Annales Univ. Sci. Budapest., Sect. Comp. 26 (2006) 171-179

SOME UNCOUNTABLE HIERARCHIES OF FORMAL LANGUAGES

G. Lischke (Jena, Germany)

Abstract. We consider several types of similarity relationships between languages, and for an arbitrary language class we define appropriate similarity classes. We show that for any of the classes of regular, linear, context-free, context-sensitive, recursive and recursively enumerable languages the appropriate similarity classes form an uncountable hierarchy of order type $\omega + \lambda + 2$. We also show that there exist linear languages which are not δ-similar to any regular language for any $\delta < \frac{1}{2}$ and we discuss this problem for $\delta \geq \frac{1}{2}$.

1. Introduction and main result

Starting with the concept of partial words, which was introduced by Berstel and Boasson [1] and was motivated by molecular biology of nucleic acids in [7], we introduced punctured languages and studied their restorations. We saw that the restoration classes of language classes coincide with similarity classes corresponding to several similarity types. Ignoring the relationship with restoration classes and their motivation here we restrict ourselves to similarity classes and their hierarchies. If $\mathcal{L}$ is a class of languages over some fixed nontrivial alphabet X, k is a natural number, and δ is a real number between 0 and 1, we define the similarity classes $\mathcal{L}_k$ and $\mathcal{L}_\delta$. Further we have the class $\mathcal{L}_{length}$. Let now $\mathcal{L}$ be one of the following classes: the class REG of all regular languages, the class LIN of all linear languages, the class CF of all context-free languages, the class CS of all context-sensitive languages, the class REC of all

This work was partly supported by the Fellowship Program of the Japan Society for Promotion of Science (JSPS) under grant S06717 and by a collaboration between Friedrich Schiller University and Eötvös Loránd University.

recursive or decidable languages, or the class RE of all recursively enumerable languages (all over the alphabet X). Then for the natural numbers k and k' and for the real numbers δ and δ', such that $0 < k < k'$ and $0 < \delta < \delta' \leq \frac{1}{2}$,

$$\mathcal{L} = \mathcal{L}_0 \subset \mathcal{L}_k \subset \mathcal{L}_{k'} \subset \mathcal{L}_\delta \subset \mathcal{L}_{\delta'} \subset \mathcal{L}_{length}$$

holds. Thus each of the language classes REG, LIN, CF, CS, REC and RE creates a hierarchy of order type $\omega + \lambda + 2$. Because in each case $\mathcal{L}_{length}$ is strictly contained in the class $\mathcal{P}(X^*)$ of all languages, we actually have the order type $\omega + \lambda + 3$.

We shall give in Section 2 our basic definitions and explain what we mean by *ind*-similarity for some index *ind*. In Section 3 we prove some lemmata from which we can conclude our main result, it is given in Section 4. Some further related results and problems are discussed in Section 5.

2. Preliminaries and definitions

Even though the following is standard in the literature (see, e.g. the textbooks [4, 5]) we briefly recall the most important notions. For the whole of our paper let X be a fixed finite nonempty alphabet. Furthermore, we assume that X is a nontrivial alphabet, which means that it has at least two symbols (in the other case all of our results become trivial or meaningless).

$\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ denotes the set of all natural numbers. X^* is the free monoid generated by X or the set of all words over X. The empty word we denote by e, and $X^+ =_{Df} X^* \setminus \{e\}$. A (formal) language (over X) is a subset L of X^*, $L \subseteq X^*$. The symbol $\subset$ between sets denotes strict inclusion. $\mathcal{P}(M)$ is the set of all subsets of a set M, and $|M|$ denotes the cardinality of M.

For a word $w \in X^*$, $|w|$ denotes the length of w, and for $1 \leq i \leq |w|$, $w[i]$ is the letter at the i-th position of w. For $x \in X$, $|w|_x =_{Df} |\{i : w[i] = x\}|$ is the number of occurences of the letter x in the word w.

For $k \in \mathbb{N}$, w^k denotes the concatenation of k copies of the word w. w^* denotes the set $\{w^k : k \in \mathbb{N}\}$, and w^*q the set $\{w^kq : k \in \mathbb{N}\}$.

Let $\mathcal{L}$ be a class of languages (over our fixed alphabet X). For some index *ind* we define $\mathcal{L}_{ind} =_{Df} \{L : \exists L'(L' \in \mathcal{L} \wedge \mathcal{L} \underset{ind}{\sim} \mathcal{L}')\}$.

The *index-similarity* $\underset{ind}{\sim}$ for languages is defined in the following way based on the index-similarity between words

$$L \underset{ind}{\sim} L' =_{Df} \forall u \exists v(u \in L \rightarrow v \in L' \wedge u \underset{ind}{\sim} v) \wedge \forall v \exists u(v \in L' \rightarrow u \in L \wedge u \underset{ind}{\sim} v).$$

The index *ind* may be a natural number k, a real number δ between 0 and 1, or the word *length*. Finally, $\underset{ind}{\sim}$ between words is defined based on the Hamming distance h known from coding theory [3]: for two words u and v of the same length let

$$h(u,v) =_{Df} |\{i : 1 \leq i \leq |u| \wedge u[i] \neq v[i]\}|.$$

For $k \in \mathbb{N}$, two words u and v are called *k-similar*, denoted by $u \underset{k}{\sim} v$, if $|u| = |v|$ and $h(u,v) \leq k$. For a nonnegative real number $\delta < 1$, words u and v are called *δ-similar*, denoted by $u \underset{\delta}{\sim} v$, if $|u| = |v|$ and $h(u,v) \leq \delta \cdot |u|$. u and v are called *length-similar* or *length-equivalent*, denoted by $u \underset{length}{\sim} v$, if $|u| = |v|$.

Finally we repeat, that by REG, LIN, CF, CS, REC and RE we will denote the classes of all regular, linear, context-free, context-sensitive, recursive and recursively enumerable languages (over X), respectively.

3. Some lemmata

Lemma 1. *For an arbitrary language class $\mathcal{L}$ and for natural numbers k and k' such that $0 \leq k \leq k'$, $\mathcal{L} = \mathcal{L}_0 \subseteq \mathcal{L}_k \subseteq \mathcal{L}_{k'} \subseteq \mathcal{L}_{length}$ holds.*

The proof is trivial by the definitions.

Lemma 2. *For an arbitrary language class $\mathcal{L}$ and for real numbers δ and δ' such that $0 \leq \delta \leq \delta' < 1$, $\mathcal{L} = \mathcal{L}_0 \subseteq \mathcal{L}_\delta \subseteq \mathcal{L}_{\delta'} \subseteq \mathcal{L}_{length}$ holds.*

The proof is trivial by the definitions.

Lemma 3. *For $ind \in \{k, \delta, length\}$, where k is a natural number and δ is a real number between 0 and 1, $REG_{ind} \subseteq LIN_{ind} \subseteq CF_{ind} \subset CS_{ind} \subset \subset REC_{ind} \subset RE_{ind}$ holds.*

This trivially follows from the corresponding relationships between REG, LIN, CF, CS, REC and RE (see, e.g. [4, 5]). For $ind = length$, $REG_{ind} = LIN_{ind} = CF_{ind}$ holds since all context-free languages over a one-letter alphabet are regular. For $ind \in \{k, \delta\}$ and $0 \leq \delta < \frac{1}{2}$, $REG_{ind} \subset LIN_{ind}$, because of Lemma 4 and Theorem 2 below. For $\delta \geq \frac{1}{2}$ see Section 5.

Lemma 4. *If $\mathcal{L}$ is a language class which is closed under union with finite sets and under difference with finite sets, then for any fixed δ, $0 < \delta < 1$: $\bigcup_{k=0}^{\infty} \mathcal{L}_k \subseteq \mathcal{L}_\delta$.*

Proof. Let $L \in \mathcal{L}_k$ for some fixed $k \in \mathbb{N}$, $n_0 \in \mathbb{N}$ such that $n_0 \geq \frac{k}{\delta}$, and define $L_1 =_{Df} \{p : p \in L \wedge |p| \leq n_0\}$, $L_2 =_{Df} L \setminus L_1$. Then $L \underset{k}{\sim} L'$ for some $L' \in \mathcal{L}$ and $L_2 \underset{k}{\sim} L_2'$ for $L_2' = L' \setminus \{p : |p| \leq n_0\} \in \mathcal{L}$. For each $w \in L_2$ there exists $w' \in L_2'$ (and vice versa) such that $w \underset{k}{\sim} w'$ and therefore $\frac{h(w,w')}{|w|} < \frac{k}{n_0} \leq \delta$. This means $L_2 \underset{\delta}{\sim} L_2'$ and, therefore, $L = L_1 \cup L_2 \underset{\delta}{\sim} L_1 \cup L_2'$ and thus $L \in \mathcal{L}_\delta$ because $L_1 \cup L_2' \in \mathcal{L}$.

Lemma 5. *Let k and k' be natural numbers such that $0 \leq k < k'$. Then there exists $L \in REG_{k'} \setminus RE_k$.*

Proof. Let T be an undecidable subset of a^* and define the following set: $L =_{Df} \{pa^{2k'} : p \in T\} \cup \{pb^{2k'} : p \in a^* \setminus T\}$. Then $L \underset{k'}{\sim} a^*a^{k'}b^{k'}$ and therefore $L \in REG_{k'}$ because of $a^*a^{k'}b^{k'} \in REG$. Assume $L \in RE_k$. Then $L \underset{k}{\sim} S$ for some $S \in RE$, and each $w \in S$ must be k-similar to some word from L and, therefore, it has the form pu, where $|u| = 2k'$ and either $|u|_a > k'$ (if $a^{|p|} \in T$) or $|u|_a < k'$ (if $a^{|p|} \notin T$). Also, for each $n \in \mathbb{N}$ there exists such a word $pu \in S$ of length $n + 2k'$. Then the enumerability of S implies the decidability of T because $T = \{a^{|p|} : \exists u(|u| = 2k' \wedge pu \in S \wedge |u|_a > k')\}$, contradicting the assumption.

It follows from the Lemmas 1, 3 and 5, that for each of the classes REG, LIN, CF, CS, REC, and RE, the k-similarity classes form a countable hierarchy.

Corollary 1. *For $\mathcal{L} \in \{REG, LIN, CF, CS, REC, RE\}$*

$$\mathcal{L} = \mathcal{L}_0 \subset \mathcal{L}_1 \subset \mathcal{L}_2 \subset \cdots \subset \mathcal{L}_n \subset \mathcal{L}_{n+1} \subset \cdots$$

holds.

By Lemma 4 the whole hierarchy is contained in $\mathcal{L}_\delta$ for arbitrary real δ, $0 < \delta < 1$, because each of our classes $\mathcal{L}$ is closed under union with finite sets and under difference with finite sets.

Lemma 6. *Let δ and δ' be real numbers such that $0 \leq \delta < \delta' < 1$ and $\delta < \frac{1}{2}$. Then there exists $L \in REG_{\delta'} \setminus RE_\delta$.*

Proof. Because of Lemma 2 it is enough to assume that $\delta' \leq \frac{1}{2}$ is a rational number, and therefore let $\delta' = \frac{r}{s}$ for natural numbers r, $s \neq 0$, where $r \leq s - r$. Let T be the same set as in the proof of Lemma 5 and define

$$L =_{Df} \{(a^{s-r}b^r)^n : a^n \in T\} \cup \{(b^ra^{s-r})^n : a^n \notin T\}.$$

Then $L \underset{\delta'}{\sim} \{a^{s \cdot n} : n \in \mathbb{N}\}$ and therefore $L \in REG_{\delta'}$. Assume $L \underset{\delta}{\sim} S$ for some $S \in RE$. Then for each $n \in \mathbb{N}$ there is a word of length ns in S, and for each $w \in S$ with length ns, either $h(w, (a^{s-r}b^r)^n) < rn$ (if $a^n \in T$) or $h(w, (b^r a^{s-r})^n) < rn$ (if $a^n \notin T$). Both at a time are impossible because of $h((a^{s-r}b^r)^n, (b^r a^{s-r})^n) = 2rn$. Then the set T would be decidable because

$$T = \{a^n : \exists w(w \in S \wedge |w| = ns \wedge h(w, (a^{s-r}b^r)^n) < rn)\}.$$

This contradicts the assumption.

It follows from the Lemmas 2, 3 and 6 that for each of the classes REG, LIN, CF, CS, REC and RE the δ-similarity classes form an uncountable hierarchy.

Corollary 2. *For $\mathcal{L} \in \{REG, LIN, CF, CS, REC, RE\}$ and real numbers δ, δ' such that $0 \leq \delta < \delta' < 1$ and $\delta < \frac{1}{2}$, $\mathcal{L}_\delta \subset \mathcal{L}_{\delta'}$ holds.*

Lemma 7. *Let δ be a real number such that $0 \leq \delta < 1$. Then there exists $L \in REG_{length} \setminus RE_\delta$.*

Proof. Let $L_1, L_2, L_3, \ldots$ be an enumeration of all recursively enumerable languages over X such that for each $n \geq 1$ there is a word of length n in L_n. For each $n \geq 1$ let w_n be a fixed word of length n in L_n (for instance, the lexicographically first word of this length in L_n). Let $\overline{w_n}$ be the word arising from w_n by changing each letter in w_n, and define $L =_{Df} \{\overline{w_n} : n \geq 1\}$. Then $L \underset{length}{\sim} X^+$ and therefore $L \in REG_{length}$. Assume $L \underset{\delta}{\sim} S$ for some $S \in RE$. Then $S = L_i$ for some suitable i, and $w_i \in S$, but there does not exist any $u \in L$ with $w_i \underset{\delta}{\sim} u$ (only $\overline{w_i} \in L$ has the same length as w_i, $|w_i| = |\overline{w_i}| = i$, but $h(w_i, \overline{w_i}) = i > \delta \cdot |w_i|$), contradicting $L \underset{\delta}{\sim} S$.

It follows from Lemmas 2, 3 and 7:

Corollary 3. *For $\mathcal{L} \in \{REG, LIN, CF, CS, REC, RE\}$ and a real number δ such that $0 \leq \delta < 1$, $\mathcal{L}_\delta \subset \mathcal{L}_{length}$ holds.*

4. The hierarchies

In set theory (see, e.g. [6]) an *order-type* is an equivalence class of isomorphic ordered sets. The most familiar ordered infinite sets in everyday life are the set $\mathbb{N}$ of natural numbers and the set $\mathbb{R}$ of real numbers with the usual order $\leq$. It is common to denote the order type of $[\mathbb{N}, \leq]$ by ω and the order type of $[\mathbb{R}, \leq]$ by λ.

Further, let $(0, \frac{1}{2}) =_{Df} \{\delta : \delta \in \mathbb{R} \wedge 0 < \delta < \frac{1}{2}\}$ and $(0, \frac{1}{2}] =_{Df} \{\delta : \delta \in \in \mathbb{R} \wedge 0 < \delta \leq \frac{1}{2}\}$. $[(0, \frac{1}{2}), \leq]$ is isomorphic to $[\mathbb{R}, \leq]$ and has the order type λ. $[(0, \frac{1}{2}], \leq]$ has the order type $\lambda + 1$. By Corollary 2, $[\{\mathcal{L}_\delta : \delta \in (0, \frac{1}{2}]\}, \subseteq]$ is isomorphic to $[(0, \frac{1}{2}], \leq]$ and therefore has also the order type $\lambda + 1$, where $\mathcal{L} \in \{REG, LIN, CF, CS, REC, RE\}$. By combining Corollary 3, Corollary 1 and Lemma 4 we get our main result.

Theorem 1. *For $\mathcal{L} \in \{REG, LIN, CF, CS, REC, RE\}$, natural numbers k, k', and real numbers δ, δ' such that $0 < k < k'$ and $0 < \delta < \delta' \leq \frac{1}{2}$,*

$$\mathcal{L} = \mathcal{L}_0 \subset \mathcal{L}_k \subset \mathcal{L}_{k'} \subset \mathcal{L}_\delta \subset \mathcal{L}_{\delta'} \subset \mathcal{L}_{length} \quad \textit{holds.}$$

These are hierarchies of order type $\omega + \lambda + 2$.

Taking into consideration the fact that in each case $\mathcal{L}_{length} \subset \mathcal{P}(X^*)$ holds, we have the order type $\omega + \lambda + 3$ (for instance, if T is a nonenumerable subset of a^* then we have $T \in \mathcal{P}(X^*) \setminus RE_{length}$). Let us further remark that all classes but $\mathcal{L}_0$ in these hierarchies have the cardinality of the continuum, and that there are continuum-many languages witnessing each of the strict inclusions. This follows from our proofs because there are continuum-many nonenumerable subsets T of a^*. In the proof of Lemma 7 we can start with such an enumeration in which for infinitely many $n \geq 1$, L_n contains at least two words of length n, and thus we can create continuum-many appropriate sets L.

5. Further results and problems

Our main result, Theorem 1, as well as Corollary 2 and Lemma 6 hold for $\delta < \frac{1}{2}$. We do not know whether they are true for $\delta \geq \frac{1}{2}$. We conjecture that this is not the case.

Conjecture 1. *$REG_\delta \subseteq RE_{\frac{1}{2}}$ holds for $\delta > \frac{1}{2}$.*

The following result describes a similar situation.

Theorem 2. $LIN \not\subseteq \bigcup\limits_{0 \leq \delta < \frac{1}{2}} REG_\delta$.

Proof. We consider $L' =_{Df} \{a^n b^n : n \in \mathbb{N}\} \in LIN$ and assume $L' \underset{\delta}{\sim} L$ for some fixed δ with $0 \leq \delta < \frac{1}{2}$ and for some $L \in REG$. By the Pumping Lemma for regular sets, every sufficiently long $w \in L$ has the form $w = w_1 w_2 w_3$, where $w_2 \neq e$ and $w_1 w_2^i w_3 \in L$ for each $i \in \mathbb{N}$. Let us define $z_i =_{Df} w_1 w_2^i w_3$. Since $L' \underset{length}{\sim} L$, w_2 has an even length $l \geq 2$ and $z_i \underset{\delta}{\sim} z_i'$ for a uniquely determined

n_i and $z'_i = a^{n_i}b^{n_i} \in L'$, where n_i is growing for growing i. Choose i so that $n_i > |w_1|$ and $n_i > |w_3|$. This means the centre of the word z_i is within w_2^i. Then for each $j \in \mathbb{N}$, the centre of z_{i+2j} is by $j \cdot l = j \cdot |w_2|$ positions to the right from the centre of z_i. The word left from the centre of z_{i+2j} must be similar to a^{n_i+jl}, and the word right from this centre must be similar to b^{n_i+jl}. Therefore,

$$h(z_{i+2j}, z'_{i+2j}) = h(z_i, z'_i) + j \cdot |w_2|_b + j \cdot |w_2|_a = h(z_i, z'_i) + j \cdot |w_2| = h(z_i, z'_i) + jl.$$

We have $|z_{i+2j}| = 2n_i + 2jl$, and therefore

$$\lim_{j \to \infty} \frac{h(z_{i+2j}, z'_{i+2j})}{|z_{i+2j}|} = \frac{1}{2} > \delta.$$

This contradicts $L' \underset{\delta}{\sim} L$.

We do not know whether we can extend the boundary for δ in Theorem 2. Especially, we do not know whether $LIN \not\subseteq REG_\delta$ is true for $\delta \geq \frac{1}{2}$. Again, we conjecture that this is not the case.

Conjecture 2. *$LIN \subseteq REG_\delta$ for $\frac{1}{2} \leq \delta < 1$.*

To prove this conjecture for the alphabet $X = \{a, b\}$, it would be sufficient to show that for an arbitrary linear language L the following language L' is regular, because $L \underset{\frac{1}{2}}{\sim} L'$:

$$L' =_{Df} \{a^{|w|} : w \in L \wedge |w|_a \geq |w|_b\} \cup \{b^{|w|} : w \in L \wedge |w|_b \geq |w|_a\}.$$

By the forthcoming Theorem 3 this would be fulfilled if both sets

$$L_a =_{Df} \{w : w \in L \wedge |w|_a \geq |w|_b\} \quad \text{and} \quad L_b =_{Df} \{w : w \in L \wedge |w|_b \geq |w|_a\}$$

were context-free. But, in general, the latter is not true as it can be seen from the following example.

The set $L =_{Df} \{a^n b^{2n} a^k : n, k \in \mathbb{N} \wedge n \geq 1\}$ is linear, but $L_a = \{a^n b^{2n} a^k : 1 \leq n \leq k\}$ is not context-free. Nevertheless, $L \underset{\frac{1}{2}}{\sim} \{a^n : n \geq 4\} \cup \{b^n : n \geq 3\}$ and therefore $L \in REG_{\frac{1}{2}}$.

For an arbitrary alphabet X, $L \subseteq X^*$ and $x \in X$ we define

$$L_x =_{Df} \{w : w \in L \wedge \forall y(y \in X \to |w|_x \geq |w|_y)\}.$$

We call the sets L_x the *majority-sets of L*.

Theorem 3. *If $L \subseteq X^*$ is a language such that all majority-sets of L are context-free, then $L \in REG_{1-\frac{1}{k}}$, where k is the cardinality of X.*

Proof. Let $L'_x =_{Df} \{x^{|w|} : w \in L_x\}$ for $x \in X$. It is a well-known fact, that every context-free language over a one-letter alphabet is regular (see, e.g. [4, 5]), and therefore, if all majority-sets of L are context-free, then L'_x is regular for each $x \in X$. Then the set $L' =_{Df} \bigcup\limits_{x \in X} L'_x$ is regular, too. Let $k = |X|$. Then $L_x \underset{1-\frac{1}{k}}{\sim} L'_x$ for each $x \in X$ and therefore $L = \bigcup\limits_{x \in X} L_x \in REG_{1-\frac{1}{k}}$.

The same result $L \in REG_{1-\frac{1}{k}}$ is true if L is a slender linear language [7], which means that there is only a constantly bounded number of words of each fixed length in L. For a two-letter alphabet X it remains to show that every linear language $L \subseteq X^*$ is $\frac{1}{2}$-similar to a regular language, where not both majority-sets of L are context-free and L is non-slender. We assume that this problem has a similar difficulty as the long-standing open problem whether the set of all primitive words over some nontrivial alphabet X is context-free or not [2].

Acknowledgment. I am very grateful to Sándor Horváth (Budapest) for our cooperation and for making corrections of my English and Péter Burcsi (Budapest) for some comments.

References

[1] **Berstel J. and Boasson L.,** Partial words and a theorem of Fine and Wilf, *Theoretical Computer Science,* **218** (1999), 135-141.

[2] **Dömösi P., Horváth S. and Ito M.,** On the connection between formal languages and primitive words, *Proc. First Session on Scientific Communication, Oradea, Romania, June 1991,* Univ. of Oradea, 1991, 59-67.

[3] **Hamming R.W.,** Error detecting and error correcting codes, *Bell System Techn. Journ.,* **29** (1950), 147-160.

[4] **Harrison M.A.,** *Introduction to formal language theory,* Addison-Wesley, Reading (Mass.), 1978.

[5] **Hopcroft J.E. and Ullman J.D.,** *Introduction to automata theory, languages and computation,* Addison-Wesley, Reading (Mass.), 1979.

[6] **Just W. and Weese M.,** *Discovering modern set theory I.,* American Mathematical Society, Providence, 1996.

[7] **Lischke G.,** Restorations of punctured languages and similarity of languages, *Math. Log. Quart.,* **52** (2006), 20-28.

(Received October 3, 2006)

G. Lischke
Faculty of Mathematics and Informatics
Friedrich Schiller University
Ernst-Abbe-Platz 1-4
D-07743 Jena, Germany
lischke@minet.uni-jena.de