UNIVERSITY OF WATERLOO
COMPUTER SCIENCE DEPARTMENT

Dense Intervals of Linguistic Families

H.A. Maurer
A. Salomaa
E. Welzl
D. Wood

CS-82-31

September, 1982

DENSE INTERVALS OF LINGUISTICAL FAMILIES[1]

by

H. A. Maurer,[2] A. Salomaa,[3]
E. Welzl[2] and D. Wood[4]

---

(1) Work carried out under the auspices of the Natural Sciences and Engineering Research Council of Canada Grant No. A-7700.

(2) Institut für Informationsverarbeitung, TU Graz,
Schießstattgasse 4a,
A-8010 Graz, Austria

(3) Department of Mathematics,
University of Turku,
SF-20500 Turku 50, Finland.

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

Abstract

Language forms, their interpretations, and their linguistical families are introduced, analogously to grammar forms. These notions enable us to prove the following three results for grammar forms:

(i) the interval of all super-regular grammatical families is not maximally dense,

(ii) there is no maximally dense interval of grammatical families which contain the family of regular languages, and

(iii) it is decidable whether or not an interval of grammatical families containing the family of regular languages is dense.

1. Introduction

In [MSW2] the study of dense hierarchies or intervals of grammatical families generated by grammar forms has been initiated. We say that a pair of grammatical families $(\mathcal{L}_1, \mathcal{L}_2)$ forms a dense interval if $\mathcal{L}_1 \subsetneq \mathcal{L}_2$ and for every pair of grammatical families $\mathcal{L}_3$ and $\mathcal{L}_4$ satisfying $\mathcal{L}_1 \subseteq \mathcal{L}_3 \subsetneq \mathcal{L}_4 \subseteq \mathcal{L}_2$ there exists a grammatical family properly in between $\mathcal{L}_3$ and $\mathcal{L}_4$.

Although the interval $(\mathcal{L}(\text{REG}), \mathcal{L}(\text{CF}))$ is shown to be dense in [MSW2], the question of its maximality within the collection of context-free grammatical families was left open. This interval is maximal if for all grammatical families $\mathcal{L} \subsetneq \mathcal{L}(\text{REG})$, $(\mathcal{L}, \mathcal{L}(\text{CF}))$ is not dense.

In the present paper we resolve this issue by demonstrating that $(\mathcal{L}(\text{REG}), \mathcal{L}(\text{CF}))$ is <u>not</u> maximal. To this end we introduce the general notion of a language form, which turns out to be crucial in the solution of this problem, enabling us to provide a characterization of those L_1 and L_2 for which $(\mathcal{L}(L_1), \mathcal{L}(L_2))$ is dense. Moreover we strengthen this result considerably by proving that there is <u>no</u> maximally dense interval $(\mathcal{L}, \mathcal{L}(\text{CF}))$. Finally, we prove that it is decidable whether or not $(\mathcal{L}, \mathcal{L}(\text{REG}))$ and, hence $(\mathcal{L}, \mathcal{L}(\text{CF}))$ is a dense interval.

The paper consists of a further three sections. Section 2 provides the basic definitions and preliminary results. Although this paper is dependent on some earlier results, we

have attempted to make it as self-contained as possible. The crucial results of earlier papers are the four propositions stated in Section 2. Section 3 states and proves two Density Characterization Theorems, while Section 4 considers various implications of these Theorems.

2. Basic Notation and Results

In this section we not only introduce the basic notions and state some propositions, but we also prove some preliminary results.

Throughout this paper we use the following convention regarding homomorphisms.

Convention: Every homomorphism $h: \Sigma^* \to \Delta^*$ is assumed to be a literal (letter-to-letter) homomorphism, that is $h(\Sigma) \subseteq \Delta$. We use the terms homomorphism and morphism interchangeably.

Let L be a language, then by alph(L) we denote its alphabet. We say L is looping if either L contains a word containing two distinct 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 $n \geq 2$ such that a_i, a_{i+1} are in w_i, $1 \leq i < n$ and a_n, a_1 are in w_n. If L is not looping we say it is nonlooping.

Let $L \subseteq \Sigma^*$ and $L' \subseteq \Sigma'^*$ then L' is an interpretation of L, denoted by $L' \lhd L$, if there exists a morphism $h: \Sigma'^* \to \Sigma^*$ such that $h(L') \subseteq L$. We say L' is a regular interpretation, finite interpretation or nonlooping interpretation of L denoted by $L' \lhd_r L$, $L' \lhd_f L$ or $L' \lhd_n L$, respectively, if $L' \lhd L$ and L' is regular, finite or nonlooping, respectively. A language interpreted in this manner is also called a language form.

Each language L defines a family of languages under interpretation, denoted by $\mathcal{L}(L)$, $\mathcal{L}_r(L)$, $\mathcal{L}_f(L)$ or $\mathcal{L}_n(L)$, we call these the linguistical, r-linguistical, etc., families of L. For example $\mathcal{L}(L) = \{L' : L' \prec L\}$. Note that $\mathcal{L}(L_1) \subseteq \mathcal{L}(L_2)$ iff $L_1 \prec L_2$ and similarly for $\prec_r$, etc.

The importance of these various notions stems from the following result.

Proposition 2.1 [MSW3]

For all languages L_1 and L_2,

$$\mathcal{L}(L_1) = \mathcal{L}(L_2) \text{ iff } \mathcal{L}_f(L_1) = \mathcal{L}_f(L_2).$$

In other words a family is characterized completely by the finite languages it generates. Of course it follows immediately that $\mathcal{L}_r(L_1) = \mathcal{L}_r(L_2)$ iff $\mathcal{L}_f(L_1) = \mathcal{L}_f(L_2)$.

We now introduce the notion of interpretation for context-free grammars, in which case we call them grammar forms, for further details see [Wo]. A (context-free) grammar G is a quadruple $G = (V,\Sigma,P,S,)$ where V is an alphabet, $P \subseteq (V-\Sigma)^*$ is a finite set of productions, usually written as $A \to \alpha$, and S in $V - \Sigma$ is the sentence symbol. It is well known that we can associate a language with each grammar G, denoted by

$L(G)$, where $L(G) \subseteq \Sigma^*$. For further details consult Harrison [H] or Hopcroft and Ullman [HU].

Let $G_i = (V_i, \Sigma_i, P_i, S_i)$ be context-free grammars, where $i = 1,2$. Then G_1 is an interpretation of G_2, denoted by $G_1 \lhd G_2$, if there exists a morphism $h: V_1^* \to V_2^*$ such that:

$$h(V_1 - \Sigma_1) \subseteq V_2 - \Sigma_2,\ h(\Sigma_1) \subseteq \Sigma_2,$$

$$h(P_1) \subseteq P_2 \text{ and } h(S_1) = S_2,$$

where $h(P_1)$ is defined as if P_1 is a finite language with words $A \to \alpha$, and $h(\to) = \to$.

With a grammar form G we associate its grammatical family $\mathcal{L}(G)$, defined by $\mathcal{L}(G) = \{L(G') : G' \lhd G\}$.

A sub-regular grammar form G and its language $L(G)$ are related in the form sense by the following proposition.

Proposition 2.2 [OSW]

For all context-free grammars G such that $\mathcal{L}(G)$ consists solely of regular languages

$$\mathcal{L}(G) = \mathcal{L}_r(L(G)).$$

Note that this result does not hold for "context-free interpretations" of languages.

We say that a language L is minimal if there is no $L' \subsetneq L$ such that $\mathcal{L}(L') = \mathcal{L}(L)$. Given L_1 and L_2 such that $\mathcal{L}(L_1) = \mathcal{L}(L_2)$ then L_1 is said to be equivalent to L_2, denoted by $L_1 \sim L_2$, otherwise L_1 is inequivalent to L_2, denoted by $L_1 \not\sim L_2$. If $L_1 \lhd L_2$ and $L_1 \not\sim L_2$ then we say L_1 is a proper interpretation of L_2, written $L_1 \lhd\!\!\!\!+\; L_2$. By $L_1 \not\lhd L_2$ we denote that L_1 is not an interpretation of L_2.

Since we also discuss regular and nonlooping interpretations in the following we add a subscript to $\lhd$, $\not\lhd$, $\lhd\!\!\!\!+$, $\sim$ giving $\lhd_r$, $\lhd_n$, $\tilde{\sim}_n$, $\not\sim_n$ for example.

The two central notions of this paper are captured by:

Definition

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two grammatical families satisfying $\mathcal{L}_1 \subsetneq \mathcal{L}_2$. Then $\mathcal{L}_1$ pred $\mathcal{L}_2$, denotes $\mathcal{L}_1$ is a predecessor of $\mathcal{L}_2$, if there is no grammatical family $\mathcal{L}_3$ with $\mathcal{L}_1 \subsetneq \mathcal{L}_3 \subsetneq \mathcal{L}_2$.

We say that $(\mathcal{L}_1, \mathcal{L}_2)$ forms a dense interval (of grammatical families) if for every pair of grammatical families $\mathcal{L}_3$ and $\mathcal{L}_4$ satisfying

$$\mathcal{L}_1 \subseteq \mathcal{L}_3 \subsetneq \mathcal{L}_4 \subseteq \mathcal{L}_2$$

there is a grammatical family $\mathcal{L}_5$ with $\mathcal{L}_3 \subsetneq \mathcal{L}_5 \subsetneq \mathcal{L}_4$. In a similar manner we define these notions for linguistical families generated under the various interpretation mechanisms. Typically we will write L_1 pred L_2 and (L_1, L_2) rather than $\mathcal{L}(L_1)$ pred $\mathcal{L}(L_2)$ and $(\mathcal{L}(L_1), \mathcal{L}(L_2))$, respectively, in this case and speak of r-denseness when we mean $(\mathcal{L}_r(L_1), \mathcal{L}_r(L_2))$ is dense with respect to $\prec_r$.

Note that Proposition 2.2 connects dense intervals of sub-regular grammatical families with r-dense intervals of the corresponding linguistical families.

It should be clear that an interval $(\mathcal{L}_1, \mathcal{L}_2)$ is dense (r-dense) if there is no $\mathcal{L}_3$ in the interval with a predecessor (r-predecessor) also in the interval. We characterize those language forms which have a predecessor in the following two theorems. However we need to define some preliminary notions first of all.

Definition

Let $L_1 \subseteq \Sigma_1^*$ and $L_2 \subseteq \Sigma_2^*$ be two languages. Then the superdisjoint union of L_1 and L_2, denoted by $L_1 \dot{\cup} L_2$, is defined as $L_1 \cup L_2$ if $\Sigma_1 \cap \Sigma_2 = \emptyset$ and is undefined otherwise. If L_1 and L_2 are language forms, we can always rename their alphabets to obtain disjointness, so in this case it is assumed that $L_1 \dot{\cup} L_2$ is always well-defined.

Let $L \subseteq \Sigma^*$ be a language. We say that L is incoherent if it can be decomposed into nontrivial L_1 and L_2 such that $L_1 \dot{\theta} L_2 = L$ (by nontrivial we mean $L_i \neq \emptyset$ and $L_i \neq \{\lambda\}$, where λ is the empty word). Otherwise we say L is coherent.

In [MSW1] the following two results have been proved:

Proposition 2.3

Let L be a finite coherent language. Then L has a predecessor iff L is nonlooping.

Proposition 2.4

A minimal finite language L has a predecessor iff $L = K \dot{\theta} N$ for some K and N, where N is nonlooping.

Theorem 2.5

Let L be a coherent language. Then L has a predecessor iff L is nonlooping.

Proof: If L is finite the result follows by Proposition 2.3. Hence consider infinite L only. Observe that every infinite L is looping, hence it only remains to show that such an L cannot have a predecessor. Therefore assume L has a predecessor P. Now $P \lhd_+ L$ and hence there exists a finite F with $F \lhd L$ but $F \not\lhd P$ by Proposition 2.1.

Consider $P \dot{\cup} F$. Clearly $P \trianglelefteq_+ P \dot{\cup} F \trianglelefteq L$. Moreover $L \not\trianglelefteq P \cup F$ since $L \not\trianglelefteq P$ by assumption, $L \not\trianglelefteq F$, since L is infinite, and L is coherent. This contradicts the assumption that P pred L. Hence L has no predecessor. □

We also have the general result.

Theorem 2.6

A minimal language L has a predecessor iff $L = K \dot{\cup} N$ for some K and N, where N is nonlooping.

Proof: This is similar to the proof of Proposition 2.4 and is therefore omitted. □

Finally we demonstrate that denseness implies r-denseness for regular linguistical families.

Lemma 2.7

Let L_1 and L_2 be two regular languages with $L_1 \trianglelefteq_+ L_2$. Then if (L_1, L_2) is dense, (L_1, L_2) is r-dense also.

Proof: Assume that (L_1, L_2) is not r-dense. Then by the remarks above there exists L and P satisfying $L_1 \trianglelefteq P \trianglelefteq_+ L \trianglelefteq_r L_2$ and P r-pred L. Now $P \trianglelefteq_{r+} L$ implies $P \trianglelefteq_+ L$ and since (L_1, L_2) is dense, there exists Q with $P \trianglelefteq_+ Q \trianglelefteq_+ L$. Now by Proposition

2.1 there exists a finite $F_Q \lhd Q$ with $F_Q \not\lhd P$ and also there exists a finite F_L with $F_L \lhd L$ and $F_L \not\lhd Q$. Putting these two facts together consider $P \dot{\cup} F_Q$. Clearly $P \underset{r+}{\lhd} P \dot{\cup} F_Q \underset{r+}{\not\lhd} L$, since $P \dot{\cup} F_Q$ is regular. This is a contradiction, hence (L_1, L_2) is r-dense. □

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.

Theorem 3.1: The First Density Characterization Theorem

Given two languages L_1 and L_2 with $L_1 \overset{\blacktriangleleft}{+} L_2$, then (L_1, L_2) is dense iff $L_1 \tilde{\underset{n}{}} L_2$.

Similarly if $L_1 \overset{\blacktriangleleft}{r+} L_2$, then (L_1, L_2) is r-dense iff $L_1 \tilde{\underset{n}{}} L_2$.

Proof: The second statement follows from the first by way of Lemma 2.7, hence we will only prove the first statement here.

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

if: Assume $L_1 \tilde{\underset{n}{}} L_2$. Observe that for all L,

$$L_1 \blacktriangleleft L \blacktriangleleft L_2, \quad L \tilde{\underset{n}{}} L_i.$$

Let $L_2 = L_2' \dot{\theta} M_1 \dot{\theta} \ldots \dot{\theta} M_m$, for coherent nonlooping M_i, $1 \leq i \leq m$ and L_2' looping, where L_2' cannot be further decomposed under $\dot{\theta}$ into a non-empty nonlooping language and a looping language. We say the above decomposition of L_2 is the maximal non-looping decomposition for L_2. Similarly, let $L_1 = L_1' \dot{\theta} K_1 \dot{\theta} \ldots \dot{\theta} K_k$ be the maximal nonlooping decomposition for L_1.

Since $L_1 \tilde{n} L_2$, $M_i \prec L_1$, $1 \leq i \leq m$. Furthermore each $M_i \not\prec L_1'$ since this would contradict the minimality of L_2. Therefore $M_i \prec K_j$ for some j. Similarly $K_j \prec M_i$, otherwise it would contradict the minimality of L_1. Hence $M_i \sim K_j$. This implies we can write L_1 as $L_1' \dot{\theta} M_1 \dot{\theta} \ldots \dot{\theta} M_m \dot{\theta} N_1 \dot{\theta} \ldots \dot{\theta} N_n$, where $n \geq 0$ and the N_i are coherent and nonlooping.

Note that $L_2' \neq \emptyset$. Otherwise $L_1 \overset{\prec}{+} L_2$ implies $L_1' = \emptyset$, since a looping language cannot be obtained from a nonlooping one. Moreover in this case $n = 0$ and hence $L_1 \sim L_2$, a contradiction.

Finally consider minimal L_3 and L_4 such that

$$L_1 \prec L_3 \overset{\prec}{+} L_4 \prec L_2.$$

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

$$L_3' \dot{\theta} M_1 \dot{\theta} \ldots \dot{\theta} M_m \dot{\theta} N_1 \dot{\theta} \ldots \dot{\theta} N_s$$

and

$$L_4 \text{ as } L_4' \dot{\cup} M_1 \dot{\cup} \ldots \dot{\cup} M_m \dot{\cup} N_1 \dot{\cup} \ldots \dot{\cup} N_t,$$

where

$$1 \leq t \leq s \leq n.$$

Moreover L_3' can be expressed as $J_1 \dot{\cup} \ldots \dot{\cup} J_p$ and L_4' as $K_1 \dot{\cup} \ldots \dot{\cup} K_q$, where each of the J_i and K_i are looping and coherent. We now show that we can always construct an L such that $L_3 \trianglelefteq^{+} L \trianglelefteq^{+} L_4$.

(i) $s = t$. In this case there exists an i such that for all j, $1 \leq j \leq p$ either $J_j \not\trianglelefteq K_i$ or $J_j \trianglelefteq^{+} K_i$. 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 $K_i' \trianglelefteq^{+} K_i$ which also satisfies $K_i' \not\sim J_j$, $1 \leq j \leq p$. Such a K_i' must exist since there are only finitely many $J_j \trianglelefteq K_i$. To conclude this subcase observe that $L_3 \dot{\cup} K_i'$ is properly between L_3 and L_4.

(ii) $s > t$. Now $N_{t+1} \dot{\cup} \ldots \dot{\cup} N_s \lhd K_1 \dot{\cup} \ldots \dot{\cup} K_q$, otherwise we would have a contradiction to the minimality of L_3. In particular this implies $N_{t+1} \lhd K_i$ for some i, $1 \leq i \leq q$. Consider K_i' such that $N_{t+1} \lhd_+ K_i' \lhd_+ K_i$. Surely such a K_i' exists and furthermore as in subcase (i) $L_3 \lhd_+ L_3 \dot{\cup} K_i' \lhd_+ L_4$.

only if: Assume (L_1, L_2) is dense. Now if $L_1 \not\eqslantless_n L_2$, then there exists a coherent nonlooping $F \lhd L_2$ such that $F \not\lhd L_1$. But this implies $L_1 \lhd_+ L_1 \dot{\cup} F \lhd L_2$ and by Theorem 2.6 $L_1 \dot{\cup} P$ is a predecessor of $L_1 \dot{\cup} F$, if P is the predecessor of F. But this implies (L_1, L_2) is not dense, a contradiction. □

Corollary 3.2

For an arbitrary regular language L, $(\mathcal{L}_r(L), \mathcal{L}(\mathrm{REG}))$ is r-dense iff L is n-complete and $\mathcal{L}_r(L) \not\subseteq \mathcal{L}(\mathrm{REG})$.

Corollary 3.3

For two arbitrary L_1 and L_2 with $L_1 \lhd_{r+} L_2$, (L_1, L_2) is not r-dense if L_1 is nonlooping.

This follows by observing that if L_2 is nonlooping then $L_2 \not\lhd L_1$ and hence $L_1 \not\eqslantless_n L_2$. On the other hand if L_2 is

looping then it can generate arbitrarily long chains of words (or broken loops, see [MSW1]) and L_1 cannot, hence once again $L_1 \not\overset{h}{=} L_2$.

Corollary 3.4

The interval $(\mathcal{L}_r(L), \mathcal{L}(REG))$ 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 $M \not\overset{h}{=} L$ and by Corollary 3.2 (L,a^*) is not dense. □

To enable us to present specific r-dense intervals of the form $(\mathcal{L}, \mathcal{L}(REG))$ 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 isolated if $alph(w) \cap alph(L'-\{w\}) = \emptyset$. We say a word w in L is an end word if

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

In this case we say a connects w and L' - {w}. We say a word w in L is an inner word if it is neither an isolated word nor an end word.

Lemma 3.5

Every coherent nontrivial nonlooping language N has at least one end word if $\#N \geq 2$.

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 iff L has a subset L' for which the following condition obtains:

> For all letters a in alph(L') and for all i, j ≥ 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 iff it has such a subset L'.

Proof: In this proof whenever an n-complete language is mentioned we always assume it is also minimal, that is 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 Theorem 3.2 we only need consider the case that L is n-complete, since a* is obviously n-complete.

if: To show that L is n-complete we need to prove that every nonlooping 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 $k \geq 1$, every N with $\#N \leq k$, 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 connecting symbol a in w. Then we can write w as $b_1 \ldots b_i a b_{i+1} \ldots b_n$, where $0 \leq i \leq n$. 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 $b_1, \ldots, b_n$ 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 $h(N) \subseteq L'$ completing this part of the proof.

only if: L is minimal 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 with $|xy| \neq 0$ otherwise L would not be minimal n-complete.

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

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

$$a_1 \ldots a_i s b_1 \ldots b_j$$

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 $M_{ij} \prec L$ for all $i, j \geq 0$. Moreover whenever $g(M_{ij}) \subseteq L$, for some morphism g, then $g(w) = xay$ by the above remarks. Moreover $g(a_1 \ldots a_i s b_1 \ldots b_j) = x_1 a y_1$ in L, for some s in alph(N) and hence 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

$L_1 = \{a,b\}^* - \{a^i, b^i : i \geq 2\}$ 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 $\Sigma_m = \{a_1, a_2, \ldots, a_m\}$ and
$K_m = (\Sigma^* - \Sigma^2) \cup \{a_1a_2, a_2a_3, \ldots, a_ma_1\}$.
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 go on to show not only is $(\mathcal{L}(REG), \mathcal{L}(CF))$ not a maximally dense interval, but also 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 iff 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:

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

where f is a new symbol.

Clearly L' satisfies the condition of Theorem 3.6 iff $M_a = \delta_a(L') \cap f^*af^* = f^*af^*$, for all a in Σ.

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

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

$$C_r \triangleleft C_m \text{ iff } r \equiv 0 \pmod{m}.$$

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

We now have:

Lemma 4.2

Let L be an n-complete language.

Then there is an m such that $C_m \subseteq L$.

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

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

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 $m \geq 1$ such that $C_m \subseteq h(P_r)$. But this implies $C_m \subseteq L' \subseteq L$ completing the proof. □

We also need:

Lemma 4.3

Let L_1 and L_2 be regular languages. Then there is a regular language L such that

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

and

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

Proof: This follows along the lines of the proof of Theorem 4.2 in [MSW4] and hence is left to the reader. □

We are now able to prove our final result:

Theorem 4.4

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 $L_0 \overset{\subset}{\neq} L$ and (L_0,a^*) is dense.

From Lemma 4.2 we know that there is an integer $m \geq 1$ such that $C_m \subseteq L$. 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 $C_{m_0} \not\subset C_{m_0+1}$.

Letting $K_{m_0+1} = (alph(L)^* - alph(L)^2 \cup \{a_1a_2, \ldots, a_{m_0+1}a1\}$. Then $C_{m_0+1} \subseteq K_{m_0+1}$ and moreover L is not an interpretation of K_{m_0+1}. Now let L_0 be a language such that

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

Note that $L_0 \overset{\subset}{\neq} L$, since L is not in $\mathcal{L}(L_0)$.

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. □

Corollary 4.5

For all context-free grammar forms G with $\mathcal{L}(G)$, $\mathcal{L}(CF)$, the interval $(\mathcal{L}(G), \mathcal{L}(CF))$ is not maximally dense.

Thus a problem posed in [MSW2] has finally been solved.

References

[H] Harrison, M.A. Introduction to Formal Language Theory. Addison-Wesley Publishing Co., Inc., Reading, Mass. (1978).

[HU] Hopcroft, J.E., and Ullman, J.D. Formal Languages and Their Relation to Automata, Second Edition. Addison-Wesley Publishing Co., Inc., Reading, Mass. (1979).

[MSW1] Maurer, H.A., Salomaa, A., and Wood, D., On predecessors of finite languages. McMaster University Computer Science Technical Report 80-CS-14 (1980).

[MSW2] Maurer, H.A., Salomaa, A., and Wood D. Dense hierarchies of grammatical families. Journal of the ACM 29 (1981), 118-126.

[MSW3] Maurer, H.A., Salomaa, A., and Wood, D., Finitary and infinitary interpretations of languages. Mathematical Systems Theory (1982), to appear.

[MSW4] Maurer, H.A., Salomaa, A., and Wood, D. On finite grammar forms. International Journal of Computer Mathematics (1982), to appear.

[OSW] Ottmann, Th., Salomaa, A., and Wood, D. Sub-regular grammar forms. Information Processing Letters 12 (1981), 184-187.

[Wo] Wood, D. Grammar and L Forms: An Introduction. Springer-Verlag Lecture Notes in Computer Science No. 91 (1980), New York.