SOME LOGICAL CHARACTERIZATIONS OF THE DOT-DEPTH HIERARCHY AND APPLICATIONS

by

Francine Blanchet-Sadri
Department of Mathematics and Statistics
McGill University, Montreal

May 1989

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Philosophiae doctor

LOGICAL CHARACTERIZATIONS OF THE DOT DEPTH HIERARCHY AND APPLICATIONS

Abstract

New connections are discovered between formal language theory and model theory. We give logical characterizations of natural subhierarchies of the Straubing hierarchy of star-free languages using logical notions such as quantifier complexity of first order sentences. A version of the Ehrenfeucht-Fraisse game is used to obtain a characterization of the star-free languages in terms of congruences.

This thesis, which studies the fine structure of the Straubing hierarchy, is concerned with applications of the above logical characterizations. Among them are: a conjecture of Pin concerning tree hierarchies of monoids is shown to be false; the studying of properties of the characterizing congruences and equation systems for the varieties of monoids corresponding to the levels of the Straubing hierarchy are closely related; upper and lower bounds on dot-depth are obtained.

Résumé

De nouvelles relations sont obtenues entre la théorie des langages formels et la théorie des modèles. Des caractérisations logiques de sous-hiérarchies naturelles de la hiérarchie de concaténation de Straubing des langages sans étoiles utilisant des notions logiques comme la complexité de quantificateurs de formules du premier ordre sont données. Une version du jeu de Ehrenfeucht-Fraisse est utilisée pour obtenir une caractérisation de ces langages en terme de congruences.

Cette thèse, qui étudie la fine structure de la hiérarchie de Straubing, contient plusieurs applications des caractérisations logiques mentionnées, parmi elles sont: une conjecture de Pin concernant les hiérarchies d'arbres de monoides est démontrée fausse; l'étude de propriétés des congruences caractéristiques et de systèmes d'équations reliées aux variétés de monoides correspondant aux niveaux de la hiérarchie de Straubing sont en rapport très étroit; des bornes supérieures et inférieures de dot-depth sont obtenues.

Acknowledgments

I wish to express my thanks to my adviser Professor Michael Makkai of the Department of Mathematics and Statistics of McGill University.

Thanks to Denis Therien and Wolfgang Thomas for introducing me to the field of star-free languages and open problems related to these languages.

Thanks to several referees of papers extracted from this thesis and also the external examiner of this thesis for their valuable comments and suggestions.

I also want to thank my spouse for his encouragement.

This research was partially supported by an FCAR Scholarship.

A mes parents
et
mes grands-parents

Table of contents

Chapter 1

INTRODUCTION

Traditionally, algebraic automata theory uses monoids as models for finite state machines. One looks at a finite state machine as processing sequences of symbols drawn from a finite input alphabet. Denoting the input alphabet by A, the universe of possible inputs is the free monoid A^* and a finite state machine can be thought of as a quotient of A^* by a finite index congruence $\sim$. $A^*/\sim$ being a finite monoid one is then led to investigate relationships between the structure of this algebraic system and the combinatorial processing of input sequences. The theory of varieties of Eilenberg constitutes an elegant framework for discussing these relationships between combinatorial descriptions of languages and algebraic properties of their recognizers. The interplay between the two points of view leads to interesting classifications of languages and finite monoids.

Let A be a given finite alphabet. The *regular*, or *recognizable*, languages over A are those subsets of A^* constructed from the finite languages over A by the boolean operations $(\cup, \cap, \sim)$ as well as the concatenation product $(.)$ and the star $(*)$ (the concatenation of L_1 and L_2, denoted L_1L_2, is the set

$\{xy \mid x$ is in L_1 and y is in $L_2\}$. Define $L^{\emptyset} = \{1\}$ where 1 is the empty word and $L^i = LL^{i-1}$ for $i \geq 1$. $L^* = \cup_{i \geq \emptyset} L^i)$. The *star-free* languages consist of those regular languages which can be obtained from the finite languages by boolean operations and the concatenation product only. According to a fundamental theorem of Schützenberger [Sc65], $L \subseteq A^*$ is star-free if and only if its syntactic monoid $M(L)$ is finite and aperiodic, that is, $M(L)$ contains only trivial subgroups. For example, $(ab)^*$ is star-free since $(ab)^* = ((aA^* \cap A^*b) \cap (\sim(A^*aaA^* \cup A^*bbA^*))) \cup \{1\}$. But $(aa)^*$ is not star-free, a consequence of the theorem of Schützenberger. General references on the star-free languages are McNaughton and Papert [MP71], Eilenberg [Ei76] or Pin [Pi84a].

Natural classifications of the star-free languages are obtained based on the alternative use of the boolean operations and the concatenation product. Let $A^+ = \sim\{1\}$. Define

$A^+\mathcal{B}_{\emptyset} = \{L \subseteq A^+ \mid L$ is finite or cofinite$\}$,

$A^+\mathcal{B}_{k+1} = \{L \subseteq A^+ \mid L$ is a boolean combination of languages of the form $L_1 \ldots L_n$ $(n \geq 1)$ with $L_1, \ldots, L_n \in A^+\mathcal{B}_k\}$.

For technical reasons, only nonempty words over A are considered to define this hierarchy; in particular, the complement operation is applied with respect to A^+. The language classes $A^+\mathcal{B}_{\emptyset}$, $A^+\mathcal{B}_1$, ... form the so-called dot-depth hierarchy introduced by Cohen and Brzozowski [CB71]. The union of the classes $A^+\mathcal{B}_{\emptyset}$, $A^+\mathcal{B}_1$, ... is the class of star-free languages.

Most of our attention will be directed toward a closely related hierarchy, this one in A^*. It was introduced by Straubing [St85]. Let

$A^*V_0 = \{\emptyset, A^*\}$,

$A^*V_{k+1} = \{L \subseteq A^* \mid L$ is a boolean combination of languages of the form $L_0a_1L_1a_2 \ldots a_nL_n$ $(n \geq 0)$ with $L_0, \ldots, L_n \in A^*V_k$ and $a_1, \ldots, a_n \in A\}$.

Let $A^*V = \bigcup_{k\geq 0} A^*V_k$. $L \subseteq A^*$ is star-free if and only if $L \in A^*V_k$ for some $k \geq 0$. The *dot-depth* of L is the smallest such k. The Straubing hierarchy appears to be the more fundamental of the two for the following reasons explained in [St86]. From the semigroup point of view, if $k \geq 1$, level k of the Straubing hierarchy corresponds to the variety of finite monoids consisting exactly of those in the variety of finite semigroups corresponding to level k of the dot-depth hierarchy. From the logical point of view, the levels of the Straubing hierarchy are exactly those defined by sentences of a first order language simpler than the one required for the levels of the dot-depth hierarchy. For more details concerning the Straubing hierarchy and its relation to the dot-depth hierarchy, see Pin [Pi84a] or [Pi84b].

In the framework of semigroup theory, Brzozowski and Knast [BK78] showed that the dot-depth hierarchy is infinite, in fact, that

$A^+\mathcal{B}_{k+1} \supset A^+\mathcal{B}_k$ but $A^+\mathcal{B}_{k+1} \neq A^+\mathcal{B}_k$ for $k \geq 0$. Thomas [Tho84] gave a new proof of this result, which shows also that the Straubing hierarchy is infinite, based on a logical characterization of the dot-depth hierarchy that he obtained in [Tho82]. His proof does not rely on semigroup theory; instead, an intuitively appealing model theoretic technique was applied: the Ehrenfeucht-Fraisse game.

It was the work of Büchi [Bü60] and Elgot [El61] that first showed how to use certain formulas of mathematical logic in order to describe properties of regular languages. These formulas, known as monadic second order formulas, are built up from variables x, y, ..., set variables X, Y, ..., a 2-place predicate symbol $<$ and a set $\{Q_a \mid a \in A\}$ of 1-place predicate symbols in one-to-one correspondence with the alphabet A. Starting with atomic formulas of the form $x = y$, $x < y$, $Q_a x$ and Xx, formulas are built up in the usual way by means of the connectives $\neg$, $\vee$, $\wedge$ and the quantifiers $\exists$ and $\forall$ binding up both types of variables. A word w on A satisfies a sentence φ if φ is true when variables are interpreted as integers, set variables as sets of integers, the predicate $<$ as the usual relation on integers and the formula $Q_a x$ as the letter in position x in w is an a.

Ladner [Lad77] and McNaughton [Mc74] were the first to consider the case where the set of formulas is restricted to first order, that is, when set variables are ignored. They proved that the languages defined in this way are precisely the star-free languages.

Thomas [Tho82] showed that the dot-depth hierarchy corresponds in a very natural way with a classical hierarchy of first order logic based on the alternation of existential and universal quantifiers. Perrin and Pin [PP86] gave a substantially different proof of the result of Thomas for the Straubing hierarchy.

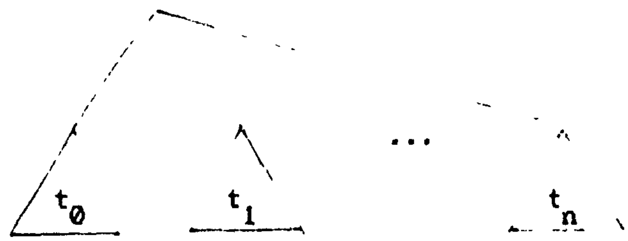
For each $k \geq 0$, there is a variety $\mathbf{V}_k$ of finite monoids such that for $L \subseteq A^*$, $L \in A^*\mathbf{V}_k$ if and only if $M(L) \in \mathbf{V}_k$. An outstanding open problem is whether one can decide if a language has dot-depth k. This is equivalent to the question "is $\mathbf{V}_k$ decidable?", i.e., does there exist an algorithm which enables us to test if a finite monoid is or is not in $\mathbf{V}_k$? The variety $\mathbf{V}_0$ consists of the trivial monoid alone. The variety $\mathbf{V}_1$ consists of all finite $\mathcal{J}$-trivial monoids [Si75]. Straubing [St86] conjectured an effective criterion, based on the syntactic monoid of the language, for the case $k = 2$. His condition is shown to be necessary in general, and sufficient in an important special case, i.e., for an alphabet of two elements. The condition is formulated in terms of a novel use of categories in semigroup theory, recently developed by Tilson [Ti87].

This thesis is concerned with the decidability problem of the Straubing hierarchy, i.e., can we effectively characterize the varieties $\mathbf{V}_k$? The aim of chapter two is to state those logical characterizations of the star-free languages which are useful in attacking the decidability question. In section one, a logical characterization of natural subhierarchies of the Straubing hierarchy refining the logical characterization of the hierarchy by Thomas is given. This logical characterization is useful when treating the question whether dot-depth is computable. As an application we can get upper bounds on the complexity of a star-free language by considering its description in the first order logical language. In section two, we state the version of the Ehrenfeucht-Fraisse game which was used in Thomas [Tho84] to prove that the Straubing hierarchy is infinite. For a sequence $\bar{m} = (m_1,...,m_k)$ of positive integers, congruences $\sim_{(m_1,...,m_k)}$ related to that version of the game are defined. Then we give a characterization of the star-free languages of level k in terms of the congruences $\sim_{(m_1,...,m_k)}$ generalizing a result of Simon [Si72]. A characterization of the varieties of monoids related to the Straubing hierarchy through Eilenberg's correspondence follows. Subclasses $\mathcal{L}_{(m_1,...,m_k)}$ of languages of level k are defined.

In chapter three, we study some properties of the characterizing congruences. Section one establishes an induction lemma. Section two gives a condition which insures $\mathcal{L}_{(m_1,...,m_k)}$ is included in $\mathcal{L}_{(m'_1,...,m'_{k'})}$.

Chapter four deals with a first application of the above logical characterizations. We show that a conjecture of Pin concerning tree hierarchies of monoids (the dot-depth and the Straubing hierarchies being particular cases) is false. More precisely, $\{\emptyset, A^*\}$ is associated to the tree reduced to a point. Then to the tree

t =



is associated the boolean algebra V_t generated by the languages of the form $L_{i_0} a_1 L_{i_1} a_2 \ldots a_r L_{i_r}$ with $0 \le i_0 < \ldots < i_r \le n$ where, for $0 \le j \le r$, $L_{i_j} \in V_{t_{i_j}}$. Pin [Pi84b] conjectured that $V_t \subseteq V_{t'}$, if and only if t is extracted from t'. Decidability and inclusion problems are discussed. $\mathcal{L}_{(m_1, \ldots, m_k)}$ are shown to be decidable.

Chapter five is concerned with a second application of the mentioned logical characterizations. Games are shown to be a way in verifying equations which are used for finding lower bounds on the dot-depth of a given star-free language or a star-free language's complexity. We define systems of equations satisfied in the monoid varieties of sublevels of level 1 of the Straubing hierarchy. ([Kn83a], [Kn83b] provide an equation system for level 1 of the dot-depth hierarchy without using Ehrenfeucht-Fraisse games). In a few cases, we show that these equation systems characterize the sublevels. In particular, the third sublevel is characterized by the equations $(xy)^3 = (yx)^3$, $xzyxvxwy = xzxyxvxwy$ and $ywxvxyzx = ywxvxyxzx$, or, equivalently, by the equations $(xy)^3 = (yx)^3$, $xuxvx = xux^2vx$, $xzyx^2wy = xzxyx^2wy$ and $ywx^2yzx = ywx^2yxzx$. We show how some of the equations can be selected for sublevels of higher levels in the hierarchy. Equations satisfied in higher levels are discussed.

Other applications of the mentioned logical characterizations are the subject of chapter six. Given any finite alphabet A, a necessary and sufficient condition is given for the monoids $A^*/\sim_{(m_1,m_2,m_3)}$ to be of dot-depth exactly 2. An equational characterization of the first sublevel of level 2 of the Straubing hierarchy for an alphabet of two letters is given.

In the following, notation and basic concepts are introduced.

1. *Algebraic preliminaries*

For more information on the matters discussed in this section, see the books by Eilenberg [Ei76], Lallement [La179] and Pin [Pi84a].

A *semigroup* is a set equipped with an associative binary operation (generally denoted multiplicatively). A *monoid* is a set M equipped with an associative binary operation and a 0-ary operation , denoted by 1, such that for all $x \in M$, $1x = x1 = x$. A *group* is a set M equipped with an associative binary operation and a 0-ary operation as above, such that for all $x \in M$, there exists $x' \in M$ satisfying $xx' = x'x = 1$. If M' is a semigroup, then $M \subseteq M'$ is a *subsemigroup* if $M^2 \subseteq M$. If M also has an identity, then M is a monoid in M'. M is a *submonoid* of M' if $M' \supseteq M$ are both monoids, with the same identity. We say that M *divides* M', $M \prec M'$, if M is a morphic image of a submonoid of M'. All the semigroups considered in this thesis are finite (except for free semigroups and free monoids). M is *aperiodic* if every group in M is a trivial one element group, or, if there exists n such that $x^n = x^{n+1}$ for all $x \in M$. If M is a monoid and m_1, $m_2 \in M$, then m_1 is said to be $\mathcal{J}$-below m_2, written $m_1 \leq_{\mathcal{J}} m_2$, if $m_1 = xm_2y$ for some x, $y \in M$; m_1 and m_2 are said to be $\mathcal{J}$-equivalent, written $m_1 \sim_{\mathcal{J}} m_2$, if $m_1 \leq_{\mathcal{J}} m_2$ and $m_2 \leq_{\mathcal{J}} m_1$. M is said to be *$\mathcal{J}$-trivial* if this equivalence relation is the identity.

Let A be a finite set. A^*, the *free monoid* generated by A, is the set of all sequences of length ≥ 0 of elements of A with concatenation being the operation (such sequences are called words). The unique string of length 0, denoted by 1 and called the empty word, acts as the identity. A *language* over A is a subset of A^*. $|w|$ denotes the length of the word w, and $|w|_a$ the number of occurrences of the letter a in w. $w\alpha$ denotes the set of letters in w. A word u is a *prefix* of w if there exists a word v such that $uv = w$. A word u is a *suffix* of w if there exists a word v such that $vu = w$. A word u is a *factor* (or *segment*) of a word v if there exist words x and y such that $v = xuy$. A word $u = a_1 \ldots a_n$ (where $a_1, \ldots, a_n$ are letters) is a *subword* of v if there exist words $v_0, \ldots, v_n$ such that $v = v_0 a_1 v_1 a_2 \ldots a_n v_n$.

An equivalence $\sim$ on A^* is a *congruence* if $x \sim y$ implies $uxv \sim uyv$ for all $u, v, x, y \in A^*$. A congruence $\sim$ is *aperiodic* if there exists $n \geq 0$ such that $x^n \sim x^{n+1}$, for all x. The $\sim$*-class* of x is $[x]_\sim = \{y \mid x \sim y\}$. The set of all $\sim$-classes is denoted by $A^*/\sim$ and the *index* of $\sim$ is defined as the cardinality of $A^*/\sim$. This set becomes a monoid by considering the operation $[x]_\sim [y]_\sim = [xy]_\sim$; $[1]_\sim$ acts as identity. There exists a surjective morphism $\sim : A^* \to A^*/\sim$, defined by $x\sim = [x]_\sim$. Conversely, any morphism $\varphi : A^* \to M$ induces a congruence on A^* defined by $x \varphi y$ if and only if $x\varphi = y\varphi$. Note that we use the same symbol to denote the congruence and the related morphism. If φ is surjective, there exists an isomorphism between A^*/φ and M. Any monoid can then be represented as a quotient of A^* by a congruence.

If $L \subseteq A^*$ is a union of $\sim$-classes, we say that L is a *$\sim$-language*. For any language L over A, the *syntactic congruence* of L is defined by $x \sim_L y$ if and only if for all $u, v \in A^*$, $uxv \in L$ if and only if $uyv \in L$. $\sim_L$ is the congruence of minimal index with the property that L is a $\sim$-language, i.e., for any congruence $\sim$ on A^*, L is a $\sim$-language if and only if $\sim \subseteq \sim_L$. The quotient monoid $A^*/\sim_L$ is denoted $M(L)$ and is called the *syntactic monoid* of L. If M is a monoid and there exists a morphism $\varphi : A^* \to M$ such that $L = S\varphi^{-1}$ for some $S \subseteq M$, we say that M *recognizes* L. We also say, in such an instance, that the morphism φ recognizes L. A language is said to be *recognizable* if it is recognized by a finite monoid. It is not difficult to see that recognition by a finite monoid is equivalent to recognition by a finite automaton, so Kleene's theorem asserts that the regular languages in A^* are exactly those recognized by finite monoids. It is well known that $M(L)$ is the monoid M of minimal cardinality with the property that M recognizes L; in fact, $M(L) \prec M$ if and only if M recognizes L. Moreover, if $\sim_1$ and $\sim_2$ are two congruences on A^* and if $\sim_1 \subseteq \sim_2$, then $A^*/\sim_2 \prec A^*/\sim_1$. Also L is regular if and only if $\sim_L$ has finite index if and only if $M(L)$ is finite.

W is a *variety of monoids* (the term variety is being used in a slightly different sense that the usual), or M-variety, if
(1) it is a class of finite monoids closed under division, i.e., if $M \in \mathbf{W}$ and $M' < M$, then $M' \in \mathbf{W}$, and
(2) it is closed under finite direct product, i.e., if $M, M' \in \mathbf{W}$, then $M \times M' \in \mathbf{W}$.

For any class C of finite monoids, we denote by $\langle C \rangle_M$ the least M-variety containing C. Clearly, $M \in \langle C \rangle_M$ if and only if there exists a finite sequence $M_1, \ldots, M_n$ of monoids of C such that $M < M_1 \times \ldots \times M_n$. We call $\langle C \rangle_M$ the M-variety generated by C.

Eilenberg [Ei76] has shown that there exists a one-to-one correspondence between M-varieties and some classes of recognizable languages called *-varieties. $\mathcal{W}$ is a **-variety of languages* if
(1) for every finite alphabet A, $A^*\mathcal{W}$ denotes a class of recognizable languages over A closed under boolean operations,
(2) if $L \in A^*\mathcal{W}$ and $a \in A$, then $a^{-1}L = \{w \in A^* \mid aw \in L\}$ and $La^{-1} = \{w \in A^* \mid wa \in L\}$ are in $A^*\mathcal{W}$, and
(3) if $L \in A^*\mathcal{W}$ and $\varphi : B^* \to A^*$ is a morphism, then $L\varphi^{-1} = \{w \in B^* \mid \varphi(w) \in L\} \in B^*\mathcal{W}$.
To a given *-variety of languages $\mathcal{W}$ corresponds the M-variety $\mathbf{W} = \{M(L) \mid L \in A^*\mathcal{W} \text{ for some } A\}$ and to a given M-variety $\mathbf{W}$ corresponds the *-variety of languages $\mathcal{W}$ where $A^*\mathcal{W} = \{L \subseteq A^* \mid$ there is $M \in \mathbf{W}$ recognizing $L\}$. The notion of variety captures the conditions under which a family of languages can be characterized by

monoids and vice versa.

The Straubing hierarchy gives examples of $*$-varieties of languages. One can show that $\mathcal{V}$ and $\mathcal{V}_k$ are $*$-varieties of languages. Let the corresponding M-varieties be denoted by $\mathbf{V}$ and $\mathbf{V}_k$ respectively. $\mathbf{V}$ is the M-variety of aperiodic monoids. We have that for $L \subseteq A^*$, $L \in A^*\mathcal{V}$ if and only if $M(L) \in \mathbf{V}$ and for each $k \geq 0$, $L \in A^*\mathcal{V}_k$ if and only if $M(L) \in \mathbf{V}_k$.

2. *Logical preliminaries*

For more information on the matters discussed in this section, see the book by Enderton [En72].

We assume that we have been given infinitely many distinct objects, which we call symbols, arranged as follows:

Logical symbols

(1) parentheses: (,),

(2) sentential connective symbols: $\neg$, $\vee$, $\wedge$,

(3) variables: x, y, ...,

(4) equality symbol (optional): $=$.

Nonlogical symbols

(1) quantifier symbols: $\exists$, $\forall$,

(2) predicate symbols: for each positive integer n, some set (possibly empty) of symbols, called n-place predicate symbols,

(3) constant symbols: some set (possibly empty) of symbols,

(4) function symbols: for each positive integer n, some set (possibly empty) of symbols, called n-place function symbols.

The equality symbol is a 2-place predicate symbol but is distinguished from the other 2-place predicate symbols by being a logical symbol rather than a nonlogical one. We do assume that some n-place predicate symbol is present for some n.

A (finite) *similarity type*, $\tau = \langle P_1, \ldots, P_i, c_1, \ldots, c_j, f_1, \ldots, f_k \rangle$, is a sequence of predicate symbols, constant symbols and function symbols. If τ is any type, then $\mathcal{L}[\tau]$, the *first order language* of τ, is the set of all formulas built up from the symbols of τ using the connective symbols, variables and the quantifiers $\exists$ and $\forall$. More precisely, the *terms* are constant symbols, variables or of the form $ft_1 \ldots t_n$, where f is an n-place function symbol and t_1, ..., t_n are terms. The *atomic formulas* are of the form $Pt_1 \ldots t_n$, where P is an n-place predicate symbol and t_1, ..., t_n are terms. The *formulas* are built up from the atomic formulas by use of the connective symbols and the quantifiers.

$\varphi(x_1, \ldots, x_n)$ will denote a formula in which x_1, ..., x_n are the only free variables. If no variable occurs free in the formula φ, then φ is a *sentence*. $\varphi \to \psi$ will abbreviate $\neg\varphi \vee \psi$ and $\varphi \leftrightarrow \psi$ will abbreviate $(\varphi \to \psi) \wedge (\psi \to \varphi)$.

A *structure* $\mathfrak{U} = \langle U, P_1^{\mathfrak{U}}, \ldots, P_i^{\mathfrak{U}}, c_1^{\mathfrak{U}}, \ldots, c_j^{\mathfrak{U}}, f_1^{\mathfrak{U}}, \ldots, f_k^{\mathfrak{U}} \rangle$ for a given first order language $\mathcal{L}[\tau]$, $\tau = \langle P_1, \ldots, P_i, c_1, \ldots, c_j, f_1, \ldots, f_k \rangle$, consists of

(1) a non-empty set U, called the *universe* of $\mathfrak{U}$,

(2) an n-ary relation $P^{\mathfrak{U}} \subseteq U^n$, i.e., $P^{\mathfrak{U}}$ is a set of n-tuples of members of the universe, for each n-place predicate symbol P,

(3) a member $c^{\mathfrak{U}}$ of the universe, for each constant symbol c,

(4) an n-ary operation $f^{\mathfrak{U}}$ on U, i.e., $f^{\mathfrak{U}} : U^n \to U$, for each

n-place function symbol f.
The idea is that $\mathfrak{U}$ assigns meaning to the nonlogical symbols. $\forall$ is to mean "for everything in U". The symbol c is to name the point $c^{\mathfrak{U}}$. The atomic formula $Pt_1 \ldots t_n$ is to mean that the n-tuple of points named by $t_1, \ldots, t_n$ is in the relation $P^{\mathfrak{U}}$. The number of elements in the universe of $\mathfrak{U}$ is abbreviated $|\mathfrak{U}|$.

A sentence φ in $\mathcal{L}[\tau]$ is given meaning by a structure $\mathfrak{U}$ of type τ as follows: the symbols from τ are interpreted by the relations, constants and operations in $\mathfrak{U}$. The quantifiers in φ range over the elements of the universe U. φ is true in $\mathfrak{U}$, or $\mathfrak{U}$ is a *model* of φ, is denoted $\mathfrak{U} \models \varphi$. Two formulas φ and ψ are called equivalent if $\mathfrak{U} \models \varphi \leftrightarrow \psi$ for all structures $\mathfrak{U}$, or equivalently, $\mathfrak{U} \models \varphi$ if and only if $\mathfrak{U} \models \psi$ for all structures $\mathfrak{U}$.

The *quantifier depth* of a sentence φ, $qd(\varphi)$, is the depth of nesting of quantifiers in φ. Inductively:

$$qd((\forall x)\varphi) = qd((\exists x)\varphi) = qd(\varphi)+1,$$

$$qd(\varphi \vee \psi) = qd(\varphi \wedge \psi) = \max(qd(\varphi),qd(\psi)), \text{ and}$$

$$qd(\neg\varphi) = qd(\varphi).$$

A formula φ is in *prenex normal form* if $\varphi = (Q\bar{x})\psi$, where $(Q\bar{x})$ is a string of quantifiers $\exists x_i$, $\forall x_i$, and ψ is quantifier-free. If the prefix $(Q\bar{x})$ consists of k alternating blocks of quantifiers such that the first block contains only existential quantifiers, the second block only universal ones, etc., and each block is nonempty, then $(Q\bar{x})\Psi$ is a Σ_k-formula. Similarly, if $(Q\bar{x})$ consists of k blocks beginning with a block of universal quantifiers, $(Q\bar{x})\Psi$ is a Π_k-formula, thus the $\Sigma_\emptyset$-formulas and the $\Pi_\emptyset$-formulas are the quantifier-free formulas. Any formula is equivalent to one in prenex normal form. The rules needed for this transformation are given by

(1) a negation of a Σ_k-formula is equivalent to a Π_k-formula,

(2) a disjunction or conjunction of Σ_k-formulas is equivalent to a Σ_k-formula,

(3) a boolean combination of Σ_k-formulas, or $B(\Sigma_k)$-formula, is equivalent to a Σ_{k+1}-formula,

(4) the statements (1)-(3) hold in dual form for Π_k-formulas.

Chapter 2

SOME LOGICAL CHARACTERIZATIONS OF THE STRAUBING HIERARCHY

1. *A quantifier complexity characterization*

Let us first state the mentioned logical characterization of the Straubing hierarchy by Thomas. One identifies any word $w \in A^*$, say of length $|w|$, with a word model $\underline{w} = \langle\{1,\ldots,|w|\}, <^w, (Q_a^w)_{a\in A}\rangle$ where the universe $\{1,\ldots,|w|\}$ represents the set of positions of letters in the word w, $<^w$ denotes the $<$-relation in w, and Q_a^w are unary relations over $\{1,\ldots,|w|\}$ containing the positions with letter a, for each $a \in A$. Sometimes it is convenient to assume that the position sets of two words u, v are disjoint; then one takes any two nonoverlapping segments of the integers as the position sets of u and v. Let $\mathcal{L}$ be the first order language with equality and nonlogical symbols $<$, Q_a, $a \in A$, i.e., $\mathcal{L} = \mathcal{L}[\tau]$ where τ is the similarity type $\langle =, <, (Q_a)_{a\in A}\rangle$. Then the satisfaction of an $\mathcal{L}$-sentence φ in a word w, written $w \models \varphi$, is defined in a natural way, and we say that $L \subseteq A^*$ is defined by the $\mathcal{L}$-sentence φ if $L = L(\varphi) = \{w \in A^* \mid w \models \varphi\}$. We also consider the formulas 0 (false) and 1 (true). Observe that $L(0) = \emptyset$ and $L(1) = A^*$.

Theorem 2.1.1 Thomas [Tho82]

A language $L \subseteq A^*$ belongs to A^*V_k if and only if L is defined by a $B(\Sigma_k)$-sentence of $\mathcal{L}$.

Corollary 2.1.2 Ladner [Lad77] and McNaughton [MP71]

A language L is star-free if and only if there exists a first order $\mathcal{L}$-sentence φ such that $L = L(\varphi)$.

For $k \geq 1$, let us define subhierarchies of A^*V as follows: for all $m \geq 1$, let

$A^*V_{k,m} = \{L \subseteq A^* \mid L$ is a boolean combination of languages of the form $L_0a_1L_1a_2\ldots a_nL_n$ $(0 \leq n \leq m)$ with $L_0, \ldots, L_n \in A^*V_{k-1}$ and $a_1, \ldots, a_n \in A\}$.

We have $A^*V_k = \bigcup_{m\geq 1} A^*V_{k,m}$. Easily, $A^*V_k \subseteq A^*V_{k+1,m}$, $A^*V_{k,m} \subseteq A^*V_{k,m+1}$. Similarly, subhierarchies of $A^+\mathcal{B}_k$ can be defined. One can show that $V_{k,m}$ is a $*$-variety of languages. Let the corresponding M-varieties be denoted by $\mathbf{V}_{k,m}$. We have that for $k \geq 1$, $m \geq 1$, $L \in A^*V_{k,m}$ if and only if $M(L) \in \mathbf{V}_{k,m}$.

In $A^+\mathcal{B}_1$ several hierarchies and classes of languages have been studied; the most prominent examples are the β-hierarchy [BS73], also called depth-one finite cofinite hierarchy, and the class of locally testable languages. In Thomas [Tho82] it was shown that both are characterized by natural restrictions on the form of Σ_1-sentences of a certain first order language extending $\mathcal{L}$.

The purpose of this section is to give a logical characterization which follows from an analysis of the proof of theorem 2.1.1 of the subhierarchies of A^*V refining the theorem of Thomas. It will be useful to extend $\mathcal{L}$ by adding constant symbols s, for every natural number s. For a word model w, the interpretation s^w of s will be the s^{th} element of w. Let $\varphi(x_1,\ldots,x_m)$ be a formula in which $x_1, \ldots, x_m$ are the only free variables. Let $s_1, \ldots, s_m$ be positive integers. The meaning and usage of $\varphi(s_1,\ldots,s_m)$ should be quite clear in what follows. $\varphi(s_1,\ldots,s_m)$ is obtained from $\varphi(x_1,\ldots,x_m)$ by replacing simultaneously all free occurrences of x_1 in φ by the constant $s_1, \ldots, x_m$ by s_m. The interpretation of the formula $\varphi(\bar{x}) = \varphi(x_1,\ldots,x_m)$ in a word model w with universe $\{1,\ldots,|w|\}$ and elements $s_1, \ldots, s_m \in \{1,\ldots,|w|\}$ is defined in the natural way; we write $w \models \varphi(s_1,\ldots,s_m)$ if φ is satisfied in w when interpreting x_i by s_i for $1 \leq i \leq m$.

A logical characterization of the subhierarchies of A^*V is based on the following two lemmas. In what follows, if $w = a_1\ldots a_n$ is a word and $1 \leq s \leq s' \leq n$, $w[s,s']$, $w(s,s')$, $w(s,s']$ and $w[s,s')$ will denote respectively the segments $a_s\ldots a_{s'}$, $a_{s+1}\ldots a_{s'-1}$, $a_{s+1}\ldots a_{s'}$ and $a_s\ldots a_{s'-1}$.

Lemma 2.1.3 Perrin and Pin [PP86]

For $k \geq 0$ and for each $B(\Sigma_k)$-sentence φ, there exist $B(\Sigma_k)$-formulas $\varphi_l(x)$, $\varphi_r(x)$, $\varphi_m(x,y)$ in which x (x,y) is (are) the only free variable(s) and such that for every n and for every word w of length n we have

(1) $w \in L(\varphi_l(s))$ if and only if $w[1,s) \in L(\varphi)$, and

(2) $w \in L(\varphi_r(s))$ if and only if $w(s,|w|] \in L(\varphi)$ for every integer s such that $1 \leq s \leq n$, and

(3) $w \in L(\varphi_m(s,s'))$ if and only if $w(s,s') \in L(\varphi)$ for every integers s, s' such that $1 \leq s < s' \leq n$.

Proof We define φ_m for every formula φ. φ_m is constructed by induction as follows (the constructions are similar for φ_l and φ_r): if φ is quantifier-free, then $\varphi_m = \varphi$. Otherwise, we set

$$(\exists z\varphi)_m = \exists z\ ((x < z < y) \wedge \varphi_m),$$

$$(\forall z\varphi)_m = \forall z\ ((x < z < y) \rightarrow \varphi_m),$$

$$(\neg\varphi)_m = \neg\varphi_m,$$

$$(\varphi \vee \psi)_m = \varphi_m \vee \psi_m,$$

$$(\varphi \wedge \psi)_m = \varphi_m \wedge \psi_m.$$

Then one can verify by induction on $k \geq 0$ the following properties:

(1) if φ and ψ are equivalent formulas, then φ_m and ψ_m are equivalent,

(2) if φ is $B(\Sigma_k)$, then φ_m is equivalent to a $B(\Sigma_k)$-formula,

(3) let φ be a sentence. If $|w| = n$ and if $1 \leq s < s' \leq n$, w satisfies $\varphi_m(s,s')$ if and only if $w(s,s')$ satisfies φ.[]

Lemma 2.1.4

Given a $B(\Sigma_k)$-formula $\varphi(x_1,\dots,x_n)$ $(n \geq 1)$, there is a system $\langle \bar{L}^j \rangle_{j<p}$ of sequences $\bar{L}^j = \langle L_\emptyset^j,\dots,L_n^j \rangle$ of languages $L_i^j \in A^*V_k$ and $\langle \bar{a}^j \rangle_{j<p}$ of sequences $\bar{a}^j = \langle a_1^j,\dots,a_n^j \rangle$, $a_i^j \in A$ such that for any w and $s_1 < \dots < s_n$ in $\{1,\dots,|w|\}$, $w \models \varphi(s_1,\dots,s_n)$ if and only if there is $j < p$ such that

(1) $w[1,s_1) \in L_\emptyset^j$ and $Q^w_{a_1^j}\, s_1$,

(2) $w(s_i,s_{i+1}) \in L_i^j$ and $Q^w_{a_{i+1}^j}\, s_{i+1}$, $1 \leq i < n$, and

(3) $w(s_n,|w|] \in L_n^j$.

Proof By induction on k (see the proof of theorem 2.1.1 [Tho82]. If $n = \emptyset$, this is just theorem 2.1.1).[]

Let φ be an $\mathcal{L}$-sentence. If φ is a boolean combination of the Σ_k-sentences φ_1, ..., φ_n, define the *quantifier rank* $qr(\varphi)$ to be the maximum number of quantifiers occurring in the leading block of one of the formulas φ_1, ..., φ_n. Let us now prove a refinement of Thomas' theorem.

Theorem 2.1.5

Let $k \geq 1$, $m \geq 1$. A language $L \subseteq A^*$ is defined by a $B(\Sigma_k)$-sentence of $\mathcal{L}$, φ, where $qr(\varphi) \leq m$ if and only if L belongs to $A^*V_{k,m}$.

Proof The case $k = 1$ is the following. Let $m \geq 1$. Let L be a language of the form $A^*a_1A^*a_2\dots a_mA^*$ where $a_i \in A$, $i = 1$, ..., m.

We have to find a boolean combination φ of Σ_1-sentences defining L such that $qr(\varphi) \leq m$. The assertion $w \in L$ can be expressed by a Σ_1-sentence as follows:

$\exists x_1 \exists x_2 \ldots \exists x_m \; (x_1 < x_2 < \ldots < x_m \wedge Q_{a_1} x_1 \wedge Q_{a_2} x_2 \wedge \ldots \wedge Q_{a_m} x_m)$. Hence L is defined by a sentence of the required form.

Conversely, we show that a given Σ_1-sentence $\exists x_1 \ldots \exists x_m \; \varphi(\bar{x})$ defines a language in $A^* V_{1,m}$, where $\varphi(\bar{x})$ is equivalent to a conjunction of atomic formulas of the form $Q_a x$, $x < y$ or $x = y$ (for x, y variables and $a \in A$) or their negation. Let $ord_1(\bar{x}), \ldots, ord_r(\bar{x})$ be the conjunctions saying $x_{i_1} \leqq \ldots \leqq x_{i_m}$, where $\{i_1, \ldots, i_m\} = \{1, \ldots, m\}$. Then $\exists \bar{x} \; \varphi(\bar{x})$ is equivalent to $\bigvee_{1 \leq i \leq r} \exists \bar{x} \; (ord_i(\bar{x}) \wedge \varphi(\bar{x}))$. Let us consider a typical member of this disjunction, say $\exists \bar{x} \; (x_1 < \ldots < x_m \wedge \varphi(\bar{x}))$ (identify variables if equalities occur between the x_i's). It suffices to show that the language L defined by $\psi = \exists \bar{x} \; (x_1 < \ldots < x_m \wedge \varphi(\bar{x}))$ is in $A^* V_{1,m}$. But ψ defines either $\emptyset$ or is equivalent to a disjunction of formulas of the form $\exists \bar{x} \; (x_1 < \ldots < x_m \wedge \varphi'(\bar{x}))$ where $\varphi'(\bar{x})$ is a conjunction of atomic formulas of the form $Q_a x$, $\neg Q_a x$ for x a variable and $a \in A$. In either case, L is easily seen to belong to $A^* V_{1,m}$. For example, $L(\exists x \; Q_a x) = A^* a A^*$, $L(\exists x \; \neg Q_a x) = \bigcup_{b \in A, b \neq a} A^* b A^*$, $L(\exists y \exists z \; (y < z \wedge Q_a y \wedge Q_b z)) = A^* a A^* b A^*$ and

$L(\exists y \exists z \; (\neg(y < z) \wedge Q_a y \wedge \neg Q_b z))$

$= L(\exists y \; (Q_a y \wedge \neg Q_b y)) \cup L(\exists y \exists z \; (z < y \wedge Q_a y \wedge \neg Q_b z))$.

Now let us assume $k > 1$, $m \geq 1$. Let L be a language of the form $L_0 a_1 L_1 a_2 \ldots a_m L_m$ where $a_i \in A$, $L_i \in A^* V_{k-1}$, $i = 0, \ldots, m$.

We have to find a boolean combination φ of Σ_k-sentences defining L such that $qr(\varphi) \le m$. By Thomas' theorem 2.1.1, let $\varphi^{\emptyset}, \varphi^1, \ldots, \varphi^m$ be $B(\Sigma_{k-1})$-sentences defining $L_{\emptyset}, L_1, \ldots, L_m$ respectively. We can find $B(\Sigma_{k-1})$-formulas $\varphi_1^{\emptyset}(x), \varphi_m^1(x,y), \varphi_m^2(x,y), \ldots, \varphi_r^m(x)$ satisfying lemma 2.1.3. Hence the assertion $w \in L$ can be expressed by the following sentence:

$$\exists x_1 \exists x_2 \ldots \exists x_m (x_1 < x_2 < \ldots < x_m \wedge Q_{a_1} x_1 \wedge$$

$$Q_{a_2} x_2 \wedge \ldots \wedge Q_{a_m} x_m \wedge \varphi_1^{\emptyset}(x_1) \wedge \varphi_m^1(x_1,x_2) \wedge \varphi_m^2(x_2,x_3) \wedge \ldots \wedge \varphi_r^m(x_m)),$$

which is easily seen to be equivalent to a $B(\Sigma_k)$-sentence of the required form.

Conversely, consider a Σ_k-sentence $\exists x_1 \ldots \exists x_m \varphi(\bar{x})$, where $\varphi(\bar{x})$ is a $B(\Sigma_{k-1})$-formula. As in the proof of the case $k = 1$, $m \ge 1$, it suffices to consider a Σ_k-sentence of the form $\varphi = \exists x_1 \ldots \exists x_m (x_1 < \ldots < x_m \wedge \varphi(\bar{x}))$. Then, by lemma 2.1.4, there is a system $\langle \bar{L}^j \rangle_{j<p}$ of sequences $\bar{L}^j = \langle L_{\emptyset}^j, \ldots, L_m^j \rangle$ of languages $L_i^j \in A^*V_{k-1}$ and $\langle \bar{a}^j \rangle_{j<p}$ of sequences $\bar{a}^j = \langle a_1^j, \ldots, a_m^j \rangle$, $a_i^j \in A$ such that for any w and $s_1 < \ldots < s_m$ in $\{1, \ldots, |w|\}$, $w \models \varphi(s_1, \ldots, s_m)$ if and only if there is $j < p$ such that $w \in L_{\emptyset}^j a_1^j L_1^j a_2^j \ldots a_m^j L_m^j$. But for every $j < p$, $L_{\emptyset}^j a_1^j L_1^j a_2^j \ldots a_m^j L_m^j \in A^*V_{k,m}$. Hence φ defines a boolean combination of languages of the required form and the proof is complete.[]

2. *A congruence characterization related to a version of the Ehrenfeucht-Fraisse game*

Thomas [Tho84], in order to show that the dot-depth hierarchy is infinite, defined some congruences which we state after describing the version of the Ehrenfeucht-Fraisse game which was used in his proof. Those congruences will be shown to characterize the star-free languages. The next three paragraphs restate [Tho84].

First we define what we mean by $\bar{m}$-formulas of $\mathcal{L}$. For a sequence $\bar{m} = (m_1,\ldots,m_k)$ of positive integers, where $k \geq 0$, let $\text{length}(\bar{m}) = k$ and $\text{sum}(\bar{m}) = m_1+\ldots+m_k$. The set of $\bar{m}$-formulas of $\mathcal{L}$ is defined by induction on $\text{length}(\bar{m})$: if $\text{length}(\bar{m}) = 0$, it is the set of quantifier-free $\mathcal{L}$-formulas; and for $\bar{m} = (m,m_1,\ldots,m_k)$, an $\bar{m}$-formula is a boolean combination of formulas $\exists x_1\ldots\exists x_m\, \varphi$ where φ is an $(m_1,\ldots,m_k)$-formula. We write $u \equiv_{\bar{m}} v$ if u and v satisfy the same $\bar{m}$-sentences of $\mathcal{L}$. For $\bar{m} = (m_1,\ldots,m_k)$, the $\bar{m}$-formulas of $\mathcal{L}$ are seen to be $B(\Sigma_k)$-formulas φ such that $qr(\varphi) \leq m_1$. Moreover, languages in $A^*V_{k,m}$ are defined by $(m,m_2,\ldots,m_k)$-formulas for some m_i, $i = 2, \ldots, k$. The following game $G_{\bar{m}}(u,v)$ is useful for showing $\equiv_{\bar{m}}$-equivalence.

The game $G_{\bar{m}}(u,v)$, where $\bar{m} = (m_1,\ldots,m_k)$, is played between two players I and II on the word models u and v. A play of the game consists of k moves. In the i^{th} move, player I chooses, in u or in v, a sequence of m_i positions; then player II chooses, in the remaining word (v or u), also a sequence of m_i positions. After k moves, by concatenating the position sequences chosen from u and from v, two sequences of positions $p_1 \ldots p_n$ from u and $q_1 \ldots q_n$ from v have been formed where $n = \mathrm{sum}(\bar{m})$. Player II has won the play if

(1) $p_i <^u p_j$ if and only if $q_i <^v q_j$,

(2) $Q_a^u p_i$ if and only if $Q_a^v q_i$, $a \in A$ for $1 \le i, j \le n$.

Equivalently, the two subwords in u and v given by the position sequences $p_1 \ldots p_n$ and $q_1 \ldots q_n$ should coincide. If there is a winning strategy for II in the game to win each play we say that player II wins $\mathcal{G}_{\bar{m}}(u,v)$ and write $u \sim_{\bar{m}} v$. $\sim_{\bar{m}}$ naturally defines a congruence on A^* which we will denote also by $\sim_{\bar{m}}$.

The standard Ehrenfeucht-Fraisse game is the special case of $\mathcal{G}_{\bar{m}}(u,v)$ where $\bar{m} = (1,\ldots,1)$. For a detailed discussion see Rosenstein [Ro82] or Fraisse [Fr72]. If $\mathrm{length}(\bar{m}) = k$ and $\bar{m} = (1,\ldots,1)$ we write $\mathcal{G}_k(u,v)$ instead of $\mathcal{G}_{\bar{m}}(u,v)$ and $u \sim_k v$ instead of $u \sim_{\bar{m}} v$. Note that in this case the $\bar{m}$-formulas are up to equivalence just the formulas of quantifier depth k (remark: one should not confuse $\mathcal{G}_k(u,v)$ and $\mathcal{G}_{(k)}(u,v)$; a play of the game $\mathcal{G}_k(u,v)$ consists of k moves but a play of the game $\mathcal{G}_{(k)}(u,v)$ of 1 move). We have the following important

Theorem 2.2.1 Ehrenfeucht and Fraisse [Eh61]

For all $\bar{m} = (m_1, \ldots, m_k)$ with $k > 0$ and $m_i > 0$ for $i = 1, \ldots, k$, we have $u \equiv_{\bar{m}} v$ if and only if $u \sim_{\bar{m}} v$.

Simon [Si72] calls $\sim_{(m)}$-languages piecewise testable languages. They constitute level 1 of the Straubing hierarchy. The purpose of this section is to characterize similarly the hierarchy, each level of it and also each subhierarchy.

To do so, we use theorem 2.1.5 and follow the technique used in [Tho82]. For a word w, we can define, by induction on length($\bar{m}$), a sentence $\varphi_w^{\bar{m}}$ which in a certain sense guarantees the satisfaction of all $\bar{m}$-sentences of $\mathcal{L}$ which are satisfied by w. The following lemma says that each equivalence class of $\sim_{\bar{m}}$ is definable by some $\bar{m}$-sentence, more precisely, $[w]_{\sim_{\bar{m}}}$ is defined by $\varphi_w^{\bar{m}}$.

Lemma 2.2.2 Thomas [Tho82]

There is a formula $\varphi_w^{\bar{m}}$ such that

(1) $w \models \varphi_w^{\bar{m}}$,

(2) $\varphi_w^{\bar{m}}$ is equivalent to a $\bar{m}$-sentence,

(3) For all w and u, if $u \models \varphi_w^{\bar{m}}$ then every $\bar{m}$-sentence satisfied in w is also satisfied in u.

We have

Lemma 2.2.3

The following are equivalent:

(1) $L = L(\varphi)$ for some $\bar{m}$-sentence φ,

(2) L is closed under $\sim_{\bar{m}}$, i.e., if $u \in L$ and $u \sim_{\bar{m}} v$, then $v \in L$, and

(3) L is a $\sim_{\bar{m}}$-language.

Proof (1) implies (2) by theorem 2.2.1. (2) implies (3) is trivial. (3) implies (1) uses theorem 2.2.1, lemma 2.2.2 and the fact that $\sim_{\bar{m}}$ has only finitely many equivalence classes.[]

We can now prove

Theorem 2.2.4

Let $k \geq 1$, $m \geq 1$. $L \in A^*V_{k,m}$ if and only if L is a $\sim_{\bar{m}}$-language for some $\bar{m} = (m, m_2, \ldots, m_k)$.

Proof $L \in A^*V_{k,m}$ if and only if $L = L(\varphi)$ for some $\bar{m}$-sentence φ with $\bar{m} = (m, m_2, \ldots, m_k)$ for some fixed $m_2, \ldots, m_k$ by theorem 2.1.5. The result follows from lemma 2.2.3.[]

Corollary 2.2.5

Let $k \geq 1$. $L \in A^*V_k$ if and only if L is a $\sim_{\bar{m}}$-language for some $\bar{m} = (m_1, \ldots, m_k)$.

Corollary 2.2.6

L is star-free if and only if L is a $\sim_{\bar{m}}$-language for some $\bar{m}$.

We end this section with a few notes on the preceding corollaries. Corollary 2.2.6 states precisely which are the important congruences related to the study of star-free languages. Kleene's theorem [Kl56], stated in terms of congruences, asserts that L is regular if and only if there exists a finite index congruence $\sim$ such that L is a $\sim$-language. Schützenberger's theorem [Sc65] states that L is star-free if and only if there exists a finite index aperiodic congruence $\sim$ such that L is a $\sim$-language. One can show that the $\sim_{\bar{m}}$ are finite index aperiodic congruences (see Rosenstein [Ro82] and results in the next chapter). Corollary 2.2.5 implies that the problem of deciding whether a language has dot-depth k is equivalent to the problem of effectively characterizing the monoids $M = A^*/\sim$ with $\sim \supseteq \sim_{\bar{m}}$ for some $\bar{m} = (m_1,\ldots,m_k)$, i.e., $V_k = \{A^*/\sim \mid \sim \supseteq \sim_{\bar{m}}$ for some $\bar{m} = (m_1,\ldots,m_k)\}$.

Later chapters will be concerned with applications of theorem 2.2.4 and its corollaries. In the sequel, $\mathcal{L}_{(m_1,\ldots,m_k)}$ will denote the class of $\sim_{(m_1,\ldots,m_k)}$-languages.

Chapter 3

SOME PROPERTIES OF THE CHARACTERIZING CONGRUENCES

1. *An induction lemma*

The following lemma is a basic result (similar to one in [Ro82] regarding $\sim_k$) which will allow us to resolve games with $k+1$ moves into games with k moves and thereby allow us to perform induction arguments. We remind the reader that $u[1,p)$ ($u(p,|u|]$) denotes the segment of u to the left (right) of position p and $u(p,q)$ the segment of u between positions p and q.

Lemma 3.1.1

Let $\bar{m} = (m_1,\ldots,m_k)$. $u \sim_{(m,m_1,\ldots,m_k)} v$ if and only if

(1) for every $p_1, \ldots, p_m \in u$ $(p_1 \leq \ldots \leq p_m)$ there are $q_1, \ldots, q_m \in v$ $(q_1 \leq \ldots \leq q_m)$ such that

(i) $Q_a^u p_i$ if and only if $Q_a^v q_i$, $a \in A$ for $1 \leq i \leq m$,

(ii) $u[1,p_1) \sim_{\bar{m}} v[1,q_1)$,

(iii) $u(p_i,p_{i+1}) \sim_{\bar{m}} v(q_i,q_{i+1})$ for $1 \leq i \leq m-1$,

(iv) $u(p_m,|u|] \sim_{\bar{m}} v(q_m,|v|]$ and

(2) for every $q_1, \ldots, q_m \in v$ $(q_1 \leq \ldots \leq q_m)$ there are $p_1, \ldots, p_m \in u$ $(p_1 \leq \ldots \leq p_m)$ such that (i), (ii), (iii) and (iv) hold.

Proof Suppose that player II has a winning strategy in $\mathcal{G}_{(m,m_1,\ldots,m_k)}(u,v)$ and suppose that $p_1, \ldots, p_m \in u$, $p_1 \leq \ldots \leq p_m$. Using the strategy we can find positions

$q_1, \ldots, q_m \in v$, $q_1 \leq \ldots \leq q_m$ such that if player I chooses $p_1, \ldots, p_m \in u$ at his first move, then player II should choose $q_1, \ldots, q_m \in v$. Moreover, $Q_a^u p_i$ if and only if $Q_a^v q_i$, $a \in A$ for $1 \leq i \leq m$. There are now k moves left in the game $\mathcal{G}_{(m,m_1,\ldots,m_k)}(u,v)$. Whenever player I chooses positions in $u[1,p_1)$ or $v[1,q_1)$, the strategy, since it produces a win for player II, will always choose positions in $v[1,q_1)$ or $u[1,p_1)$. Thus player II's winning strategy for $\mathcal{G}_{(m,m_1,\ldots,m_k)}(u,v)$ includes within it a winning strategy for $\mathcal{G}_m^-(u[1,p_1),v[1,q_1))$, and similarly it includes a winning strategy for $\mathcal{G}_m^-(u(p_i,p_{i+1}),v(q_i,q_{i+1}))$ for $1 \leq i \leq m-1$, and $\mathcal{G}_m^-(u(p_m,|u|],v(q_m,|v|])$. This proves (1). By symmetry, (2) also holds.

Conversely, assuming that (1) and (2) hold, we describe a winning strategy for player II in $\mathcal{G}_{(m,m_1,\ldots,m_k)}(u,v)$. If player I chooses positions $p_1, \ldots, p_m \in u$ $(p_1 \leq \ldots \leq p_m)$ on his first move, then player II uses (1) to find positions $q_1, \ldots, q_m \in v$ $(q_1 \leq \ldots \leq q_m)$. Thereafter, whenever player I chooses positions of $u[1,p_1)$ or $v[1,q_1)$, player II uses his winning strategy in $\mathcal{G}_m^-(u[1,p_1),v[1,q_1))$ to respond; and similarly, whenever player I chooses positions of $u(p_i,p_{i+1})$ or $v(q_i,q_{i+1})$ $(u(p_m,|u|]$ or $v(q_m,|v|])$, player II uses his winning strategy in $\mathcal{G}_m^-(u(p_i,p_{i+1}),v(q_i,q_{i+1}))$ $(\mathcal{G}_m^-(u(p_m,|u|],v(q_m,|v|]))$ to reply. Since there are only k subsequent moves in the game and $\sim_{(m_1,\ldots,m_k)}$ implies $\sim_{(m_1',\ldots,m_k')}$ for all $m_i' \leq m_i$, player I can choose no more than k times from $u[1,p_1)$ or $v[1,q_1)$, $(u(p_i,p_{i+1})$ or $v(q_i,q_{i+1}))$ $(u(p_m,|u|]$ or $v(q_m,|v|])$ and no more than m_i positions each time. Hence player II's winning strategies in

$\mathcal{G}^-_m(u[1,p_1),v[1,q_1))$, $(\mathcal{G}^-_m(u(p_i,p_{i+1}),v(q_i,q_{i+1})))$ $(\mathcal{G}^-_m(u(p_m,|u|],v(q_m,|v|]))$ provides him with moves in all contingencies. If, on the other hand, player I chooses positions $q_1, \ldots, q_m \in v$, then player II uses (2) to find his correct first move and then proceeds analogously to the above. Thus player II has a winning strategy in $\mathcal{G}_{(m,m_1,\ldots,m_k)}(u,v)$. []

2. *A condition for inclusion*

Let us find a condition which insures $\mathcal{L}_{(m_1,\ldots,m_k)} \subseteq \mathcal{L}_{(m'_1,\ldots,m'_{k'})}$. A trivial condition is the following:

$$k \le k' \quad \text{and} \quad \exists\ 1 \le i_1 < \ldots < i_k \le k' \quad \text{such that}$$

$$m_1 \le m'_{i_1}$$

$$\vdots$$

$$m_k \le m'_{i_k}.$$

Define $N(m_1,\ldots,m_k) =$

$$m_1 + \ldots + m_k$$

$$+ \Sigma_{1\le i_1<i_2\le k}\, m_{i_1} m_{i_2} + \ldots + \Sigma_{1\le i_1<\ldots<i_{k-1}\le k}\, m_{i_1}\ldots m_{i_{k-1}} +$$

$$m_1 \ldots m_k.$$

A simpler definition is $N(m_1,\ldots,m_k) = (m_1+1)\ldots(m_k+1)-1$.

Proposition 3.2.1

$xyx^n zx \sim_{(m_1,\ldots,m_k)} xyx^{n'} zx$ for n, $n' \ge N-2$, where $N = N(m_1,\ldots,m_k)$.

Proof The proof is similar to the one of a property of $\sim_k$ in

[Tho84]. First, $x^n \sim_{(m_1,\dots,m_k)} x^{n'}$ for n, $n' \geq N$. To see this, consider the natural decompositions of $u = x^n$ and $v = x^{n'}$ into x-segments. Before each move we have in u and v certain segments in which positions have been chosen, and others where none have been. A maximal segment of succeeding x-segments without chosen positions will be called a gap which may be empty. Before each move there is a natural correspondence between the gaps in u and v given by their order. By induction on $k-i$, II chooses his segments in the following manner: for $i = k$, when m_k elements are still to be chosen by both players, two corresponding gaps both consist of any

number $\geq m_k = N(m_k)$ of x-segments, or else both consist of the same

number $< m_k = N(m_k)$ of x-segments. For $2 \leq i+1 \leq k$, when

$m_i+\dots+m_k$ elements are still to be chosen by both players, two corresponding gaps both consist of any

number $\geq m_i+(m_i+1)N(m_{i+1},\dots,m_k) = N(m_i,\dots,m_k)$ of x-segments, or else both consist of the same

number $< m_i+(m_i+1)N(m_{i+1},\dots,m_k) = N(m_i,\dots,m_k)$ of x-segments. In

$m_i+(m_i+1)N(m_{i+1},\dots,m_k)$, the first m_i corresponds to the number of elements chosen in the i^{th} move, m_i+1 is the number of gaps formed by those m_i positions, and $N(m_{i+1},\dots,m_k)$ is the minimum number of x-segments in any of those gaps that are necessary for player II to win in case the numbers of x-segments in two such corresponding gaps were different. Of course, inside his segments, II picks exactly those positions which match the ones chosen by I in the corresponding segments. Next, we apply the strategy above to $u = xyx^nzx$ and

$v = xyx^{n'}zx$ with $n,\ n' \geq N-2$ ignoring y and z (except when player I forces considering y and z). When player I chooses either of the two end x-segments, the strategy that we have tells us to choose the same end x-segment.[]

Note that $N(1,\ldots,1) = 2^k-1$. By putting $y = z = 1$ in the above proposition, we get as a corollary that if $m,\ m' \geq 2^k-1$, then $(w)^m \sim_k (w)^{m'}$. $N = N(m_1,\ldots,m_k)$ is seen to be the smallest n such that $x^n \sim_{(m_1,\ldots,m_k)} x^{n+1}$ for $|x| = 1$ (remark: $y = z = 1$ implies $x^N \sim_{(m_1,\ldots,m_k)} x^{N+1}$ and $x^{N-1} \not\sim_{(m_1,\ldots,m_k)} x^N$ as is easily seen by considering the play of the game $\mathcal{G}_{(m_1,\ldots,m_k)}(x^{N-1},x^N)$ where player I in the first move chooses m_1 x's in x^N separated by $N(m_2,\ldots,m_k)$ x's starting with the $N(m_2,\ldots,m_k)+1^{th}$ x. Then in the second move, I chooses a gap in x^N corresponding to a gap in x^{N-1} with a number of x's $< N(m_2,\ldots,m_k)$ and so on). Moreover, we see that if $u,\ v \in A^*$ and $u \sim_{(m_1,\ldots,m_k)} v$, then $|u|_a = |v|_a < N(m_1,\ldots,m_k)$ or $|u|_a,\ |v|_a \geq N(m_1,\ldots,m_k)$. Also, similarly to the above proof, one can show that if $u \sim_{(m_1,\ldots,m_k)} v$ and $k \geq 2$, then either $u = v$ or u and v have a common prefix and suffix of length $\geq m_1\ldots m_k$.

Proposition 3.2.2

(1) $\sim_{(m_1,\ldots,m_k)} \subseteq \sim_{(N(m_1,\ldots,m_k))}$, and

(2) $\sim_{(m_1,\ldots,m_k)} \not\subseteq \sim_{(N(m_1,\ldots,m_k)+1)}$.

Proof By the preceding proposition, choosing $|x| = 1$, we have $u = x^{N(m_1,\ldots,m_k)} \sim_{(m_1,\ldots,m_k)} x^{N(m_1,\ldots,m_k)+1} = v$. $x^{N(m_1,\ldots,m_k)+1}$ is a subword of length $N(m_1,\ldots,m_k)+1$ of v but not of u. This gives (2). (1) follows easily from lemma 3.1.1.[]

Another condition for $\mathcal{L}_{(m_1,\ldots,m_k)}$ to be included in $\mathcal{L}_{(m'_1,\ldots,m'_{k'})}$ is stated in

Proposition 3.2.3

If $k \le k'$ and $\exists\ 0 = j_0 < \ldots < j_{k-1} < j_k = k'$ such that $m_i \le N(m'_{j_{i-1}+1},\ldots,m'_{j_i})$ for $1 \le i \le k$, then $\sim_{(m'_1,\ldots,m'_{k'})} \subseteq \sim_{(m_1,\ldots,m_k)}$.

Proof The result comes from the following observation: for $1 \le i \le j \le k'$, we have $\sim_{(m'_1,\ldots,m'_i,\ldots,m'_j,\ldots,m'_{k'})} \subseteq \sim_{(m'_1,\ldots,m'_{i-1},N(m'_i,\ldots,m'_j),m'_{j+1},\ldots,m'_{k'})}$, which is a consequence of the preceding proposition (1).[]

Proposition 3.2.3 implies that if $n \ge \mathrm{sum}(\bar{m})$ and $u \sim_n v$, then $u \sim_{\bar{m}} v$. Moreover, if $\sim_{(m'_1,\ldots,m'_{k'})} \subseteq \sim_{(m_1,\ldots,m_k)}$, then $\sim_{(m'_1,\ldots,m'_{k'})} \subseteq \sim_{(N(m_1,\ldots,m_k))}$. Hence by proposition 3.2.2, $N(m_1,\ldots,m_k) \le N(m'_1,\ldots,m'_{k'})$. The next chapters include other results of inclusion. $N(m_1,\ldots,m_k)$ will appear several times in the sequel.

Chapter 4

AN ANSWER TO A CONJECTURE OF PIN

First we introduce some terminology. The study of the concatenation product leads to the definition of the Schützenberger product of finite monoids. The reader is referred to [St81] for the important properties of this construction. Let $M_1, \ldots, M_n$ be finite monoids. The *Schützenberger product* of $M_1, \ldots, M_n$, denoted by $\Diamond_n(M_1,\ldots,M_n)$, is the submonoid of upper triangular $n \times n$ matrices with the usual product of matrices, of the form $p = (p_{ij})$ $1 \le i,\ j \le n$ in which the (i,j)-entry is a subset of $M_1 \times \ldots \times M_n$ and all of whose diagonal entries are singletons, i.e.,

(1) $p_{ij} = \emptyset$ if $i > j$,

(2) $p_{ii} = \{(1,\ldots,1,m_i,1,\ldots,1)\}$ for some $m_i \in M_i$,
$\uparrow i^{th}$

(3) $p_{ij} \subseteq \{(m_1,\ldots,m_n) \in M_1 \times \ldots \times M_n \mid m_1 = \ldots = m_{i-1} = 1$
$= m_{j+1} = \ldots = m_n\}$.

Condition (2) allows to identify the coefficient p_{ii} with an element of M_i and condition (3) p_{ij} with a subset of $M_i \times \ldots \times M_j$. If $\mu = (m_i,\ldots,m_j) \in M_i \times \ldots \times M_j$ and
$\mu' = (m'_j,\ldots,m'_k) \in M_j \times \ldots \times M_k$, then we define
$\mu\mu' = (m_i,\ldots,m_{j-1},m_j m'_j,m'_{j+1},\ldots,m'_k)$. This product is extended to

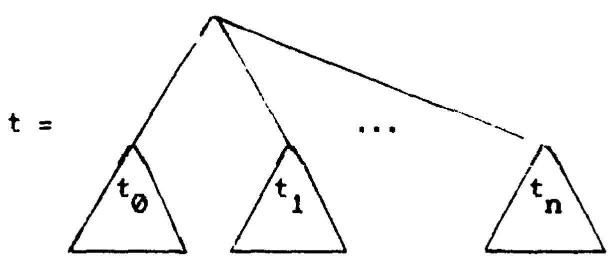
sets in the usual fashion; addition is given by set union.

Straubing [St81] has demonstrated that if the languages $L_i \subseteq A^*$ $(0 \leq i \leq n)$ are recognized by the monoids M_i, then the language $L_0a_1L_1a_2\ldots a_nL_n$, where the a_i are letters, is recognized by the monoid $\Diamond_{n+1}(M_0,\ldots,M_n)$. It is easy to verify that if $0 \leq i_0 < \ldots < i_r \leq n$, then $\Diamond_{r+1}(M_{i_0},\ldots,M_{i_r})$ is a submonoid of $\Diamond_{n+1}(M_0,\ldots,M_n)$ $(\Diamond_{r+1}(M_{i_0},\ldots,M_{i_r}) < \Diamond_{n+1}(1,\ldots,1,M_{i_0},1,\ldots,1,M_{i_1},\ldots,M_{i_r},1,\ldots,1))$. This implies that the monoid $\Diamond_{n+1}(M_0,\ldots,M_n)$ recognizes all languages of the form $L_{i_0}a_1L_{i_1}a_2\ldots a_rL_{i_r}$, where L_{i_k} is recognized by M_{i_k}, in particular, boolean combinations of languages of the form $L_0a_1L_1a_2\ldots a_nL_n$ where the M_i's recognize the L_i's. A converse has been established. The case $n = 1$ has been treated by Reutenauer [Re79] and the general case by Pin [Pi84b]. We have that if a language $L \subseteq A^*$ is recognized by $\Diamond_{n+1}(M_0,\ldots,M_n)$ then L is in the boolean algebra generated by the languages of the form $L_{i_0}a_1L_{i_1}a_2\ldots a_rL_{i_r}$ where $0 \leq i_0 < \ldots < i_r \leq n$ where for $0 \leq k \leq r$, $a_k \in A$ and L_{i_k} is a language recognized by M_{i_k}.

Let $\mathbf{W}$ be a M-variety. We define $\Diamond\mathbf{W}$, to be the variety of all finite monoids that divide some Schützenberger product $\Diamond_n(M_1,\ldots,M_n)$ for some n, where $M_i \in \mathbf{W}$ for $i = 1, \ldots, n$. From the above discussion, we have that for $k \geq 0$, $\mathbf{V}_{k+1} = \Diamond\mathbf{V}_k$. In particular, $\mathbf{V}_1 = \mathbf{J} = \Diamond\mathbf{I}$ and $\mathbf{V}_2 = \Diamond\mathbf{J}$ where $\mathbf{I}$ denotes the variety consisting of the trivial monoid alone and $\mathbf{J}$ of all finite $\mathcal{J}$-trivial monoids.

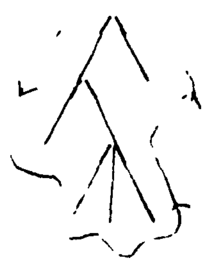
1. *Decidability and inclusion problems*

Pin [Pi84b] demonstrated that the Straubing hierarchy is a particular case of a more general construction obtained in associating varieties of languages not to integers but to trees in the following fashion. A variety of languages is associated by definition to the tree reduced to a point. Then to the tree



is associated the boolean algebra generated by the languages of the form $L_{i_0} a_1 L_{i_1} a_2 \ldots a_r L_{i_r}$ with $0 \le i_0 < \ldots < i_r \le n$ where, for $0 \le j \le r$, L_{i_j} is member of the variety of languages associated to the tree t_{i_j}. Since the Schützenberger product is perfectly adapted to the operation $(L_0, \ldots, L_n) \to L_0 a_1 L_1 a_2 \ldots a_n L_n$, it permits us to construct, without reference to languages, hierarchies of varieties of monoids corresponding, via Eilenberg's theorem, to the hierarchies of languages previously constructed. Starting with a variety of monoids **W**, we associate with each tree t, respectively with each set of trees T, a variety of monoids $\langle\rangle_t(W)$ $(\langle\rangle_T(W))$. Descriptions of the hierarchies of monoids are given after a few definitions.

We will denote by $\mathcal{T}$ the set of trees on the alphabet $\{a,\bar{a}\}$. Formally, $\mathcal{T}$ is the set of words in $\{a,\bar{a}\}^*$ congruent to 1 in the congruence generated by the relation $a\bar{a} = 1$. Intuitively, the words in $\mathcal{T}$ are obtained as follows: we draw a tree and starting from the root we code a for going down and $\bar{a}$ for going up. For example,



is coded by $aa\bar{a}aa\bar{a}a\bar{a}a\bar{a}\bar{a}\bar{a}a\bar{a}$. The number of leaves of a word t in $\{a,\bar{a}\}^*$, denoted by $l(t)$, is by definition the number of occurrences of the factor $a\bar{a}$ in t. Each tree t factors uniquely into $t = at_1\bar{a}at_2\bar{a}\ldots at_n\bar{a}$ where $n \geq 0$ and where the t_i's are trees. Let t be a tree and let $t = t_1at_2\bar{a}t_3$ be a factorization of t. We say that the occurrences of a and $\bar{a}$ defined by this factorization are *related* if t_2 is a tree. Let t and t' be two trees. We say that t is *extracted* from t' if t is obtained from t' by removing in t' a certain number of related occurrences of a and $\bar{a}$. We now state the algebraic interpretation of the above stated hierarchy construction using the Schützenberger product.

To each tree t and to each sequence $\mathbf{W}_1, \ldots, \mathbf{W}_{l(t)}$ of varieties of monoids, we associate a variety of monoids $\langle\rangle_t(\mathbf{W}_1,\ldots,\mathbf{W}_{l(t)})$ defined recursively by:

(1) $\langle\rangle_1(\mathbf{W}) = \mathbf{W}$ for every $\mathcal{M}$-variety $\mathbf{W}$,

(2) if $t = at_1\bar{a}at_2\bar{a}\ldots at_n\bar{a}$ with $n \geq 0$ and $t_1, \ldots, t_n \in \mathcal{T}$, $\langle\rangle_t(\mathbf{W}_1,\ldots,\mathbf{W}_{l(t)})$ is the variety of monoids M such that M divides some $\langle\rangle_n(M_1,\ldots,M_n)$ with $M_1 \in \langle\rangle_{t_1}(\mathbf{W}_1,\ldots,\mathbf{W}_{l(t_1)})$, $\ldots$, $M_n \in \langle\rangle_{t_n}(\mathbf{W}_{l(t_1)+\ldots+l(t_{n-1})+1},\ldots,\mathbf{W}_{l(t_1)+\ldots+l(t_n)})$.

When $\mathbf{W}_1 = \ldots = \mathbf{W}_{l(t)} = \mathbf{W}$, we denote simply $\langle\rangle_t(\mathbf{W})$ the variety $\langle\rangle_t(\mathbf{W}_1,\ldots,\mathbf{W}_{l(t)})$. More generally, if T is a language contained in $\mathcal{T}$, we denote $\langle\rangle_T(\mathbf{W})$ the smallest variety containing the varieties $\langle\rangle_t(\mathbf{W})$ with $t \in T$.

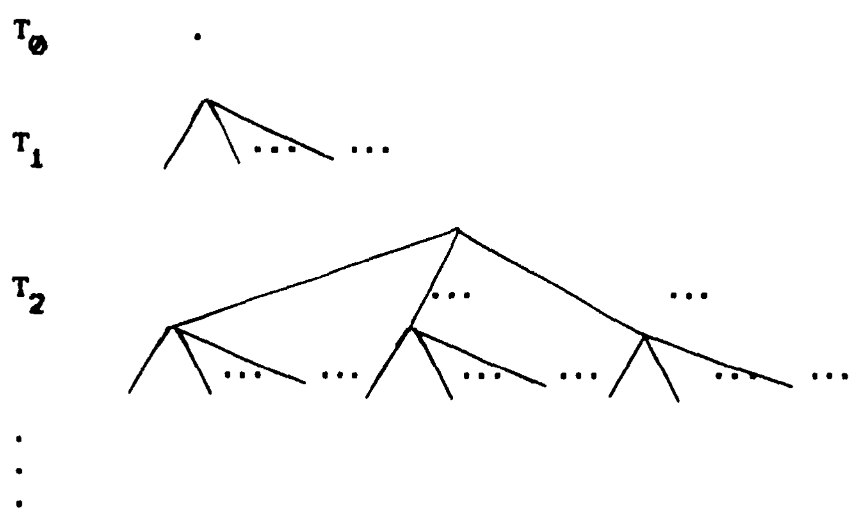
A consequence of the above definition is that if $t = at_1\bar{a}at_2\bar{a}\ldots at_n\bar{a}$ with $t_1, \ldots, t_n \in \mathcal{T}$, we have $\langle\rangle_t(\mathbf{W}) = \langle\rangle_{(a\bar{a})^n}(\langle\rangle_{t_1}(\mathbf{W}),\ldots,\langle\rangle_{t_n}(\mathbf{W}))$.

The following proposition allows us, by recursion, to describe the languages associated to the varieties $\langle\rangle_t(\mathbf{W}_1,\ldots,\mathbf{W}_{l(t)})$ for each tree t.

Proposition 4.1.1 Pin [Pi84b]

Let n be a positive integer and $\mathbf{W}_0, \ldots, \mathbf{W}_n$ be M-varieties. We denote respectively by $\mathcal{W}_j$ and $\mathcal{W}$ the *-varieties of languages corresponding to $\mathbf{W}_j$ $(0 \le j \le n)$ and to $\Diamond_{(a\bar{a})}{}^{n+1}(\mathbf{W}_0,\ldots,\mathbf{W}_n)$. Then for each alphabet A, $A^*\mathcal{W}$ is the boolean algebra generated by the languages of the form $L_{i_0}a_1L_{i_1}a_2\ldots a_rL_{i_r}$ where $0 \le i_0 < \ldots < i_r \le n$ where for $0 \le j \le r$, $a_j \in A$ and $L_{i_j} \in A^*\mathcal{W}_{i_j}$.

The Straubing hierarchy $\mathbf{V}_k$ can be described in the following fashion. Let T_k be the sequence of languages defined by $T_0 = \{1\}$ and $T_{k+1} = (aT_k\bar{a})^*$. Intuitively, we can represent the languages by trees infinite in width:



Proposition 4.1.2

For $k \geq 0$, $\mathcal{V}_k = \langle\rangle_{T_k}(I)$. In particular, $\langle\rangle_{T_0}(I) = I$, $\langle\rangle_{T_1}(I) = J$, $\langle\rangle_{T_2}(I) = \langle\rangle J$.

Proof This is an immediate consequence of proposition 4.1.1.[]

More precisely,

Proposition 4.1.3

For $k \geq 1$, $m \geq 1$, $\mathcal{V}_{k,m} = \langle\rangle_{(aT_{k-1}\bar{a})^{m+1}}(I)$.

Proof Let $\mathcal{W}_{k,m}$ be the $*$-variety of languages corresponding to $\langle\rangle_{(aT_{k-1}\bar{a})^{m+1}}(I) = \langle\rangle_{(a\bar{a})^{m+1}}(\langle\rangle_{T_{k-1}}(I))$. We have to establish the equality $\mathcal{W}_{k,m} = \mathcal{V}_{k,m}$. Proposition 4.1.1 and $\mathcal{V}_k = \langle\rangle_{T_k}(I)$ of the preceding proposition show that for each alphabet A, $A^*\mathcal{W}_{k,m}$ is the boolean algebra generated by the languages of the form $L_0a_1L_1a_2\ldots a_nL_n$ where $0 \leq n \leq m$, $L_0, \ldots, L_n \in A^*\mathcal{V}_{k-1}$ and $a_1, \ldots, a_n \in A$. The result clearly follows.[]

Let $\bar{m} = (m_1,\ldots,m_k)$. By induction on k, we define a tree $t_{\bar{m}}$ as follows:

if length$(\bar{m}) = 1$, then $t_{\bar{m}} = (a\bar{a})^{m_1+1}$,

for $\bar{m} = (m,m_1,\ldots,m_k)$, $t_{\bar{m}} = (at_{(m_1,\ldots,m_k)}\bar{a})^{m+1}$.

It is easy to see that $l(t_{(m_1,\ldots,m_k)})$ is $N(m_1,\ldots,m_k)+1 = (m_1+1)\ldots(m_k+1)$.

Let t be a tree and let V_t be the *-variety of languages associated with $\langle\rangle_t(I)$. We have

Proposition 4.1.4

$$V_{t_{(m_1,\ldots,m_k)}} = \mathcal{L}_{(m_1,\ldots,m_k)}.$$

Proof The proof is by induction on k. If $k = 1$, then $\langle\rangle_{t_{(m_1)}}(I) = \nabla_{1,m_1}$ by proposition 4.1.3. The result then follows from theorem 2.2.4. Suppose true for k, i.e., letting $\bar{m} = (m_1,\ldots,m_k)$, $V_{t_{\bar{m}}} = \mathcal{L}_{\bar{m}}$. Let us show that $V_{t_{(m,m_1,\ldots,m_k)}} = \mathcal{L}_{(m,m_1,\ldots,m_k)}$. From $\langle\rangle_{t_{(m,m_1,\ldots,m_k)}}(I) = \langle\rangle_{(at_{\bar{m}}\bar{a})^{m+1}}(I) = \langle\rangle_{(a\bar{a})^{m+1}}(\langle\rangle_{t_{\bar{m}}}(I))$, using the induction hypothesis and proposition 4.1.1, we can conclude that for each alphabet A, $A^*V_{t_{(m,m_1,\ldots,m_k)}}$ is the boolean algebra generated by the languages of the form $L_0a_1L_1a_2\ldots a_nL_n$ with $n \le m$ and where for $0 \le j \le n$, $a_j \in A$ and $L_j \in A^*\mathcal{L}_{(m_1,\ldots,m_k)}$. The result follows since each $\sim_{(m,m_1,\ldots,m_k)}$-class is a boolean combination of sets of the form $L_0a_1L_1a_2\ldots a_nL_n$, with $n \le m$ and where each L_j is a $\sim_{(m_1,\ldots,m_k)}$-class.[]

The following result perhaps constitutes a first step towards the general solution of the decidability problem.

Proposition 4.1.5 Pin [Pi84b]

For each tree t, the variety $\langle\rangle_t(I)$ is decidable.

Using propositions 4.1.4 and 4.1.5, we get

Proposition 4.1.6

For fixed $(m_1,\dots,m_k)$ the M-variety $\langle\rangle_{t_{(m_1,\dots,m_k)}}(I)$ is decidable, so the *-variety of languages $\mathcal{L}_{(m_1,\dots,m_k)}$ is decidable.

Among the many problems concerning these tree hierarchies, is the comparison between the varieties inside a hierarchy. More precisely, the problem consists in comparing the different varieties $\langle\rangle_t(W)$ (or even $\langle\rangle_T(W)$). A partial result and a conjecture on this problem was given in Pin [Pi84b]. It was shown that for every variety W, if t is extracted from t', then $\langle\rangle_t(W) \subseteq \langle\rangle_{t'}(W)$, and it was conjectured that if t, $t' \in T'$, $\langle\rangle_t(I)$ is contained in $\langle\rangle_{t'}(I)$ if and only if t is extracted from t'. Here, T' denotes the set of trees in which each node is of arity different from 1.

Theorem 4.1.7

The above conjecture is false.

To see this, $\mathcal{L}_{(1,2)} \subseteq \mathcal{L}_{(2,1)}$ by lemma 4.2.7 of the next section. Hence $\langle\rangle_{t_{(1,2)}}(I) \subseteq \langle\rangle_{t_{(2,1)}}(I)$ by proposition 4.1.4. But it is easy to verify that the tree $t_{(1,2)}$ is not extracted from the tree $t_{(2,1)}$. The main step of the proof of theorem 4.1.7 is given in the next section.

2. *The conjecture is false*

This section is devoted to the proof of theorem 4.1.7 of the preceding section. The proof goes through seven lemmas, lemmas 4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5, 4.2.6 and 4.2.7.

When is $\sim(2,m'_2) \subseteq \sim(1,m_2)$? Of course, if $m'_2 \geq m_2$, it is true. We will be considering the case when $m'_2 < m_2$, or, $m'_2+1 \leq m_2$. Assume $u \sim_{(2,1)} v$ and $|u|_a$, $|v|_a > 0$. Let $u = u_\emptyset au_1 \ldots au_n$, $v = v_\emptyset av_1 \ldots av_m$ where $n = |u|_a$, $m = |v|_a$. If $Q_a^u p_i$, $Q_a^v q_j$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$, then $u_i = u(p_i, p_{i+1})$, $i = 1, \ldots, n-1$, $v_j = v(q_j, q_{j+1})$, $j = 1, \ldots, m-1$. $u_\emptyset = u[1,p_1)$, $v_\emptyset = v[1,q_1)$, $u_n = u(p_n, |u|]$, $v_m = v(q_m, |v|]$.

Lemma 4.2.1

(1) $u_\emptyset \sim_1 v_\emptyset$, $u_1 \sim_1 v_1$,

$u_{n-1} \sim_1 v_{m-1}$, $u_n \sim_1 v_m$,

(2) $u_2au_3 \ldots au_{n-2} \sim_1 v_2av_3 \ldots av_{m-2}$.

Proof (1) Player I, in the first move chooses 2 consecutive a's among the first or the last 2 ones (of u or v). Since $u \sim_{(2,1)} v$, Player II chooses 2 consecutive a's, the same occurrences among the first or the last 2 ones (of v or u). The result follows from lemma 3.1.1.

(2) Let w be a subword of length ≤ 1 of $u_2au_3...au_{n-2}$ (or of $v_2av_3...av_{m-2}$). Hence w is a subword of $v_2av_3...av_{m-2}$ (or of $u_2au_3...au_{n-2}$) because $aawaa$ is a subword of length $\leq N(2,1) = 5$ of u (or of v) ($\sim_{(2,1)} \subseteq \sim_{(N(2,1))}$ by proposition 3.2.2(1)).[]

Lemma 4.2.2

(1) $u_1au_2...au_n \sim_{(2)} v_1av_2...av_m$,

$u_2au_3...au_n \sim_{(2)} v_2av_3...av_m$,

$u_3au_4...au_n \sim_{(2)} v_3av_4...av_m$,

(2) $u_\emptyset au_1...au_{n-1} \sim_{(2)} v_\emptyset av_1...av_{m-1}$,

$u_\emptyset au_1...au_{n-2} \sim_{(2)} v_\emptyset av_1...av_{m-2}$,

$u_\emptyset au_1...au_{n-3} \sim_{(2)} v_\emptyset av_1...av_{m-3}$.

Proof (1) Let $1 \leq i \leq 3$. Let w be a subword of length ≤ 2 in $u_iau_{i+1}...au_n$. Consider $w' = a^iw$ of length $\leq i+2 \leq N(2,1)$. $u \sim_{(N(2,1))} v$ (proposition 3.2.2(1)) and the fact that w' is a subword of u of length $\leq N(2,1)$ imply that w' is also a subword of v, and hence w a subword in $v_iav_{i+1}...av_m$. Similarly, for subwords of $v_iav_{i+1}...av_m$. For (2), we consider wa^i.[]

Lemma 4.2.3

(1) $u_\emptyset \sim_{(2)} v_\emptyset$,

(2) $u_n \sim_{(2)} v_m$.

Proof (1) Let $w = w_1...w_{|w|}$ be a subword of length ≤ 2 in $u_\emptyset$. Let p, $p' \in u$ be such that $p \leq p' < p_1$ and $Q^u_{w_1}p$, $Q^u_{w_{|w|}}p'$.

Consider the following play of the game $\mathcal{G}_{(2,1)}(u,v)$. In the first move, player I chooses p and p_1. Using lemma 3.1.1, there is $q \in v$, $q < q_1$, $Q^{v}_{w_1}q$ and $u(p,p_1) \sim_1 v(q,q_1)$. Since $w_{|w|}$ is a subword of length ≤ 1 in $u(p,p_1)$ and $u(p,p_1) \sim_1 v(q,q_1)$, $w_{|w|}$ is a subword of length ≤ 1 in $v(q,q_1)$. Hence w is also a subword in $v_\emptyset$. Similarly, for subwords of $v_\emptyset$. For (2), let $w = w_1 \dots w_{|w|}$ be a subword of length ≤ 2 in u_n. Let p, $p' \in u$ be such that $p_n < p' \leq p$ and $Q^{u}_{w_{|w|}}p$, $Q^{u}_{w_1}p'$. In the first move, player I chooses p_n and p. The result follows similarly as (1).[]

Lemma 4.2.4

(1) $u_\emptyset a u_1 \sim_{(2)} v_\emptyset a v_1$,

(2) $u_{n-1} a u_n \sim_{(2)} v_{m-1} a v_m$.

Proof (1) We will show that $u_\emptyset a u_1 \sim_{(2)} v_\emptyset a v_1$. The proof is similar for (2). Let $w = w_1 \dots w_{|w|}$ be a subword of length ≤ 2 in $u_\emptyset a u_1$ (similar if starting with $v_\emptyset a v_1$). We want to show that w is a subword of $v_\emptyset a v_1$. If w is a subword of $u_\emptyset$, w is also a subword of $v_\emptyset$ by lemma 4.2.3(1). If not, let j, $1 \leq j \leq |w|$, be the first index such that $w_1 \dots w_j$ is not a subword of $u_\emptyset$ but $w_1 \dots w_{j-1}$ is a subword of $u_\emptyset$. We have that $w_1 \dots w_{j-1}$ is a subword of $v_\emptyset$ by lemma 4.2.3(1) but we do not have that $w_1 \dots w_j$ is a subword of $v_\emptyset$ (if we had, $w_1 \dots w_j$ would be in $u_\emptyset$ for the same reason). If $w_j = a$, $w_1 \dots w_j$ is a subword of $u_\emptyset a$ and $v_\emptyset a$, and since $u_1 \sim_1 v_1$ by lemma 4.2.1(1) and $1 \leq j \leq |w|$, w is a subword of $v_\emptyset a v_1$. If $w_j \neq a$, let p be the first position in u after p_1 such that $Q^u_{w_j}p$. Now, since $u_1 \sim_1 v_1$ by lemma 4.2.1(1), w_j occurs between q_1 and q_2. Let q be the first position in v after q_1 such that $Q^v_{w_j}q$. If $|w_j \ldots w_{|w|}| \le 1$, the proof is complete. If not, i.e., $|w_j \ldots w_{|w|}| > 1$ then $j = 1$, $|w| = 2$. Consider the following play of the game $\mathcal{G}_{(2,1)}(u,v)$. Player I in the first move, chooses positions p and p_2 in u. Player II should choose q in v. If not, II would choose a position q' in v such that $q' > q$ because he needs at least 1 a before q', and q is the first position in v after q_1 such that $Q^v_{w_1}q$. But then, player I, in the second move could choose an occurrence of w_1 from $v[1,q')$ (not possible for II in $u[1,p)$ from the choice of j and the fact that $w_j \neq a$). Player II cannot choose a position q'' such that $Q^v_a q''$ before q_2 because he needs at least 1 a before q. Since there is no a between p and p_2, there should not be any between q and q''. Hence player II should choose q and q_2. Hence $u(p,p_2) \sim_1 v(q,q_2)$ and (1) follows.[]

Lemma 4.2.5

Let $p'_1, \ldots, p'_s$ in u $(p'_1 < \ldots < p'_s)$ $(q'_1, \ldots, q'_{s'}$ in v $(q'_1 < \ldots < q'_{s'}))$ be the positions which spell the first and the last occurrences of every letter in u (v). Then

(1) $s = s'$,

(2) $Q^u_b p'_i$ if and only if $Q^v_b q'_i$, $b \in A$ for $1 \le i \le s$,

(3) $u[1,p'_i) \sim_{(2)} v[1,q'_i)$ and $u(p'_i,|u|] \sim_{(2)} v(q'_i,|v|]$ for $1 \le i \le s$,

(4) $u(p'_i,p'_{i+1}) \sim_1 v(q'_i,q'_{i+1})$ for $1 \le i \le s-1$,

(5) for $1 \le i \le s-1$ and for every $p' \in u(p'_i,p'_{i+1})$, there exists $q' \in v(q'_i,q'_{i+1})$ such that

(1') $Q^u_b p'$ if and only if $Q^v_b q'$, $b \in A$,

(2') $u(p'_i,p') \sim_1 v(q'_i,q')$.

Also, there exists $q' \in v(q'_i,q'_{i+1})$ (which may be different from the one which satisfies (1'), (2')) such that (1'),

(2'') $u(p',p'_{i+1}) \sim_1 v(q',q'_{i+1})$.

Similarly, for every $q' \in v(q'_i,q'_{i+1})$, there exists $p' \in u(p'_i,p'_{i+1})$ such that (1'), (2') hold (also (1'), (2'') hold) and

(6) for $1 \le i \le s-1$ and for every p''_1, $p''_2 \in u(p'_i,p'_{i+1})$ $(p''_1 < p''_2)$, there exist q''_1, $q''_2 \in v(q'_i,q'_{i+1})$ $(q''_1 < q''_2)$ such that

(1''') $Q^u_b p''_j$ if and only if $Q^v_b q''_j$, $b \in A$ for $1 \le j \le 2$,

(2''') $u(p''_1,p''_2) \sim_1 v(q''_1,q''_2)$.

Similarly, for every q''_1, $q''_2 \in v(q'_i,q'_{i+1})$ $(q''_1 < q''_2)$, there exist p''_1, $p''_2 \in u(p'_i,p'_{i+1})$ $(p''_1 < p''_2)$ such that (1''') and (2''') hold.

Proof (1) holds since $u \sim_{(2,1)} v$, by chapter three, implies $|u|_b = |v|_b < N(2,1) = 5$ or $|u|_b$, $|v|_b \ge N(2,1)$ for every $b \in A$.

(2) holds since $\sim_{(2,1)} \subseteq \sim_{(1,1)}$ and we may consider the plays of the game $\mathcal{G}_{(1,1)}(u,v)$ where player I in the first move chooses p'_i

for some i, $1 \le i \le s$.

(3) follows from the arguments in the proofs of lemmas 4.2.2 and 4.2.3 since p'_i (q'_i) is either the first or the last occurrence of a letter in u (v) (in lemmas 4.2.2 and 4.2.3 we were considering p_1 (q_1) which are the first occurrences of the letter a in u (v) and p_n (q_m) which are the last occurrences of that letter in u (v)).

(4), (5) and (6) follow by considering different plays of the game $G_{(2,1)}(u,v)$. First, from the choice of the p'_r's and the q'_r's and lemma 3.1.1, if p'_i (q'_i) is among the positions chosen in u (v) by player I in the first move, then q'_i (p'_i) should be among the ones chosen in v (u) by player II in the first move. Second, if the positions chosen by player I in the first move are in $u(p'_i,p'_{i+1})$ $(v(q'_i,q'_{i+1}))$, then the positions chosen by player II in the first move should be in $v(q'_i,q'_{i+1})$ $(u(p'_i,p'_{i+1}))$ for the same reasons. For (4), consider the play of the game $G_{(2,1)}(u,v)$ where player I, in the first move, chooses p'_i and p'_{i+1}; for (5), I chooses p'_i and p', or p' and p'_{i+1}; for (6), he chooses p''_1 and p''_2.[]

Lemma 4.2.6

Let $p'_1, \ldots, p'_s$ in u $(p'_1 < \ldots < p'_s)$ $(q'_1, \ldots, q'_s$ in v $(q'_1 < \ldots < q'_s))$ be the positions which spell the first and last occurrences of every letter in u (v) so (satisfying) (2), (3), (4), (5) and (6) of lemma 4.2.5. For i fixed between 1 and s-1, let $p''_1, \ldots, p''_{s_i}$ in $u(p'_i,p'_{i+1})$ $(p''_1 < \ldots < p''_{s_i})$

$(q''_1, \ldots, q''_{s'_i}$ in $v(q'_i,q'_{i+1})$ $(q''_1 < \ldots < q''_{s'_i}))$ be the positions which spell the first and the last occurrences of every letter in $u(p'_i,p'_{i+1})$ $(v(q'_i,q'_{i+1}))$. Then

(1'''') $s_i = s'_i$,

(2'''') $Q_b^u p''_j$ if and only if $Q_b^v q''_j$, $b \in A$ for $1 \le j \le s_i$ and

(3'''') $u[1,p''_j) \sim_{(2)} v[1,q''_j)$ and $u(p''_j,|u|] \sim_{(2)} v(q''_j,|v|]$ for $1 \le j \le s_i$.

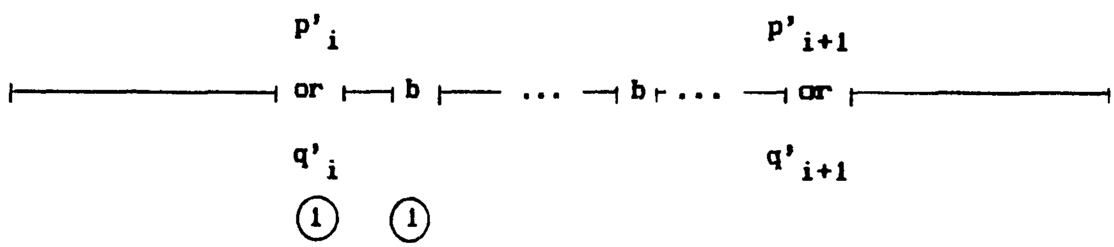
Proof By (4) of lemma 4.2.5 we have $u(p'_i,p'_{i+1}) \sim_1 v(q'_i,q'_{i+1})$. Now, if in one of these segments, either $u(p'_i,p'_{i+1})$ or $v(q'_i,q'_{i+1})$, there is only one occurrence of some letter and in the other segment there are two or more occurrences of that same letter, then player I in the first move could choose two of these occurrences (not possible for II in the remaining segment contradicting (6) of the preceding lemma). Hence (1'''') holds.

For (2''''), consider any two letters, $b \neq c$, in $u(p'_i,p'_{i+1})$ (and hence in $v(q'_i,q'_{i+1})$ by lemma 4.2.5(4)) and consider their first and last occurrences in $u(p'_i,p'_{i+1})$ and $v(q'_i,q'_{i+1})$ (by (1''''), the numbers of these occurrences agree). We claim that we have the same pattern: there are six possibilities, namely, pattern 1: bbcc, or, pattern 2: bcbc, or, pattern 3: bccb, or, pattern 4: cbbc, or, pattern 5: cbcb, or, pattern 6: ccbb. Expressed differently, the subwords formed by these occurrences are the same (similar proof if only one occurrence of a letter instead of a first and a last: the patterns would be shorter words). Let us separate different patterns by considering plays of the game

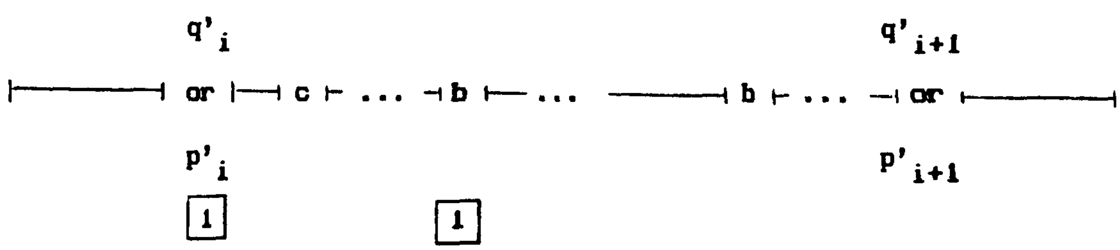
$\mathcal{G}_{(2,1)}(u,v)$. We will illustrate the plays by diagrams. The first move of I will be indicated by (1) and the first move of II by [1]. In each diagram, the segment between the positions chosen by I in move 1 $\neq_1$ the segment between the positions chosen by II in move 1, in contradiction with lemma 4.2.5 (5) or (6). We show how to separate patterns 1-2-3 from patterns 4-5-6, pattern 1 from patterns 2 and 3, pattern 2 from pattern 3. The separation of the patterns 4, 5 and 6 is similar to the separation of 1, 2 and 3.

To separate patterns 1-2-3 from patterns 4-5-6:

pattern 1, 2 or 3

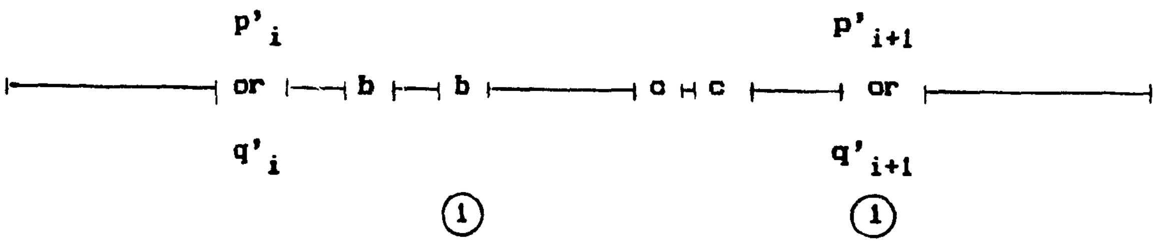


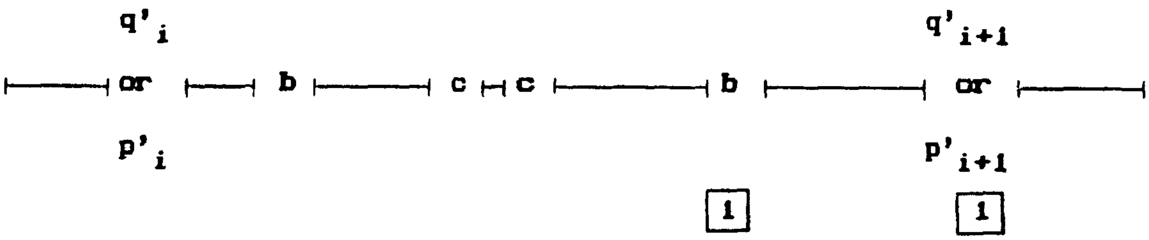
pattern 4, 5 or 6



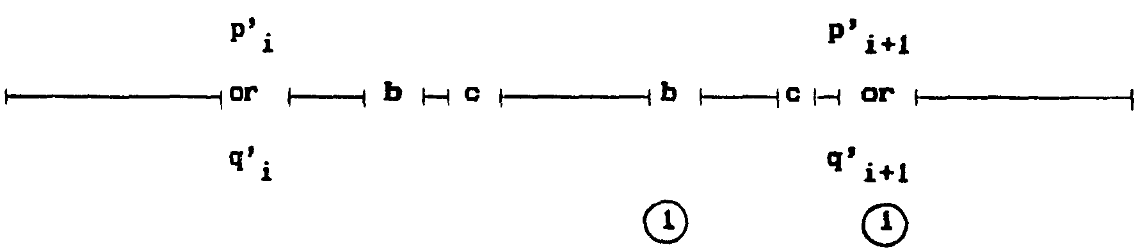
The above diagram is in contradiction with lemma 4.2.5(5) (II has to choose the first occurrence of b but there is an occurrence of c between the positions that he chooses which is not the case for I).

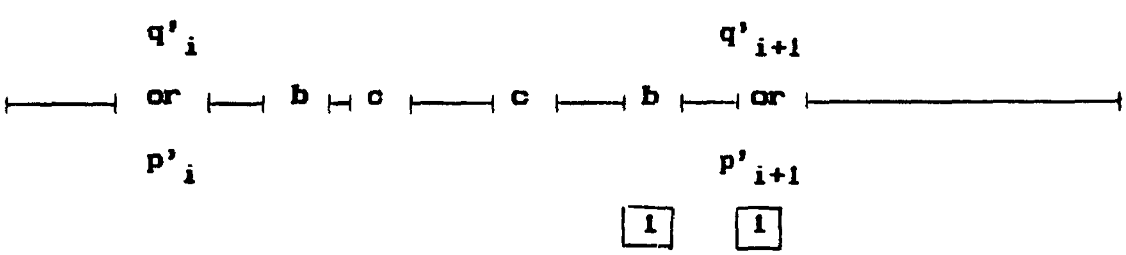
To separate patterns 1 and 3:



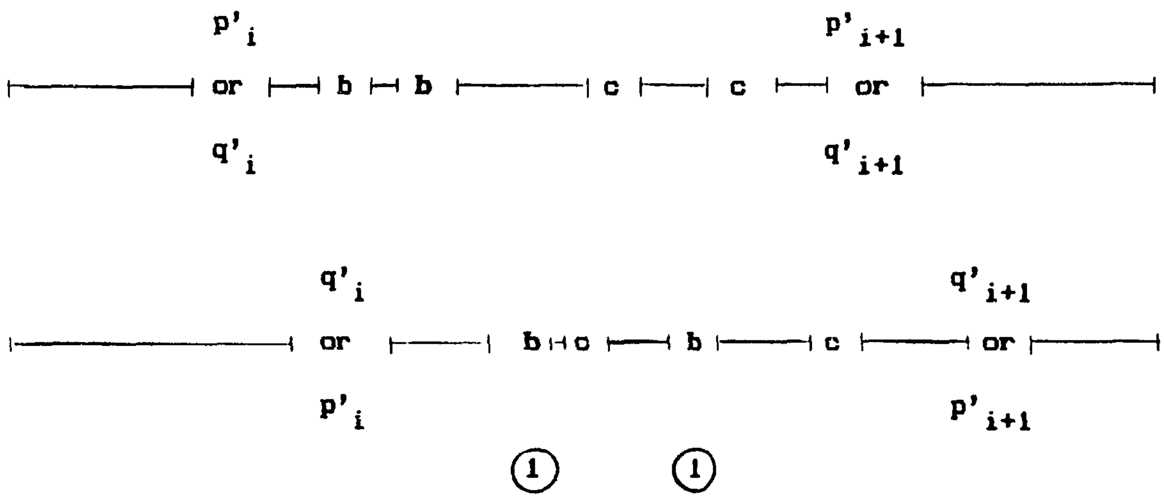


To separate patterns 2 and 3:





To separate patterns 1 and 2:



Here, player II cannot choose two b's separated by a c (in contradiction with 4.2.5(6)).

The diagrams above show that any two letters obey the same pattern. $Q_b^u p''_1$ if and only if $Q_b^v q''_1$ is clear. Now, by induction on j, assume $Q_b^u p''_k$ if and only if $Q_b^v q''_k$ for $1 \le k \le j$. Suppose, say $Q_b^u p''_{j+1}$ and $Q_c^v q''_{j+1}$ with $b \neq c$. But b and c have the same pattern in $u(p'_i, p''_j]$ and in $v(q'_i, q''_j]$ by induction hypothesis and the result follows.

We now prove (3''''). Let $1 \le j \le s_i$. We will show that $u[1,p''_j) \sim_{(2)} v[1,q''_j)$ (the proof is similar for $u(p''_j, |u|] \sim_{(2)} v(q''_j, |v|]$). Let $w = w_1 \ldots w_{|w|}$ be a subword of length ≤ 2 in $u[1,p''_j)$ (similar if in $v[1,q''_j)$). We want to show that w is a subword of $v[1,q''_j)$. If $|w| = 1$, then there is an occurrence of w_1 in $u[1,p'_i]$ (and hence in $v[1,q'_i]$) from the choice of the p'_r's and the q'_r's and lemma 4.2.5(1,2) and the proof is complete. If $|w| = 2$, and w is in $u[1,p'_i)$, then w is in $v[1,q'_i)$ by lemma 4.2.5(3). If there is an occurrence of w_1 in $u[1,p'_i)$ (and hence in $v[1,q'_i)$ by lemma 4.2.5(3)) and $Q^u_{w_2}p'_i$ (and hence $Q^v_{w_2}q'_i$ by lemma 4.2.5(2)) the proof is complete. Otherwise, there is an occurrence of w_1 in $u[1,p'_i]$ (and hence in $v[1,q'_i])$ from the choice of the p'_r's and q'_r's and lemma 4.2.5(1,2) and also an occurrence of w_2 in $u(p'_i,p''_j)$. From the choice of the p''_r's, there exists k, $k < j$, such that $Q^u_{w_2}p''_k$. Hence, from the choice of the q''_r's and $(1'''',2'''')$, $Q^v_{w_2}q''_k$.[]

Lemma 4.2.7

$\sim_{(2,1)} \subseteq \sim_{(1,2)}$.

Proof Suppose $u \sim_{(2,1)} v$. Then there is a winning strategy for player II in the game $\mathcal{G}_{(2,1)}(u,v)$ to win each play. Let us describe a winning strategy for player II in the game $\mathcal{G}_{(1,2)}(u,v)$ to win each play. Let p be a position in u chosen by player I in the first move. Suppose $Q^u_a p$ for some $a \in A$.

Case 1: $|u|_a = |v|_a < 5 = N(1,2) = N(2,1)$.

If p is the i^{th} occurrence of a in u chosen by player I in the first move, then player II chooses the same occurrence of a in v, say position q. The fact that $u[1,p) \sim_{(2)} v[1,q)$ and $u(p,|u|] \sim_{(2)} v(q,|v|]$ follows from lemmas 4.2.2, 4.2.3 and 4.2.4.

Case 2: $|u|_a = |v|_a = 5$.

Same as case 1.

Case 3: $|u|_a = 5$, $|v|_a > 5$.

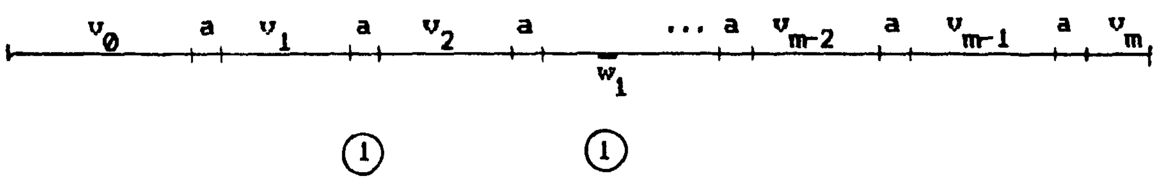
We include this case because the strategy here for player II is very easy but the arguments in case 4 are enough to prove the lemma. If p is the i^{th} occurrence of a in u $(1 \le i \le 2)$ chosen by player I in the first move, then player II chooses the same occurrence of a in v, say position q. If p is the $6-i^{th}$ occurrence of a in u $(1 \le i \le 2)$, player II chooses the $m-i+1^{th}$ occurrence of a in v. The fact that $u[1,p) \sim_{(2)} v[1,q)$ and $u(p,|u|] \sim_{(2)} v(q,|v|]$ follows from lemmas 4.2.2, 4.2.3 and 4.2.4. If $p = p_3$, then player II chooses position q, an a, among the middle ones in v, i.e., among $q_3, \ldots, q_{m-2}$. Lemma 4.2.2 implies that $u_3au_4au_5 \sim_{(2)} v_3av_4\ldots av_m$ and $u_\emptyset au_1au_2 \sim_{(2)} v_\emptyset av_1\ldots av_{m-3}$. Observe that if we show $u_\emptyset au_1au_2 \sim_{(2)} v_\emptyset av_1av_2$ and $u_3au_4au_5 \sim_{(2)} v_{m-2}av_{m-1}av_m$ the proof is complete since we will have $u_\emptyset au_1au_2 \sim_{(2)} v[1,q)$ and $u_3au_4au_5 \sim_{(2)} v(q,|v|]$ for any position q among $q_3, \ldots, q_{m-2}$. If player I had chosen p among the middle positions in v, then player II would choose p_3 in u. So let us show that $u_\emptyset au_1au_2 \sim_{(2)} v_\emptyset av_1av_2$. The proof of $u_3au_4au_5 \sim_{(2)} v_{m-2}av_{m-1}av_m$ is similar.

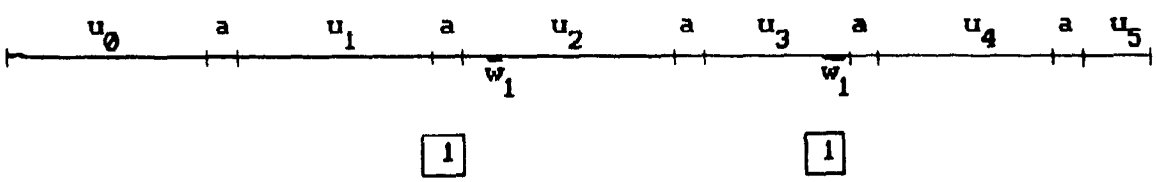
First, let w be a subword of length ≤ 2 in $v_\emptyset av_1av_2$. Then w is a subword of length ≤ 2 in $v_\emptyset av_1\ldots av_{m-3}$. But since $u_\emptyset au_1au_2 \sim_{(2)} v_\emptyset av_1\ldots av_{m-3}$, w is a subword of $u_\emptyset au_1au_2$.

Now, let $w = w_1\ldots w_{|w|}$ be a subword of length ≤ 2 in $u_\emptyset au_1au_2$. We want to show that w is a subword of $v_\emptyset av_1av_2$. If w

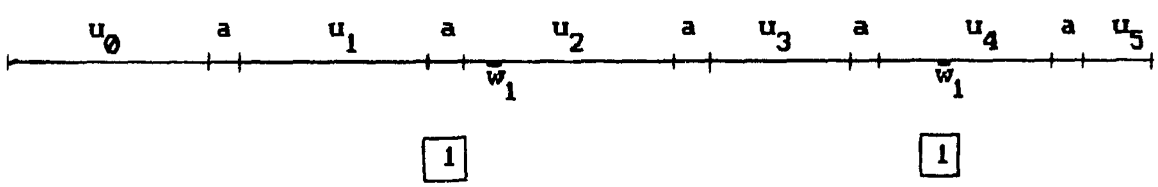
is a subword of $u_\emptyset a u_1$, w is a subword of $v_\emptyset a v_1$ by lemma 4.2.4(1). If not, let j be the first index such that $w_1 \ldots w_j$ is not a subword of $u_\emptyset a u_1$ but $w_1 \ldots w_{j-1}$ is a subword of $u_\emptyset a u_1$. We have to consider the case where $j = 1$ and the case where $j = 2$. In each case, $u_\emptyset a u_1 a u_2 \sim_{(2)} v_\emptyset a v_1 a v_2$ will follow by considering different plays of the game $\mathcal{G}_{(2,1)}(u,v)$. We will illustrate the plays by diagrams. The first move of I will be indicated by (1) and the first move of II by [1].

$j = 1$: We have that w_1 is not a subword of $v_\emptyset a v_1$. $w_1 \neq a$ since otherwise w_1 would be in $u_\emptyset a u_1$ contradicting the choice of j. So let p' be the first position in u after p_2 such that $Q^u_{w_1} p'$. Now, since $u_\emptyset a u_1 a u_2 \sim_{(2)} v_\emptyset a v_1 \ldots a v_{m-3}$ and w_1 is not in $v_\emptyset a v_1$, w_1 occurs between q_2 and q_{m-2}. Let q' be the first position in v after q_2 such that $Q^v_{w_1} q'$. q' is not between q_2 and q_3 in v because then we would have $w_1 aaaa$ in v but not in u. Hence q' is between q_3 and q_{m-2}. Consider the following play of the game $\mathcal{G}_{(2,1)}(u,v)$ (illustrated in the diagram below). Player I in the first move chooses q_2 and q'. Player II should choose an occurrence of a before the first occurrence of w_1 in u (which is in u_2) because in $v_\emptyset a v_1$ there is no occurrence of w_1 and since he needs at least 1 a before the occurrence of a that he chooses, he has to choose p_2. II also needs at least 1 a between and after the positions that he chooses. Player II cannot win this play of the game, a contradiction on the fact that $u \sim_{(2,1)} v$ (II cannot win since there is no occurrence of w_1 between the positions chosen by player I in the first move but there is an occurrence of w_1 between the positions chosen by player II in the first move). Hence $j = 1$ is eliminated. (remark: $j = 1$ is eliminated can also be seen by considering the play of the game $\mathcal{G}_{(2,1)}(u,v)$ where player I in the first move chooses q_1 and q_3. There is no occurrence of w_1 between q_1 and q_3 but there is one between p_1 and p_3 or p_1 and p_4).



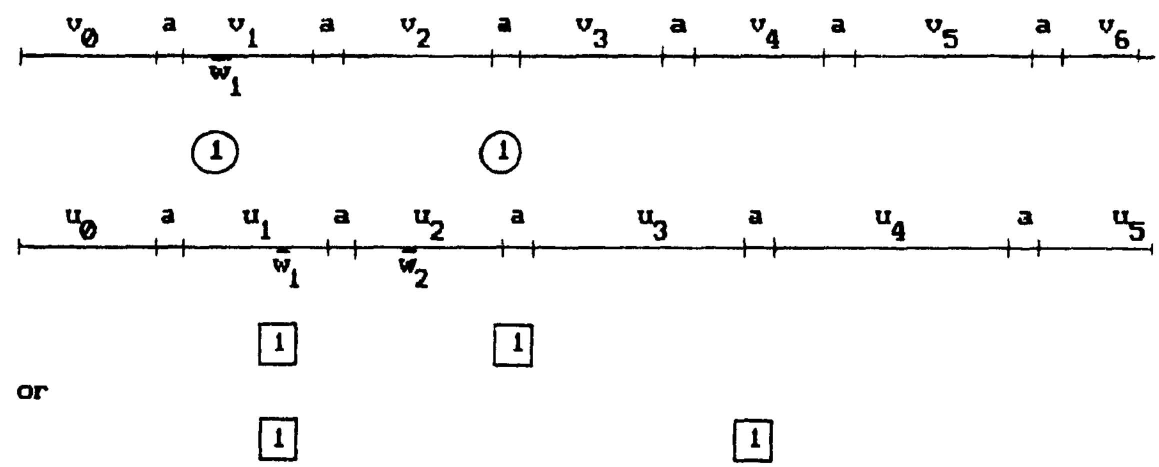


or



$j = 2$: We have that w_1 is a subword of $v_\emptyset av_1$, but we do not have that w_1w_2 is a subword of $v_\emptyset av_1$. If $w_2 = a$, w_1w_2 is a subword of $v_\emptyset av_1a$ and hence of $v_\emptyset av_1av_2$. So, assume $w_2 \neq a$ and let p' be the first position in u after p_2 such that $Q^u_{w_2}p'$. Now, since $u_\emptyset au_1au_2 \sim_{(2)} v_\emptyset av_1\ldots av_{m-3}$, w_2 occurs between q_2 and q_{m-2}. Let q' be the first position in v after q_2 such that $Q^v_{w_2}q'$. Suppose q' is not between q_2 and q_3 in v. If the first occurrence of w_1

in v is in v_1 (and hence in u_1 by lemma 4.2.1(1)), consider the following play of the game $\mathcal{G}_{(2,1)}(u,v)$ (illustrated in the diagram below). Player I in the first move chooses the first occurrence of w_1 in v and q_3 in v. Player II cannot win this play of the game, a contradiction on the fact that $u \sim_{(2,1)} v$ (II cannot win since there is no w_2 between the positions chosen by player I in the first move but there is an occurrence of w_2 between the positions chosen by player II in the first move).



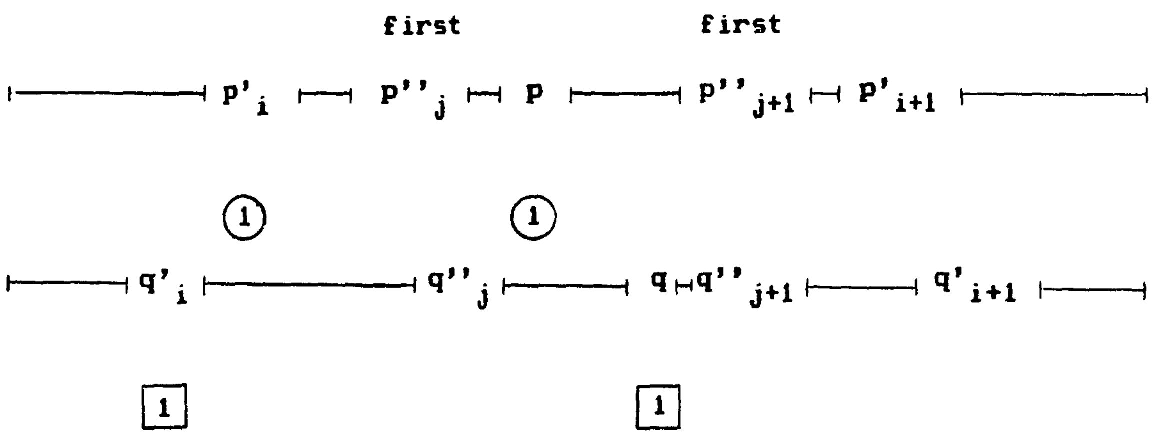
If the first occurrence of w_1 in v is in $v_0 a$, player I in the first move chooses q_1 and q_3 in v. Player II cannot win this play of the game, for the same reason as above. Hence q' should be between q_2 and q_3.

Case 4: $|u|_a > 5$, $|v|_a > 5$.

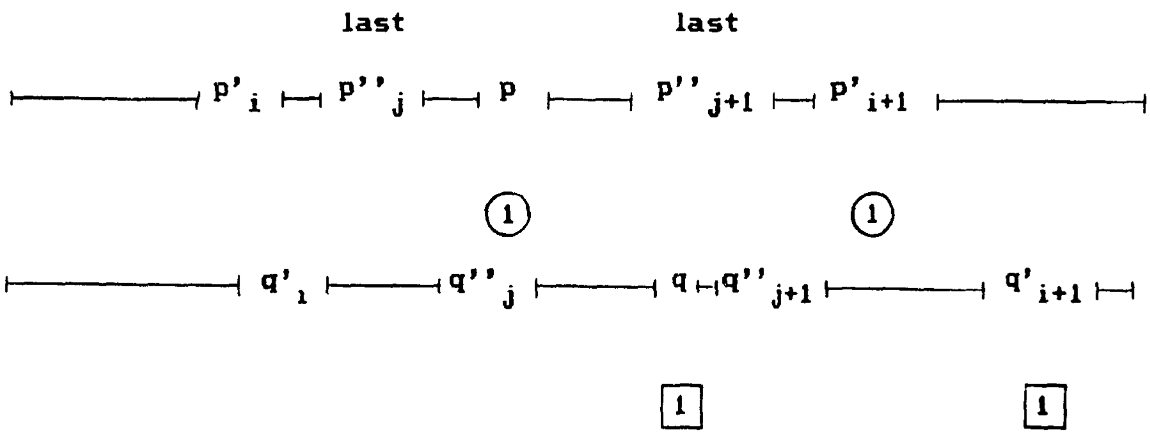
Let $p'_1, \ldots, p'_s$ in u $(p'_1 < \ldots < p'_s)$ $(q'_1, \ldots, q'_s$ in v $(q'_1 < \ldots < q'_s))$ be the positions which spell the first and the last occurrences of every letter in u (v) satisfying (2,3,4,5,6) of lemma 4.2.5. Now if p is any middle position in u (among $p_3, \ldots, p_{n-2}$) chosen by player I in the first move, then $p \in u(p'_i, p'_{i+1})$ for some i, $1 \le i \le s-1$. Then player II chooses a middle position q in v (among $q_3, \ldots, q_{m-2}$) as follows. Let $p''_1, \ldots, p''_{s_i}$ in $u(p'_i, p'_{i+1})$ $(p''_1 < \ldots < p''_{s_i})$ $(q''_1, \ldots, q''_{s_i}$ in $v(q'_i, q'_{i+1})$ $(q''_1 < \ldots < q''_{s_i}))$ be the positions which spell the first and the last occurrences of every letter in $u(p'_i, p'_{i+1})$ $(v(q'_i, q'_{i+1}))$ satisfying (2'''',3'''') of lemma 4.2.6. First, if $p = p''_j$ for some j, $1 \le j \le s_i$, then let $q = q''_j$. $u[1,p) \sim_{(2)} v[1,q)$ and $u(p,|u|] \sim_{(2)} v(q,|v|]$ follow from lemma 4.2.6(3''''). Second, if $p \in u(p''_j, p''_{j+1})$ for some j, $1 \le j \le s_i - 1$, then q will be chosen according to the following rules, rules 1 to 4, which describe different plays of the game $\mathcal{G}_{(2,1)}(u,v)$. Rules 1 to 4 depend on p''_j and p''_{j+1} being first or last occurrences of letters in $u(p'_i, p'_{i+1})$ (remark: it can happen that, for example, p''_j is both a first and a last occurrence of a letter; in such a case, q will be chosen according to any of the rules that apply). We will illustrate the plays by diagrams. The first move of I will be indicated as before by (1) and the first move of II by [1].

Rule 1: Rule 1 is an application of lemma 4.2.5(5). If p''_j and p''_{j+1} are first occurrences of letters in $u(p'_i, p'_{i+1})$, then consider the play of the game $\mathcal{G}_{(2,1)}(u,v)$ where, in move 1, player I chooses p'_i and p. Player II should choose q'_i and a position q in $v(q'_i, q'_{i+1})$ such that $Q^v_a q$ and $u(p'_i, p) \sim_1 v(q'_i, q)$. Since

p''_j and p''_{j+1} (and hence q''_j and q''_{j+1}) are first occurrences of letters in $u(p'_i,p'_{i+1})$ ($v(q'_i,q'_{i+1})$), q must be in $v(q''_j,q''_{j+1})$ (otherwise there would be contradiction with $u(p'_i,p) \sim_1 v(q'_i,q)$). More precisely, q is not in $v(q'_i,q''_j)$ and $q \neq q''_j$ since otherwise there would be an occurrence of the letter of p''_j in $u(p'_i,p)$ but not in $v(q'_i,q)$; q is not in $v(q''_{j+1},q'_{i+1})$ since otherwise there would be an occurrence of the letter of q''_{j+1} in $v(q'_i,q)$ but not in $u(p'_i,p)$; $q \neq q''_{j+1}$ since otherwise $Q^v_a q''_{j+1}$ and hence $Q^u_a p''_{j+1}$ contradicting the fact that p''_{j+1} is the first occurrence of a letter in $u(p'_i,p'_{i+1})$ ($Q^u_a p$ and $p < p''_{j+1}$).

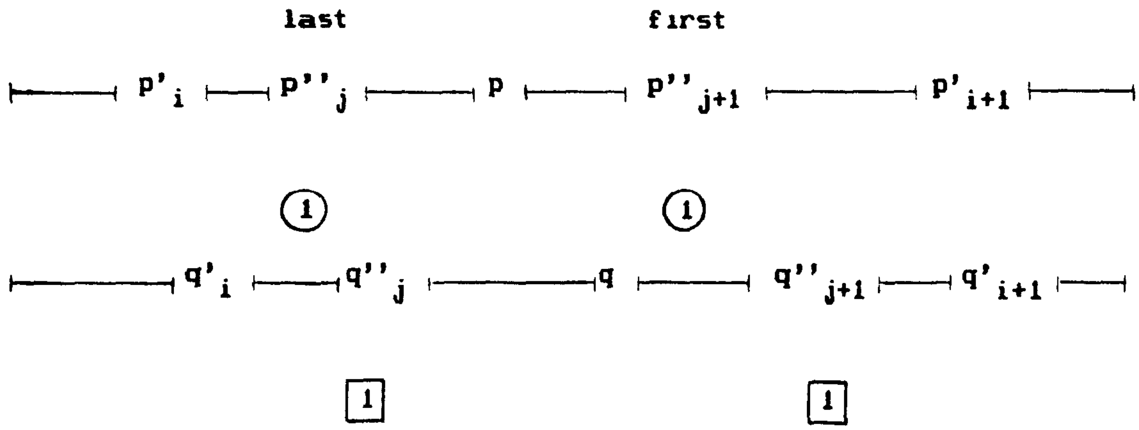


Rule 2: Rule 2 is an application of lemma 4.2.5(5). If p''_j and p''_{j+1} are last occurrences of letters in $u(p'_i,p'_{i+1})$, then player I, in the first move chooses p and p'_{i+1}. Player II should choose q'_{i+1} and a position q in $v(q'_i,q'_{i+1})$ such that $Q^v_a q$ and $u(p,p'_{i+1}) \sim_1 v(q,q'_{i+1})$. Similarly as in rule 1, q must be in $v(q''_j,q''_{j+1})$.

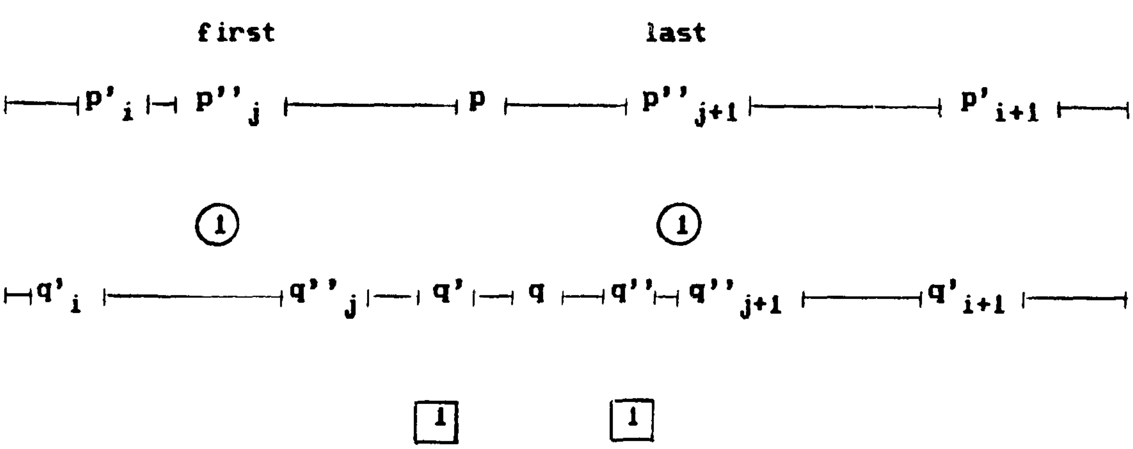


Rules 3 and 4 are applications of lemma 4.2.5(6).

Rule 3: If p''_j is the last occurrence of a letter in $u(p'_i,p'_{i+1})$ and p''_{j+1} is the first occurrence of a letter in $u(p'_i,p'_{i+1})$, then player I, in the first move chooses p''_j and p''_{j+1}. Hence there exist q' and q'' in $v(q'_i,q'_{i+1})$ $(q' < q'')$ such that $Q^v_b q'$ if and only if $Q^u_b p''_j$ if and only if $Q^v_b q''_j$, $Q^v_b q''$ if and only if $Q^u_b p''_{j+1}$ if and only if $Q^v_b q''_{j+1}$, $b \in A$ and $u(p''_j,p''_{j+1}) \sim_1 v(q',q'')$. $q' \leq q''_j$ (since q''_j is the last occurrence of the letter of q' and q''_j in $v(q'_i,q'_{i+1})$) and $q''_{j+1} \leq q''$ (since q''_{j+1} is the first occurrence of the letter of q'' and q''_{j+1} in $v(q'_i,q'_{i+1})$). $q' < q''_j$ or $q''_{j+1} < q''$ would contradict $u(p''_j,p''_{j+1}) \sim_1 v(q',q'')$. More precisely, $q' < q''_j$ $(q''_{j+1} < q'')$ would imply an occurrence of the letter of q''_j (q''_{j+1}) in $v(q',q'')$ but there is no such occurrence in $u(p''_j,p''_{j+1})$. Hence $q' = q''_j$ and $q'' = q''_{j+1}$. Since $u(p''_j,p''_{j+1}) \sim_1 v(q''_j,q''_{j+1})$, there exists q in $v(q''_j,q''_{j+1})$ such that $Q^v_a q$.



Rule 4: If p''_j is the first occurrence of a letter in $u(p'_i,p'_{i+1})$ and p''_{j+1} is the last occurrence of a letter in $u(p'_i,p'_{i+1})$, then player I, in the first move chooses p''_j and p''_{j+1}. Hence there exist q' and q'' such that $q''_j \le q' < q'' \le q''_{j+1}$ and satisfying $Q_b^v q'$ if and only if $Q_b^u p''_j$ if and only if $Q_b^v q''_j$, $Q_b^v q''$ if and only if $Q_b^u p''_{j+1}$ if and only if $Q_b^v q''_{j+1}$, $b \in A$ and $u(p''_j,p''_{j+1}) \sim_1 v(q',q'')$. Since $u(p''_j,p''_{j+1}) \sim_1 v(q',q'')$, there exists q in $v(q',q'')$ such that $Q_a^v q$.



In rules 1 to 4, the facts that $u[1,p) \sim_{(2)} v[1,q)$ and $u(p,|u|] \sim_{(2)} v(q,|v|]$ will follow similarly as lemma 4.2.6(3''''). We show $u(p,|u|] \sim_{(2)} v(q,|v|]$ for rule 4. Let $w = w_1 \ldots w_{|w|}$ be a subword of length ≤ 2 in $v(q,|v|]$ (similar if in $u(p,|u|]$). We want to show that w is a subword of $u(p,|u|]$. If $|w| = 1$, then there is an occurrence of w_1 in $v[q'_{i+1},|v|]$ (and hence in $u[p'_{i+1},|u|]$) from the choice of the p'_r's and the q'_r's and lemma 4.2.5(1,2) and the proof is complete. If $|w| = 2$, and w is in $v(q'_{i+1},|v|]$, then w is in $u(p'_{i+1},|u|]$ by lemma 4.2.5(3). If there is an occurrence of w_2 in $v(q'_{i+1},|v|]$ (and hence in $u(p'_{i+1},|u|]$ by lemma 4.2.5(3)) and $Q^v_{w_1} q'_{i+1}$ (and hence $Q^u_{w_1} p'_{i+1}$ by lemma 4.2.5(2)) the proof is complete. Otherwise, there is an occurrence of w_2 in $v[q'_{i+1},|v|]$ (and hence in $u[p'_{i+1},|u|]$) from the choice of the p'_r's and the q'_r's and lemma 4.2.5(1,2) and there is also an occurrence of w_1 in $v(q,q'_{i+1})$. From the choice of the q''_r's, there exists k, $k \geq j+1$, such that $Q^v_{w_1} q''_k$. Hence, from the choice of the p''_r's and lemma 4.2.6(1'''',2''''), $Q^u_{w_1} p''_k$. The result follows.[]

Chapter 5

EQUATIONS

The problem of finding equations satisfied in the M-varieties $\mathbf{V}_k$, problem related to the decidability of the $\mathbf{V}_k$'s, is the subject of this chapter. Studying properties of the recognizers $A^*/\sim_m$ sheds some light on the syntactic monoids of the star-free languages. The material of this chapter appears in [B188a].

Let u, $v \in A^*$. A monoid M satisfies the equation $u = v$ if and only if $u\varphi = v\varphi$ for all morphisms $\varphi : A^* \to M$. One can show that the class of monoids M satisfying the equation $u = v$ is an M-variety, denoted by $\mathbf{W}(u,v)$. Let $(u_n,v_n)_{n>0}$ a sequence of pairs of words of A^*. Consider the following M-varieties:
$\mathbf{W}' = \bigcap_{n>0} \mathbf{W}(u_n,v_n)$ and $\mathbf{W}'' = \bigcup_{m>0} \bigcap_{n\geq m} \mathbf{W}(u_n,v_n)$. We say that $\mathbf{W}'$ ($\mathbf{W}''$) is *defined* (*ultimately defined*) by the equations $u_n = v_n$ ($n > 0$): this corresponds to the fact that a monoid M is in $\mathbf{W}'$ ($\mathbf{W}''$) if and only if M satisfies the equations $u_n = v_n$ for all $n > 0$ (for all n sufficiently large). The equational approach to varieties is discussed in Eilenberg [Ei76]. Eilenberg showed that every M-variety is ultimately defined by a sequence of equations. For example, the M-variety $\mathbf{V}$ of aperiodic monoids is ultimately defined

by the equations $x^n = x^{n+1}$ $(n > 0)$. One can show that every M-variety generated by a single monoid is defined by a sequence of equations. $V_{1,m}$ being generated by $A^*/\sim_{(m)}$, are the M-varieties $V_{1,m}$ defined by a finite sequence of equations? An attempt to answer this open problem is made in the following section.

1. *Equations related to the first level of the Straubing hierarchy*

An attempt to generalize the following proposition is made in this section. A proof of part(2) appears nowhere in the literature. We include a proof based on combinatorial properties of the congruences $\sim_{(m)}$. We remind the reader that from corollary 2.2.5, we have for $k \geq 1$, $M \in V_k$ if and only if for every morphism $\varphi : A^* \to M$ there exists $\bar{m} = (m_1, \ldots, m_k)$ such that $\sim_{\bar{m}}$ refines φ, or, more precisely, using theorem 2.2.4, for $k \geq 1$, $m \geq 1$, $M \in V_{k,m}$ if and only if for every morphism $\varphi : A^* \to M$ there exists $\bar{m} = (m, m_2, \ldots, m_k)$ such that $\sim_{\bar{m}}$ refines φ.

Proposition 5.1.1 Simon [Si72]

(1) The M-variety $V_{1,1}$ is defined by the equations $x = x^2$ and $xy = yx$, i.e., $V_{1,1}$ is the M-variety of idempotent and commutative monoids.

(2) The M-variety $V_{1,2}$ is defined by the equations $xyzx = xyxzx$ and $(xy)^2 = (yx)^2$.

The above proposition follows from the following combinatorial properties of the congruences $\sim_{(m)}$.

Lemma 5.1.2 Simon [Si75]

Let $m \geq 1$. Let $u, v \in A^*$. If $u \sim_{(m)} v$, then there exists w such that u is a subword of w, v is a subword of w and $u \sim_{(m)} w \sim_{(m)} v$.

Lemma 5.1.3 Simon [Si75]

Let $m \geq 1$. Let $u, v \in A^*$. Then

(1) $u \sim_{(m)} uv$ if and only if there exist $u_1, \ldots, u_m \in A^*$ such that $u = u_1 \ldots u_m$ and $v\alpha \subseteq u_m\alpha \subseteq \ldots \subseteq u_1\alpha$.

(2) $u \sim_{(m)} vu$ if and only if there exist $u_1, \ldots, u_m \in A^*$ such that $u = u_1 \ldots u_m$ and $v\alpha \subseteq u_1\alpha \subseteq \ldots \subseteq u_m\alpha$.

Lemma 5.1.4 Simon [Si75]

Let $m \geq 1$. Let $a \in A$ and $u, v \in A^*$. Then $uv \sim_{(m)} uav$ if and only if there exist nonnegative integers m_1, m_2, $m_1+m_2 \geq m$ such that $u \sim_{(m_1)} ua$ and $v \sim_{(m_2)} av$.

Proof of proposition 5.1.1

(1) We have to prove that $M \in \mathbf{V}_{1,1}$ if and only if it satisfies the equations $xy = yx$ and $x = x^2$ or $M \in \mathbf{V}_{1,1}$ if and only if for every morphism $\varphi : A^* \to M$, $xy\varphi = yx\varphi$ and $x\varphi = x^2\varphi$. Suppose $M \in \mathbf{V}_{1,1}$ and let $\varphi : A^* \to M$ be a morphism. Then $\sim_{(1)} \subseteq \varphi$. Now $xy \sim_{(1)} yx$ and $x \sim_{(1)} x^2$. Hence we have the result.

Conversely, let $\varphi : A^* \to M$ be a surjective morphism satisfying $xy\varphi = yx\varphi$ and $x\varphi = x^2\varphi$. We want to show that $\sim_{(1)} \subseteq \varphi$. Let

$f \sim_{(1)} g$. Lemma 5.1.2 permits to consider only the case where f is a subword of g. We observe also that if f is a subword of h and h is a subword of g, we have also $f \sim_{(1)} h$. Hence we have only to consider the case where $f = uv$ and $g = uav$. So we have $uv \sim_{(1)} uav$.

Case 1: $u \sim_{(1)} ua$ or $u = u_1au_2$ for some $u_1, u_2 \in A^*$. Hence $uv\varphi = u_1au_2v\varphi = u_1a^2u_2v\varphi$ (by using $x^2\varphi = x\varphi$) $= u_1au_2av\varphi$ (by using $xy\varphi = yx\varphi$) $= uav\varphi$.

Case 2: $v \sim_{(1)} av$ or $v = v_1av_2$ for some $v_1, v_2 \in A^*$. Similar to Case 1.

(2) If $M \in V_{1,2}$, let $\varphi : A^* \to M$ be a morphism. Then $\sim_{(2)} \subseteq \varphi$. Now $(xy)^2 \sim_{(2)} (yx)^2$ and $xyxzx \sim_{(2)} xyzx$. Hence $(xy)^2\varphi = (yx)^2\varphi$ and $xyxzx\varphi = xyzx\varphi$. Now let $\varphi : A^* \to M$ be a surjective morphism satisfying $(xy)^2\varphi = (yx)^2\varphi$ and $xyxzx\varphi = xyzx\varphi$. We want to show that $\sim_{(2)} \subseteq \varphi$. Let $f \sim_{(2)} g$. Similarly to (1), by lemma 5.1.2 we have only to consider the case where $f = uv$ and $g = uav$. So we have $uv \sim_{(2)} uav$. Lemma 5.1.4 implies the existence of m_1 and m_2 such that $m_1+m_2 \geq 2$, $u \sim_{(m_1)} ua$ and $v \sim_{(m_2)} av$. We have the following cases.

Case 1: $u \sim_{(1)} ua$ and $v \sim_{(1)} av$. Lemma 5.1.3 implies that $u = u_1au_2$, $v = v_1av_2$ for some $u_1, u_2, v_1, v_2 \in A^*$. $uv = u_1au_2v_1av_2$ and $uav = u_1au_2av_1av_2$. Hence $uv\varphi = uav\varphi$ by using $xyxzx\varphi = xyzx\varphi$.

Case 2: $u \sim_{(2)} ua$ and $v \sim_{(0)} av$. Lemma 5.1.3 implies the existence of u_1, u_2 and $u_3 \in A^*$ such that $u = u_1au_2au_3$, u_3 does not contain any a and every letter of u_3 is in either u_1 or u_2. If $u_3 = 1$, then $uv\varphi = u_1au_2au_3v\varphi = u_1au_2av\varphi = u_1au_2aav\varphi = uav\varphi$, (by using $xyxzx\varphi = xyzx\varphi$). If $u_3 = a_1 \ldots a_n$, $n \geq 1$, $(a_1, \ldots, a_n \neq a)$, then we have $uav\varphi = u_1au_2aa_1 \ldots a_nav\varphi$

$= u_1au_2aa_1 \ldots a_{n-1}a_naa_nav\varphi$ (by using $xyxzx\varphi = xyzx\varphi$ two times)

$= u_1au_2aa_1 \ldots a_{n-1}aa_naa_nv\varphi$ (by using $(xy)^2\varphi = (yx)^2\varphi$)

$= u_1au_2aa_1 \ldots a_{n-1}aa_nv\varphi$ (by using $xyxzx\varphi = xyzx\varphi$ two times)

$= u_1au_2aa_1 \ldots a_{n-2}a_{n-1}aa_{n-1}aa_nv\varphi = u_1au_2aa_1 \ldots a_{n-2}aa_{n-1}aa_{n-1}a_nv\varphi$

$= u_1au_2aa_1 \ldots a_{n-2}aa_{n-1}a_nv\varphi = \ldots = u_1au_2aa_1aa_2 \ldots a_nv\varphi$

$= u_1au_2a_1aa_1aa_2 \ldots a_nv\varphi = u_1au_2aa_1aa_1a_2 \ldots a_nv\varphi = u_1au_2aa_1 \ldots a_nv\varphi = uv\varphi$.

Case 3: $u \sim_{(0)} ua$ and $v \sim_{(2)} av$. Similar to Case 2.[]

We would like to generalize the above proposition 5.1.1. In order to do this, let us define classes of equations as follows. For $m \geq 1$, $C^1_{(m)}$ consists of the equations $(xy)^m = u$ where u is any word consisting of m blocks, each block being xy or yx. These equations describe different ways of permuting an equal number of x and y. The equation $(xy)^m = (yx)^m$ is such an example. It is easily seen that monoids in $V_{1,m}$ satisfy $C^1_{(m)}$. This comes from the fact that if $M \in V_{1,m}$, then $M < A^*/\sim_{(m)}$ for a suitable A. Since $A^*/\sim_{(m)}$ satisfies $C^1_{(m)}$, M satisfies $C^1_{(m)}$.

For $m = 1$, $C^2_{(m)}$ consists of the equation $x = x^2$, for $m \geq 2$, of the following equation

$$xyx^{m-2}zx = xyx^{m-1}zx.$$

The above equation generalizes $x^m = x^{m+1}$ and is easily seen to be satisfied in $\mathbf{V}_{1,m}$, a consequence of proposition 3.2.1. The equations in $\cup_{r\leq 2} C^r_{(2)}$ can be reduced to the equations defining $\mathbf{V}_{1,2}$ of proposition 5.1.1(2). We have

Proposition 5.1.5

(1) $\mathbf{V}_{1,1}$ is defined by $C^1_{(1)} \cup C^2_{(1)}$,

(2) $\mathbf{V}_{1,2}$ is defined by $C^1_{(2)} \cup C^2_{(2)}$.

Let us now define the class $C^3_{(m)}$. For $3 \leq m$, $C^3_{(m)}$ consists of the following equations

$$xzx^{m-3}y^e xvxwy = xzx^{m-2}y^e xvxwy$$
$$ywxvxy^e x^{m-3}zx = ywxvxy^e x^{m-2}zx$$
$$xzx^{m-3}y^e xvywx = xzx^{m-2}y^e xvywx$$
$$xwyvxy^e x^{m-3}zx = xwyvxy^e x^{m-2}zx$$

where $e = 1, \ldots, m-1$.

The class $C^4_{(m)}$, for $4 \leq m$, will consist of the equations

$$xzx^{m-4}yxy^e xvxwy = xzx^{m-3}yxy^e xvxwy$$
$$ywxvxy^e xyx^{m-4}zx = ywxvxy^e xyx^{m-3}zx$$
$$xzx^{m-4}y^e x^2 yvxwy = xzx^{m-3}y^e x^2 yvxwy$$
$$ywxvyx^2 y^e x^{m-4}zx = ywxvyx^2 y^e x^{m-3}zx$$

$$xzx^{m-4}y^ex^2yvywx = xzx^{m-3}y^ex^2yvywx$$

$$xwyvyx^2y^ex^{m-4}zx = xwyvyx^2y^ex^{m-3}zx$$

$$xzx^{m-4}yxy^exvywx = xzx^{m-3}yxy^exvywx$$

$$xwyvxy^exyx^{m-4}zx = xwyvxy^exyx^{m-3}zx$$

where $e = 1, \ldots, m-2$.

These are easy exercises on the games $\mathcal{G}_{(m)}$ to verify that for $3 \le m$ $(4 \le m)$, every monoid in $\mathbf{V}_{1,m}$ satisfies $C^3_{(m)}$ $(C^4_{(m)})$. In $C^3_{(m)}$, the instances with $e > m-1$ follow from those with $1 \le e \le m-1$ and $C^2_{(m)}$. Similarly, in $C^4_{(m)}$, the instances with $e > m-2$ follow from those with $1 \le e \le m-2$ and $C^3_{(m)}$. Further classes of equations $C^r_{(m)}$, for $r \le m$, can be described, each containing equations generalizing $x^m = x^{m+1}$ and satisfied in $\mathbf{V}_{1,m}$, each equation involving powers of x not less than $m-r$. More precisely, the class $C^r_{(m)}$, for $5 \le r \le m$, will consist of the equations

$$xzx^{m-r}yxy^fxu_1 = xzx^{m-(r-1)}yxy^fxu_1$$

$$u_2xy^fxyx^{m-r}zx = u_2xy^fxyx^{m-(r-1)}zx$$

$$xzx^{m-r}y^fxxyu_1 = xzx^{m-(r-1)}y^fxxyu_1$$

$$u_2yxxy^fx^{m-r}zx = u_2yxxy^fx^{m-(r-1)}zx$$

where $f = 1, \ldots, m-(r-2)$, and where
$xzx^{m-(r-1)}y^exu_1 = xzx^{m-(r-2)}y^exu_1$ and
$u_2xy^ex^{m-(r-1)}zx = u_2xy^ex^{m-(r-2)}zx$ are in $C^{r-1}_{(m)}$ for some u_1, u_2 and e between 1 and $m-(r-3)$. We have the following

Theorem 5.1.6

Every monoid in $\mathbf{V}_{1,m}$ satisfies $\bigcup_{r\le m} C^{r}_{(m)}$.

Proof The result follows from the congruence characterization of $\mathbf{V}_{1,m}$ and the properties of $\sim_{(m)}$ stated in lemmas 5.1.3 and 5.1.4.[]

Simplifications occur. For example,

Proposition 5.1.7

The equations in $\bigcup_{r\le 3} C^{r}_{(3)}$ reduce to the following system

$$(yx)^3 = (xy)^3$$

$$xzyxvxwy = xzxyxvxwy$$

$$ywxvxyzx = ywxvxyxzx.$$

Proof Let us show how the equation $(xy)^3 = xy^2x^2y$ in $C^{1}_{(3)}$ comes from the above system. $xy^2x^2y = xyxyx^2y$ (using the second equation with $z := y$, $v := 1$ and $w := 1$) $= xyx^2yx^2y$ (second equation with $y := 1$, $z := y$, $v := y$ and $w := 1$) $= xyx^2yxy$ (third equation with $v := 1$, $w := 1$ and $z := 1$) $= xyxyxy$ (second equation with $y := 1$, $z := y$, $v := y$ and $w := 1$) $= (xy)^3$. The equation $xzy^2xvxwy = xzxy^2xvxwy$ in $C^{3}_{(3)}$ comes from the above system as follows: $xzy^2xvxwy = x(zy)yxvxwy = x(zy)xyxvxwy$ (using the second equation with $z := zy$) $= xzyx^2yxvxwy$ (second equation with $y := 1$, $z := zy$, $v := y$ and $w := 1$) $= xzxyx^2yxvxwy$ (second equation with $v := 1$ and $w := 1$) $= xzxyxyxvxwy$ (second equation with $y := 1$, $z := y$, $v := y$ and $w := 1$) $= xzxyyxvxwy$ (second equation with $z := y$) $= xzxy^2xvxwy$. The other equations follow similarly.[]

Proposition 5.1.8

The M-variety $V_{1,3}$ is defined by the equations in proposition 5.1.7.

Proof If $M \in V_{1,3}$, let $\varphi : A^* \to M$ be a morphism. Then $\sim_{(3)} \subseteq \varphi$. Now $(yx)^3 \sim_{(3)} (xy)^3$, $xzyxvxwy \sim_{(3)} xzxyxvxwy$ and $ywxvxyzx \sim_{(3)} ywxvxyxzx$. Hence $(yx)^3\varphi = (xy)^3\varphi$, $xzyxvxwy\varphi = xzxyxvxwy\varphi$ and $ywxvxyzx\varphi = ywxvxyxzx\varphi$. Now let $\varphi : A^* \to M$ be a surjective morphism satisfying $(yx)^3\varphi = (xy)^3\varphi$, $xzyxvxwy\varphi = xzxyxvxwy\varphi$ and $ywxvxyzx\varphi = ywxvxyxzx\varphi$. Let us state first some useful consequences of the equations, like (1) $xuxvx\varphi = xux^2vx\varphi$, (2) $xzyx^2wy\varphi = xzxyx^2wy\varphi$ and (3) $ywx^2yzx\varphi = ywx^2yxzx\varphi$. We want to show that $\sim_{(3)} \subseteq \varphi$. Let $f \sim_{(3)} g$. Similarly to proposition 5.1.1, by lemma 5.1.2 we have only to consider the case where $f = uv$ and $g = uav$. So we have $uv \sim_{(3)} uav$. Lemma 5.1.4 implies the existence of m_1 and m_2 such that $m_1+m_2 \geq 3$, $u \sim_{(m_1)} ua$ and $v \sim_{(m_2)} av$. We have the following cases.

Case 1: $u \sim_{(2)} ua$ and $v \sim_{(1)} av$. Lemma 5.1.3 implies the existence of u_1, u_2, u_3, v_1 and $v_2 \in A^*$ such that $u = u_1au_2au_3$, $v = v_1av_2$, v_1 and u_3 do not contain any a and every letter of u_3 is in either u_1 or u_2. If $u_3 = 1$, then $uv\varphi = u_1au_2av_1av_2\varphi$ $= u_1au_2a^2v_1av_2\varphi = uav\varphi$ (by using (1)). If $u_3 = a_1 \ldots a_n$, $n \geq 1$, $(a_1, \ldots, a_n \neq a)$, then we have $uv\varphi = u_1au_2aa_1\ldots a_nv_1av_2\varphi$ $= u_1au_2a^2a_1a_2\ldots a_nv_1av_2\varphi$ (by using (1)) $= u_1au_2a^2a_1aa_2\ldots a_nv_1av_2\varphi$ (by using (3) and the fact that a_1 is in u_1 or u_2)

$= u_1au_2a^2a_1a^2a_2a_3\ldots a_nv_1av_2\varphi$ (1) $= u_1au_2a^2a_1a^2a_2aa_3\ldots a_nv_1av_2\varphi$ ((3) and the fact that a_2 is in u_1 or u_2)

$= u_1au_2a^2a_1a^2a_2a^2a_3\ldots a_nv_1av_2\varphi$ (1)

$= \ldots = u_1au_2a^2a_1a^2a_2a^2a_3a^2\ldots a^2a_nv_1av_2\varphi$

$= u_1au_2a^2a_1a^2a_2a^2a_3a^2\ldots a^2a_nav_1av_2\varphi = u_1au_2a^2a_1aa_2a^2a_3a^2\ldots a^2a_nav_1av_2\varphi$

$= u_1au_2a^2a_1a_2a^2a_3a^2\ldots a^2a_nav_1av_2\varphi = \ldots = u_1au_2a^2a_1a_2a_3\ldots a_nav_1av_2\varphi$

$= u_1au_2aa_1a_2a_3\ldots a_nav_1av_2\varphi = uav\varphi$.

Case 2: $u \sim_{(1)} ua$ and $v \sim_{(2)} av$. Similar to case 1.

Case 3: $u \sim_{(3)} ua$ and $v \sim_{(\emptyset)} av$. Lemma 5.1.3 implies the existence of u_1, u_2, u_3, u_4 and $u_5 \in A^*$ such that $u = u_1au_2u_3au_4au_5$, u_5 does not contain any a, every letter of u_5 is in either u_3 or u_4, every letter of u_3 and u_4 is in u_1au_2. If $u_5 = 1$ and $u_4 = 1$, then $uv\varphi = u_1au_2u_3a^2v\varphi = u_1au_2u_3a^3v\varphi$ (by using (1)) $= uav\varphi$. If $u_5 = 1$ and $u_4 = b_1\ldots b_n$, $n \geq 1$, then

$uv\varphi = u_1au_2u_3ab_1\ldots b_{n-1}b_nav\varphi = u_1au_2u_3a^2b_1b_2\ldots b_nav\varphi$ (by using (1))

$= u_1au_2u_3a^2b_1ab_2\ldots b_nav\varphi$ ((3) and b_1 is in u_1au_2)

$= u_1au_2u_3a^2b_1a^2b_2b_3\ldots b_nav\varphi$ (1) $= u_1au_2u_3a^2b_1a^2b_2ab_3\ldots b_nav\varphi$ ((3) and b_2 is in u_1au_2) $= \ldots = u_1au_2u_3a^2b_1a^2b_2a^2b_3\ldots b_{n-1}a^2b_nav\varphi$

$= u_1au_2u_3a^2b_1a^2b_2a^2b_3\ldots b_{n-1}a^2b_na^2v\varphi$

$= u_1au_2u_3a^2b_1ab_2a^2b_3\ldots b_{n-1}a^2b_na^2v\varphi$ (1)

$= u_1au_2u_3a^2b_1b_2a^2b_3\ldots b_{n-1}a^2b_na^2v\varphi$ ((3) and b_1 is in u_1au_2)

$= \ldots = u_1au_2u_3a^2b_1b_2\ldots b_na^2v\varphi = u_1au_2u_3ab_1\ldots b_na^2v\varphi = uav\varphi$. Now, let $u_5 = c_1\ldots c_t$, $t \geq 1$, (c_1, ..., $c_t \neq a$). We have

$uav\varphi = u_1au_2u_3au_4ac_1c_2\ldots c_tav\varphi = u_1au_2u_3au_4a^2c_1c_2\ldots c_tav\varphi$ (by using (1)) $= u_1au_2u_3au_4a^2c_1ac_2\ldots c_tav\varphi$ (by using (3) and the fact that

c_1 is in u_3 or u_4) $= u_1au_2u_3au_4a^2c_1a^2c_2\ldots c_tav^{\varphi}$ (using (1))

$= u_1au_2u_3au_4a^2c_1a^2c_2ac_3\ldots c_tav^{\varphi}$ (using (3) and the fact that c_2 is in u_3 or u_4)

.

. (using (1), (3) and the fact that u_5 is in u_3 or u_4)

.

$= u_1au_2u_3au_4a^2c_1a^2c_2a^2c_3a^2\ldots a^2c_tav^{\varphi}$

$= u_1au_2u^tc_tb_1^t\ldots b_{s_t}^t\, a^2c_1a^2c_2a^2c_3a^2\ldots a^2c_{t-1}a^2c_tav^{\varphi}$ (c_t being in u_3 or u_4 implies $u_3au_4 = u^tc_tb_1^t\ldots b_{s_t}^t$ where u^t is in A^*, $b_1^t\ldots b_{s_t}^t$ is 1 or b_i^t is in A, $i = 1, \ldots, s_t$)

$= u_1au_2u^tc_t^2b_1^t\ldots b_{s_t}^t\, a^2c_1a^2c_2a^2c_3a^2\ldots a^2c_{t-1}a^2c_tav^{\varphi}$ (using (1) and the fact that c_t is in u_1 or u_2)

$= u_1au_2u^tc_t^2b_1^tc_tb_2^t\ldots b_{s_t}^t\, a^2c_1a^2c_2a^2c_3a^2\ldots a^2c_{t-1}a^2c_tav^{\varphi}$ (using (3) and the fact that b_1^t is in u_1au_2)

. (using (1) and the fact that c_t is in u_1

. or u_2, and using (3) and the fact that

. b_1^t, a, c_1, ..., c_t are in u_1au_2)

$= u_1au_2u^tc_t^2b_1^tc_t^2b_2^t\ldots c_t^2b_{s_t}^t\, c_t^2ac_t^2ac_t^2c_1c_t^2ac_t^2ac_t^2c_2c_t^2ac_t^2a\ldots c_t^2c_{t-1}c_t^2ac_tac_tav^{\varphi}$

$= u_1au_2u^tc_t^2b_1^tc_t^2\ldots c_t^2b_{s_t}^t\, c_t^2ac_t^2ac_t^2c_1c_t^2ac_t^2ac_t^2c_2c_t^2ac_t^2a\ldots c_t^2c_{t-1}c_tac_tac_tac_tv^{\varphi}$

(using $(yx)^{3\varphi} = (xy)^{3\varphi}$)

. (using (1) and the fact that c_t is in u_1

. or u_2, and using (3) and the fact that

. b_i^t, a, c_1, ..., c_t are in u_1au_2)

$= u_1au_2u^tc_tb_1^tb_2^t\ldots b_{s_t}^t a^2c_1a^2c_2a^2c_3\ldots c_{t-1}a^2ac_tv\varphi$

$= u_1au_2u_3au_4a^2c_1a^2c_2a^2c_3\ldots c_{t-1}a^2ac_tv\varphi = u_1au_2u_3au_4ac_1c_2\ldots c_{t-1}ac_tv\varphi$ (using (1) and (3) and the fact that u_5 is in u_3 or u_4). Repeating (using c_{t-1} instead of c_t) the steps that showed that $u_1au_2u_3au_4ac_1c_2\ldots c_tav\varphi = u_1au_2u_3au_4ac_1c_2\ldots c_{t-1}ac_tv\varphi$ leads to $uav\varphi = u_1au_2u_3au_4ac_1c_2\ldots c_{t-2}ac_{t-1}c_tv\varphi$. Repeating again (using $c_{t-2}, \ldots, c_1$) leads to $uav\varphi = u_1au_2u_3au_4a^2c_1c_2\ldots c_{t-1}c_tv\varphi$. If $u_4 = 1$, then $uav\varphi = u_1au_2u_3a^3c_1\ldots c_tv\varphi = u_1au_2u_3a^2c_1\ldots c_tv\varphi$ (using (1)) $= u_1au_2u_3aac_1\ldots c_tv\varphi = uv\varphi$. If $u_4 = a_1\ldots a_r$, $r \geq 1$, then $uav\varphi = u_1au_2u_3aa_1\ldots a_ra^2c_1\ldots c_tv\varphi = u_1au_2u_3a^2a_1\ldots a_ra^2c_1\ldots c_tv\varphi$ (1)
$= u_1au_2u_3a^2a_1aa_2\ldots a_ra^2c_1\ldots c_tv\varphi$ ((3) and a_i in u_1au_2)
$= u_1au_2u_3a^2a_1a^2a_2\ldots a_ra^2c_1\ldots c_tv\varphi$ (1)
$= u_1au_2u_3a^2a_1a^2a_2a^2\ldots a^2a_ra^2c_1\ldots c_tv\varphi$ ((1) and (3) and a_i in u_1au_2) $= u_1au_2u_3a^2a_1a^2a_2a^2\ldots a^2a_rac_1\ldots c_tv\varphi$ ((3) and a_r in u_1au_2)
$= u_1au_2u_3a^2a_1a_2\ldots a_rac_1\ldots c_tv\varphi$ ((1) and (3) and a_i in u_1au_2)
$= u_1au_2u_3aa_1\ldots a_rac_1\ldots c_tv\varphi$ (1) $= u_1au_2u_3au_4ac_1\ldots c_tv\varphi = uv\varphi$.

Case 4: $u \sim_{(\emptyset)} ua$ and $v \sim_{(3)} av$. Similar to case 3.[]

Remark: In the proof of the above proposition, we have used only (1) $xuxvx = xux^2vx$, (2) $xzyx^2wy = xzxyx^2wy$, (3) $ywx^2yzx = ywx^2yxzx$ and (4) $(xy)^3 = (yx)^3$ ((2) is used in cases 2 and 4). The set of equations in proposition 5.1.7 can be deduced from (1), (2), (3) and (4). For example, the equation $xzyxvxwy = xzxyxvxwy$ can be deduced as follows: $xzyxvxwy = xzyx^2vxwy$ (1) $= xzxyx^2vxwy$ (2) $= xzxyxvxwy$ (1). Similarly for ywxvxyzx = ywxvxyxzx. (1), (2), (3) and (4) gives another interesting set of equations for $V_{1,3}$ since it uses at most four variables.

A better understanding of the equations related to the first level of the Straubing hierarchy is useful in finding equations for the higher levels as the following shows. We are now interested in the varieties $V_{k,m}$ for $k > 1$. We would like to find equations satisfied in them. Some of these equations may be selected from the classes previously defined since every $\sim_m$ refines some $\sim_{(m)}$ by proposition 3.2.2. We get the following properties of the congruences $\sim_{(1,m)}$, for $m \geq 1$ related to the variety $V_{2,1}$ or the classes $\mathcal{L}_{(1,m)}$ by theorem 2.2.4.

Proposition 5.1.9

Every monoid in $V_{2,1}$ satisfies

(1) $(xy)^m x(xy)^m = (xy)^m x^2 (xy)^m$ and

(2) $(xy)^m xy(xy)^m = (xy)^m yx(xy)^m$ for some $m > 0$.

Proof The result follows from the congruence characterization of $V_{1,m}$ and lemmas 3.1.1, 5.1.3.[]

The equation $(xy)^m xy(xy)^m = (xy)^m yx(xy)^m$ belongs to $C^1_{(N(1,m))}$. In chapter six, section two, it will be shown that (1) and (2) of the above proposition are part of a finite system of equations ultimately defining $V_{2,1}$ for an alphabet of two letters. Note that the latter equations are of the form $u_1xv_1 = u_1x^2v_1$ and $u_2xyv_2 = u_2yxv_2$ where $x = x^2$ and $xy = yx$ are the defining equations for $V_{1,1}$. This type of equations is called equations in context and has been studied by Therien [The80]. Equations satisfied by $A^*/\sim_{(m_1,m_2)}$ may be selected from the classes $C^r_{(N(m_1,m_2))}$. It is easy to check that $A^*/\sim_{(m_1,m_2)}$ satisfies $C^2_{(N(m_1,m_2))}$, $C^3_{(N(m_1,m_2))}$ for $m_2 > 1$ and $C^4_{(N(m_1,m_2))}$ for $m_2 > 2$. In general, for $\sim_{(m_1,\ldots,m_k)}$, we may be able to choose from $C^r_{(N(m_1,\ldots,m_k))}$. Equations similar to the one in the above proposition (2) will be studied in the next section.

2. *Equations related to higher levels of the Straubing hierarchy*

The M-variety $\mathbf{V}_1$ of $\mathcal{J}$-trivial monoids is ultimately defined by the equations $x^m = x^{m+1}$ and $(xy)^m = (yx)^m$, or $(xy)^m x = (xy)^m = y(xy)^m$ [Pi84a]. This gives a decision procedure for $\mathbf{V}_1$ based on an algebraic characterization of the monoids $M = A^*/\sim$ with $\sim \supseteq \sim_{(m)}$ for some m. $M \in \mathbf{V}_1$ if and only if for all $x, y \in M$, $(xy)^m = (yx)^m$ and $x^m = x^{m+1}$ with m the cardinality of M. The necessity of the condition is immediate since $A^*/\sim_{(m)}$ satisfies the equations. A generalization of the above equations follows.

Let $m \geq 1$. A sequence of equations is defined inductively as follows:

$$Eq_{1,m} \text{ is } (xy)^m = (yx)^m.$$

$Eq_{k+1,m}$ is obtained from $Eq_{k,m}$ in the following manner: $Eq_{k+1,m}$ is obtained by replacing each occurrence of x in $Eq_{k,m}$ by $(xy)^m x(xy)^m$, and each occurrence of y by $(xy)^m y(xy)^m$. For example,

$$Eq_{2,m} \text{ is } ((xy)^m x(xy)^{2m} y(xy)^m)^m = ((xy)^m y(xy)^{2m} x(xy)^m)^m.$$

For all $k, m \geq 1$, let $J_m^{(k)}$ be the collection of all finite monoids which satisfy the pair of equations $Eq_{k,m}$ and $x^m = x^{m+1}$. Easily, $J_m^{(k)} \subseteq J_m^{(k+1)}$ and $J_m^{(k)} \subseteq J_{m+1}^{(k)}$. $J^{(k)} = \cup_{m\geq 1} J_m^{(k)}$ is a M-variety and $J = J^{(1)} \subseteq J^{(2)} \subseteq \ldots$. The M-varieties $J^{(k)}$ were introduced in [BK78]. Members of $J^{(k)}$ are called aperiodic k-mutative monoids. In [St81] it was shown that $V = \cup_{k\geq 1} J^{(k)}$.

The above sequence of equations can also be written as:

$$Eq_{k,m} \text{ is } (x^{(k)}y^{(k)})^m = (y^{(k)}x^{(k)})^m$$

where $x^{(1)} = x$, $y^{(1)} = y$ and $x^{(k+1)} = (x^{(k)}y^{(k)})^m x^{(k)} (x^{(k)}y^{(k)})^m$, $y^{(k+1)} = (x^{(k)}y^{(k)})^m y^{(k)} (x^{(k)}y^{(k)})^m$.

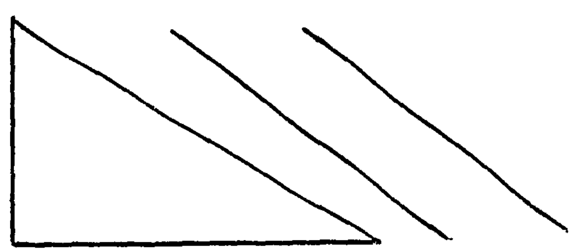
A result of Straubing [St81] states that $V_k \subseteq J^{(k)}$. We include another proof of this result based on the game.

Theorem 5.2.1

$V_k \subseteq J^{(k)}$.

Proof Let $M \in V_k$. We have $M = A^*/\sim$ with $\sim \supseteq \sim_{(m_1,\ldots,m_k)}$ for some $(m_1,\ldots,m_k)$. Hence $A^*/\sim \prec A^*/\sim_{(m_1,\ldots,m_k)}$. $A^*/\sim$ satisfies the equations $Eq_{k,m}$ and $x^m = x^{m+1}$ with $m = N(m_1,\ldots,m_k)$ since they are satisfied by $A^*/\sim_{(m_1,\ldots,m_k)}$. $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies $x^m = x^{m+1}$ since $x^{N(m_1,\ldots,m_k)+1} \sim_{(m_1,\ldots,m_k)} x^{N(m_1,\ldots,m_k)}$ by proposition 3.2.1. By induction on k, $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies $Eq_{k,m}$. We now show that $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies $(x^{(k)}y^{(k)})^{N(m_1,\ldots,m_k)} = (y^{(k)}x^{(k)})^{N(m_1,\ldots,m_k)}$. For $k = 1$, it is true since $A^*/\sim_{(m)}$ satisfies $(xy)^m = (yx)^m$, i.e., $(x^{(1)}y^{(1)})^{N(m)} = (y^{(1)}x^{(1)})^{N(m)}$. Suppose true for k, i.e., $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies $(x^{(k)}y^{(k)})^{N(m_1,\ldots,m_k)} = (y^{(k)}x^{(k)})^{N(m_1,\ldots,m_k)}$. Put $N = N(m,m_1,\ldots,m_k)$. Let us show that $A^*/\sim_{(m,m_1,\ldots,m_k)}$ satisfies $u = (x^{(k+1)}y^{(k+1)})^N = (y^{(k+1)}x^{(k+1)})^N = v$. To see this, consider the natural decompositions of u and v into $x^{(k+1)}$- and $y^{(k+1)}$-segments. II should play according to the following strategy. In the first move, suppose player I chooses from u (the strategy is similar if player I chooses from v). I chooses from at most m segments in u. There is a correspondence between the chosen segments in u and some corresponding segments in v (shown by triangles or lines in the diagram below). We have

$$u = (x^{(k)}y^{(k)})^N x^{(k)} (x^{(k)}y^{(k)})^N (x^{(k)}y^{(k)})^N y^{(k)} (x^{(k)}y^{(k)})^N \ldots$$



$$v = (x^{(k)}y^{(k)})^N y^{(k)} (x^{(k)}y^{(k)})^N (x^{(k)}y^{(k)})^N x^{(k)} (x^{(k)}y^{(k)})^N \ldots$$

The positions chosen from the first $x^{(k+1)}$-segment (or last $y^{(k+1)}$-segment) in u should be played in the first (last) $y^{(k+1)}$- and $x^{(k+1)}$-segments in v. Call a $x^{(k+1)}$- (or $y^{(k+1)}$-) segment in u, a middle segment, if it is not the first (the last) $x^{(k+1)}$- (or $y^{(k+1)}$-) segment. If player I chooses some of his first m positions from middle segments in u, then II should pick exactly those positions which match the positions chosen by I in corresponding segments. Now, by the induction hypothesis and $z^N \sim_{(m_1,\dots,m_k)} z^{N+1}$ we can conclude that

$(x^{(k)}y^{(k)})^N \sim_{(m_1,\dots,m_k)} (x^{(k)}y^{(k)})^N y^{(k)}(x^{(k)}y^{(k)})^N(x^{(k)}y^{(k)})^N$. To see this,

$(x^{(k)}y^{(k)})^N \sim_{(m_1,\dots,m_k)} (x^{(k)}y^{(k)})^{N+1} \sim_{(m_1,\dots,m_k)} (y^{(k)}x^{(k)})^{N+1}$

$\sim_{(m_1,\dots,m_k)} y^{(k)}(x^{(k)}y^{(k)})^N x^{(k)}$. This implies

$(x^{(k)}y^{(k)})^N \sim_{(m_1,\dots,m_k)} (y^{(k)})^N(x^{(k)}y^{(k)})^N(x^{(k)})^N$

$\sim_{(m_1,\dots,m_k)} (y^{(k)})^N(x^{(k)}y^{(k)})^N(x^{(k)})^{N+1} \sim_{(m_1,\dots,m_k)} (x^{(k)}y^{(k)})^N x^{(k)}$.

Similarly, $(x^{(k)}y^{(k)})^N \sim_{(m_1,\dots,m_k)} y^{(k)}(x^{(k)}y^{(k)})^N$. Hence

$(x^{(k)}y^{(k)})^N \sim_{(m_1,\dots,m_k)} (x^{(k)}y^{(k)})^{3N} \sim_{(m_1,\dots,m_k)}$

$(x^{(k)}y^{(k)})^N y^{(k)}(x^{(k)}y^{(k)})^N(x^{(k)}y^{(k)})^N$. Hence the result follows by the induction lemma 3.1.1 and the proof is complete.[]

Similarly to the above proof, one can show that for every monoid M in V_k, there exists $m > 0$ such that M satisfies $(x^{(k)}y^{(k)})^m x^{(k)} = (x^{(k)}y^{(k)})^m = y^{(k)}(x^{(k)}y^{(k)})^m$.

The complexity of a congruence is related to its power of discriminating between words. For example, for m_1, $m_2 \geq 1$, $\sim_{(m_1,m_2)}$ distinguishes $(xy)^n$ and $(yx)^n$ but $\sim_{(n)}$ does not. Hence $(xy)^n = (yx)^n$ is characteristic to the first level V_1. More generally, for sufficiently large m_i, $\sim_{(m_1,\ldots,m_k)}$ distinguishes the words in $Eq_{k-1,m}$. The following theorem proves that $V_k \nsubseteq J^{(k-1)}$, thus proving the infinity of the Straubing hierarchy for an alphabet of at least two letters.

Theorem 5.2.2

$V_k \nsubseteq J^{(k-1)}$.

Proof First, it is easy to see that $V_2 \nsubseteq J^{(1)}$. For $k \geq 3$, we show that for sufficiently large m_i, there is no $m > 0$ such that $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies the equation $u_m = (x^{(k-1)}y^{(k-1)})^m = (y^{(k-1)}x^{(k-1)})^m = v_m$. We illustrate a winning strategy for player I. (I,i) ((II,i)) denotes a position chosen by player I (II) in the i^{th} move, $i = 1, \ldots, k$. Let $N \geq N(m_1,\ldots,m_k)$. Using $x^N \sim_{(m_1,\ldots,m_k)} x^{N+1}$ (proposition 3.2.1), one sees that

$$u_N \sim_{(m_1,\ldots,m_k)} \ldots(x^{(k-2)}y^{(k-2)})^N x$$

$$\uparrow$$

(II,1)

$$(x^{(k-3)}y^{(k-3)})^N x (x^{(k-3)}y^{(k-3)})^N y (x^{(k-3)}y^{(k-3)})^N (x^{(k-2)}y^{(k-2)})^{N-2}$$

$$(x^{(k-3)}y^{(k-3)})^N x (x^{(k-3)}y^{(k-3)})^N y (x^{(k-3)}y^{(k-3)})^N y (x^{(k-2)}y^{(k-2)})^N$$

$$\uparrow \qquad \uparrow$$

(I,2)

Similarly,

$$v_N \sim_{(m_1,\ldots,m_k)} \ldots(x^{(k-2)}y^{(k-2)})^N x$$

$$\uparrow$$

(I,1)

$$(x^{(k-3)}y^{(k-3)})^N x (x^{(k-3)}y^{(k-3)})^N y (x^{(k-3)}y^{(k-3)})^N$$

$$(x^{(k-2)}y^{(k-2)})^{M_1-1} (x^{(k-3)}y^{(k-3)})^N x (x^{(k-3)}y^{(k-3)})^N$$

$$y (x^{(k-3)}y^{(k-3)})^N x (x^{(k-3)}y^{(k-3)})^N y (x^{(k-3)}y^{(k-3)})^N (x^{(k-2)}y^{(k-2)})^{M_2}$$

$$\uparrow \qquad \uparrow$$

(II,2)

where $M_1+M_2 = N-2$. Player I, in the first move, chooses the middle x of the last $x^{(k-2)}$ followed immediately by an $x^{(k-2)}$ in v_N. Player II, in the first move, has to choose the middle x of the last $x^{(k-2)}$ followed immediately by an $x^{(k-2)}$ in u_N (if not, player I in the next $k-1$ moves could win by choosing in the second move the middle x of the last two consecutive $x^{(k-2)}$'s in u_N). Player I, in the second move, chooses the middle y of the last two consecutive $y^{(k-2)}$'s in u_N. Player II, in the second move, cannot choose the middle y of the last two consecutive $y^{(k-2)}$'s in v_N to the right of the previously chosen position. Hence he is forced to choose two $y^{(k-2)}$'s separated by an $x^{(k-2)}$. Player I, in the third move, chooses the middle x of the last two consecutive $x^{(k-3)}$'s in v_N between the positions chosen in the preceding move by II. Player II, in the third move, cannot choose the middle x of the last two consecutive $x^{(k-3)}$'s in u_N between the previously chosen positions by I. Hence he is forced to choose two $x^{(k-3)}$'s separated by an $y^{(k-3)}$ and so on. Player I, in the $k-1^{th}$ move, chooses the last two consecutive x's (or y's) in v_N (or u_N) between the chosen positions in the preceding move by II. Player II, in the $k-1^{th}$ move, is forced to choose two x's (or y's) in u_N (or v_N) separated by a y (or an x). Player I, in the last move, selects that y (or x). Player II loses since he cannot choose a y (or x) between the two consecutive x's chosen in the $k-1^{th}$ move by I. The result follows.[]

Note that similarly to the proof of the preceding theorem, one can show that for sufficiently large m_1, there is no $m > 0$ such that $A^*/\sim_{(m_1,\ldots,m_k)}$ satisfies

$$(x^{(k-1)}y^{(k-1)})^m = (x^{(k-1)}y^{(k-1)})^m x^{(k-1)} \quad \text{and}$$

$$(x^{(k-1)}y^{(k-1)})^m = y^{(k-1)}(x^{(k-1)}y^{(k-1)})^m.$$

The two preceding theorems provide examples of equations that can characterize a level, i.e., monoids of that level satisfy the equations and some monoids of the next level do not satisfy the equations. Although they may not be sufficient to characterize completely a level, they at least form a subset of equations that characterize a level.

3. *Lower bounds on dot-depth*

In this section, Ehrenfeucht-Fraisse games are used to prove lower bounds on a language's complexity through equations. Upper bounds on a language's complexity are obtained by using Thomas' theorem 2.1.1 or theorem 2.1.4. Lower bounds can be demonstrated by using the following criterion:

Criterion for lower bounds

Given any alphabet A, any language $L \subseteq A^*$, to show that L is of dot-depth $\geq k$, it suffices to show that for all $\bar{m} = (m_1,\ldots,m_{k-1})$, there exist $u_{\bar{m}} \in L$, $v_{\bar{m}} \notin L$ such that $u_{\bar{m}} \sim_{\bar{m}} v_{\bar{m}}$. More precisely, to show that L is not in A^*V_{k-1} $(A^*V_{k-1,m})$, it suffices to show that for all $\bar{m} = (m_1,\ldots,m_{k-1})$ $(= (m,m_2,\ldots,m_{k-1}))$, there exist $u_{\bar{m}} \in L$, $v_{\bar{m}} \notin L$ such that $u_{\bar{m}} \sim_{\bar{m}} v_{\bar{m}}$.

A criterion like the above one is useful as long as we know what kind of words $u_{\bar{m}}$ and $v_{\bar{m}}$ can be used. Equations give words $u_{\bar{m}}$ and $v_{\bar{m}}$ in $\sim_{\bar{m}}$-relation. We give some examples.

Example 5.3.1

Let L be the set of all words such that the 10^{th} symbol from the right end is b. One easily can write a $B(\Sigma_2)$-sentence of $\mathcal{L}$

defining L. Hence by Thomas' theorem 2.1.1, we can conclude that the dot-depth of L is smaller than or equal to 2. Define $u_m = (ba)^m baaaaaaaaa \in L$, $v_m = (ba)^m aaaaaaaaa \notin L$. But $u_m \sim_{(m)} v_m$ since $(ba)^m b \sim_{(m)} (ba)^m$ and $\sim_{(m)}$ is a congruence. Hence by the above criterion the dot-depth of L is 2.

Example 5.3.2

Let L be the set of all words in which every pair of adjacent a's appears before any pair of adjacent b's. One easily can write a $B(\Sigma_2)$-sentence of $\mathcal{L}$ defining L. Hence by Thomas' theorem , we can conclude that the dot-depth of L is smaller than or equal to 2. Define $u_{(1,m)} = (ab)^m(ab)(ab)^m \in L$, $v_{(1,m)} = (ab)^m(ba)(ab)^m \notin L$. But $u_{(1,m)} \sim_{(1,m)} v_{(1,m)}$ by proposition 5.1.9(2). Hence by the above criterion $L \notin \{a,b\}^* V_{2,1}$. It implies that the $B(\Sigma_2)$-sentence of $\mathcal{L}$ defining L is not equivalent to a (1,m)-sentence of $\mathcal{L}$.

In the preceding examples, the Ehrenfeucht-Fraisse games have been used to prove lower bounds on the dot-depth of a star-free language or a star-free language's complexity through equations. A conjecture of an effective criterion for $\mathbf{V}_{2,1}$ is the following: for A a fixed alphabet, $L \subseteq A^*$, if M(L) does not satisfy the equations in proposition 5.1.9 (with m the cardinality of M(L)), then $M(L) \notin \mathbf{V}_{2,1}$.

Chapter 6

ON DOT-DEPTH TWO

1. *A sequence of monoids of dot-depth two*

The material of this section appears in [Bl88b]. We show that for positive integers m_1, m_2 and m_3, $A^*/\sim_{(m_1,m_2,m_3)}$ is of dot-depth exactly 2 if and only if $m_2 = 1$. The following lemma shows the necessity of the condition.

Lemma 6.1.1

Let A be an alphabet of at least two letters. Let m_1 and m_3 be positive integers. Then $A^*/\sim_{(m_1,2,m_3)}$ is of dot-depth exactly 3.

Proof Let $m > 0$. Consider $u_m = ((xy)^m x(xy)^{2m} y(xy)^m)^m$, $v_m = ((xy)^m y(xy)^{2m} x(xy)^m)^m$. Theorem 5.2.1 implies that monoids in $\mathbf{V}_2$ are 2-mutative and hence satisfy $u_m = v_m$ for all sufficiently large m. However, for every $N \geq N_{(1,2,1)}$, $u_N \not\sim_{(1,2,1)} v_N$. A winning strategy for player I in the game $\mathcal{G}_{(1,2,1)}(u_N, v_N)$ appears in the proof of theorem 5.2.2. The result follows.[]

Assume $|u|_a$, $|v|_a > 0$. Let $u = u_\emptyset a u_1 \ldots a u_{|u|_a}$, $v = v_\emptyset a v_1 \ldots a v_{|v|_a}$. If $Q_a^u p_i$, $Q_a^v q_j$ for $i = 1, \ldots, |u|_a$, $j = 1, \ldots, |v|_a$, then $u_i = u(p_i, p_{i+1})$, $i = 1, \ldots, |u|_a - 1$, $v_j = v(q_j, q_{j+1})$, $j = 1, \ldots, |v|_a - 1$. $u_\emptyset = u[1, p_1)$, $v_\emptyset = v[1, q_1)$,

$u_{|u|_a} = u(p_{|u|_a}, |v|]$, $v_{|v|_a} = v(q_{|v|_a}, |v|]$.

The next two lemmas will be used in showing that for positive integers m_1 and m_3, $A^*/\sim_{(m_1,1,m_3)}$ is of dot-depth exactly 2.

Lemma 6.1.2

Assume $u \sim_{(m_1,m_2)} v$. Then

(1) $u[1,p_{(s-1)m_2+i}) \sim_{(m_1-s,m_2)} v[1,q_{(s-1)m_2+i})$,

(2) $u(p_{|u|_a+1-(s-1)m_2-i}, |u|] \sim_{(m_1-s,m_2)} v(q_{|v|_a+1-(s-1)m_2-i}, |v|]$ for $i = 1, \ldots, m_2$ and $s = 1, \ldots, m_1-1$.

Proof (1) Let $1 \le i \le m_2$ and $1 \le s \le m_1-1$. Let $p'_1, \ldots, p'_{m_1-s}$ $(p'_1 \le \ldots \le p'_{m_1-s})$ be positions in $u[1,p_{(s-1)m_2+i})$. Consider the following play of the game $\mathcal{G}_{(m_1,m_2)}(u,v)$. Player I, in the first move, chooses $p_{m_2}, p_{2m_2}, \ldots, p_{(s-1)m_2}, p_{(s-1)m_2+i}, p'_1, \ldots, p'_{m_1-s}$. Hence by the lemma of induction 3.1.1, there exist positions $q'_1, \ldots, q'_{m_1-s}$ $(q'_1 \le \ldots \le q'_{m_1-s})$ in $v[1,q_{(s-1)m_2+i})$ such that player II, by choosing $q_{m_2}, q_{2m_2}, \ldots, q_{(s-1)m_2}, q_{(s-1)m_2+i}, q'_1, \ldots, q'_{m_1-s}$ for the corresponding positions, wins this play of the game. It is clear that

$$u[1,p'_1) \sim_{(m_2)} v[1,q'_1),$$

$$u(p'_j,p'_{j+1}) \sim_{(m_2)} v(q'_j,q'_{j+1}) \text{ for } 1 \le j \le m_1-s-1, \text{ and}$$

$$u(p'_{m_1-s}, p_{(s-1)m_2+i}) \sim_{(m_2)} v(q'_{m_1-s}, q_{(s-1)m_2+i}).$$

Note that player II has to choose q_{m_2}, q_{2m_2}, ..., $q_{(s-1)m_2}$, $q_{(s-1)m_2+i}$ because there is a number of a's $< m_2$ between any two consecutive positions among p_{m_2}, p_{2m_2}, ..., $p_{(s-1)m_2}$, $p_{(s-1)m_2+i}$. The proof is similar, when starting with positions in $v[1, q_{(s-1)m_2+i})$. For (2), we consider $p_{|u|_a+1-m_2}$, $p_{|u|_a+1-2m_2}$, ..., $p_{|u|_a+1-(s-1)m_2}$, $p_{|u|_a+1-(s-1)m_2-i}$, p'_1, ..., p'_{m_1-s}.[]

Lemma 6.1.3

Assume $u \sim_{(m_1,m_2)} v$. Then

(1) $u(p_{(s-1)m_2+i}, |u|] \sim_{(m_1-s,m_2)} v(q_{(s-1)m_2+i}, |v|]$,

(2) $u[1, p_{|u|_a+1-(s-1)m_2-i}) \sim_{(m_1-s,m_2)} v[1, q_{|v|_a+1-(s-1)m_2-i})$

$i = 1, \ldots, m_2$ and $s = 1, \ldots, m_1-1$.

Proof Similar to lemma 6.1.2.[]

In the following theorem, we talk about positions spelling the first and last occurrences of every subword of length $\le m$ of a word w. We illustrate what we mean by this with the following example. Let $A = \{a,b,c\}$ and u = abccccaabbabbacccabababccaaaabbaa....

↑↑↑↑ ↑ ↑ ↑↑↑↑ ↑↑↑↑↑

p

The six arrows on the left point to the positions which spell the first occurrences of every subword of length ≤ 2 in $u[1,p)$ and the eight arrows on the right (before the one pointing to p) to the positions which spell the last occurrences of every subword of length ≤ 2 in $u[1,p)$.

Theorem 6.1.4

Let $A = \{a_1,\ldots,a_r\}$, $r > 1$. Let m_1, m_2 and m_3 be positive integers. Then $A^*/\sim_{(m_1,m_2,m_3)}$ is of dot-depth exactly 2 if and only if $m_2 = 1$.

Proof If $A^*/\sim_{(m_1,m_2,m_3)}$ is of dot-depth exactly 2, then $m_2 < 2$ by lemma 6.1.1. Conversely, we show that for any positive integers m_1 and m_2, $\sim_{(m_1+(m_1+1)2m_2(r+1)^{m_2},m_2)} \subseteq \sim_{(m_1,1,m_2)}$. To see this, suppose $u \sim_{(m_1+(m_1+1)2m_2(r+1)^{m_2},m_2)} v$. Then there is a winning strategy for player II in the game $\mathcal{G}_{(m_1+(m_1+1)2m_2(r+1)^{m_2},m_2)}(u,v)$ to win each play. A winning strategy for player II in the game $\mathcal{G}_{(m_1,1,m_2)}(u,v)$ to win each play is described as follows. Let $p'_1, \ldots, p'_{m_1}$ $(p'_1 \leq \ldots \leq p'_{m_1})$ be positions in u chosen by player I in the first move. Player II chooses positions $q'_1, \ldots, q'_{m_1}$ $(q'_1 \leq \ldots \leq q'_{m_1})$ by considering the following play of the game $\mathcal{G}_{(m_1+(m_1+1)2m_2(r+1)^{m_2},m_2)}(u,v)$. In the first move, player I chooses $p'_1, \ldots, p'_{m_1}$ and the positions which spell the first and last occurrences of every subword of length $\leq m_2$ in $u[1,p'_1)$,

$u(p'_1,p'_2)$, ..., $u(p'_{m_1-1},p'_{m_1})$ and $u(p'_{m_1},|u|]$ for a total of no more than $m_1+(m_1+1)2m_2(r+1)^{m_2}$ positions (there are r^{m_2} possible words of length m_2 for a total of no more than $m_2(r+1)^{m_2}$ positions to spell the first (last) occurrences of every subword of length $\leq m_2$). More details follow for the special case $u \sim_{(1+4m_2(r+1)^{m_2},m_2)} v$. We have a winning strategy for player II in the game $\mathcal{G}_{(1+4m_2(r+1)^{m_2},m_2)}(u,v)$ to win each play. Let us describe a winning strategy for player II in the game $\mathcal{G}_{(1,1,m_2)}(u,v)$ to win each play. Let p be a position in u chosen by player I in the first move. Suppose $Q^u_a p$ for some $a \in A$. If p is the i^{th} occurrence of a in u $(1 \leq i \leq N(1,m_2) = 2m_2+1)$, then player II chooses the same occurrence of a in v, say position q. The fact that $u[1,p] \sim_{(1,m_2)} v[1,q]$ and $u(p,|u|] \sim_{(1,m_2)} v(q,|v|]$ follows from lemmas 6.1.2 and 6.1.3 $(N(1,m_2) \leq (4m_2(r+1)^{m_2})m_2)$. If p is the $|u|_a+1-i^{th}$ occurrence of a in u $(1 \leq i \leq N(1,m_2))$, player II chooses the $|v|_a+1-i^{th}$ occurrence of a in v. If p is among p_{2m_2+2}, ..., $p_{|u|_a-2m_2-1}$, then player II chooses position q, an a, among q_{2m_2+2}, ..., $q_{|v|_a-2m_2-1}$ by considering the following play of the game $\mathcal{G}_{(1+4m_2(r+1)^{m_2},m_2)}(u,v)$. In the first move, player I chooses p, the positions which spell the first and last occurrences of every subword of length $\leq m_2$ in $u[1,p)$ and in $u(p,|u|]$. Hence there exists a position q in v such that player II, by choosing q, the positions which spell the first and last occurrences of every subword of length $\leq m_2$ in $v[1,q)$ and in $v(q,|v|]$, wins the play of the game. Let us show that $u[1,p) \sim_{(1,m_2)} v[1,q)$ (the proof that $u(p,|u|] \sim_{(1,m_2)} v(q,|v|]$ is similar). Let p' be a position in $u[1,p)$ (the proof is similar when starting with a position in $v[1,q)$). Assume $Q^u_{a_i} p'$.

Case 1: p' is among the positions which spell the first occurrences of every subword of length $\leq m_2$ in $u[1,p)$. Let q' be the corresponding position among the ones chosen by II in $v[1,q)$. It is clear that $u(p',p) \sim_{(m_2)} v(q',q)$ and $u[1,p') \sim_{(m_2)} v[1,q')$.

Case 2: p' is among the positions which spell the last occurrences of every subword of length $\leq m_2$ in $u[1,p)$. Similar to case 1.

Case 3: Otherwise, let p'' and p''' $(p'' < p''')$ be the closest positions to p' in $u[1,p')$ and $u(p',p)$ respectively among the chosen positions by player I. Let q'' and q''' $(q'' < q''')$ be the corresponding positions chosen by player II. Since $u(p'',p''') \sim_{(m_2)} v(q'',q''')$, there is q' in $v(q'',q''')$ such that $Q^v_{a_i} q'$. Let us show that $u(p',p) \sim_{(m_2)} v(q',q)$. $u[1,p') \sim_{(m_2)} v[1,q')$ follows similarly. Let $w = w_1 \ldots w_{|w|}$, $|w| \leq m_2$ in $v(q',q)$. The proof is similar when starting with w in $u(p',p)$. If w is a subword of $v[q''',q)$, it is clear that w is a subword of $u[p''',p)$, hence in $u(p',p)$. So let us assume w is not a subword of $v[q''',q)$. Let $p_{w_1}, \ldots, p_{w_{|w|}}$ in $v(q',q)$, at least p_{w_1} being in $v(q',q''')$, be positions which spell $w_1 \ldots w_{|w|}$. $p_{w_1}, \ldots, p_{w_{|w|}}$

are hence positions which spell an occurrence of a subword of length $\leq m_2$ in $v[1,q)$. Hence they are smaller than or equal to those positions which spell the last occurrence of w in $v[1,q)$ which are in $v[q''',q)$. Hence w is a subword of $u(p',p)$.[]

The following corollary gives another result for inclusion (one was proposition 3.2.3).

Corollary 6.1.5

Let $|A| = r$. Then $\sim_{(m_1+(m_1+1)2m_2(r+1)^{m_2},m_2)} \subseteq \sim_{(m_1,N(1,m_2))}$.

Proof From theorem 6.1.4 and proposition 3.2.3.[]

2. *An equational characterization of the first sublevel of the second level of the Straubing hierarchy*

In this section, we show that the equations in proposition 5.1.9 are part of a system ultimately defining $V_{2,1}$ for an alphabet of two letters.

Lemma 6.2.1

Let $m \geq 1$. Let $u, v \in A^+$ and let $p_1, \ldots, p_s$ in u $(p_1 < \ldots < p_s)$ $(q_1, \ldots, q_{s'}$ in v $(q_1 < \ldots < q_{s'}))$ be the positions which spell the first and last occurrences of every subword of length $\leq m$ in u (v). $u \sim_{(1,m)} v$ if and only if

(1) $s = s'$,

(2) $Q^u_a p_i$ if and only if $Q^v_a q_i$ for $i = 1, \ldots, s$ and some $a \in A$ and

(3) $u(p_i, p_{i+1}) \sim_1 v(q_i, q_{i+1})$ for $i = 1, \ldots, s-1$.

Proof Assume (1), (2) and (3) hold. A winning strategy for player II in the game $\mathcal{G}_{(1,m)}(u,v)$ to win each play is described as follows. Let p be a position in u chosen by player I in the first move (the proof is similar when starting with a position in v). Assume $Q^u_a p$.
Case 1: p is among $p_1, \ldots, p_s$, i.e., $p = p_i$ for some i, $1 \leq i \leq s$. Since (1) holds, we can consider $q = q_i$. (2) implies that $Q^v_a q$.

Case 2: $p \in u(p_i, p_{i+1})$ for some i, $1 \le i \le s-1$. From (3), there is $q \in v(q_i, q_{i+1})$ such that $Q_a^v q$. In either case, (1), (2), (3) and the choice of q imply that $u(p, |u|] \sim_{(m)} v(q, |v|]$ and $u[1,p) \sim_{(m)} v[1,q)$.

Conversely, assume $u \sim_{(1,m)} v$. (1) and (2) obviously hold. Also, $u(p_i, p_{i+1}) \sim_{(1)} v(q_i, q_{i+1})$ for $i = 1, \ldots, s-1$. To see this, let p be in $u(p_i, p_{i+1})$ (the proof is similar when starting with q in $v(q_i, q_{i+1})$). Consider the following play of the game $\mathcal{G}_{(1,m)}(u,v)$. Player I, in the first move, chooses p. Hence there exists q in v such that $u(p, |u|] \sim_{(m)} v(q, |v|]$ and $u[1,p) \sim_{(m)} v[1,q)$. Assume that $q \notin v(q_i, q_{i+1})$. Hence $q \in v[1, q_i]$ or $q \in v[q_{i+1}, |v|]$. From the choice of the p_i's and the q_i's, either $u(p, |u|] \not\sim_{(m)} v(q, |v|]$ or $u[1,p) \not\sim_{(m)} v[1,q)$. Contradiction. The result follows.[]

Lemma 6.2.2

Let $m \ge 1$. Let $u, v \in A^*$. If $u \sim_{(1,m)} v$, then there exists $w \in A^*$ such that u is a subword of w, v is a subword of w and $u \sim_{(1,m)} w \sim_{(1,m)} v$.

Proof Let $A = \{a_1, \ldots, a_r\}$. If $r = 1$, $u = v$ or $|u|$, $|v| \ge N(1,m)$ by chapter three. Choose w such that $|w| = \max\{|u|, |v|\}$. For $r > 1$, let $p_1, \ldots, p_s$ $(p_1 < \ldots < p_s)$ be the positions which spell the first and last occurrences of every subword of length $\le m$ in u. s is no more than $2m(r+1)^m$. Assume $Q^u_{a_{j_i}} p_i$. Since $u \sim_{(1,m)} v$, by lemma 6.2.1, the positions $q_1, \ldots, q_s$

$(q_1 < \ldots < q_s)$ in v which spell the first and last occurrences of every subword of length $\le m$ in v are such that $Q^v_{a_{j_i}} q_i$ and $u(p_i,p_{i+1}) \sim_{(1)} v(q_i,q_{i+1})$ for $i = 1, \ldots, s-1$. Hence by lemma 5.1.2, since $u(p_i,p_{i+1}) \sim_{(1)} v(q_i,q_{i+1})$, there exists w_i such that $u(p_i,p_{i+1})$ is a subword of w_i, $v(q_i,q_{i+1})$ is a subword of w_i and $u(p_i,p_{i+1}) \sim_{(1)} w_i \sim_{(1)} v(q_i,q_{i+1})$. Let $w = a_{j_1} w_1 a_{j_2} w_2 \ldots a_{j_{s-1}} w_{s-1} a_{j_s}$. u is a subword of w, v is a subword of w and $u \sim_{(1,m)} w \sim_{(1,m)} v$ by lemma 6.2.1.[]

Now, let us define classes of equations as follows. For $m \ge 1$, $C^1_{(1,m)}$ consists of the equations

$$u_1 \ldots u_m xy v_1 \ldots v_m = u_1 \ldots u_m yx v_1 \ldots v_m$$

where the u's and the v's are of the form $x^e y$, $y^e x$, xy^e or yx^e for some e, $1 \le e \le N(1,m)$. The equation $(xy)^m xy(xy)^m = (xy)^m yx(xy)^m$ is such an example.

$C^2_{(1,m)}$ consists of the equations

$$u_1 \ldots u_i x^{m-i} x x^{m-j} v_1 \ldots v_j = u_1 \ldots u_i x^{m-i} x^2 x^{m-j} v_1 \ldots v_j$$

where the u's and the v's are as above and $0 \le i, \; j \le m$. The equation $(xy)^m x(xy)^m = (xy)^m x^2 (xy)^m$ is an example.

Lemma 6.2.3

The monoids in $V_{2,1}$ satisfy $C^1_{(1,m)} \cup C^2_{(1,m)}$ for all sufficiently large m.

Proof It is easily seen, using lemma 6.2.1, that monoids in $V_{2,1}$ satisfy $C^1_{(1,m)} \cup C^2_{(1,m)}$ for some $m \geq 1$. This comes from the fact that if $M \in V_{2,1}$, then $M < A^*/\sim_{(1,m)}$ for some $m \geq 1$. Since $A^*/\sim_{(1,m)}$ satisfies $C^1_{(1,m)} \cup C^2_{(1,m)}$, M satisfies $C^1_{(1,m)} \cup C^2_{(1,m)}$. Moreover, if M in $V_{2,1}$ satisfies $C^1_{(1,m)} \cup C^2_{(1,m)}$ for some $m \geq 1$, then it satisfies $C^1_{(1,n)} \cup C^2_{(1,n)}$ for all $n \geq m$ since $\sim_{(1,n)} \subseteq \sim_{(1,m)}$ for those n.[]

Theorem 6.2.4

Let M be a monoid generated by two elements. Then M belongs to $V_{2,1}$ if and only if it ultimately satisfies the equations $\bigcup_{m\geq 1} C^1_{(1,m)} \cup C^2_{(1,m)}$.

Proof We have to prove that $M \in V_{2,1}$ if and only if it satisfies the equations in $C^1_{(1,m)} \cup C^2_{(1,m)}$ for all m sufficiently large. By lemma 6.2.3, monoids in $V_{2,1}$ satisfy $C^1_{(1,m)} \cup C^2_{(1,m)}$ for all sufficiently large m.

Conversely, let $\varphi : A^* \to M$ be a surjective morphism satisfying $u\varphi = v\varphi$ for every equation $u = v$ in $\bigcup_{n\geq m} C^1_{(1,n)} \cup C^2_{(1,n)}$ for some $m \geq 1$. Let us show that $M \in V_{2,1}$. It is sufficient to prove that for all x and y in A^*, $x \sim_{(1,m)} y$ implies $x\varphi = y\varphi$. For $x = y = 1$, it is certainly true. Assume x, $y \neq 1$. Let $p_1, \ldots, p_s$ $(p_1 < \ldots < p_s)$ $(q_1, \ldots, q_s$ $(q_1 < \ldots < q_s))$ be the positions which spell the first and last occurrences of every subword of length $\le m$ in x (y). By lemma 6.2.1, they satisfy the following

$$Q^x_{a_{j_i}} p_i \text{ if and only if } Q^y_{a_{j_i}} q_i, \quad 1 \le i \le s, \text{ and}$$

$$u(p_i,p_{i+1}) \sim_{(1)} v(q_i,q_{i+1}) \text{ for } i = 1, \ldots, s-1.$$

Lemma 6.2.2 implies the existence of $z = a_{j_1} z_1 a_{j_2} z_2 \ldots a_{j_{s-1}} z_{s-1} a_{j_s}$ satisfying

$$x \sim_{(1,m)} z \sim_{(1,m)} y,$$

$$x(p_i,p_{i+1}) \sim_{(1)} z_i \sim_{(1)} y(q_i,q_{i+1}) \text{ and}$$

$$x(p_i,p_{i+1}) \text{ and } y(q_i,q_{i+1}) \text{ are subwords of } z_i$$

$$\text{for } i = 1, \ldots, s-1.$$

Hence, lemma 6.2.2 allows us to consider only the case where $x = a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_{s-1}} x_{s-1} a_{j_s}$ and $y = a_{j_1} y_1 a_{j_2} y_2 \ldots a_{j_{s-1}} y_{s-1} a_{j_s}$ where x_i is a subword of y_i and $x_i \sim_{(1)} y_i$ for $i = 1, \ldots, s-1$. We observe also that if x_i is a subword of w_i and w_i a subword of y_i, we have also $x_i \sim_{(1)} w_i$. Hence we have only to consider the case where

$x = a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_i} uv a_{j_{i+1}} \ldots a_{j_{s-1}} x_{s-1} a_{j_s}$,

$y = a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_i} uav a_{j_{i+1}} \ldots a_{j_{s-1}} x_{s-1} a_{j_s}$ for some i between 1 and $s-1$, some a in u or in v. We have the following cases.

Case 1: If p_i is the last position among the ones which spell a first occurrence of a subword of length $\le m$ in x and p_{i+1} the first position among the ones which spell a last occurrence of a subword of length $\le m$ in x, then using a particular case of $C^2_{(1,m)}$, i.e., $x^{N(1,m)} = x^{N(1,m)+1}$ enables us to assume that x and

y do not contain more than $N(1,m)$ consecutive occurrences of a letter. Hence we are able to write $x\varphi$ and $y\varphi$ as $x\varphi = u_1 \ldots u_m uvv_1 \ldots v_m \varphi$, $y\varphi = u_1 \ldots u_m uavv_1 \ldots v_m \varphi$ where the u's and the v's satisfy the properties stated in $C^1_{(1,m)}$. Then using $C^1_{(1,m)}$ and $C^2_{(1,m)}$ enables us to write $y\varphi$ as $u_1 \ldots u_m uvv_1 \ldots v_m \varphi = x\varphi$ since a is in u or in v.

Case 2: Otherwise, uv contains only a's. Assume $uv = u'_\emptyset a u_\emptyset v$ for some $u'_\emptyset$, $u_\emptyset \in A^*$. The case where a is in v is similar. Using $x^{N(1,m)} = x^{N(1,m)+1}$ enables us as in Case 1 to assume that x and y do not contain more than $N(1,m)$ consecutive occurrences of a letter. From the choice of the a_{j_i}'s,

$a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_i} u'_\emptyset \sim_{(m)} a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_i} u'_\emptyset a$ and

$u_\emptyset v a_{j_{i+1}} \ldots a_{j_{s-1}} x_{s-1} a_{j_s} \sim_{(m)} a u_\emptyset v a_{j_{i+1}} \ldots a_{j_{s-1}} x_{s-1} a_{j_s}$. Lemma 5.1.3 hence implies the existence of $u_1, \ldots, u_m \in A^*$, $v_1, \ldots, v_m \in A^*$ such that $a_{j_1} x_1 a_{j_2} x_2 \ldots a_{j_i} u'_\emptyset = u_1 \ldots u_m$,

$u_\emptyset v a_{j_{i+1}} \ldots a_{j_{s-1}} x_{s-1} a_{j_s} = v_1 \ldots v_m$, $\{a\} \subseteq u_m\alpha \subseteq \ldots \subseteq u_1\alpha$ and $\{a\} \subseteq v_1\alpha \subseteq \ldots \subseteq v_m\alpha$. Moreover, it is easy to see that one may assume that there exist k and l between $\emptyset$ and m such that $u_{k+1} = \ldots = u_m = a = v_1 = \ldots = v_{m-l}$, and such that the u's and the v's are of the form $x^e y$, $y^e x$, xy^e or yx^e, for some e, $1 \leq e \leq N(1,m)$. $C^2_{(1,m)}$ gives $x\varphi = u_1 \ldots u_k a^{N(1,m)-k-l} v_{m-l+1} \ldots v_m \varphi = u_1 \ldots u_k a^{N(1,m)+1-k-l} v_{m-l+1} \ldots v_m \varphi = y\varphi$. The result follows.[]

Chapter 7

CONCLUSION

Analogously to $*$-varieties of languages and M-varieties of monoids, $+$-varieties of languages and S-varieties of semigroups are defined by replacing $*$ by $+$ and monoid by semigroup. The correspondence between $+$-varieties and S-varieties holds. $\mathcal{B}_k$ are examples of $+$-varieties of languages and let us denote by $\mathbf{B}_k$ the corresponding S-varieties. A result of Straubing [St85] states that if $\mathbf{V}_k$ is decidable, then $\mathbf{B}_k$ is decidable.

Simon's characterization of the recognizable languages whose syntactic monoids are $\mathcal{J}$-trivial, i.e., $M(L) \in \mathbf{J}$ if and only if L is a $\sim_{(m)}$-language for some m, or $\mathbf{V}_1 = \mathbf{J}$, gives an algorithm to decide if a recognizable language is of dot-depth 1. If $\mathbf{W}$ is decidable, is $\langle\rangle\mathbf{W}$ decidable? The solution of this open problem could provide an algorithm to test if a language is of dot-depth k since $\mathbf{V}_{k+1} = \langle\rangle\mathbf{V}_k$. Simon's result is the basis for much recent research, for example, the effective characterization of $A^+\mathcal{B}_1$, the level 1 of the dot-depth hierarchy. Knast [Kn83a], [Kn83b] demonstrated the decidability of $\mathbf{B}_1$. A simpler proof was obtained by Therien [The85] using categories. A number of other consequences of

Simon's theorem are: a general theory of congruence varieties [The81], the study of languages whose syntactic monoids are p-groups or nilpotent groups [Ei76], [The84], some purely combinatorial investigations [Lo83]. Some other consequences are given in [St80].

Most of the proofs of Simon's theorem that have been published so far, [Ei76], [La179], [Pi84a], [Si75] for example, depend on a detailed study of combinatorial properties of the congruences $\sim_{(m)}$. In [ST85], semigroup expansions were used to show the result that every finite $\mathcal{J}$-trivial monoid is a quotient of a finite monoid admitting a partial order that is compatible with multiplication. As a consequence, a radically new proof of Simon's theorem was obtained.

Our future research is concerned with more applications of the logical characterizations stated in chapter two. For example, we would like to settle some open problems such as a generalization of Simon's theorem. The following open questions concerning the decidability of the Straubing hierarchy remain other goals for later investigation:

(1) Find a necessary and sufficient condition for $\mathcal{L}_{(m_1,\ldots,m_k)}$ to be included in $\mathcal{L}_{(m'_1,\ldots,m'_{k'})}$. Chapters three, four, five and six include partial results. A necessary condition is $N(m_1,\ldots,m_k) \leq N(m'_1,\ldots,m'_{k'})$.

(2) Do the equation systems in chapter five, section one, completely characterize the M-varieties $\mathbf{V}_{1,m}$ for $m \geq 4$? There, it was shown

that they do for $V_{1,1}$, $V_{1,2}$ and $V_{1,3}$ for any alphabet A.

(3) Let $k \geq 1$. Let $m_1, \ldots, m_k$ be positive integers. Let A contain at least two letters. Find a necessary and sufficient condition for $A^*/\sim_{(m_1,\ldots,m_k)}$ to be of dot-depth exactly d. It is easy to see that $A^*/\sim_{(m_1,\ldots,m_k)}$ is of dot-depth exactly 1 if and only if $k = 1$. Using theorems 5.2.2 and 6.1.4, $A^*/\sim_{(m_1,\ldots,m_k)}$ is of dot-depth exactly 2 if and only if $k = 2$ or ($k = 3$ and $m_2 = 1$). Also, similarly to lemma 6.1.1, for $k \geq 3$, m_i positive integers, and $m_i \geq 2$ for $i = 2, \ldots, k-1$, we have that $A^*/\sim_{(m_1,\ldots,m_k)}$ is of dot-depth exactly k.

(4) Generalize the equation systems of chapter six, section two, to equation systems that characterize $V_{2,1}$ for any alphabet. A generalization of these systems for $V_{2,m}$, $m > 1$, would provide an equational characterization of dot-depth 2 monoids.

[Bl88a] F.Blanchet-Sadri, *Games, equations and the dot-depth hierarchy*, preprint 1988. To appear in Computers & Mathematics with applications.

[Bl88b] F.Blanchet-Sadri, *On dot-depth two*, preprint 1988. To appear in RAIRO Informatique Theorique et applications.

[Br76] J.A.Brzozowski, *Hierarchies of aperiodic languages*, RAIRO Informatique Theorique 10, 1976, p.33-49.

[BK78] J.A.Brzozowski and R.Knast, *The dot-depth hierarchy of star-free languages is infinite*, Journal of Computer and System Sciences 16, 1978, p.37-55.

[BS73] J.A.Brzozowski and I.Simon, *Characterizations of locally testable events*, Discrete Mathematics 4, 1973, p.243-271.

[Bü60] J.R.Büchi, *Weak second order arithmetic and finite automata*, Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 6, 1960, p.66-92.

[CB71] R.S.Cohen and J.A.Brzozowski, *Dot-depth of star-free events*, Journal of Computer and System Sciences 5, 1971, p.1-16.

[Eh61] A.Ehrenfeucht, *An application of games to the completeness problem for formalized theories*, Fundamenta Mathematicae 49, 1961, p.129-141.

[Ei76] S.Eilenberg, *Automata, languages and machines*, vol.B, Academic Press, New York, 1976.

[El61] C.C.Elgot, *Decision problems of finite automata design and related arithmetics*, Transactions of the American Mathematical Society 98, 1961, p.21-52.

[En72] H.B.Enderton, *A mathematical introduction to logic*, Academic Press, New York, 1972.

[Fr72] R.Fraisse, *Cours de logique mathematique*, Tome 2, Gauthier Villars, Paris, 1972.

[Kl56] S.C.Kleene, *Representation of events in nerve nets and finite automata*, Automata Studies, Princeton University Press, Princeton, New Jersey, 1956, p.3-42.

[Kn83a] R.Knast, *A semigroup characterization of dot-depth one languages*, RAIRO Informatique Theorique 17, 1983, p.321-330.

[Kn83b] R.Knast, *Some theorems on graph congruences*, RAIRO Informatique Theorique 17, 1983, p.331-342.

[Lad77] R.E.Ladner, *Application of model theoretic games to discrete linear orders and finite automata*, Information and Control 33, 1977, p.281-303.

[Lal79] G.Lallement, *Semigroups and combinatorial applications*, Wiley, New York, 1979.

[Lo83] M.Lothaire, *Combinatorics on words*, Addison Wesley, Boston, 1983.

[Mc74] R.McNaughton, *Algebraic decision procedures for local testability*, Mathematical Systems Theory 8, 1974, p.60-76.

[MP71] R.McNaughton and S.Papert, *Counter-free automata*, MIT Press, Cambridge, Massachussets, 1971.

[PP86] D.Perrin and J.E.Pin, *First order logic and star-free sets*, Journal of Computer and System Sciences 32, 1986, p.393-406.

[Pi83] J.E.Pin, *Concatenation hierarchies: decidability results and problems*, in: L.J.Cummings ed., Combinatorics on words, Progress and Perspectives, Academic Press, New York, 1983, p.195-228.

[Pi84a] J.E.Pin, *Varietes de langages formels*, Masson, Paris, 1984. (*Varieties of formal languages*, Plenum, London, 1986).

[Pi84b] J.E.Pin, *Hierarchies de contatenation*, RAIRO Informatique Theorique 18, 1984, p.23-46.

[PS81] J.E.Pin and H.Straubing, Monoids of upper triangular boolean matrices, Colloquia Mathematica Societatis Janos Bolyai 39, 1981, p.259-272.

[Re79] C.Reutenauer, *Sur les varietes de langages et de monoides*, Lecture Notes in Computer Science 67, Springer Verlag, Berlin, 1979, p.260-265.

[Ro82] J.G.Rosenstein, *Linear orderings*, Academic Press, New York, 1982.

[Sc65] M.P.Schützenberger, *On finite monoids having only trivial subgroups*, Information and Control 8, 1965, p.190-194.

[Si72] I.Simon, *Hierarchies of events of dot-depth one*, Ph.D. Thesis, University of Waterloo, 1972.

[Si75] I.Simon, *Piecewise testable events*, Proceedings of the 2nd GI Conference, Lecture Notes in Computer Science 33, Springer Verlag, Berlin, 1975, p.214-222.

[Ste85] J.Stern, *Characterizations of some classes of regular events*, Theoretical Computer Science 35, 1985, p.17-42.

[St80] H.Straubing, *On finite J-trivial monoids*, Semigroup Forum 19, 1980, p.107-110.

[St81] H.Straubing, *A generalization of the Schützenberger product of finite monoids*, Theoretical Computer Science 13, 1981, p.137-150.

[St85] H.Straubing, *Finite semigroup varieties of the form V * D*, Journal of Pure and Applied Algebra 36, 1985, p.53-94.

[St86] H.Straubing, *Semigroups and languages of dot-depth two*, Proceedings of the 13th ICALP, Lecture Notes in Computer Science, Springer Verlag, New York, 1986, p.416-423.

[ST85] H.Straubing and D.Therien, *Partially ordered finite monoids and a theorem of I. Simon*, Journal of Pure and Applied Algebra, 1985.

[The80] D.Therien, *Classification of regular languages by congruences*, Ph.D. Thesis, University of Waterloo, 1980.

[The81] D.Therien, *Classification of finite monoids: the language approach*, Theoretical Computer Science 14, 1981, p.195-208.

[The84] D.Therien, *Subword counting and nilpotent groups*, in: L.J.Cummings ed., Combinatorics on words, Progress and Perspectives, Academic Press, New York, 1983, p.297-305.

[The85] D.Therien, *Categories et langages de dot-depth un*, preprint 1985.

[Tho82] W.Thomas, *Classifying regular events in symbolic logic*, Journal of Computer and System Sciences 25, 1982, p.360-376.

[Tho84] W.Thomas, *An application of the Ehrenfeucht-Fraisse game in formal language theory*, Bulletin de la Societe mathematique de France, Memoire (nouvelle serie) 16, 1984, p.11-21.

[Ti87] B.Tilson, *Categories as algebra*, Journal of Pure and Applied Algebra 48, 1987, p.83-198.