1 \subsetneqq L_2$ means that L_1 is strictly included in L_2;
- $|w|$ denotes the length of a word w;
- $\#(w_1, w_2)$ denotes the number of times w_1 occurs as a substring in w_2.

In the sequel we use an arbitrary but fixed Thue-sequence, i.e. an infinite sequence over an also arbitrary but fixed alphabet V_E (with at least three different letters) which does not contain a substring vv, with v non-empty. Prefixes of this sequence will be denoted by $x_1 x_2 \ldots x_i$.

Define: $\mathscr{P}(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that any word $z \in L$ with $|z| > p(L)$ can be written as $z = uvw$ such that $0 < |v| \leqq p(L)$ and $\{ uv^i w \mid i \geqq k \} \subseteqq L$.

The case $k=0$ in the above definition gives the class of all languages which are proof against the pumping as stated in the classical pumping lemma for regular languages. It is clear from the definition that for all $k \geqq 0$ we have $\mathscr{P}(k) \subseteqq \mathscr{P}(k+1)$. It is also obvious that $\mathscr{P}(1) = \mathscr{P}(2)$. This proves to be the only case where $\mathscr{P}(k) = \mathscr{P}(k+1)$, whence we have the inclusion diagram shown below.

There are various languages known which show that $\mathscr{R} \subsetneqq \mathscr{P}(0)$, one of those is

$$L_1 = \{ w \in \{ a, b \}^* \mid \#(a, w) = \#(b, w) \}.$$

It is also known that $\mathscr{P}(1) - \mathscr{P}(0)$ is non-empty. This is exemplified by $L_2 = \{ a^n b^m \mid n > m \geqq 1 \}$.

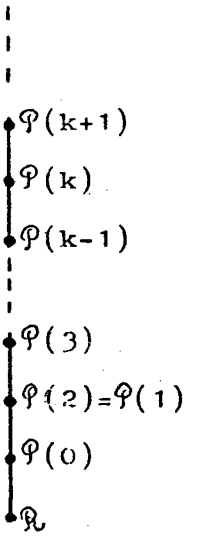

Figure 1

Fact 1: $\mathscr{P}(k)$ is strictly included in $\mathscr{P}(k+1)$ for all $k \geqq 2$.

To see this consider the languages $LP(k)$ which, using our Thue-sequence, are defined as follows

$$LP(k) = \{ x_1 x_2 \ldots x_i^j \mid i \geqq 1 \text{ and } j = 1 \text{ or } j \geqq k \}, \qquad k \geqq 2.$$

Obviously $LP(k) \in \mathscr{P}(k)$, just take $p(LP(k))$ equal to one and pump the last letter in any given word. Furthermore $LP(k+1) \notin \mathscr{P}(k)$ whence $\mathscr{P}(k)$ is strictly included in $\mathscr{P}(k+1)$. To see that $LP(k+1) \notin \mathscr{P}(k)$ observe that for al $LP(k)$ we have $z \in LP(k)$ contains a subword vv (with v non-empty) if and only if there exists an $x \in V_E$ such that z ends in xx. Suppose that $LP(k+1) \in \mathscr{P}(k)$ and let $p(LP(k+1))$ be equal to p. The prefix $z = x_1 x_2 \ldots x_{p+1}$ of our Thue-sequence belongs to $LP(k+1)$ whence by assumption there exist u, v, $w \in V_E^*$ such that $z = uvw$ and $z_i = uv^i w \in LP(k+1)$ for all $i \geqq k \geqq 2$. Therefore z_i must end in xx for some $x \in V_E$. If $|vw| \geqq 2$ then z itself must end in xx, contrary to the choice of z. Therefore $|vw| \leqq 1$ whence $|v| = 1$ and $w = \varepsilon$. But then if follows that $z_k = uv^k \in LP(k+1)$, contradicting the definition of $LP(k+1)$.

The argument above uses languages over an alphabet with at least three different letters. It should be obvious that the given hierarchy also exists if we restrict ourselves to two-letter languages.

Fact 2: If we restrict ourselves to one-letter languages we have $\mathscr{R} = \mathscr{P}(k)$ for all $k \geqq 0$.

Suppose $L \subseteq \{a\}^*$ and $L \in \mathscr{P}(k)$ for some $k \geqq 2$. *Let* p_k be the required constant and let $p_1 = k.p_k$. *If* $z \in L$ and $|z| > p_1 > p_k$ there exist u and v such that $0 < |v| \leqq p_k$ and $uv^i \in L$ for all $i \geqq k$. Let $v_1 = v^k$. Then $0 < |v_1| \leqq p_1$ and there exists a u_1 such that $z = u_1 v_1$ and $z_i = u_1 v_1^i \in L$ for all $i \geqq 1$. Therefore $L \in \mathscr{P}(1)$ and $\mathscr{P}(k) = \mathscr{P}(1)$ for all $k \geqq 1$. It remains to show that $\mathscr{R} = \mathscr{P}(1)$. Suppose once more that $L \subseteq \{a\}^*$ belongs to $\mathscr{P}(1)$. Let p be the required constant and let q be the least common multiple of $1, 2, 3, \ldots, p$. Now if $z \in L$ and $|z| > p$ then $za^{q.r} \in L$ for all $r \geqq 0$ because $q.r$ is a multiple of any number less than or equal to p. For any i such $1 \leqq i \leqq q$ define the set A_i as follows

$$A_i = L \cap \{a^n \mid n = p + i + q.r,\ r \geqq 0\}.$$

Every A_i is a regular set. L is the union of the A_i's and some finite set whence L is regular too and $\mathscr{R} = \mathscr{P}(0) = \mathscr{P}(1) = \mathscr{P}(2) = \ldots$

Fact 3: $\mathscr{P}(k)$ is closed with respect to ε-free substitution.

Let V be some alphabet and suppose that the language $L \subseteq V^*$ as well as the languages L_a, $a \in V$, belong to $\mathscr{P}(k)$. Consider the language $s(L)$ obtained by substituting L_a for a in L. Take p' to be the maximum of $\{p(L_a) \mid a \in V\}$ and set $p = p(L).p'$. Let $z \in s(L)$ be such that $|z| > p$. There exists a word $z_1 = a_1 a_2 \ldots a_n \in L$ and words $y_i \in L_{a_i}$ such that $z = y_1 y_2 \ldots y_n$. Two cases arise:

– there exists an i such that $|y_i| > p' \geqq p(L_{a_i})$. Then there exist u, v, w such that

$$0 < |v| \leqq p(L_{a_i}) \leqq p' \leqq p \qquad \text{and} \qquad \{uv^j w \mid j \geqq k\} \subseteq L_{a_i},$$

whence

$$\{y_1 y_2 \ldots y_{i-1} uv^j w y_{i+1} \ldots y_n \mid j \geqq k\} \subseteq s(L).$$

– $|y_i| \leqq p'$ for all $i \leqq n$. But then $p = p(L).p' < |z| \leqq |z_1|.p'$ and therefore $|z_1| > p(L)$. Thus there exist $u = a_1 \ldots a_g$, $v = a_{g+1} \ldots a_h$, and $w = a_{h+1} \ldots a_n$ such that $0 < |v| \leqq p(L)$ and $\{uv^j w \mid j \geqq k\} \subseteq L$. Since

$$0 < |y_{g+1} \ldots y_h| \leqq |v|.p' \leqq p(L).p' = p$$

we have

$$\{y_1 \ldots y_g (y_{g+1} \ldots y_h)^j . y_{h+1} \ldots y_n \mid j \geqq k\} \subseteq s(L).$$

Since there are no other possibilities, the language $s(L) \in \mathscr{P}(k)$.

Fact 4: $\mathscr{P}(k)$ is closed with respect to union, concatenation and iteration.

Since $\mathscr{R} \subseteq \mathscr{P}(k)$ fact 4 follows from fact 3 by paying tribute to the nuisance of the empty word.

Fact 5: $\mathscr{P}(k)$ is not closed with respect to erasing homomorphism and hence not closed with respect to arbitrary substitution. However, $\mathscr{P}(0)$ is closed with respect to arbitrary substitution.

The negative part of this observation follows by an example. Let $L=\{a^n c^m b^n \mid n, m \geqq 1\}$. Then $L \in \mathscr{P}(1)-\mathscr{P}(0)$. Take as an erasing homomorphism $h(a)=a$, $h(b)=b$ and $h(c)=\varepsilon$. Then $h(L)=\{a^n b^n \mid n \geqq 1\}$, which language does not belong to any $\mathscr{P}(k)$, $k \geqq 0$. The proof of the positive closure result is almost identical to the proof of fact 3. Given the word $z \in s(L)$ take $z_1=a_1 a_2 \ldots a_n$ to be a word of minimal length among those which can produce $z: \{z_1' \in L \mid z \in s(z_1')\}$.

The argument is now the same as above, apart from the assertion $0<|y_{g+1} \ldots y_h|$. This previously followed from the assumption that each $y_i \neq \varepsilon$. Now the assertion is true by the minimality of z_1. For if $y_{g+1} \ldots y_h=\varepsilon$ then $z=y_1 \ldots y_g y_{h+1} \ldots y_n$ and $z \in s(uw)$ but $|uw|<|uvw|$ and uvw would not have been of minimal length.

Fact 6: None of the classes $\mathscr{P}(k)$ is closed with respect to intersecting with a regular set and thus neither with respect to arbitrary intersection. Consequently, none of the classes $\mathscr{P}(k)$ is closed with respect to complement.

Consider the language

$$L_1=\{w \in \{a, b\}^* \mid \#(a, w)=\#(b, w)\}, \qquad L_1 \in \mathscr{P}(0) \subseteqq \mathscr{P}(k).$$

The intersection of L_1 with the regular set denoted by $a^* b^*$ equals $\{a^n b^n \mid n \geqq 0\} \notin P(k)$. The other statements are now immediate.

We will now formulate two well-known stronger versions of the classical pumping lemma and define classes of languages which are proof against pumping in this stronger sense.

Define: $\mathscr{S}(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that any subword z of a word $z_1=y_1 z y_2 \in L$ with $|z|>p(L)$ can be written as $z=uvw$ such that $0<|v| \leqq p(L)$ and $\{y_1 u v^i w y_2 \mid i \geqq k\} \subseteqq L$.

Define: $\mathscr{Q}(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that if in a word $z \in L$ more than $p(L)$ occurrences of letters are marked then z can be written as $z=uvw$ such that v contains at least one and at most $p(L)$ marked occurrences and $\{uv^i w \mid i \geqq k\} \subseteqq L$.

In the sequel it will be shown that the classes $\mathscr{P}(k)$, $\mathscr{S}(k)$ and $\mathscr{Q}(k)$ are related as pictured in the diagram below. Note that the family $\mathscr{Q}(k)$ is not identical to the "regular-Ogden-like" family since we have put no restrictions on u and w.

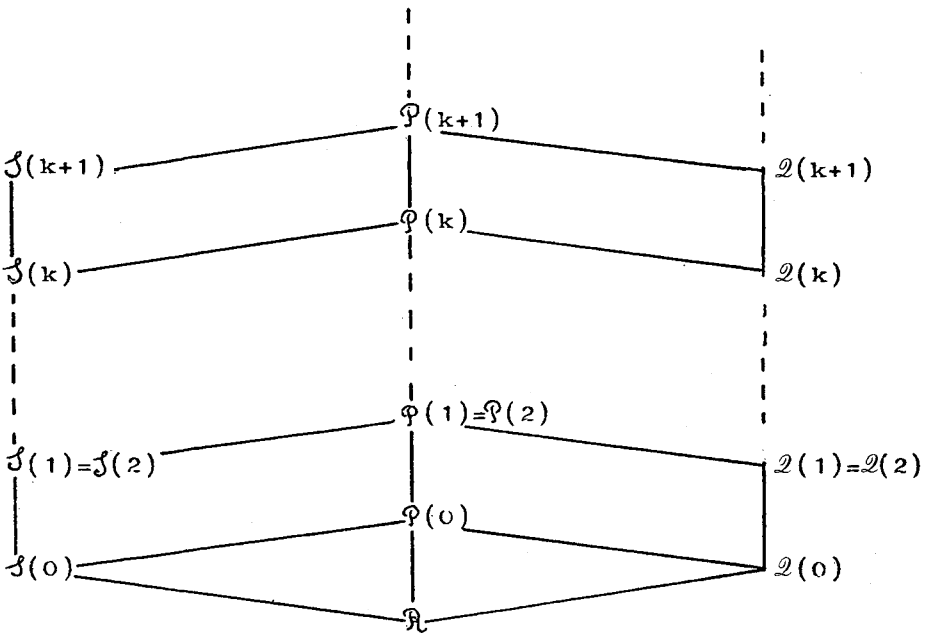

Figure 2

The inclusions shown in this diagram are immediate consequences of the definitions. The fact that they are strict will be shown by considering a set of examples.

– $\mathcal{S}(0) \not\subseteq \mathcal{S}(1)$.

Let $L_3 = \{ f^n g^n \mid n \geq 1 \}$ and let the substitution s be defined by $s(f) = \{ ab^n \mid n \geq 1 \}$ and $s(g) = \{ cb^n \mid n \geq 1 \}$. Let $L_4 = s(L_3)$. Then $L_4 \in \mathcal{S}(1) - \mathcal{S}(0)$. To see that $L_4 \in \mathcal{S}(1)$ take $p(L_4) = 2$. Any subword of length at least three of a word of L_4 contains at least one b. Repeating this b one or more times produces a word also in L_4. To see that $L_4 \notin \mathcal{S}(0)$ verify that a word of the form $(ab)^n (cb)^n$ can not be pumped. Therefore $\mathcal{S}(0)$ is strictly included in $\mathcal{S}(1)$.

– $\mathcal{Q}(0) \not\subseteq \mathcal{Q}(1)$, $\mathcal{R} \not\subseteq \mathcal{S}(0)$, $\mathcal{Q}(0) \not\subseteq \mathcal{P}(0)$ and $\mathcal{S}(0) - \mathcal{Q}(0) \neq \emptyset$.

Let

$$K = \{ w \in \{ a, b, c \}^* \mid \#(aa, w) + \#(ac, w) + \#(cc, w) \geq 1 \}$$

and let $L_5 = L_4 \cup K$. Now $L_5 \in \mathcal{Q}(1) - \mathcal{Q}(0)$. Suppose that $L_5 \in \mathcal{Q}(0)$ and that p is the required constant. Consider the word $(ab)^{p+1} (cb)^{p+1}$ and mark the first $p+1$ occurrences of the letter a. There must exist u, v, w, such that $uv^i w \in L_5$ for all $i \geq 0$. Since v contains a marked occurrence, the number of a's can not be

stable. Furthermore, v can not be chosen such that $uv^2w \in L_4$. Therefore uv^2w must belong to K. But then $v^2 \in K$ since $uv \notin K$ and $vw \notin K$. Thus both the first and the last letter of v must be a. But in this case $uw \notin L_5$. Hence $L_5 \notin \mathcal{Q}(0)$. Since obviously $L_5 \in \mathcal{Q}(1)$ we have that $\mathcal{Q}(0)$ is strictly included in $\mathcal{Q}(1)$. Since L_5 also belongs to $\mathcal{S}(0)$ and is non-regular (1), this same language shows that $\mathcal{S}(0) - \mathcal{Q}(0)$ and $\mathcal{S}(0) - \mathcal{R}$ and $\mathcal{P}(0) - \mathcal{Q}(0)$ are non-empty.

– $\mathcal{R} \not\subseteq \mathcal{Q}(0)$, $\mathcal{Q}(0) - \mathcal{S}(0) \neq \emptyset$ and $\mathcal{S}(0) \not\subseteq \mathcal{P}(0)$.

Consider the language L_6 consisting of all words over the alfabet $\{a, b, c\}$ in which the number of a's, the number of b's and the number of c's are not all equal to one another

$$L_6 = \{w \in \{a, b, c\}^* \mid \neg(\#(a, w) = \#(b, w) = \#(c, w))\}.$$

To show that $L_6 \in \mathcal{Q}(0)$ take $p(L_6) = 2$ and then verify that a word can be pumped as required by analyzing the different cases which arise considering the number of occurrences of the various letters. Since obviously L_6 is non-regular we have that $\mathcal{R}$ is strictly included in $\mathcal{Q}(0)$. Suppose now that $L_6 \in \mathcal{S}(0)$ and that p is the required constant. The subword a^{p+1} in $a^k b^j c^j$ with $k = p+1$ and $j = p+1+(p+1)!$ can not be written as uvw such that $uv^i wb^j c^j \in L_6$ for all $i \geqq 0$.

Therefore $L_6 \notin \mathcal{S}(0)$ and $\mathcal{Q}(0) - \bigcup_{k \geqq 0} \mathcal{S}(k)$ is non-empty. This same language shows that $\mathcal{S}(0)$ is strictly included in $\mathcal{P}(0)$.

The examples above show the correctness of the bottom part of the inclusion diagram given in figire 2.

Fact 7: $\mathcal{S}(k)$ is strictly included in $\mathcal{S}(k+1)$ and $\mathcal{Q}(k)$ is strictly included in $\mathcal{Q}(k+1)$ for all $k \geqq 2$.

To see this consider the languages $LSQ(k)$ which, using our Thue-sequence, are defined as follows

$$LSQ(k) = \{x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n} \mid n \geqq 1 \text{ and for all } j \leqq n,\ i_j = 1 \text{ or } i_j \geqq k\},$$

Clearly $LSQ(k) \in \mathcal{S}(k) \cap \mathcal{Q}(k)$. Just take $p(LSQ(k)) = 1$ and repeat any (marked) letter. Furthermore $LSQ(k+1) \notin \mathcal{S}(k) \cup \mathcal{Q}(k)$ whence $\mathcal{S}(k) \not\subseteq \mathcal{S}(k+1)$ and $\mathcal{Q}(k) \not\subseteq \mathcal{Q}(k+1)$. Suppose that $LSQ(k+1)$ does belong to $\mathcal{S}(k) \cup \mathcal{Q}(k)$ and let the required constant be p. The prefix $x_1 x_2 \ldots x_{p+1}$ of our Thue-sequence belongs to $LSQ(k+1)$ whence by assumption there exist u, v, $w \in V_E^*$ such that $0 < |v| \leqq p$ (or the number of marked occurrences is within these bounds) and $z_i = uv^i w \in LSQ(k+1)$ for all $i \geqq k \geqq 2$. Since $uv^k w \in LSQ(k+1)$, v can not be just a single letter. If v is not of the form $v = xv'x$ for some $x \in V_E$, the sequence $uv^i w$ must be a Thue-sequence, which is impossible since $v \neq \varepsilon$. But if

$v=xv'x$ then v' is non-empty because uvw is a Thue-sequence. But then $uv^k w \notin LSQ(i)$ for $i>2$ and thus the assumption that $LSQ(k+1) \in \mathscr{S}(k) \cup \mathscr{Q}(k)$ is falsified.

Fact 8: $\mathscr{S}(k)$ is strictly included in $\mathscr{P}(k)$, $\mathscr{Q}(k)$ is strictly included in $\mathscr{P}(k)$ and the classes $\mathscr{S}(k)$, $\mathscr{Q}(k)$ and $\mathscr{P}(k-1)$ are mutually incomparable.

Reconsider the family of languages $LP(k)$ as defined following fact 1. It should be obvious that

$$LP(k) \in \mathscr{P}(k) - \bigcup_{i \geqq 0} (\mathscr{S}(i) \cup \mathscr{Q}(i)),$$

which proves the strictness results. This family also shows that $\mathscr{P}(k-1)-\mathscr{S}(k)$ and $\mathscr{P}(k-1)-\mathscr{Q}(k)$ are non-empty. Furthermore $LSQ(k) \in \mathscr{S}(k) \cap \mathscr{Q}(k)$ whence $\mathscr{P}(k-1)$ is incomparable with $\mathscr{S}(k)$ and also incomparable with $\mathscr{Q}(k)$. We have already seen that the language L_6 belongs to $\mathscr{Q}(k)-\mathscr{S}(k)$, which set is thus non-empty. It remains to show that $\mathscr{S}(k)-\mathscr{Q}(k)$ is non-empty. To this end we define with the aid of our Thue-sequence the family of languages $LS(k)$:

$$LS(k) = \{ x_1^{i_1} x_2 x_3^{i_3} x_4 \ldots x_{2n-1}^{i_{2n-1}} x_{2n} \mid n \geqq 1 \text{ and } i_j = 1 \text{ or } i_j \geqq k \}.$$

To see that $LS(k) \in \mathscr{S}(k)$, take $p(LS(k))=2$ and note that any subword of length two or more of some word of $LS(k)$ contains at least one letter which may be repeated k or more times. The fact that $LS(k)$ does not belong to $\bigcup_{k \geqq 0} \mathscr{Q}(k)$ follows by marking only even-place letters in a prefix of our Thue-sequence. This shows that $\mathscr{S}(k)-\mathscr{Q}(k)$ is non-empty. Therefore the classes $\mathscr{S}(k)$ and $\mathscr{Q}(k)$ are incomparable.

Fact 9: Both $\mathscr{S}(k)$ and $\mathscr{Q}(k)$ are closed with respect to ε-free substitution, $\mathscr{S}(0)$ and $\mathscr{Q}(0)$ are closed with respect to arbitrary substitution.

The proof of this fact is similar to the concatenation of the proof of fact 3 and the proof of fact 5 and will be omitted.

Fact 10: Both $\mathscr{S}(k)$ and $\mathscr{Q}(k)$ are closed with respect to union, concatenation and iteration.

Fact 11: Neither $\mathscr{S}(k)$ nor $\mathscr{Q}(k)$ is closed with respect to intersecting with a regular set. Neither of these classes is closed with respect to arbitrary intersection, or with respect to complementation.

Consider the previously defined languages K and L_5. Since K is regular, so is $\overline{K} \cap \{ (ab)^n (cb)^m \mid n, m \geqq 0 \}$. Intersecting L_5 with this set produces

$\{(ab)^n(cb)^n \mid n \geqq 1\}$ which does not belong to $\bigcup_{k \geqq 0} \mathscr{S}(k)$. As to $\mathscr{Q}(k)$, consider the previously defined language L_6. Intersecting L_6 with the regular set denoted by $a^*b^*c^*$ produces $\{a^h b^i c^i \mid \neg(h=i=j)\}$, which does not belong to $\bigcup_{k \geqq 0} \mathscr{Q}(k)$. The other statements are now immediate.

The pumping possibilities of regular languages have an aspect which has not been brought to bear upon. This is the fact that the way in which a word z is split in parts u, v, and w such that $uv^i w$ belongs to the language should be locally determined. That is to say that some other word zz' of the language should allow the same splitting of z. This observation leads us to the following:

Define: $\mathscr{T}(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that any word z of length greater than $p(L)$ can be written as $z=uvw$ where $0<|v| \leqq p(L)$ such that for all z' for which $zz' \in L$ we have $\{uv^i wz' \mid i \geqq k\} \subseteq L$.

The above definition gives rise to an infinite hierarchy of classes of languages. These classes prove to be incomparable to the previously defined $\mathscr{S}$- and $\mathscr{Q}$-classes.

Fact 12: $\mathscr{R} \not\subseteq \mathscr{T}(0) \not\subseteq \mathscr{T}(1)$ and $\mathscr{T}(1)=\mathscr{T}(2)$.

The fact that $\mathscr{T}(1)=\mathscr{T}(2)$ is immediate from the definition. Let $L_7=\{a^n b^p \mid n \geqq 1$ and p is a prime$\}$ and let $L_8=L_7 \cup \{b^n \mid n \geqq 1\}$. Then $L_8 \in \mathscr{T}(0)-\mathscr{R}$. It is clear that L_8 is non-regular. To see that L_8 does belong to $\mathscr{T}(0)$ take $p(L_8)=1$ and always repeat the first letter zero or more times. It is also clear that $L_7 \in \mathscr{T}(1)-\mathscr{T}(0)$, whence $\mathscr{R} \not\subseteq \mathscr{T}(0) \not\subseteq \mathscr{T}(1)$.

Fact 13: $\mathscr{T}(k)$ is strictly included in $\mathscr{T}(k+1)$ for all $k \geqq 2$.

Consider the family of languages $LT(k)$ which, using our Thue-sequence, are defined as follows

$$LT(k)=\{x_1^j x_2 x_3 \ldots x_i \mid i \geqq 1 \text{ and } j=1 \text{ or } j \geqq k\}.$$

Then $LT(k+1) \in \mathscr{T}(k+1)-\mathscr{T}(k)$. This is shown by an argument analogous to the one given for the $\mathscr{P}$-classes.

Fact 14: $\mathscr{T}(k)$ is incomparable with $\mathscr{S}(k)$ and with $\mathscr{Q}(k)$.

The previously defined language L_6 belongs to $\mathscr{Q}(0)-\bigcup_{k \geqq 0} \mathscr{T}(k)$ and the also earlier defined language L_8 belongs to $\mathscr{T}(0)-\bigcup_{i \geqq 0} \mathscr{Q}(i)$, so the $\mathscr{Q}$- and the $\mathscr{T}$-classes are incomparable. Since L_8 does not belong to any $\mathscr{S}(k)$, it remains to be

shown that $\mathscr{S}(k)-\mathscr{T}(k)\neq\emptyset$. To this end consider the languages $L1ST(k)$ and $L2ST(k)$ which, using our Thue-sequence, are defined as follows

$$L1ST(k)=\{x_1^{i_1}x_2^2x_3^{i_3}x_4^2\ldots x_{2j}^2x_{2j+1}^{i_{2j+1}} \mid j\geqq 1 \text{ and for all } n\leqq j,\ i_{2n+1}=2 \text{ or } i_{2n+1}\geqq 2k\},$$

$$L2ST(k)=\{x_1^2x_2^{i_2}x_3^2x_4^{i_4}\ldots x_{2j-1}^2x_{2j}^{i_{2j}} \mid j\geqq 1 \text{ and for all } n\leqq j,\ i_{2n}=2 \text{ or } i_{2n}\geqq 2k\}.$$

$L1ST(k)$, $L2ST(k)\in\mathscr{T}(k)\cap\mathscr{S}(k)$ for all $k\geqq 1$. To see this, take $p(L1ST(k))=p(L2ST(k))=4$. Every subword of length four or more contains a group of two letters corresponding to odd places, which may therefore be repeated one or more times (respectively corresponding to even places). The language $L1ST(k)\cup L2ST(k)$ however, belongs to $\mathscr{S}(k)-\mathscr{T}(k)$.

This leaves the case $k=0$ which is settled by considering $L1ST(1)\cup L2ST(1)\cup L_9$ where

$$L_9=\bigcup_{x\in V_E}V_E^*(V_E-x)x(V_E-x)V_E^*.$$

This language clearly belongs to $\mathscr{S}(0)-\mathscr{T}(0)$ since every subword of length four or more contains at least one letter which may be repeated zero or more times; which letter this is however, depends on the right context. Therefore the $\mathscr{T}-$, the $\mathscr{Q}-$ and the $\mathscr{S}$-classes are mutually incomparable.

Fact 15: None of the classes $\mathscr{T}(k)$ is closed with respect to union, intersection, intersection with a regular set or complement.

We have seen that $L1ST(k)$ and $L2ST(k)$ belong to $\mathscr{T}(k)$ but that $L1ST(k)\cup L2ST(k)$ does not. Let L_{10} be the language consisting of all words over the alfabet V_E of the form $aabbcc\ldots ff\ldots$ which do not contain a substring ggg for some $g\in V_E$. Since L_{10} is a regular set and $L1ST(k)\cap L_{10}\notin\bigcup_{i\geqq 0}\mathscr{P}(i)$, $\mathscr{T}(k)$ is not closed with respect to intersection with a regular set and thus also not closed with respect to arbitrary intersection. The case $k=0$ is dealt with by adding the regular set L_9 where necessary in the above argument. As to the complement, let V be an alphabet with at least three letters and let

$$L_{11}=\{w_1vvw_2 \mid w_1,\ w_2\in V^* \text{ and } v\in V^+\}.$$

Then obviously

$$L_{11}\in\mathscr{T}(0)\cap\mathscr{S}(0)\cap\mathscr{Q}(0)-\mathscr{R} \qquad \text{and} \qquad \overline{L}_{11}\notin\bigcup_{k\geqq 0}\mathscr{P}(k).$$

It is possible to divide the $\mathscr{T}$-classes in $\mathscr{T}_S$- and $\mathscr{T}_Q$-subclasses:

Define: $\mathscr{T}_S(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that any subword z of some word $z_1 = y_1 z y_2$ with $|z| > p(L)$ can be written as $z = uvw$ such that $0 < |v| \leqq p(L)$ and for all z' such that $z_1 z' \in L : \{ y_1 u v^i w y_2 z' \mid i \geqq k \} \not\subseteq L$.

Define: $\mathscr{T}_Q(k)$ as to consist of all and only those languages L for which there exist constants $p(L)$ such that if in some word z more than $p(L)$ occurrences of letters are marked then z can be written as $z = uvw$ such that v contains at least one and at most $p(L)$ marked occurrences and for all z' such that $zz' \in L$: $\{ uv^i w z' \mid i \geqq k \} \subseteq L$.

These definitions result in infinite hierarchies since

$$LSQ(k) \in (\mathscr{T}_S(k) \cap \mathscr{T}_Q(k)) - (\mathscr{T}_S(k-1) \cup \mathscr{T}_Q(k-1)).$$

The $\mathscr{T}_S$- and the $\mathscr{T}_Q$-classes are strictly included in the $\mathscr{T}$-classes since

$$L_8 \in \mathscr{T}(k) - \bigcup_{i \geqq 0} (\mathscr{T}_S(i) \cup \mathscr{T}_Q(i)).$$

An other example of this is provided by $LT(k)$.

The $\mathscr{T}_S$-classes are strictly included in the S-classes since

$$L1ST(k) \cup L2ST(k) \in \mathscr{S}(k) - \mathscr{T}_S(k).$$

The $\mathscr{T}_Q$-classes are strictly included in the Q-classes since

$$L_6 \in \mathscr{Q}(0) - \bigcup_{k \geqq 0} \mathscr{T}_Q(k).$$

The $\mathscr{T}_S$- and $\mathscr{T}_Q$-classes are mutually incomparable. The languages $L1ST(k)$ show that $\mathscr{T}_S(k) - \mathscr{T}_Q(k)$ is non-empty. Consider the language $L_{12} = \{a, b\}^* . L_7 \cup \{b\}^*$. This L_{12} belongs to $\mathscr{T}_Q(0)$. Just take $p(L_{12}) = 1$ and always repeat the smallest prefix which ends in a marked letter.

It is clear that $L_{12} \notin \bigcup_{k \geqq 0} \mathscr{S}(k)$ and thus does not belong to any $\mathscr{T}_S(k)$. This shows that $\mathscr{T}_Q(k) - \mathscr{T}_S(k)$ is non-empty too, whence the $\mathscr{T}_S$- and the $\mathscr{T}_Q$-classes are incomparable.

The above definitions of the classes $\mathscr{T}_S(k)$ and $\mathscr{T}_Q(k)$ do not add much to the already available tools. An other possibility to sharpen the definition of pumping is to require that the pumping of an initial subword of a word of the language will produce only words also belonging to the language and that pumping of an initial subword of a word not belonging to the language will produce only words also not belonging to the language. More formally:

The class $\mathcal{U}$ consists of all and only those languages L for which there exist constants $p(L)$ such that any word z of length larger than $p(L)$ can be written as $z=uvw$ where $0<|v|\leqq p(L)$ such that for all $z' : zz' \in L$ if and only if $\{uv^i wz' \mid i \geqq 0\} \subseteq L$.

It is easy to prove that the Nerode equivalence of a language belonging to $\mathcal{U}$ has finite index. It is then obvious that $\mathcal{U}=\mathcal{R}$. The fact that the pumping property used above to define the class U characterizes the regular sets has independently been observed by Jaffe [3]. For the purpose of characterizing the regular sets the requirements in this definition are a little too strict in the sense that it is not necessary to require that $\{uv^i wz' \mid i \geqq 0\} \subseteq L$ but that $uwz' \in L$ suffices.

Fact 16: A language L is regular if and only if there exists a constant p such that any word z of length larger than p can be written as $z=uvw$ where $0<|v|\leqq p$ such that for all words z' we have $uvwz' \in L$ iff $uwz' \in L$.

Fact 16 should be considered a useful tool because it will not let you down as the classical pumping lemma in its various disguises sometimes does. Furthermore it is at least as easy to use as the pumping lemma and more importantly the non-regularity proofs obtained with it are very close to intuition.

There are lots of questions left open in this note. To mention a few of them:

– do there exist languages L such that both L and $\overline{L}$ belong to $\mathcal{S}(k)-\mathcal{R}$, $\mathcal{Q}(k)-\mathcal{R}$, $\mathcal{T}(k)-\mathcal{R}$;

– find further closure properties of the various classes of languages defined above;

– find a convincing example to show that fact 16 can be used non-trivially to show the regularity of some set.

REFERENCES

Apart from [4] the following are to be considered as general references. These papers are also concerned with pumping although not mainly with regular pumping.

The idea of using Thue-sequences belongs to [4].

1. L. BOASSON, *Un critère de rationnalité des langages algébriques. In* M. NIVAT, Ed., Automata, Languages and Programming, 1972, pp. 359-365.
2. S. HORVÀTH, *The Family of Languages Satisfying Bar-Hillel's Lemma*, R.A.I.R.O., Informatique Theorique, Vol. 12, No. 3, 1978, pp. 193-199.
3. J. JAFFE, *A Necessary and Sufficient Pumping Lemma for Regular Languages*, S.I.G.A.C.T. News summer, 1978, pp. 48-49.
4. J. VAN LEEUWEN, Private communication, 1978.