Annales Academiæ Scientiarum Fennicæ
Series A. I. Mathematica
Volumen 11, 1986, 167—178

DENSENESS, MAXIMALITY, AND DECIDABILITY OF GRAMMATICAL FAMILIES[1]

H. A. MAURER, A. SALOMAA, E. WELZL and D. WOOD

We demonstrate that there is no sub-regular maximally dense interval of grammatical families by way of two characterizations of sub-regular dense intervals. Moreover we prove that it is decidable whether or not a given sub-regular interval is dense. These results are proved using the twin notions of language forms and linguistical families that are of interest in their own right.

1. Introduction and overview

The study of grammatical similarity via the tool of grammar forms now forms a substantial chapter in the development of formal language theory. Not only has grammar form theory contributed to our understanding of similarity, but it has also raised many challenging and interesting problems. It is the purpose of this paper to present the solution to one of these problems. The problem we tackle is found when trying to refine some basic hierarchy results for language families. To explore this further we need to first introduce grammar forms and their related language families. A (context-free) *grammar form* is simply a context-free grammar $G=(V, \Sigma, P, S)$, where, as usual, V is a finite alphabet, $\Sigma\subseteq V$ is a terminal alphabet and $V-\Sigma$ is the nonterminal alphabet, $P\subseteq(V-\Sigma)\times V^*$ is a finite set of productions, where a production (A, α) is usually written as $A\to\alpha$, and S in $V-\Sigma$ is a sentence symbol. We use $L(G)$ to denote the language generated by G, as usual.

Given two grammars $G'=(V', \Sigma', P', S')$ and $G=(V, \Sigma, P, S)$ we say G' is an *interpretation of* G, denoted by $G'\leqq G$ if there is a (strict alphabetic) morphism $h\colon V'\to V$ such that $h(V'-\Sigma')\subseteq V-\Sigma$, $h(\Sigma')\subseteq\Sigma$, $h(P')\subseteq P$, $h(S')=S$, where $h(P')=\{h(A)\to h(\alpha)\colon A\to\alpha \text{ is in } P'\}$. A morphism is strict-alphabetic if it maps letters to letters; all morphisms considered in this paper are strict alphabetic. Associated with each grammar G under interpretation is a family of languages called the *grammatical family* of G. It is denoted by $\boldsymbol{L}(G)$ and is defined as

[1] Work carried out under the auspices of the Natural Sciences and Engineering Research Council of Canada Grant No. A-5192.

doi:10.5186/aasfm.1986.1114

$\mathcal{L}(G)=\{L(G'):\ G'\leqq G\}$. When a grammar is interpreted in this way it is often called a grammar form. Since the relation $\leqq$ is reflexive and transitive $\mathcal{L}(G')\subseteq\mathcal{L}(G)$ whenever $G'\leqq G$. Thus it is natural to consider the partially-ordered set of all grammatical families ordered with respect to containment. Such investigations are traditional in formal language theory, leading to numerous hierarchy results.

For $i\geqq 1$, let F_i be $S\to a^j$, $1\leqq j\leqq i$. Then $L(F_i)$ is finite as is $L(F_i')$, for all $F_i'\leqq F_i$. Moreover $L(F_i)\subseteq L(F_{i+1})$. It is not difficult to show that

$$\mathcal{L}(F_1)\subset\mathcal{L}(F_2)\subset\mathcal{L}(F_3)\subset\ldots\subset\mathcal{L}(REG).$$

In a similar manner, based on deeper results in the theory it is possible to demonstrate infinite hierarchies of regular families, linear families, and context-free families. Showing the existence of such hierarchies, which are paths in the poset of grammatical families, is only a first step in obtaining a better understanding of the structure of this poset. It should be noted that the coarser interpretation relation, the first one to be introduced and studied by [CG] leads to a much simpler poset structure as the recent papers [GGS1] and [GGS2] demonstrate. In our setting a reasonable question is: whenever $\mathcal{L}(G_1)\subset\mathcal{L}(G_2)$ for two grammars G_1 and G_2 does there exist G_3 with $\mathcal{L}(G_1)\subset\mathcal{L}(G_3)\subset\mathcal{L}(G_2)$? That such is not always the case is seen by considering the following pair of grammars:

$$G_1\colon\ S\to ab\qquad G_2\colon\ S\to ab|cde.$$

Clearly $\mathcal{L}(G_1)\subset\mathcal{L}(G_2)$ by the obvious length argument. That there is no G_3 properly in between is demonstrated as follows.

First observe that for finite forms G and H with S as their only nonterminal $\mathcal{L}(G)\subset\mathcal{L}(H)$ if and only if $G\leqq H$ and $H\not\leqq G$, where $\not\leqq$ means 'is not an interpretation of', that is $G<H$. Clearly $G\leqq H$ implies $\mathcal{L}(G)\subseteq\mathcal{L}(H)$. However if $\mathcal{L}(G)\subseteq\mathcal{L}(H)$ then $L(G)$ is in $\mathcal{L}(H)$ and hence there is a grammar $F\leqq H$ with $L(F)=L(G)$. But G and H have the same simple form therefore $G\leqq F$ and, hence, $G\leqq H$. Finally proper inclusion implies $H\not\leqq G$ by a similar argument.

Other examples of this kind are easily obtained, however what happens when there is no difference in the lengths of words generated by the two grammars? For example let G_1: $S\to ab$; G_2: $S\to aa$ then $\mathcal{L}(G_1)\subset\mathcal{L}(G_2)$ and all words are of length two. In [MSW1] this led to the notion of interpretations of directed graphs and hence to directed graph families, see [S]. Basically each word specifies an edge, so ab is an edge between nodes a and b. It was demonstrated in [MSW1] that there are infinitely many grammatical families between $\mathcal{L}(G_1)$ and $\mathcal{L}(G_2)$. Moreover for any two families G_3 and G_4 satisfying $\mathcal{L}(G_1)\subseteq\mathcal{L}(G_3)\subset\mathcal{L}(G_4)\subseteq\mathcal{L}(G_2)$ there is a G_5 properly in between G_3 and G_4, that is $\mathcal{L}(G_3)\subset\mathcal{L}(G_5)\subset\mathcal{L}(G_4)$. For this reason we say that the *interval* defined by $\mathcal{L}(G_1)$ and $\mathcal{L}(G_2)$, denoted by (G_1, G_2), is *dense*. In [MSW3] a quite surprising result is proved, namely, the interval (G', G) is dense, whenever $\mathcal{L}(G')=\mathcal{L}(REG)$ and $\mathcal{L}(G)=\mathcal{L}(CF)$. Thus there are dense intervals of sub-regular grammatical families and also dense intervals of super-regular grammatical families. One basic question about such intervals is: Are there maximal dense intervals? That is are there dense intervals which cannot be extended either above or below while retaining density. In this paper we partially solve this problem for regular intervals by demonstrating that there are *no* maximal dense regular intervals whose upper family is $\mathbf{L}_{REG}$. Extending this result to all regular dense intervals is not immediate, even if it holds, whereas for context-free dense intervals it probably does not hold.

Apart from this partial solution to the maximality question we also demonstrate that denseness is decidable for regular intervals. It has recently been shown that denseness is undecidable for context-free intervals [N].

These solutions are obtained by way of language forms and linguistical families, concepts introduced in [MSW4] and further investigated in [MSW5]. For a regular grammar form G is it well-known [OSW] that $\mathbf{L}(G)$ is characterized completely by $L(G)$, in the following sense. Consider a regular language $L' \subseteq \Sigma'^*$ and let $L = L(G)$ with Σ the alphabet of L. We write $L' \leqq L$ if there is a strict alphabetic morphism $h: \Sigma'^* \to \Sigma^*$ such that $h(L') \subseteq L$. In analogy with the introduction of the grammatical family of a grammar form we define the *regular linguistical family* of the *regular language form* L by: $\mathbf{L}_r(L) = \{L' : L' \leqq L$ and L' is regular$\}$. It is proved in [OSW] that if $\mathbf{L}(G) \subseteq \mathbf{L}(REG)$ then $\mathbf{L}(G) = \mathbf{L}_r(L(G))$. This characterization implies that we need only treat regular language forms and regular linguistical families, rather than the more indirect (regular) grammar forms and regular grammatical families.

2. Some definitional and theoretical preliminaries

Given a language L and a language L' we say L' is an *interpretation* of L if there is a strict alphabetic morphism h such that $h(L') \subseteq L$. We denote this by $L' \leqq L$. We say L' is a *regular interpretation* of L if $L' \leqq L$ and L' is regular, this is denoted by $L' \leqq_r L$. Note that L itself need not be regular. Similarly we say L' is a *finite interpretation* of L, denoted by $L' \leqq_f L$, if $L' \leqq L$ and L' is finite. Moreover, we write $L' < L (L' <_r L, L' <_f L)$ if $L' \leqq L$ but L is not an interpretation of L' (and L' is regular, finite, respectively). If $L' \leqq L$ and $L \leqq L'$ then we say that L and L' are equivalent, denoted by $L \sim L'$.

The corresponding linguistical families are denoted by $\mathbf{L}(L)$, $\mathbf{L}_r(L)$, and $\mathbf{L}_f(L)$, respectively. These notions are tied together in the following theorem, see [MSW4].

Theorem 2.1. *For all languages L_1 and L_2 the following statements are equivalent:*

(1) $\mathbf{L}(L_1) = \mathbf{L}(L_2)$

(2) $\mathbf{L}_r(L_1) = \mathbf{L}_r(L_2)$

(3) $\mathbf{L}_f(L_1) = \mathbf{L}_f(L_2)$.

The above theorem has the obvious implication that to obtain distinct linguistical families we only need obtain distinct regular-linguistical families or, even, distinct finite-linguistical families. These, it is assumed, will be easier to handle. Note that $\boldsymbol{L}(L_1) \subseteqq \boldsymbol{L}(L_2)$ if and only if $L_1 \leqq L_2$ if and only if $\boldsymbol{L}_r(L_1) \subseteqq \boldsymbol{L}_r(L_2)$ if and only if $\boldsymbol{L}_f(L_1) \subseteqq \boldsymbol{L}_f(L_2)$.

In analogy with the definition of dense interval for grammar forms we say that (L_1, L_2) denotes an *interval* if $L_1 < L_2$ and hence $\boldsymbol{L}(L_1) \subset \boldsymbol{L}(L_2)$. The interval (L_1, L_2) is *dense* if for all languages L_3 and L_4 that satisfy $L_1 \leqq L_3 < L_4 \leqq L_2$ there is an L_5 with $L_3 < L_5 < L_4$. Similarly we say that an interval (L_1, L_2) is *regular* if both L_1 and L_2 are regular and it is *regular dense, r-dense* for short, if it is regular and for all regular languages L_3 and L_4 that satisfy $L_1 \leqq_r L_3 <_r L_4 \leqq_r L_2$ there is a regular language L_5 with $L_3 <_r L_5 <_r L_4$.

We have defined these notions in terms of interpretations rather than in terms of linguistical families, but since $L_1 \leqq L_2$ if and only if $\boldsymbol{L}(L_1) \subseteqq \boldsymbol{L}(L_2)$ this is only a matter of convenience.

Density and regular density are somewhat related as we will show below, but we first need to define super-disjoint union.

Let $L_1 \subseteqq \Sigma_1^*$ and $L_2 \subseteqq \Sigma_2^*$ be two languages. Then *the super-disjoint union* of L_1 and L_2, denoted by $L_1 \cup L_2$, is their union if $\Sigma_1 \cap \Sigma_2 = \emptyset$ and is undefined otherwise. We call it super-disjoint union since it is not only a disjoint union $(L_1 \cap L_2 = \emptyset)$, but also $\Sigma_1 \cap \Sigma_2 = \emptyset$. If L_1 and L_2 are arbitrary *language forms*, then we can always rename the alphabet of L_1, say, to obtain disjoint alphabets, hence, in this case, we assume that $L_1 \cup L_2$ is always well-defined. We use the ordinary union sign for super-disjoint unions, because of typographical reasons.

We now relate dense and regular dense intervals.

Theorem 2.2. *Let (L_1, L_2) be a regular interval. If (L_1, L_2) is dense then (L_1, L_2) is regular dense.*

Proof. Consider an arbitrary regular interval (L_3, L_4) that satisfies $L_1 \leqq_r L_3$ and $L_4 \leqq_r L_2$; clearly such an interval always exists. Since (L_1, L_2) is dense there is an L_5 with $L_3 <_r L_5 < L_4$. Now by Theorem 2.1 this implies there is a finite language F which is an interpretation of L_5 but not of L_3. Consider $L = L_3 \cup F$. Clearly $L_3 <_r L$, L is regular, therefore $L \leqq_r L_5$, and hence $L <_r L_4$. In other words (L_1, L_2) is a regular dense interval. □

If an interval (L_1, L_2) contains no language L properly in between L_1 and L_2, then we say that L_1 is a *predecessor* of L_2 and L_2 has a predecessor.

Predecessors and density are complementary notions, since we have:

Proposition 2.3. *Let (L_1, L_2) be an interval. Then (L_1, L_2) is dense if and only if it contains no language L having a predecessor in the interval.*

It turns out that characterizing those languages which have predecessors is one step on the way to characterizing those intervals which are dense. For this purpose we require three auxiliary notions.

Let L be a language and Σ be its alphabet. We say L is *coherent* if for all nonempty disjoint alphabets Σ_1 and Σ_2 with $\Sigma_1 \cup \Sigma_2 = \Sigma$, there is a word x in L with x in $\Sigma^* \Sigma_1 \Sigma^* \Sigma_2 \Sigma^* \cup \Sigma^* \Sigma_2 \Sigma^* \Sigma_1 \Sigma^*$. We say L is *incoherent* otherwise. Observe that if L is incoherent then there are L_1 and L_2 with $L = L_1 \cup L_2$, where $\emptyset \neq L_i \neq \{\lambda\}$, $i = 1, 2$.

A language form L is *minimal* if there is no language form $L' \subset L$ with $\boldsymbol{L}(L') = \boldsymbol{L}(L)$. If L is finite, then minimality is clearly decidable and if L is finite and non-minimal then the construction of an equivalent minimal $L' \subset L$ is straightforward.

We now introduce our third notion, looping languages. A language L is *looping* if either L contains a word containing two appearances of the same letter, or there exist distinct words $w_1, \ldots, w_n$ in L and distinct letters $a_1, \ldots, a_n$ in $alph(L)$, for $n \geqq 2$, such that a_i and a_{i+1} are in w_i, $1 \leqq i < n$ and a_n and a_1 are in w_n. If L is not looping we say it is *nonlooping*. ($alph(L)$ is the smallest alphabet Σ such that $L \subseteq \Sigma^*$.)

Given a language form L, L' is a nonlooping interpretation of L, denoted by $L' \leqq_n L$ if $L' \leqq L$ and L' is nonlooping. We therefore have $\boldsymbol{L}_n(L)$ as well.

In [MSW2] the following result is to be found.

Proposition 2.4. *Let L be a finite language.*

(i) *If L is minimal and coherent, then L has a predecessor if and only if L is nonlooping.*

(ii) *If L is minimal, then L has a predecessor if and only if $L = K \cup N$ for some K and nontrivial N, where N is nonlooping.*

We extend this result to arbitrary languages, by first treating the coherent case.

Theorem 2.5. *Let L be a coherent minimal language. Then L has a predecessor if and only if L is nonlooping.*

Proof. If L is finite the result follows by Proposition 2.4, therefore assume L is infinite. Since each language is over a finite alphabet an infinite language is always looping. Therefore we only need demonstrate that an infinite language never has a predecessor to complete the Theorem. Assume L has a predecessor P. We argue by contradiction demonstrating that there is always a language properly in between P and L. By Theorem 2.1, there exists a finite F with $F \leqq L$ and $F \nleqq P$. This implies $P < P \cup F \leqq L$. We also have $L \nleqq P \cup F$. This follows from the coherence of L, the finiteness of F, and $P < L$. Thus $P < P \cup F < L$ and we have obtained a language properly in between P and L as required, therefore L has no predecessor. □

We now generalize the second part of Proposition 2.4.

Theorem 2.6. *Let L be a minimal language. Then L has a predecessor if and only if $L = K \cup N$, for some language K and some nontrivial, nonlooping N.*

Proof. The proof for finite L is to be found in [MSW2]. The infinite case follows analogously, we merely give a brief proof sketch. Assume L is infinite. If L is coherent then L has no predecessor by Theorem 2.5 and it has no decomposition of the required form. Thus the Theorem holds in this case. Therefore assume L is incoherent. If $L=K\cup N$, where N is nontrivial and nonlooping, then N has a predecessor P and we need to show that $K\cup P$ is a predecessor of L. On the other hand if L has no nontrivial, nonlooping component N, then it only remains to demonstrate that there is a language properly in between P and L for any $P<L$. In both cases we make heavy use of the observation that if a coherent language Q satisfies $Q\leqq L$, then Q is an interpretation of some coherent component of L. □

3. The density characterization theorems

One of the major obstacles to proving decidability results for intervals of grammatical families has been the lack of a density characterization theorem for such intervals. In the present section we provide such theorems which are then used to provide examples of dense intervals.

First we need to introduce some additional notation and terminology concerning nonlooping languages. We say that two languages L_1 and L_2 are *nonlooping equivalent,* denoted by $L_1\sim_n L_2$, if $\mathbf{L}_n(L_1)=\mathbf{L}_n(L_2)$ and are *nonlooping inequivalent,* denoted by $L_1\nsim_n L_2$, if $\mathbf{L}_n(L_1)\neq\mathbf{L}_n(L_2)$. We also say that a language L is *nonlooping complete* or *n-complete* if $\mathbf{L}_n(L)$ is the family of all nonlooping languages.

Theorem 3.1. The first density characterization theorem. *Given two languages L_1 and L_2 with $L_1<L_2$, then (L_1, L_2) is dense if and only if $L_1\sim_n L_2$. Similarly if L_1 and L_2 are regular, then (L_1, L_2) is r-dense if and only if $L_1\sim_n L_2$.*

Proof. The second statement follows from the first by way of Theorem 2.2, hence we will only prove the first statement here.

Without loss of generality assume both L_1 and L_2 are minimal.

If: Assume $L_1\sim_n L_2$. Observe that for all L satisfying

$$L_1\leqq L\leqq L_2$$

we have $L\sim_n L_i$, $i=1, 2$. Hence, if we show that for $L_1\sim_n L_2$ and $L_1<L_2$ there is an L such that $L_1<L<L_2$, then the "if-part" follows immediately.

Let $L_2=L_2'\cup M_1\cup\ldots\cup M_m$, for distinct, nontrivial coherent minimal nonlooping M_i, $1\leqq i\leqq m$ and L_2' looping, where L_2' cannot be further decomposed under $\cup$ into a nontrivial nonlooping language and a looping language. We say the above decomposition of L_2 is a *maximal nonlooping decomposition of* L_2. Similarly, let $L_1=L_1'\cup K_1\cup\ldots\cup K_k$ be a maximal nonlooping decomposition of L_1. Note that $L_1'\leqq L_2'$, since a looping language cannot be an interpretation of a nonlooping one.

Since $L_1\sim_n L_2$, $M_i\leqq L_1$, $1\leqq i\leqq m$. Furthermore $M_i\nleqq L_1'$ since if it were, then $M_i\leqq L_2'$ which contradicts the minimality of L_2. Therefore $M_i\leqq K_j$ for some j. Similarly K_j is an interpretation of the same M_i, otherwise L_1 is not minimal. Hence

$M_i \sim K_j$. This implies we can write L_1 as $L_1' \cup M_1 \cup \ldots \cup M_m \cup N_1 \cup \ldots \cup N_n$, where $n \geqq 0$ and the N_i are nontrivial, coherent, minimal, and nonlooping.

Note that $L_2' \neq \emptyset$. Otherwise $L_1' = \emptyset$ and $n=0$, hence $L_1 \sim L_2$, a contradiction. Finally consider minimal L_3 and L_4 such that

$$L_1 \leqq L_3 < L_4 \leqq L_2.$$

Then by similar arguments to those for L_1 above we can express L_3 as

$$L_3' \cup M_1 \cup \ldots \cup M_m \cup N_1 \cup \ldots \cup N_s$$

and L_4 as

$$L_4' \cup M_1 \cup \ldots \cup M_m \cup N_1 \cup \ldots \cup N_t, \text{ where } 1 \leqq t \leqq s \leqq n.$$

Moreover L_3' can be expressed as $J_1 \cup \ldots \cup J_p$ and L_4' as $K_1 \cup \ldots \cup K_q$, where each of the J_j and K_i are looping and coherent. We now show that we can always construct an L such that $L_3 < L < L_4$, that is (L_1, L_2) is dense.

(i) $s=t$. In this case there exists an i such that for all j, $1 \leqq j \leqq p$ either $J_j \not\equiv K_i$ or $J_j < K_i$. For otherwise $L_3' \sim L_4'$ and hence $L_3 \sim L_4$. Since K_i is looping it has no predecessor (by Theorem 2.5). Therefore consider a $K_i' < K_i$ which also satisfies $K_i' \not< J_j$, $1 \leqq j \leqq p$. Such a K_i' must exist since there are only finitely many $J_j \leqq K_i$, but infinitely many inequivalent K_i' with $K_i' < K_i$. To conclude this subcase observe that $L_3 \cup K_i'$ is properly between L_3 and L_4.

(ii) $s > t$. Now $N_{t+1} \cup \ldots \cup N_s \leqq K_1 \cup \ldots \cup K_q$, otherwise L_3 would not be minimal. In particular this implies $N_{t+1} \leqq K_i$ for some i, $1 \leqq i \leqq q$. Consider a K_i' such that $N_{t+1} < K_i' < K_i$. Surely such a K_i' exists and furthermore as in subcase (i) $L_3 < L_3 \cup K_i' < L_4$.

Only if: Assume (L_1, L_2) is dense. If $L_1 \not\sim_n L_2$, then there exists a coherent nonlooping N with $N \leqq L_2$ such that $N \not\equiv L_1$. But this implies $L_1 < L_1 \cup N \leqq L_2$ and by Theorem 2.6 $L_1 \cup P$ is a predecessor of $L_1 \cup N$, if P is a predecessor of N. But this implies (L_1, L_2) is not dense, a contradiction. □

Corollary 3.2. *For an arbitrary regular language L, (L, a^*) is r-dense if and only if L is n-complete and $\mathbf{L}_r(L) \subset \mathbf{L}(REG)$.*

Corollary 3.3. *For two arbitrary languages L_1 and L_2 with $L_1 <_r L_2$, (L_1, L_2) is not r-dense if L_1 is nonlooping.*

This follows by observing that if L_2 is nonlooping then $L_2 \not\equiv L_1$ and hence $L_1 \not\sim_n L_2$. On the other hand if L_2 is looping then it can generate arbitrarily long chains of words (or broken loops, see [MSW2]) and L_1 cannot. Hence once again $L_1 \not\sim_n L_2$.

Corollary 3.4. *The interval (L, a^*) is not r-dense, where $L = (a^* - \{a^2\}) \cup \{ab, ba, b\}$.*

Proof. Consider the language $M = \{ab, acd, bef\}$. Clearly M is nonlooping and M is minimal and coherent. Now both a and b appear in a word of length 3. Therefore letting h be a morphism such that $h(M) \subseteq L$, it follows that $h(acd) = h(bef) = aaa$ and hence $h(ab) = aa$. But aa is not in L, hence $M \nleqq L$ and by Corollary 3.2 (L, a^*) is not dense. □

To enable us to present specific r-dense intervals of the form (L, a^*) we need to strengthen Theorem 3.1 for the case of n-completeness. This we now do by way of the following definitions.

Let $L \subseteq \Sigma^*$ be an arbitrary nonlooping language and let $L' = L - \Sigma$. We say a word w in L is an *end word* if

$$alph(w) \cap alph(L' - \{w\}) = \{a\}, \quad \text{for some} \quad a \text{ in } \Sigma.$$

In this case we say *a connects w* and $L' - \{w\}$.

Lemma 3.5.

(i) *Every nontrivial, coherent, nonlooping language N has at least one end word if $\# N \geqq 2$.*

(ii) *If N is a coherent, nonlooping language and w is an end word in N, then $N - \{w\}$ is coherent.*

Proof. Immediate. □

We are now ready to state and prove our second characterization theorem.

Theorem 3.6. The second density characterization theorem. *Let L be an arbitrary language.*

Then (L, a^*) is r-dense if and only if L has a nontrivial subset L' for which the following condition obtains:

For all letters a in $alph(L')$ and for all i, $j \geqq 0$ there is a word x in $(alph(L'))^i$ and a word y in $(alph(L'))^j$ such that xay is in L'.
In other words L is n-complete if and only if it has such a subset L'.

Proof. In this proof whenever an n-complete language is mentioned we always assume it is also minimal in the sense that every proper subset of it is not n-complete. Clearly this is no loss of generality since each n-complete language has a minimal n-complete subset.

Because of Corollary 3.2 we only need consider the case that L is n-complete, since a^* is obviously n-complete. Moreover we observe that L is n-complete if and only if $N \leqq L$ for every coherent nonlooping language N.

If: To show that L is n-complete we need to prove that every nonlooping coherent language N has a morphic image in L' and hence in L. We prove this by induction on the cardinality of N. Note that L' contains words of all lengths. For $\# N = 1$, since the only word must consist of distinct letters it trivially has a morphic image in L'.

Now assume that for some $k \geqq 1$, every coherent, nonlooping N with $\# N \leqq k$, has a morphic image which is a subset of L'.

Let N be a nonlooping language with $\# N=k+1$. For w an end word in N there is a morphism h such that $h(N-\{w\})$ is a subset of L.

Consider the symbol a which connects w and L. Then we can write w as $b_1 \dots b_i a b_{i+1} \dots b_n$, where $0 \leqq i \leqq n$. Clearly there is a word v in L' satisfying

$$v = x_1 h(a) x_2,$$

where $|x_1|=i$ and $|x_2|=n-i$. Note that the letters $b_1, \dots, b_n$ are distinct from each other and from *alph* $(N-\{w\})$. Hence we can extend h to these new symbols such that $h(w)=v$. In other words $h(N) \subseteqq L'$ completing this part of the proof.

Only if: L is minimal and n-complete by assumption, hence we prove it satisfies the property in the Theorem statement.

Let a be a letter in *alph* (L) and let xay be a word in L. Clearly there must be at least one such word otherwise a would not be in *alph* (L).

Now there is a nonlooping language N such that whenever $h(N) \subseteqq L$, then there is a word w in N with $h(w)=xay$. If this is not the case $L-\{xay\}$ is also n-complete, contradicting the minimality of L. We define nonlooping languages M_{ij} for all $i, j \geqq 0$ by:

For every symbol s in *alph* (N) add a word

$$a_1 \dots a_i s b_1 \dots b_j$$

to N, where a_l and b_m are new symbols for every symbol s in *alph* (N).

Now since each M_{ij} is nonlooping we have $M_{ij} \leqq L$ for all $i, j \geqq 0$. Moreover whenever $g(M_{ij}) \subseteqq L$, for some morphism g, then $g(w)=xay$ by the above remarks Hence $g(a_1 \dots a_i s b_1 \dots b_j)=x_1 a y_1$, for some s in *alph* (N), where $|x_1|=i$ and $|y_1|=j$. Since $x_1 a y_1$ is in L, L satisfies the property in the Theorem statement, completing the proof. □

This leads immediately to some specific examples of n-complete languages and hence dense intervals.

Corollary 3.7. *$L_1=\{a, b\}^*-\{a^i, b^i\colon i \geqq 2\}$ is n-complete and hence (L_1, a^*) is an r-dense interval.*

Proof. L_1 clearly satisfies the condition of Theorem 3.6. □

Corollary 3.8. *$L_2=\{a, b, c\}^*-\{a^3, b^3, c^3, aab, aac, aba, aca, baa, caa, bbc, bcb, cbb\}$ is n-complete.*

More importantly:

Corollary 3.9. *Let $\Sigma_m=\{a_1, a_2, \dots, a_m\}$ and $K_m=(\Sigma_m^*-\Sigma_m^2) \cup \{a_1 a_2, a_2 a_3, \dots, a_m a_1\}$. Then K_m is n-complete.*

4. Decidability and maximality

In this section we first prove that n-completeness is decidable for context-free languages, and then show that there is no maximally r-dense interval (L, a^*).

Theorem 4.1. *N-completeness is decidable for context-free languages.*

Proof. L is n-complete if and only if it has a subset L', which satisfies the condition of Theorem 3.6, that is $L'=L\cap\Sigma^*$ for some $\Sigma\subseteq alph(L)$. Now define finite substitutions δ_a for all a in Σ by:

$$\delta_a(a) = \{f, a\}$$

$$\delta_a(b) = \{f\}, \quad \text{for all } b \quad \text{in} \quad \Sigma,\ b \neq a,$$

where f is a new symbol. Clearly L' satisfies the condition of Theorem 3.6. if and only if $M_a=\delta_a(L')\cap f^*af^*$ equals f^*af^*, for all a in Σ.

This is decidable since f^*af^* is a bounded regular set and M_a is context-free. □

This together with Theorem 3.1 immediately gives:

Corollary 4.2. *Given a context-free (regular) language L it is decidable whether or not (L, a^*) is dense (r-dense).*

In order to prove the maximality result we need to consider *directed cycles of length m*, denoted by C_m. Letting $\Sigma_m=\{a_1, a_2, \ldots, a_m\}$ we define C_m by:

$$C_m = \{a_1a_2, a_2a_3, \ldots, a_ma_1\}.$$

It is a straightforward observation that

$$C_r \leqq C_m \quad \text{if and only if} \quad r = 0\,(mod\ m).$$

On the other hand every nonlooping language $N\subset\Sigma^2$ is an interpretation of C_m for all $m\geqq 1$.

We now have:

Lemma 4.3. *Let L be an n-complete language. Then there is an m and a bijection g such that $g(C_m)\subseteq L$.*

Proof. We only need consider $L'=\{w \text{ is in } L\colon |w|=2\}$. Let $\#L'=r$. Now since all nonlooping languages are interpretations of L, then in particular

$$P_r = \{a_1a_2, a_2a_3, \ldots, a_ra_{r+1}, a_{r+1}a_{r+2}\}$$

where the a_i's are different letters for different i's, is an interpretation of L', that is there is a morphism h such that $h(P_r)\subseteq L'$. Now h cannot be one-to-one, since $\#P=r+1>\#L'$. Therefore h merges at least two letters and hence there is an $m\geqq 1$ such that $C_m\subseteq h(P_r)$. But this implies $g(C_m)\subseteq L'\subseteq L$ for some bijection g completing the proof. □

We also need:

Lemma 4.4. *Let L_1 and L_2 be (regular) languages. Then there is a (regular) language L such that*

$$\boldsymbol{L}(L) = \boldsymbol{L}(L_1) \cap \boldsymbol{L}(L_2)$$

and

$$\boldsymbol{L}_r(L) = \boldsymbol{L}_r(L_1) \cap \boldsymbol{L}_r(L_2).$$

Proof. This follows along the lines of the proof of Theorem 4.2 in [MSW5] and therefore it is left to the reader. □

We are now able to prove our final result:

Theorem 4.5. *There is no (regular) language L such that (L, a^*) is maximally dense (r-dense).*

Proof. We show that every dense interval (L, a^*) can be extended. In other words that there exists an L_0 such that $L_0 < L$ and (L_0, a^*) is dense.

From Lemma 4.3 we know that there is an integer $m \geqq 1$ and a bijection g such that $g(C_m) \subseteqq L$. Let m_0 be the greatest such m.

Immediately $L' = \{w \text{ is in } L: |w| = 2\}$ is not an interpretation of C_{m_0+1}, since $C_{m_0} \not\equiv C_{m_0+1}$.

Now consider K_{m_0+1} from Corollary 3.9. Then $C_{m_0+1} \subseteqq K_{m_0+1}$ and moreover is not an interpretation of K_{m_0+1}. Now let L_0 be a language such that

$$\boldsymbol{L}(L_0) = \boldsymbol{L}(L) \cap \boldsymbol{L}(K_{m_0+1}).$$

Note that $L_0 < L$, since L is not in $\boldsymbol{L}(K_{m_0+1})$ and so it is not in $\boldsymbol{L}(L_0)$.

It remains to demonstrate that L_0 is n-complete. However L is n-complete by assumption and K_{m_0+1} is n-complete by Corollary 3.9. Hence L_0 is n-complete and (L_0, a^*) is both dense and an extension of (L, a^*) as required.

If L is regular, then L_0 can be chosen to be regular (Lemma 4.4) since K_{m_0+1} is regular. Hence by Theorem 2.2, the "regular" version of the theorem follows. □

References

[CG] Cremers, A. B., and S. Ginsburg: Context-free grammar forms. - J. Comput. System Sci. 11, 1975, 86—116.

[GGS1] Ginsburg, S., J. Goldstine, and E. H. Spanier: A prime decomposition theorem for grammatical families. - J. Comput. System Sci. 24, 1982, 315—361.

[GGS2] Ginsburg, S., J. Goldstine, and E. H. Spanier: On the equality of grammatical families. - J. Comput. System Sci. 26, 1983, 171—196.

[H] Harrison, M. A.: Introduction to formal language theory. - Addison-Wesley Publishing Co., Inc., Reading, Mass., 1978.

[HU] Hopcroft, J. E., and J. D. Ullman: Formal languages and their relation to automata, Second edition. - Addison-Wesley Publishing Co., Inc., Reading, Mass., 1979.

[MSW1] MAURER, H. A., A. SALOMAA, and D. WOOD: Colorings and interpretations: A connection between graphs and grammar forms. - Discrete Appl.. Math. 9, 1981, 119—135.

[MSW2] MAURER, H. A., A. SALOMAA, and D. WOOD: On predecessors of finite languages. - Inform. and Control 50, 1981, 259—275.

[MSW3] MAURER, H. A., A. SALOMAA, and D. WOOD: Dense hierarchies of grammatical families.- J. Assoc. Comput. Mach. 29, 1982, 118—126.

[MSW4] MAURER, H. A., A. SALOMAA, and D. WOOD: Finitary and infinitary interpretations of languages. - Math. Systems Theory 15, 1982, 251—265.

[MSW5] MAURER, H. A., A. SALOMAA, and D. WOOD: On finite grammar forms. - Internat. J. Comput. Math. 12, 1983, 227—240.

[N] NIEMI, V.: Personal communication, 1984.

[OSW] OTTMANN, TH., A. SALOMAA, and D. WOOD: Sub-regular grammar forms. - Inform. Process. Lett. 12, 1981, 184—187.

[S] SALOMAA, A.: Jewels of formal language theory. - Computer Science Press, Inc., Rockville, Maryland, 1981.

[Wo] WOOD, D.: Grammar and L forms; An introduction. - Lecture Notes in Computer Science 91, Springer-Verlag, New York, 1980.

TU Graz
Institute für Informationsverarbeitung
Schießstattgasse 4a
A-8010 Graz
Austria

University of Turku
Department of Mathematics
SF-20500 Turku
Finland

TU Graz
Institute für Informationsverarbeitung
Schießstattgasse 4a
A-8010 Graz
Austria

University of Waterloo
Department of Computer Science
Waterloo, Ontario N2L 3G1
Canada

Received 11 October 1984