The equational theory of regular words

Stephen L. Bloom [a,*] , Zoltán Ésik [b,1]

[a] *Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030*

[b] *Institute for Informatics, University of Szeged, Szeged, Hungary*

Abstract

Courcelle introduced the study of regular words, i.e., words isomorphic to frontiers of regular trees. Heilbrunner showed that a nonempty word is regular iff it can be generated from the singletons by the operations of concatenation, omega power, omega-op power, and the infinite family of shuffle operations. We prove that the algebra of nonempty regular words on the set A, equipped with these operations, is freely generated by A in a variety which is axiomatizable by an infinite collection of some natural equations. We also show that this variety has no finite equational basis and that its equational theory is decidable in polynomial time.

Key words: Word, arrangement, regular, linear order, equational theory.

---

1 Introduction

By "word" we understand a labeled linear order, extending the familiar notion of a labeling of $\{1, 2, \ldots, n\}$, for some $n \geq 0$. Courcelle [Cour78] introduced the study of such words ("arrangements", in his terminology). He showed that every finite or countable word is isomorphic to the frontier of a complete binary tree, where the linear order on the leaves of the tree is the lexicographic (left to right) order. He introduced several operations on such words, including concatenation (or product), omega and omega-op powers. He proved that initial solutions of finite systems of fixed point equations

$$x_i = u_i, \quad i = 1, 2, \ldots, k, \tag{1}$$

---

[*] Corresponding author.
Email address: `bloom@cs.stevens.edu` (Stephen L. Bloom).

[1] Partially supported by the National Foundation of Hungary for Scientific Research, grant T46686. Currently visiting Rovira i Virgili University, Tarragona, Spain.

(where $x_1, \dots, x_k$ are variables and the u_i are finite words on the letters in a set A and the variables) are isomorphic to frontiers of *regular* trees. Further, he showed that the solutions of certain kinds of systems can be expressed by "quasi-rational" expressions, formed from single letters in A using the operations of concatenation, omega and omega-op power. Courcelle asked for a complete set of axioms for these operations. In [BlCho01], a complete set of axioms for just the concatenation and omega power operation on words was given, and in [BlEs03a] Courcelle's question was answered.

We call a word which is isomorphic to the frontier of a regular binary tree a **regular word**. Several results on regular words have been obtained by Heilbrunner [Heil80], Thomas [Th86], and the authors [BlEs03]. Heilbrunner showed that all nonempty regular words on the set A can be generated from single letters by means of the above mentioned operations, namely concatenation, omega and omega-op power, together with (infinitely many) "shuffle" operations. Terms formed from letters in A and these operations are called "terms on A". Heilbrunner gave an algorithm which, given a finite system of fixed point equations of the form (1) such that the first component (i.e., x_1) of the initial solution is nonempty, produces a term denoting it. Thomas gave an algorithm to determine when two terms denote isomorphic words. His algorithm is based on Rabin's theorem on automata for infinite trees.

Heilbrunner discussed several identities involving the terms with both Courcelle's operations, as well as the shuffle operations, but did not obtain a completeness result. Our paper gives a set Ax of axioms and

- in Theorem 76, shows them to be complete. This result implies that
- for any alphabet A, the algebra of regular words on an alphabet A is freely generated by A in the variety defined by these equations.
- We show also that the equational theory of this variety is decidable in polynomial time, (see Theorem 79), and is not finitely based, (see Theorem 82).

The completeness theorem, and the corresponding complexity result, provide a solution to a problem that has been open for over twenty years.

We describe our method, which may be of independent interest.

We find an appropriate "condensation" [Ro82] of the linear order of a regular word u, and replace certain subwords by appropriately labeled points. Given a word u on the alphabet A, we show that its underlying linear order L_u is partitioned into **blocks** of an equivalence relation: two points $p < q$ are in the same block iff they both belong to some "uniform" subword (defined below) or neither does and the interval $\{x \in L_u : p \leq x \leq q\}$ is finite. The blocks of L_u are also linearly ordered in the obvious way, and we denote this linearly ordered set by $\widehat{L}_u$. The blocks of regular words are denoted by what we call the "primitive terms" below. If two primitive terms s, t denote isomorphic words, then our axioms Ax are strong enough so that $\mathsf{Ax} \vdash s = t$.

Now, we let $\widehat{u}$ denote the word on $\widehat{L}_u$, obtained from u by replacing each block v

of u by a point labeled by a new letter determined by v. In order to show that $\widehat{u}$ is regular when u is, we make use of "proper terms".

From [Cour78] and [Heil80], we know that if u is regular, there is some term denoting u. In Section 9, we show how to obtain a substitution σ and, for each term t, a proper term s such that if t denotes the word u, then $\widehat{u}$ is denoted by s, and we show $\mathsf{Ax} \vdash \sigma(s) = t$.

Unfortunately, it takes some time to show how to obtain the proper terms, and we don't get to prove the completeness theorem until Section 10.

In Section 11, the properties of the construction of the proper terms leads to a polynomial time decision algorithm to decide whether two terms denote isomorphic words.

The algebra of all (regular) countable words on some alphabet, equipped with the associative operation of concatenation, allows one to solve systems of fixed point equations (1). We end the paper by describing an open problem that may be roughly phrased as follows: what is the most general theory of an associative operation admitting fixed points?

In order to find decision procedures, several people have studied automata on infinite trees and on words of various kinds, for example: Choueka [Chou78], who studied automata on ω^n tapes, and Bruyère and Carton [BruyCar,BruyCar2] consider automata on all countable scattered words. In [BlEs03], the authors use finite automata to decide whether the frontier of regular tree is scattered. See also [GTW02] for a thorough study.

2 Preliminaries

A linearly ordered set $(L, \leq)$ is a set equipped with a reflexive, transitive, antisymmetric and total order. Sometimes, we say just "L is a linearly ordered set", to mean $(L, \leq)$ is a linearly ordered set.

A **morphism** $h : (L, \leq) \to (L', \leq')$ of linearly ordered sets is a function $h : L \to L'$ such that if $x \leq y$ in L, then $h(x) \leq' h(y)$ in L'. The following common linearly ordered sets are used frequently: $\mathbb{Z}$ denotes the usual ordering on all integers; $\mathbb{Q}$, the standard ordering on all rationals; ω, the linearly ordered set of the nonnegative integers, and ω^{op}, the linearly ordered set of the negative integers. For $n \in \omega$, $[n]$ denotes the n element set $\{1, 2, \ldots, n\}$, ordered as usual.

Suppose that p, q are points in a linearly ordered set. We will say that q is the *successor* of p, or that p is the *predecessor* of q, if $p < q$ and there is no element x with $p < x < q$. (When the successor or predecessor of a point exists, it is unique.)

A linear order $(L, \leq)$ is **quasi-dense** if there is an injective morphism

$$(\mathbb{Q}, \leq) \longrightarrow (L, \leq).$$

If $(L, \leq)$ is not quasi-dense, it is **scattered** [Ro82] or **discrete** [BlEs03].

An **interval of** L is a subset I of L such that if $x < y < z$ in L, and if $x, z \in I$, then $y \in I$. (Thus, in particular, the empty set is an interval, as is any singleton subset.)

An interval with no greatest element is said to be **right-open**; an interval with no least element is **left-open**; an interval with neither a least or greatest element is **open**.

An interval I is **dense** if I has at least two distinct points and for all $x, y \in I$, if $x < y$, then there is some $z \in I$ with $x < z < y$. (Thus, every dense interval is infinite.)

We note some elementary properties of intervals.

Proposition 1 *The intersection of two open intervals is an open interval; the intersection of two dense open intervals is either empty or a dense open interval. If* I, J *are open intervals whose intersection is nonempty, then* $I \cup J$ *is an open interval, and if both are dense, so is* $I \cup J$. *If* $(I_j)_{j \in J}$ *is a chain of open (dense, resp.) intervals, then* $\bigcup_{j \in J} I_j$ *is an open (dense, resp.) interval.*

3 Words

A **word** $(L_u, \leq, u, A)$ (a "generalized word" in Thomas [Th86], an "arrangement" in Courcelle [Cour78]) consists of a linearly ordered set $(L_u, \leq)$, a nonempty set A of *labels*, and a *labeling function* $u : L_u \to A$.

Usually, we write $\leq$ for the order relation in any linearly ordered set. Also, we use just the labeling function to name the word, and let $(L_u, \leq)$ denote the underlying linear order of the word u. We say that u is a word *on* A, (or u is an "A-labeled word") *over* the linear order $(L_u, \leq)$. The linear order $(L_u, \leq)$ is the *underlying linear order* of u. The range of the labeling function is called the *alphabet* of the word, denoted **alph**(u). When L_u is empty, we have the **empty word** on A, written $\mathbf{1}$.

Suppose that u and v are words over A with underlying linear orders $(L_u, \leq)$ and $(L_v, \leq)$, respectively. A **morphism** $h : u \to v$ is a morphism $h : (L_u, \leq) \to (L_v, \leq)$ which preserves the labeling:

$$u(x) = v(h(x)), \quad x \in L_u.$$

Thus, for any set A, the collection of words on A forms a category. Two words u, v on A are **isomorphic** when they are isomorphic in this category, i.e., when there are morphisms $h : u \to v$, $g : v \to u$ such that $u \xrightarrow{h} v \xrightarrow{g} u$ and $v \xrightarrow{g} u \xrightarrow{h} v$ are the respective identities. We write

$$u \cong v$$

to indicate that u and v are isomorphic. *We usually identify isomorphic words.*

If u, v are words on A, we say v is a **subword** of u if L_v is an interval of $(L_u, \leq)$ and for $p \in L_v$, $v(p) = u(p)$. (Some people say v is a factor of u, rather than subword.) And since any interval of L_u determines a unique subword of u, we sometimes identify an interval with a subword.

4 The regular operations on words

We will define the collection of "regular operations" on words by means of **word substitution**. First, we define sum and generalized sum [Ro82] for linear orders.

Definition 2 (sum) *Suppose that L_1 and L_2 are linear orders with disjoint underlying sets. Then $L_1 + L_2$ is the linear order on the union of the two sets defined by:*

$$x \leq y \iff (x \leq y \textit{ in } L_1 \textit{ or in } L_2) \textit{ or } x \in L_1 \textit{ and } y \in L_2.$$

Definition 3 (generalized sum) *Suppose that $(L, \leq)$ is a linear order, and for each $x \in L$, let $(K_x, \leq)$ be a linear order. The ordering*

$$\sum_{x \in L} K_x$$

obtained by substitution of K_x for $x \in L$, is defined as follows: the underlying set of $\sum_{x \in L} K_x$ is the set of pairs (k, x) with $x \in L$ and $k \in K_x$ ordered by:

$$(k, x) \leq (k', x') \iff x < x' \textit{ or } (x = x' \textit{ and } k \leq k').$$

Definition 4 *Let u be a word with $\mathbf{alph}(u) \subseteq A = \{a_1, \ldots, a_n\}$, and let v_a be a word on the set B, for each $a \in A$. The sets A, B need not be the same. We define **the word** w **obtained by substituting** v_a **for each occurrence of** a **in** u, as follows. The underlying order of w is the linear order $\sum_{x \in L_u} L_{v_{u(x)}}$, defined just above, labeled as follows:*

$$w(k, x) := v_{u(x)}(k), \quad x \in L_u,\ k \in L_{v_{u(x)}}.$$

When $\mathbf{alph}(u)$ is a subset of the infinite set $A = \{a_1, a_2, \ldots\}$ and v_i is a word on the set B, for each $i \geq 1$, then we use

$$u(a_1/v_1, a_2/v_2, \ldots)$$

to denote the word obtained by substituting v_i for a_i in u.

We define the words $c, p_\omega, r_{\omega^{op}}, \rho_n$, $n \geq 1$, on the countable set $a_1, a_2, \ldots$, as follows.

- $c := a_1 a_2$, the word $([2], \leq, u)$ with $u(1) = a_1,\ u(2) = a_2$.
- $p_\omega := a_1 a_1 \ldots$, the word whose underlying linear order is ω, each point of which is labeled a_1.

- $r_{\omega^{op}} := \ldots a_1 a_1$, the word whose underlying linear order is ω^{op}, each point of which is labeled a_1.
- For $1 \leq n < \omega$, ρ_n is the word whose underlying linear order is $\mathbb{Q}$, every point labeled by some a_i, $i \in [n]$, and between any two points $q < q'$ in $\mathbb{Q}$, for each $j \in [n]$ there is a point labeled a_j. (There is a unique such word, up to isomorphism. See [Ro82], pp 116.)

We now define the **regular operations** of concatenation (or product or composition) $u \cdot v$, omega power u^{ω}, omega-op power $u^{\omega^{op}}$, and shuffle, $[\![u_1, \ldots, u_n]\!]^{\eta}$.

Definition 5 (Regular operations) *For any words $u, v, u_1, \ldots, u_n$ on A:*

$$\begin{aligned} u \cdot v &:= c(a_1/u, a_2/v) \\ u^{\omega} &:= p_{\omega}(a_1/u) \\ u^{\omega^{op}} &:= r_{\omega^{op}}(a_1/u) \\ [\![u_1, \ldots, u_n]\!]^{\eta} &:= \rho_n(a_1/u_1, \ldots, a_n/u_n). \end{aligned}$$

Note that there is one shuffle operation $(u_1, \ldots, u_n) \mapsto [\![u_1, \ldots, u_n]\!]^{\eta}$, for each positive integer n. Also,

$$\rho_n = [\![a_1, \ldots, a_n]\!]^{\eta}.$$

Sometimes we write just uv instead of $u \cdot v$. The operation of concatenation may be extended to any (countable) number of arguments. When $v_1, v_2, \ldots$ are words on A, we define $v_1 \cdot v_2 \cdots$ as the word $p'_{\omega}(a_1/v_1, a_2/v_2, \ldots)$ where $p'_{\omega} = a_1 a_2 \ldots$. In a later section, we will also be considering the **reverse** operation $u \mapsto u^r$. The reverse of the word $(L, \leq, u, A)$ is $(L, \geq, u, A)$. The labeling function of the reverse is the same as that of u.

5 Blocks

Definition 6 *Suppose that u is a word on the set A and let $X \subseteq A$, $X \neq \emptyset$. A subword v of u is X-**uniform** if*

- *$L_v \subseteq L_u$ is an **open interval** with at least two points and*
- *for each $p \in L_v$, the label of p, $u(p)$, belongs to X, and*
- *for each $p, q \in L_v$ and $a \in X$, if $p < q$, then there is some $r \in L_v$ with $p < r < q$ and $u(r) = a$.*

*An **extended** X-uniform word is a word of the form v, xv, vx or xvx', where v is X-uniform and x, x' are singletons in X.*

*We say a subword of u is **uniform** if it is X-uniform for some $X \subseteq A$.*

Remark 7 *Note that if v is an (extended) X-uniform subword of u, then* $\mathbf{alph}(v) =$

X. Thus, if v is a uniform subword of u, then there is a unique X such that v is X-uniform. Also, if $X = \{a_1, \ldots, a_n\}$, then any X-uniform subword of u is isomorphic to $[\![a_1, \ldots, a_n]\!]^\eta$. Every uniform word is dense.

An obvious fact is the following.

Lemma 8 *Suppose that v is an X-uniform subword of u. If w is a subword of u determined by a nonempty open interval contained in v, then w is an X-uniform subword of u.*

Proposition 9 *If v is an X-uniform subword of u, and if w is a Y-uniform subword of u, and if the underlying intervals of v and w intersect, then $X = Y$.*

Proof. Since both intervals L_v, L_w are open and dense in L_u, their intersection is open and dense by Proposition 1, and hence there is a nonempty interval (p, q) which is a subset of each. Thus, all of the points in this interval are labeled with exactly the letters in X and exactly the letters in Y. Thus $X = Y$. □

Lemma 10 *If v and v' are X-uniform subwords of u and if $L_v \cap L_{v'} \neq \emptyset$, then $(L_v \cup L_{v'}, \leq, u, A)$ is an X-uniform subword of u.*

Proof. Suppose $L_v \cap L_{v'} \neq \emptyset$. Clearly, $L_v \cup L_{v'}$ is open and dense. We show $(L_v \cup L_{v'}, \leq, u)$ is an X-uniform subword of u. Suppose that $z < y$ in $L_v \cup L_{v'}$. We have to show that for each letter a in X, there is some element $w \in L_v \cup L_{v'}$ between z and y with $u(w) = a$.

If both z, y are in the same interval, there is no problem. So assume $z \in L_v$ and $y \in L_{v'}$. Let $p \in L_v \cap L_{v'}$. There are three possibilities: $p \leq z < y$, $z < y < p$ or $z < p \leq y$. In the first two cases, z, y are forced to be in the same interval. In the last case, since $p, z \in L_v$, for each $a \in X$ there is some w with $z < w < p \leq y$ and $u(w) = a$. □

If v, v' are subwords of u, write $v \subseteq v'$ if the underlying interval of v is contained in the underlying interval of v'. We say that a subword v of the word u is a **maximal** X-uniform subword of u if v is X-uniform, and whenever $v \subseteq v'$, for an X-uniform subword v' of u, then $v = v'$.

Proposition 11 *Any X-uniform subword of a word u is contained in a unique maximal X-uniform subword.*

Proof. Suppose that w is an X-uniform subword. We claim the union of all X-uniform subwords containing w is the maximal X-uniform subword containing w. Indeed, this fact follows from the same argument given in Lemma 10 above. □

Corollary 12 *Any X-uniform subword of a word u is contained in a unique maximal uniform subword. This maximal uniform subword is also X-uniform.*

Proof. Suppose that v is an X-uniform subword of u. By Proposition 11, it is contained in a unique maximal X-uniform subword w. We show that w is the unique maximal uniform subword over v. Suppose that w' is a Y-uniform subword containing v. Then, by Proposition 9, $X = Y$. Thus, by Proposition 11, w' is a subword of w. $\square$

Proposition 13 *Suppose u is a word on A that contains a dense subword. Then if A is finite, u contains a uniform subword.*

Proof. Let v be an open dense subword of u such that the size of $X = \mathbf{alph}(v)$ is minimum. Since A is finite, such a word must exist. Then if $a \in X$, every nonempty open subinterval of v must contain a point labeled a, i.e., v is X-uniform. $\square$

We sometimes write $\langle a_i\rangle [\![a_1, \ldots, a_n]\!]^\eta$ to mean either $a_i \cdot [\![a_1, \ldots, a_n]\!]^\eta$ or $[\![a_1, \ldots, a_n]\!]^\eta$. Similarly, $[\![a_1, \ldots, a_n]\!]^\eta \langle a_j\rangle$ means either $[\![a_1, \ldots, a_n]\!]^\eta$ or $[\![a_1, \ldots, a_n]\!]^\eta \cdot a_j$. Last,

$$\langle a_i\rangle [\![a_1, \ldots, a_n]\!]^\eta \langle a_j\rangle$$

means either $a_i \cdot [\![a_1, \ldots, a_n]\!]^\eta \cdot a_j$ or $\langle a_i\rangle [\![a_1, \ldots, a_n]\!]^\eta$ or $[\![a_1, \ldots, a_n]\!]^\eta \langle a_j\rangle$.

The following proposition characterizes those words isomorphic to a word of the form

$$\langle a_i\rangle [\![a_1, \ldots, a_n]\!]^\eta \langle a_j\rangle.$$

Proposition 14 *Let $w = [\![a_1, \ldots, a_p]\!]^\eta$, and suppose that u, v are nonempty words.*

- *$u \cdot v \cong w$ iff either*
(1) $u \cong w$ and $v \cong w$, or for some $i \in [p]$,
(2) $u \cong w \cdot a_i$ and $v \cong w$, or
(3) $u \cong w$ and $v \cong a_i \cdot w$.
- *$u^\omega \cong w$ iff either $u \cong w$ or $u \cong w \cdot a_i$, for some $i \in [p]$.*
- *$u^{\omega^{op}} \cong w$ iff either $u \cong w$ or $u \cong a_i \cdot w$, for some $i \in [p]$.*
- *$[\![u_1, \ldots, u_n]\!]^\eta \cong w$, for nonempty words $u_1, \ldots, u_n$ iff either the set $\{u_1, \ldots, u_n\}$ can be partitioned into $\{d_1, \ldots, d_k\}$ and $\{e_1, \ldots, e_m\}$, where $k > 0$ and the words d_i are extended $\{a_1, \ldots, a_p\}$-uniform words, and the words e_t are letters in $\{a_1, \ldots, a_p\}$, or all u_i are letters and for each $j \in [p]$ there is some $i \in [n]$ with $u_i = a_j$.*

For $i, j \in [p]$,

- *$u \cdot v \cong a_i \cdot w$ iff*
(1) $u = a_i$ and $v \cong w$, or,
(2) $u \cong a_i \cdot w \cdot \langle a_\ell\rangle$ and $v \cong \langle a_k\rangle \cdot w$, for some $\ell, k \in [p]$, where not both a_ℓ and a_k appear.
- *$u^\omega \cong a_i \cdot w$ iff $u \cong a_i \cdot w$.*
- *$u^{\omega^{op}} \cong w \cdot a_j$ iff $u \cong w \cdot a_j$.*
- *$u \cdot v \cong a_i \cdot w \cdot a_j$ iff*
(1) $u = a_i$ and $v \cong w \cdot a_j$, or
(2) $u \cong a_i \cdot w$ and $v = a_j$, or
(3) $u \cong a_i \cdot w \cdot \langle a_k\rangle$ and $v \cong \langle a_s\rangle \cdot w \cdot a_j$, and not both a_k, a_s appear, $k, s \in [p]$.
- *No word $u^{\omega^{op}}$ or $[\![u_1, \ldots, u_n]\!]^\eta$ is isomorphic to $a_i \cdot w$. No word u^ω or $[\![u_1, \ldots, u_n]\!]^\eta$ is isomorphic to $w \cdot a_j$. No word $u^\omega, u^{\omega^{op}}$ or $[\![u_1, \ldots, u_n]\!]^\eta$ is isomorphic to $a_i \cdot w \cdot a_j$.*

Proof. We prove only the first statement. If $u \cdot v \cong w$, then u cannot have a least element, and v cannot have a greatest; also, since both are nonempty, both are necessarily extended $\{a_1, \ldots, a_p\}$-uniform. □

Suppose that u is a word and $p, q \in L_u$. We write $q = S(p)$ (and $p = P(q)$) if q is the successor of p in L_u.

For any word u and any point $p \in L_u$, either there is some $q = S(p)$, or not. If so, we say $S(p)$ **exists**; similarly, if there is some $q = P(p)$, we say $P(p)$ **exists**. For $n \geq 0$, define:

$$\begin{aligned} S_0(p) &:= p \\ S_{n+1}(p) &:= S(S_n(p)), \text{ if both } S_n(p) \text{ and } S(S_n(p)) \text{ exist} \\ S_{-(n+1)}(p) &:= P(S_{-n}(p)), \text{ if both } S_{-n}(p) \text{ and } P(S_{-n}(p)) \text{ exist.} \end{aligned}$$

The set

$$\{S_k(p) : k \in \mathbb{Z}, S_k(p) \text{ exists}\}$$

is a nonempty interval, since $S_0(p)$ always exists. We will only be interested in the case that the point p is not contained in a uniform subword. Note that if $q = S(p)$, neither p nor q belongs to a uniform subword.

Definition 15 (s-closed) *Suppose that u is a word and v is a subword of u such that **no point of** v **is contained in a uniform subword**. We say that v is **s-closed** (in u) if v is nonempty and for each p in L_v, if $S(p)$ exists then $S(p)$ is also in L_v, and similarly, if $P(p)$ exists, then $P(p)$ is in L_v. A **minimal s-closed subword of** u is a nonempty s-closed subword that contains no other s-closed subword.*

Proposition 16 *Suppose that u is a word. A subword v of u is a minimal s-closed subword of u iff for some point $p \in L_u$ which does not belong to any uniform subword, v is the collection of points $\{S_k(p) : k \in \mathbb{Z}, S_k(p)$ exists$\}$, ordered and labeled as in u.* □

We now show how any word may be subdivided (by a particular "condensation" [Ro82]), into blocks.

Definition 17 (blocks) *A **block** of the word u is either a maximal uniform subword of u, called a **dense block**, or a minimal s-closed subword of u, called a **scattered block**.*

Note that if $S(p)$ or $P(p)$ exists, then p does not belong to any uniform word; a scattered block does not contain any point that belongs to a uniform subword.

Proposition 18 *If v, w are blocks in the word u, then either $v = w$ or L_v and L_w are disjoint.*

Proof. This follows from Corollary 12. □

Proposition 19 *Let u be a word. Each point $p \in L_u$ belongs to some block, denoted* $\mathsf{Bl}(p)$, *and hence a unique block.*

Proof. If p belongs to some uniform subword v, then $\mathsf{Bl}(p)$ is the unique maximal uniform subword over v, which exists by Corollary 12. Otherwise, $\mathsf{Bl}(p)$ is the minimal s-closed subword containing p. This exists by Proposition 16. Uniqueness follows. $\square$

Note that for any word u, the **blocks of** u are the subwords $\mathsf{Bl}(p)$, for $p \in L_u$.

Proposition 20 *Let u be a word and suppose that $p < q$ are points in L_u. Then $\mathsf{Bl}(p) = \mathsf{Bl}(q)$ iff there is some uniform subword containing both p and q, or neither p nor q belongs to a uniform subword and the interval (p, q) is finite, i.e., $q = S_n(p)$, for some $n > 0$.*

Proof. First suppose that $\mathsf{Bl}(p) = \mathsf{Bl}(q)$. If p belongs to a uniform subword, then q belongs to the maximal uniform subword containing p. If p does not belong to any uniform subword, then neither does q, furthermore, $q = S_n(p)$ for some positive $n \in \mathbb{Z}$, and thus the interval (p, q) has $n - 1$ points in it.

Conversely, suppose that p, q belong to the same uniform subword. Then $\mathsf{Bl}(p) = \mathsf{Bl}(q)$, as shown in the proof of Proposition 19. If neither belongs to a uniform subword and the interval (p, q) is finite, then $q = S_n(p)$ for some positive integer n, so here, too, $\mathsf{Bl}(p) = \mathsf{Bl}(q)$. $\square$

Example 21
- *Let w be the word denoted by the term $[\![a]\!]^\eta \cdot a \cdot b \cdot [\![b]\!]^\eta$. Then w has three blocks, denoted $[\![a]\!]^\eta$, ab, and $[\![b]\!]^\eta$, respectively.*
- *Let w be the word denoted by $[\![a]\!]^\eta \cdot a$. Then w has two blocks.*
- *Suppose that w is $[\![a]\!]^\eta \cdot [\![a, b]\!]^\eta$. Then w has two blocks, denoted $[\![a]\!]^\eta$, and $[\![a, b]\!]^\eta$, respectively.*
- *Suppose that $w = a^\omega \cdot a \cdot [\![a]\!]^\eta$. Then w has three blocks.*
- *The word $[\![a, b]\!]^\eta \cdot a \cdot [\![a, b]\!]^\eta$ has one block.*
- *The word $[\![[\![a]\!]^\eta \cdot a]\!]^\eta$ has one block.*

We note the fact that the blocks of a word partition the underlying linear order.

Proposition 22 *The blocks of a word u are linearly ordered by the relation:*

$$\mathsf{Bl}(p) < \mathsf{Bl}(q) \iff x < y, \quad \textit{in } L_u, \textit{ for all } x \in \mathsf{Bl}(p),\ y \in \mathsf{Bl}(q).$$

(When $\mathsf{Bl}(p) \neq \mathsf{Bl}(q)$, then $\mathsf{Bl}(p) < \mathsf{Bl}(q)$ if $x < y$ for some $x \in \mathsf{Bl}(p)$, $y \in \mathsf{Bl}(q)$.) $\square$

Definition 23 *The underlying order on the blocks of the word u is denoted*

$$(\mathsf{Bl}(u), \leq).$$

In Corollary 70, we will show that each block of any regular word is denoted by a "primitive term", defined below.

The next lemma explains what happens when two or three blocks are concatenated.

Lemma 24 *Suppose that each of the A-labeled words u and v has one block (so that neither word is empty). Then uv has one or two blocks.*

(1) uv has a single block iff
  - *u and v are both scattered, u is right-closed and v is left-closed, or*
  - *u and v are both X-uniform, for some $X \subseteq A$.*

(2) uv has the two blocks u, v iff one of the following cases occurs.
  - *Both u and v are scattered and either u is right-open or v is left-open.*
  - *u is scattered, v is uniform.*
  - *v is scattered, u is uniform.*
  - *u is X-uniform, v is Y-uniform for some $X, Y \subseteq A$, and $X \neq Y$.*

(3) If w also has one block, then uvw has one, two or three blocks. The word uvw has one block iff one of the following cases occurs.
  - *Each of u, v, w is scattered, u is right-closed, v is finite and w is left-closed.*
  - *u, v, w are X-uniform, for some $X \subseteq A$.*
  - *u, w are X-uniform, for some $X \subseteq A$, and v is a singleton, labeled $a \in X$.*

  The word uvw has two blocks iff one of the following cases occurs.
  - *uv has a single block which is scattered, w is scattered, and either uv is right-open or w is left-open.*
  - *vw has a single block which is scattered, u is scattered, and either vw is left-open or u is right-open.*
  - *uv has a single block which is scattered and w is uniform.*
  - *vw has a single block which is scattered and u is uniform.*
  - *uv is X-uniform and w is scattered or Y-uniform with $X \neq Y$.*
  - *u is X-uniform or scattered and vw is Y-uniform with $X \neq Y$.*

  Otherwise, uvw has three blocks, namely u, v, w.

Definition 25 *Suppose that u, v, w are nonempty words. We say u, v* ***merge*** *if the blocks of uv are not the same as the blocks of u together with the blocks of v. We say u, v, w* ***merge*** *if the blocks of uvw are not the same as the blocks of u, together with the blocks of v and the blocks of w.*

Example 26 *If $u = a^\eta,\ v = a^\eta$ or $u = a,\ v = a^\omega$, then u, v merge. If $u = a^\eta,\ v = a,\ w = a^\eta$, then neither u, v nor v, w merge, but u, v, w merge.*

Remark 27 *Suppose that u, v, w are each words consisting of a single block. Then u, v* ***merge*** *iff uv has a single block; if either u, v do not merge, or v, w do not merge, then u, v, w merge iff uvw consists of a single block.*

Corollary 28 (Merge Corollary) *Suppose that u, v, w are each words consisting of a single block. Then u, v merge iff one of the following cases occurs.*

- *Both u and v are scattered, u is right-closed and v is left-closed, or*
- *both u and v are X-uniform, for some $X \subseteq A$.*

Assume that neither u, v, nor v, w merge. Then u, v, w merge iff u, w are X-uniform, for some $X \subseteq A$, and v is a singleton, labeled $a \in X$.

From Lemma 24, we immediately derive the following facts.

Corollary 29 *Suppose that* p *and* q *are points of the words* u *and* v*, respectively. If* $\mathsf{Bl}(p)$ *in* u *is not the last or second to last block of* u*, then it is also a block of* uv*. Symmetrically, if* $\mathsf{Bl}(q)$ *in* v *is not the first or second block of* v*, then it is also a block of* uv*.*

Corollary 30 (First block concatenation corollary) *Suppose that* u *and* v *are nonempty words on the set* A*.*

1. *If* u *has no last block, or* v *has no first block, then* u, v *do not merge.*
2. *If* u *has a last block* u_1*, but no second to last block, and* v *has a first block* v_1 *but no second block, then* u, v *merge iff* u_1, v_1 *merge; when* u_1, v_1 *merge, the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_1 *and the block* u_1v_1*.*
3. *If* u *has a last block* u_1 *and a second to last block* u_2*, and* v *has a first block* v_1*, but no second block, then* u, v *merge iff* u_2, u_1, v_1 *merge or* u_1, v_1 *merge, and exactly one of these cases can occur. If* u_2, u_1, v_1 *merge, then the blocks of* uv *are the blocks of* u *less than* u_2*, the blocks of* v *greater than* v_1 *and the one block* $u_2u_1v_1$*; if* u_1, v_1*merge, the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_1 *and the block* u_1v_1*.*
4. *If* u *has a last block* u_1*, but no second to last block, and* v *has a first block* v_1 *and second block* v_2*, then* u, v *merge iff either* u_1, v_1, v_2 *merge or* u_1, v_1 *merge, and exactly one of these cases can occur. If* u_1, v_1, v_2 *merge, then the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_2 *and the block* $u_1v_1v_2$*. If* u_1, v_1 *merge, the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_1 *and the block* u_1v_1*.*
5. *If* u *has a last block* u_1 *and a second to last block* u_2*, and* v *has a first block* v_1 *and second block* v_2*, then* u, v *merge iff either* u_2, u_1, v_1 *merge, or* u_1, v_1, v_2 *merge or* u_1, v_1 *merge, and exactly one of these cases can occur. If* u_2, u_1, v_1 *merge, then the blocks of* uv *are the blocks of* u *less than* u_2*, the blocks of* v *greater than* v_1 *and the block of* $u_2u_1v_1$*. If* u_1, v_1, v_2 *merge, the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_1 *and the block* $u_1v_1v_2$*. Finally, if* u_1, v_1 *merge, the blocks of* uv *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* v_2 *and the block* u_1v_1*.*

Corollary 31 (Second block concatenation corollary) *Suppose that* u, v, w *are nonempty*
words on the set A *such that* u, v *and* v, w *do not merge.*

1. *If* u *has no last block or* w *has no first block or* v *has more than one block, then* u, v, w *do not merge.*
2. *If* u *has a last block* u_1*,* w *has a first block* w_1 *and* v *has a single block, then* u, v, w *merge iff* u_1, v, w_1 *merge; when* u_1, v, w_1 *merge, the blocks of* uvw *are the blocks of* u *less than* u_1*, the blocks of* v *greater than* w_1 *and the block* u_1vw_1*.*

We state yet another corollary of these facts.

Corollary 32 (General block concatenation corollary) *Let* $u_1, u_2, \ldots$ *be a finite or infinite sequence of nonempty words on* A*. The blocks of* $u_1 u_2 \cdots$ *are the blocks in each* u_i *iff there is no consecutive pair* u_i, u_{i+1}*, or triple of words* u_i, u_{i+1}, u_{i+2} *that merge. Similarly, the blocks of* $\cdots u_3 u_2 u_1$ *are those in each* u_i *iff there is no consecutive pair* u_{i+1}, u_i *or triple* u_{i+2}, u_{i+1}, u_i *that merge.*

We turn to blocks of a shuffle.

Proposition 33 *Suppose that* $U = \{u_1, \ldots, u_n\}$*,* $n \geq 1$*, is a set of nonempty* A*-labeled words at least one of which is not a singleton. Then the blocks of*

$$v = [\![u_1, \ldots, u_n]\!]^\eta$$

are copies of the blocks of the u_i *iff* U *cannot be partitioned into two parts such that, for some* $X \subseteq A$*, the words in one part are all extended* X*-uniform and the words in the other part are all singletons labeled in* X*.*

Proof. The meaning of the long statement is that blocks of the words u_i are not destroyed by the shuffle operation except in one case: when U can be so partitioned, v has one block, by Proposition 14.

When U cannot be so partitioned, first we show that any block of v is a block of a copy of some u_i. So suppose that w is a block of v. If w is not a block of some copy of a word in U, w contains a point p in w that belongs to one copy of a word u_i in v and a point q in w that belongs to a different copy of the same u_i, or to a copy of some u_j with $j \neq i$. Assume $p < q$. By the definition of the word v, for each k there exists a copy of u_k between p and q. It follows that w is not a scattered block and thus X-dense, for some $X \subseteq A$. Moreover, each u_k is either extended X-uniform or a singleton labeled in X, contradicting our assumptions.

Next, we show that for any copy of any word u_i in v, and for any block w of this copy of u_i, it holds that w is a block of v. There are two cases, w is scattered or dense. Suppose first that w is scattered. Then clearly, w is a block of v.

Suppose now that w is X-uniform, for some X. If w is not a block of v, it is properly included in an X-uniform subword w' of v. But then w' contains a point not included in the same copy u_i' of the same u_i of which w is a block. Thus a contradiction results, as above. ☐

6 Terms

For us, the structures of interest are algebras

$$(X, \cdot, {}^\omega, {}^{\omega^{op}}, {}^\eta)$$

which are enrichments of a semigroup $(X, \cdot)$ by two unary operations: $x \mapsto x^\omega$ and $x \mapsto x^{\omega^{op}}$, and an n-ary operation $(x_1, \ldots, x_n) \mapsto [\![x_1, \ldots, x_n]\!]^\eta$, for each $n \geq 1$. The intended models are the algebras $(AW, \cdot, ^{\omega}, ^{\omega^{op}}, ^{\eta})$ of all finite and countable words on the set A, enriched with the indicated operations. For each such algebra, we let AR denote the least subalgebra of AW containing the singletons, i.e., the subalgebra of AW consisting of the nonempty *regular words* (as proved by Heilbrunner [Heil80], see Theorem 38 below).

Proposition 34 *Suppose that A and B are sets and $\mathbf{B}$ is any algebra of words on B equipped with the operations $\cdot, ^{\omega}, ^{\omega^{op}}, ^{\eta}$. Then any function $A \to B$ can be extended to a homomorphism $AW \to \mathbf{B}$.*

Proof. Given $h : A \to B$, for each word u in AW define $h^{\sharp}(u)$ as the word obtained by substituting a disjoint copy of $h(a)$ for each $x \in L_u$, where $u(x) = a$. It is a routine matter to show that $h^{\sharp}$ is a homomorphism. □

Definition 35 *Let A be a fixed nonempty set.*

*(1) A **term** s **on the set** A is either some letter $a \in A$, or s is $t \cdot t'$, t^{ω}, $t^{\omega^{op}}$, or $[\![t_1, \ldots, t_n]\!]^{\eta}$, where $t, t', t_1, \ldots, t_n$ are terms on A.*

*(2) A **scattered term** is a term that has no occurrence of the function symbol $^{\eta}$.*

*(3) The **height** of a term t, denoted $\mathsf{ht}(t)$, is the maximum number of nested $^{\omega}$, $^{\omega^{op}}$ and $^{\eta}$ operations in t. Recursively,*

$$\mathsf{ht}(t) := \begin{cases} 0 & \textit{if } t \in A \\ \max\{\mathsf{ht}(t_1), \mathsf{ht}(t_2)\} & \textit{if } t = t_1 \cdot t_2 \\ 1 + \mathsf{ht}(s) & \textit{if } t = s^{\omega} \textit{ or } t = s^{\omega^{op}} \\ 1 + \max\{\mathsf{ht}(t_1), \ldots, \mathsf{ht}(t_n)\} & \textit{if } t = [\![t_1, \ldots, t_n]\!]^{\eta} \end{cases}$$

(4) When t is a term on the set A, we let $|t|$ denote the word on A denoted by t. More precisely, for $a \in A$, $|a|$ is a singleton set, labeled a, and, inductively, we define

$$\begin{aligned} |t \cdot t'| &:= |t| \cdot |t'| \\ |t^{\omega}| &:= |t|^{\omega} \\ |t^{\omega^{op}}| &:= |t|^{\omega^{op}} \\ |[\![t_1, \ldots, t_n]\!]^{\eta}| &:= [\![|t_1|, \ldots, |t_n|]\!]^{\eta}. \end{aligned}$$

*(5) An equation $t = t'$ between two terms on A is **valid**, and t and t' are **equivalent**, if $|t| \cong |t'|$.*

*(6) The **size** of a term t on A is defined as follows:*

$$\mathsf{size}(t) := \begin{cases} 1 & \textit{if } t \in A \\ \mathsf{size}(t_1) + \mathsf{size}(t_2) + 1 & \textit{if } t = t_1 \cdot t_2 \\ \mathsf{size}(t_1) + 1 & \textit{if } t = t_1^{\omega} \textit{ or } t = t_1^{\omega^{op}} \\ 1 + \sum_{i=1}^{n} \mathsf{size}(t_i) & \textit{if } t = [\![t_1, \ldots, t_n]\!]^{\eta} \end{cases} \tag{2}$$

We sometimes add a term **1** of height 0 and size 1 denoting the empty word. Terms of height zero are called "finite". Those of positive height are "infinite".

In the usual way, each term t on A induces a function $X^A \to X$, for any algebra X equipped with the operations $\cdot, ^\omega, ^{\omega^{op}}, ^\eta$. In fact, for a term t, the word $|t|$ is just the value of the function induced by t over the algebra AW when each letter a is evaluated as the singleton word labeled a.

From Proposition 34 we immediately infer the following fact.

Proposition 36 *For any terms t, t' over A, $t = t'$ is valid iff $t = t'$ holds in all word algebras $(BW, \cdot, ^\omega, ^{\omega^{op}}, ^\eta)$ under any evaluation of the letters in A as words on B, i.e., when $t = t'$ holds in the variety generated by all word algebras.*

A binary tree *on A* consists of a finite or infinite binary tree (considered as a nonempty, prefix closed set of binary strings, as usual) and a labeling function mapping the set of leaves of the tree to A. The leaves, when ordered lexicographically, form a linearly ordered set. When T is a binary tree on A and u is a node of T, the subtree rooted at u is also a binary tree on A. A tree is **regular** if up to isomorphism it has a finite number of subtrees. The *frontier* of a tree on A is the A-labeled linearly ordered set of the leaves.

Definition 37 (Regular word) *A word u on A is called **regular** if it is isomorphic to the frontier of a regular binary tree on A.*

Note that if u is a regular word on A, then $\mathbf{alph}(u)$ is a finite subset of A.

The following theorem summarizes some of the results in Heilbrunner [Heil80] and Courcelle [Cour78].

Theorem 38 *The following are equivalent for a nonempty word u on A.*

- *u is regular.*
- *u is a component of the initial solution to a system of fixed point equations of the form (1) above.*
- *u belongs to the least collection of words containing the singletons a, $a \in A$, closed under the regular operations.*
- *$u = |t|$, for some term t on A.*

Also, if t is a term on A, then the word $|t|$ has a scattered underlying linear order iff t is a scattered term. □

Remark 39 *The least class of linear orders containing the empty and singleton orders, closed under the regular operations is denoted* M *in [Ro82].*

We note that the regular words are closed under substitution.

Proposition 40 *If v_i is a regular word on the set B, for each $i \in [n]$, and u is a regular word on A, where $\mathbf{alph}(u) \subseteq \{a_1, \ldots, a_n\}$, then $u(a_1/v_1, \ldots, a_n/v_n)$ is a regular word on B.* □

For any terms $u_1, \ldots, u_n$, the terms

$$[\![u_1, \ldots, u_n]\!]^\eta \text{ and } [\![u_{f(1)}, \ldots, u_{f(n)}]\!]^\eta$$

are equivalent, for any permutation f of $[n]$. Our axioms will guarantee that this fact is provable. Thus, we will usually implicitly identify the terms $[\![u_1, \ldots, u_n]\!]^\eta$ and $[\![u_{f(1)}, \ldots, u_{f(n)}]\!]^\eta$, for a permutation f of $[n]$.

Note: We will sometimes omit syntactically needed parentheses in expressions such as $a_1 \cdots a_n$ when no confusion will result.

We now single out an important class of terms.

Definition 41 *A term on A is* ***primitive*** *if it has one of the following forms:*

- $a_1 \cdots a_n$, $\quad n > 0$, $a_i \in A$, *(a finite term)*
- $a_1 \cdots a_n(b_1 \cdots b_k)^\omega$, $\quad n \geq 0$, $k > 0$, $a_i, b_j \in A$ *(a left-closed, right-open scattered primitive term)*
- $(c_m \cdots c_1)^{\omega^{op}} a_1 \cdots a_n$, $\quad n \geq 0$, $m > 0$, $a_i, c_j \in A$ *(a left-open, right-closed scattered primitive term)*
- $(c_m \cdots c_1)^{\omega^{op}} a_1 \cdots a_n(b_1 \cdots b_k)^\omega$, $\quad a_i, b_j, c_k \in A$, $m, k > 0$, $n \geq 0$ *(an open scattered primitive term)*
- $[\![a_1, \ldots, a_n]\!]^\eta$, $\quad n > 0$ *(an $\{a_1, \ldots, a_n\}$-uniform primitive term)*

An ***infinite scattered primitive term*** *is a scattered primitive term that is not finite. A* ***dense primitive term*** *is a primitive term of the last kind. An* ***extended primitive term*** *is either the term* **1** *or a primitive term.*

We say that a word u on A is **primitive** if u is isomorphic to $|t|$, for some primitive term t. Note that a primitive term t is finite, infinite, scattered, dense, etc. iff the primitive word $|t|$ it denotes has the corresponding property. The following fact is clear.

Proposition 42 *A word is primitive iff it is regular and has a single block.*

There are efficient algorithms to determine whether a term is primitive, and if so, exactly what kind.

Proposition 43 *There is a $O(\mathsf{size}(t))$ algorithm to determine, given a term t, whether t is primitive, and if so, finite, scattered, left-closed, etc.* □

Definition 44 *We say that a pair of terms l, r merge when the corresponding words $|l|, |r|$ merge. Similarly, we say the terms l, m, r merge when the corresponding words $|l|, |m|, |r|$ do.*

We apply Lemma 24 to obtain an explicit description of those primitive terms that merge.

Proposition 45 *Suppose that l, m, r are primitive terms on A.*

(1) l, r merge iff

- *both l and r are scattered, l is right-closed and r is left-closed, or*
- *both l and r are X-uniform, for some $X \subseteq A$.*

(2) If l, m do not merge, and m, r do not merge, then l, m, r merge iff l, r are X-uniform, for some $X \subseteq A$, and m is a letter in X.

Example 46 *If r is an infinite scattered primitive term, then the pair r, r does not merge, since $|r|$ does not have both a least and greatest element.*

Corollary 47 *For primitive terms l, m, r on A, there is a $O(n \log n)$ algorithm to determine whether l, m, r merge, or whether l, r merge, where n is sum of the sizes of l, m, r or l, r, respectively.*

Proof. For two scattered terms, we need only check whether l is right-closed and m is left-closed. This takes $O(n)$ time. To tell whether two dense terms $[\![a_1, \ldots, a_k]\!]^\eta$ and $[\![b_1, \ldots, b_j]\!]^\eta$ are X-uniform for the same set X, we assume that the underlying alphabet is sorted, and then sort the two sequences $(a_1, \ldots, a_k)$, $(b_1, \ldots, b_j)$ to determine whether they consist of the same elements. This may be done in $O(n \log n)$ time. □

7 The axioms

In this section we list the axioms Ax used in our completeness theorem in Section 10. The axioms are divided into several groups.

Definition 48 (scattered axioms)

$$(x \cdot y) \cdot z = x \cdot (y \cdot z) \tag{3}$$
$$(x \cdot y)^\omega = x \cdot (y \cdot x)^\omega \tag{4}$$
$$(x \cdot y)^{\omega^{op}} = (y \cdot x)^{\omega^{op}} \cdot y \tag{5}$$
$$(x^n)^\omega = x^\omega, \quad n \geq 2 \tag{6}$$
$$(x^n)^{\omega^{op}} = x^{\omega^{op}}, \quad n \geq 2 \tag{7}$$

We note two consequences of the scattered axioms.

$$x \cdot x^\omega = x^\omega \tag{8}$$
$$x^{\omega^{op}} \cdot x = x^{\omega^{op}}. \tag{9}$$

We prove (8).

$$\begin{aligned} x^\omega &= (x^2)^\omega, \quad \text{by Axiom 6, } n = 2, \\ &= x \cdot (x^2)^\omega, \quad \text{by Axiom 4, } x = y, \\ &= x \cdot x^\omega, \quad \text{by Axiom 6 again.} \quad \square \end{aligned}$$

The remaining axioms concern the shuffle operation. We call the first group the logical axioms.

Definition 49 (logical axioms)

$$[\![x_{f(1)},\ldots,x_{f(n)}]\!]^\eta = [\![x_1,\ldots,x_p]\!]^\eta, \tag{10}$$

where $f:[n]\to[p]$ is any set-theoretic surjection.

The logical axioms say that the shuffle operation $[\![a_1,\ldots,a_n]\!]^\eta$ is a function whose value is determined by the set $\{a_1,\ldots,a_n\}$, not the sequence $(a_1,\ldots,a_n)$; for example, using these axioms one may derive the facts that $[\![a,a,b]\!]^\eta = [\![b,b,a]\!]^\eta = [\![a,b]\!]^\eta = [\![b,a]\!]^\eta$.

The logical axioms may be replaced by the following collection of somewhat simpler looking identities:

$$\begin{aligned}
[\![x_1,x_2,\ldots,x_p]\!]^\eta &= [\![x_2,x_1,x_3,\ldots,x_p]\!]^\eta, \quad p\geq 2\\
[\![x_1,x_2,\ldots,x_p]\!]^\eta &= [\![x_2,x_3,\ldots,x_p,x_1]\!]^\eta, \quad p\geq 3\\
[\![x_1,x_2,\ldots,x_p]\!]^\eta &= [\![x_1,x_1,x_2,\ldots,x_p]\!]^\eta, \quad p\geq 1.
\end{aligned}$$

The remaining axioms show how the shuffle operation interacts with the concatenation, omega, omega-op operations, and with itself.

Definition 50 (concatenation/shuffle axioms)

$$[\![x_1,\ldots,x_p]\!]^\eta\cdot[\![x_1,\ldots,x_p]\!]^\eta = [\![x_1,\ldots,x_p]\!]^\eta \tag{11}$$

$$[\![x_1,\ldots,x_p]\!]^\eta\cdot x_i\cdot[\![x_1,\ldots,x_p]\!]^\eta = [\![x_1,\ldots,x_p]\!]^\eta, \quad i\in[p]. \tag{12}$$

Definition 51 (omega/shuffle axioms)

$$([\![x_1,\ldots,x_p]\!]^\eta)^\omega = [\![x_1,\ldots,x_p]\!]^\eta \tag{13}$$

$$([\![x_1,\ldots,x_p]\!]^\eta\cdot x_i)^\omega = [\![x_1,\ldots,x_p]\!]^\eta, \quad i\in[p]. \tag{14}$$

Definition 52 (omega-op/shuffle axioms)

$$([\![x_1,\ldots,x_p]\!]^\eta)^{\omega^{op}} = [\![x_1,\ldots,x_p]\!]^\eta \tag{15}$$

$$(x_i\cdot[\![x_1,\ldots,x_p]\!]^\eta)^{\omega^{op}} = [\![x_1,\ldots,x_p]\!]^\eta, \quad i\in[p]. \tag{16}$$

Definition 53 (shuffle/shuffle axioms)

$$[\![u_1,\ldots,u_k,v_1,\ldots,v_s]\!]^\eta = [\![x_1,\ldots,x_p]\!]^\eta, \quad k\geq 0,\ s>0, \tag{17}$$

where in (17), the terms u_i are letters in the set $\{x_1,\ldots,x_p\}$, and each term v_j is one of the following:

$$[\![x_1,\ldots,x_p]\!]^\eta,\ x_i[\![x_1,\ldots,x_p]\!]^\eta,\ [\![x_1,\ldots,x_p]\!]^\eta x_j,\ \text{or}\ \ x_i[\![x_1,\ldots,x_p]\!]^\eta x_j.$$

Note that a special case of the shuffle/shuffle axioms is the identity

$$(\llbracket x_1,\dots,x_p\rrbracket^\eta)^\eta = \llbracket x_1,\dots,x_p\rrbracket^\eta.$$

Let Ax denote the collection of all axioms mentioned above: the scattered axioms, the logical axioms, concatenation/shuffle, omega/shuffle, omega-op/shuffle and shuffle/shuffle axioms.

Proposition 54 (Soundness) *Each axiom is valid.* □

For later use, we note the following standard fact about equational logic.

Proposition 55 *If γ is a term morphism from terms on A to terms on B, s, t are terms on A, and $\mathsf{Ax} \vdash s = t$, then $\mathsf{Ax} \vdash \gamma(s) = \gamma(t)$.* □

Remark 56 *The shuffle/shuffle and concatenation/shuffle axioms were discussed in [Heil80], as were some of the scattered axioms.*

8 Normal forms of primitive terms

Primitive terms on A may be put into a normal form. In order to do so, we assume that there is a linear order on the set $A' = A \cup \{), (, \cdot, {}^\omega, {}^{\omega^{op}}, {}^\eta\}$ which induces the lexicographic order $\leq_\ell$ on all finite words on A': for finite words u, v on A,

$$u \leq_\ell v \iff \exists u_1, u_2, w \ ((v = uw) \text{ or } (u = wau_1 \text{ and } v = wa'u_2 \text{ and } a < a')).$$

Definition 57 *A primitive term t is in **normal form** if either*

- *t is finite, or*
- *$t = uv^\omega$, where u and v are finite words such that v is not empty, and if u is not empty then its last letter is different from the last letter of v, moreover, v is not a proper power, i.e., w^j for some $j \geq 2$, or*
- *$t = v^{\omega^{op}}u$, where u and v are finite words such that v is not empty and not a proper power, and if u is not empty then its first letter is different from the first letter of v, or*
- *$t = u^{\omega^{op}}vw^\omega$, where the finite, nonempty words u, w are not proper powers, moreover, v is a finite nonempty word whose first letter is different from the first letter of u and whose last letter is different from the last letter of w, or*
- *$t = u^{\omega^{op}}v^\omega$, where the finite, nonempty words u, v are not proper powers, moreover, there exist no nonempty words x, y, z with $u = xy$, $v = xz$ and $yxzx <_\ell xyxz$, or with $u = yx$, $v = zx$ and $xyxz <_\ell yxzx$, or*
- *$t = \llbracket a_1,\dots,a_n\rrbracket^\eta$, $\quad n > 0$ and $a_1 < \ldots < a_n$ in A.*

Later, we will need the fact that normal forms may be found in quadratic time.

Lemma 58 (Single block lemma) *There is a quadratic time algorithm which, given a term t determines whether $|t|$ has a single block, and if so, produces a primitive term p in normal form such that*

$$\mathsf{Ax} \vdash t = p,$$

and $\mathsf{ht}(p) \leq \mathsf{ht}(t)$*. Moreover, the size of* p *is at most the size of* t*.*

Proof. In linear time, one may determine whether a term is scattered or dense, but to tell if there is only one block, one needs to tell if the underlying set of two lists is the same, to know if $[\![a_1, \ldots, a_m]\!]^\eta$ and $[\![b_1, \ldots, b_k]\!]^\eta$ denote the same word, for example. This may be done in $O(n \log n)$ steps by sorting the two lists, where $n = m + k$.

If t is a scattered term, then we apply the results in [BlEs03a] to produce an essentially unique primitive term l such that the scattered axioms $\vdash t = p$. The size of p is linear in the size of t.

Thus, assume that the single block of the word $|t|$ is dense, and is denoted by the term

$$p = [\![a_1, \ldots, a_k]\!]^\eta$$

in normal form. The construction will show the height of p may be less than the height of t. In fact, we will prove using induction that if $|t| \cong u$, where u is denoted either by p or by $[\![a_1, \ldots, a_k]\!]^\eta \cdot a_j$ or by $a_i \cdot [\![a_1, \ldots, a_k]\!]^\eta$, or by $a_i \cdot [\![a_1, \ldots, a_k]\!]^\eta \cdot a_j$, then $\mathsf{Ax} \vdash t = p$ or $\mathsf{Ax} \vdash t = p \cdot a_j$ or $\mathsf{Ax} \vdash t = a_i \cdot p$, or $\mathsf{Ax} \vdash t = a_i \cdot p \cdot a_j$, respectively.

The proof is based on Proposition 14, and uses induction on the number of operation symbols in the term t.

If t_0 is a term with the fewest number of operation symbols such that $|t_0| \cong [\![a_1, \ldots, a_k]\!]^\eta$, then $t_0 = [\![b_1, \ldots, b_n]\!]^\eta$, where the sets $\{b_1, \ldots, b_n\}$ and $\{a_1, \ldots, a_k\}$ are the same. Hence $\mathsf{Ax} \vdash t_0 = p$ by the logical axioms. If t_r is a term with the fewest number of operation symbols such that $|t_r| \cong [\![a_1, \ldots, a_k]\!]^\eta \cdot a_j$, then $t_r = [\![b_1, \ldots, b_n]\!]^\eta \cdot a_j$, where, again, the sets $\{b_1, \ldots, b_n\}$ and $\{a_1, \ldots, a_k\}$ are the same. Again, by the logical axioms $\mathsf{Ax} \vdash t_r = p \cdot a_j$.

Similarly, for $a_i \cdot p \cdot a_j$.

(1) Suppose that $t = t_1 \cdot t_2$ and $|t| \cong |p|$. Then by Proposition 14,

$$\begin{aligned} |t_1| &\cong [\![a_1, \ldots, a_k]\!]^\eta \langle a_\ell \rangle \quad \text{and} \\ |t_2| &\cong \langle a_m \rangle [\![a_1, \ldots, a_k]\!]^\eta, \end{aligned}$$

where not both a_ℓ and a_m occur. Then, by induction and the associativity and concatenation/shuffle axioms,

$$\begin{aligned} \mathsf{Ax} &\vdash t_1 = [\![a_1, \ldots, a_k]\!]^\eta \langle a_\ell \rangle \\ \mathsf{Ax} &\vdash t_2 = \langle a_m \rangle [\![a_1, \ldots, a_k]\!]^\eta \\ \mathsf{Ax} &\vdash t_1 \cdot t_2 = [\![a_1, \ldots, a_k]\!]^\eta. \end{aligned}$$

Similarly, if $|t| \cong |a_i \cdot p|$ or $|t \cong |p \cdot a_j|$, or $|t| \cong |a_i \cdot p \cdot a_j|$.

(2) If $t = t_1^\omega$, and $|t| \cong |p|$, then, by Proposition 14, either

$$|t_1| \cong [\![a_1,\ldots,a_k]\!]^\eta, \text{ or}$$
$$|t_1| \cong [\![a_1,\ldots,a_k]\!]^\eta a_\ell,$$

for some $\ell \in [p]$. Thus, by induction,

$$\mathsf{Ax} \vdash t_1 = [\![a_1,\ldots,a_k]\!]^\eta,$$

or

$$\mathsf{Ax} \vdash t_1 = [\![a_1,\ldots,a_k]\!]^\eta a_\ell,$$

Thus, in either case, by the omega/shuffle axioms,

$$\mathsf{Ax} \vdash t_1^\omega = [\![a_1,\ldots,a_k]\!]^\eta.$$

Similarly, when $|t| \cong |a_\ell \cdot p|$. By Proposition 14, $|t|$ is not isomorphic to $|a_m \cdot p \cdot a_\ell|$. Note: $|t|$ cannot be $|p \cdot a_i|$.

(3) Similarly, when $t = t_1^{\omega^{op}}$, and $|t| \cong |p|$ or $|t| \cong |p \cdot a_i|$.

(4) Last, if $t = [\![t_1,\ldots,t_n]\!]^\eta$, there are two cases, by Proposition 14. Either each of $t_1,\ldots,t_n$ is a letter in A or not. If so, the logical axioms imply $\mathsf{Ax} \vdash t = [\![a_1,\ldots,a_k]\!]^\eta$. Otherwise, the subterms $t_1,\ldots,t_n$ can be divided into two sets $u_1,\ldots,u_{s_1},v_1,\ldots,v_{s_2}$, $s_1 \geq 0$, $s_2 > 0$, where the terms $u_1,\ldots$ are letters in $\{a_1,\ldots,a_k\}$, and each $v_{j'}$, $j' \in [s_2]$, satisfies

$$|v_{j'}| \cong [\![a_1,\ldots,a_k]\!]^\eta, \quad \text{or}$$
$$|v_{j'}| \cong a_\ell[\![a_1,\ldots,a_k]\!]^\eta, \quad \text{or}$$
$$|v_{j'}| \cong [\![a_1,\ldots,a_k]\!]^\eta a_\ell, \quad \text{or}$$
$$|v_{j'}| \cong a_\ell[\![a_1,\ldots,a_k]\!]^\eta a_m,$$

for some $\ell, m \in [p]$. Hence, induction and the shuffle/shuffle axioms imply that

$$\mathsf{Ax} \vdash [\![t_1,\ldots,t_n]\!]^\eta = [\![a_1,\ldots,a_k]\!]^\eta.$$

The determination of which case applies and how to produce the desired primitive term p may be carried out in quadratic time. It is clear that the size of p is linear in the size of t. □

Corollary 59 *There is a quadratic time algorithm which, given a term t determines whether $|t|$ has finitely many blocks.*

Proof. $|t|$ has finitely many blocks iff either $|t|$ has a single block, which can be determined in quadratic time by the Single block lemma, or if $t = t_1 \cdot t_2 \cdots t_n$, $n \geq 2$, where no t_i can be written as a product of two terms, and each of the words $|t_i|$ has finitely many blocks.

When $t = t_1^\omega$ or $t = t_1^{\omega^{op}}$ or $t = [\![t_1,\ldots,t_k]\!]^\eta$, and $|t|$ has more than one block, then $|t|$ has infinitely many blocks. □

9 Proper Terms

Definition 60 *We let $D = D(A)$ be a new alphabet containing a letter, written $\langle\!|t|\!\rangle$, for each primitive term t on A in normal form. We let σ denote the term morphism from terms on $D(A)$ to terms on A determined by the function*

$$\langle\!|t|\!\rangle \mapsto t.$$

Definition 61 (The condensation) *Suppose that u is a word on A all of whose blocks are primitive words. We define the word*

$$\widehat{u}$$

as the word on $D(A)$ whose underlying linear order is $(\mathsf{Bl}(u), \leq)$ (Definition 23). For $p \in L_u$, the label of the point $\mathsf{Bl}(p)$ is $\langle\!|v|\!\rangle$, where v is the primitive term in normal form denoting the subword $\mathsf{Bl}(p)$ of u.

We note that condensations are preserved and reflected by isomorphisms.

Proposition 62 *If u, v are words on A all of whose blocks are primitive words, then*

$$u \cong v \iff \widehat{u} \cong \widehat{v}.$$

Proof sketch. If $\varphi : u \to v$ is an isomorphism, define $\widehat{\varphi} : L_{\widehat{u}} \to L_{\widehat{v}}$ by:

$$\widehat{\varphi}(\mathsf{Bl}(p)) := \mathsf{Bl}(\varphi(p)),$$

for $p \in L_u$. It is easy to see that $\widehat{\varphi}$ is an isomorphism. Conversely, we may extend any isomorphism $\widehat{u} \to \widehat{v}$ to an isomorphism $u \to v$ by substituting the word $|t|$ for each occurrence of the letter $\langle\!|t|\!\rangle$ in $\widehat{u}, \widehat{v}$. □

Definition 63 (Proper term) *A term t **on** $D(A)$ is proper if one of the following cases holds.*

- *t is a single letter.*
- *t is $t_1 \cdots t_n$, $n \geq 2$, each t_i is proper and cannot be written as the product of two terms, and, most importantly, for each $i < n$, $\sigma(t_i), \sigma(t_{i+1})$ do not merge, and for each $i < n-1$, $\sigma(t_i), \sigma(t_{i+1}), \sigma(t_{i+2})$ do not merge.*
- *$t = (t_1)^{\omega}$, and $t_1 \cdot t_1$ is proper.*
- *$t = (t_1)^{\omega^{op}}$, and $t_1 \cdot t_1$ is proper.*
- *$t = [\![t_1, \ldots, t_k]\!]^{\eta}$, and each of $t_1, \ldots, t_k$ is proper, and, furthermore, the subterms $t_1, \ldots, t_k$ **cannot** be divided into two sets $u_1, \ldots, u_n$, $v_1, \ldots, v_m$, $n \geq 0$, $m > 0$, where there are distinct letters $a_1, \ldots, a_p$ in A such that the terms $\sigma(u_i)$ are letters in $\{a_1, \ldots, a_p\}$, and each term $\sigma(v_j)$ is $[\![a_1, \ldots, a_p]\!]^{\eta}$, $a_i[\![a_1, \ldots, a_p]\!]^{\eta}$, $[\![a_1, \ldots, a_p]\!]^{\eta}a_j$, or $a_i[\![a_1, \ldots, a_p]\!]^{\eta}a_j$, for some $i, j \in [p]$.*

*An **extended proper term** is either* **1** *or a proper term.*

Recall the definition of the size of a term on A from Definition 35, part 6. For a term t in $D(A)$ we define the **size** differently:

$$\mathsf{size}(t) := \mathsf{size}(p) \text{ if } t = \langle\!|p|\!\rangle \in D(A), \tag{18}$$

and otherwise, as in Definition 35, part 6.

Thus, if $a \in A$, the term $(a^\omega)^\omega$ on A, and the term $(\langle\!|a^\omega|\!\rangle)^\omega$ on $D(A)$ and the term $\langle\!|\langle\!|a^\omega|\!\rangle^\omega|\!\rangle$ on $D(D(A))$ all have size 3.

In this section, we will prove the following theorem.

Theorem 64 *There is a quadratic time algorithm which, given a term t on A, produces a proper term s on $D(A)$ with the following properties:*

- $\mathsf{Ax} \vdash t = \sigma(s)$.
- *If* $\mathsf{ht}(t) > 0$ *, then* $\mathsf{ht}(s) < \mathsf{ht}(t)$*; otherwise* $\mathsf{ht}(s) = \mathsf{ht}(t) = 0$.

We will say more about the size of s below in Corollary 69.

For a given term t on A we must produce a proper term s on $D(A)$ such that $\sigma(s)$ and t denote the same word and this fact is provable from the axioms. In order to do this, since pairs or triples of blocks may merge, we need names (proper terms) from which we can find primitive terms denoting the first, second, last and next to last blocks of a word, when these blocks exist, as well as the remaining 'middle' blocks.

Thus, we restate the theorem as follows.

Theorem 65 *There is a quadratic time algorithm which, given a term t on A, produces a proper term s on $D(A)$ and a sequence of one of two forms: when $|t|$ has $k \leq 4$ blocks,* $\mathsf{seq}(t) = (b_1, \ldots, b_k)$*, or* $\mathsf{seq}(t) = (l_1, l_2, m, r_2, r_1)$*, when $|t|$ has more than 4 blocks. The term s and the sequences have the following properties:*

- *b_i are letters in $D(A)$.*
- *l_i, r_i, $i = 1, 2$ are either letters in $D(A)$ or $\mathbf{1}$, and m is a proper term on $D(A)$.*
- *If* $\mathsf{seq}(t) = (b_1, \ldots, b_k)$*, $k \leq 4$, then $s = b_1 \cdots b_k$.*
- *If* $\mathsf{seq}(t) = (l_1, l_2, m, r_2, r_1)$*, then $s = l_1 \cdot l_2 \cdot m_1 \cdots m_k \cdot r_2 \cdot r_1$, where $m = m_1 \cdots m_k$ and none of the m_i can be written as the product of two terms and we understand that $\mathbf{1} \cdot x = x \cdot \mathbf{1} = x$, for any term x. (Thus, there are no occurrences of $\mathbf{1}$ in s.)*
- *$l_1 \neq \mathbf{1}$ iff $|t|$ has a first block; $l_2 \neq \mathbf{1}$ iff $|t|$ has a second block; $r_1 \neq \mathbf{1}$ iff $|t|$ has a last block, and $r_2 \neq \mathbf{1}$ iff $|t|$ has a next to last block.*
- $\mathsf{Ax} \vdash t = \sigma(s)$.
- *If* $\mathsf{ht}(t) > 0$ *, then* $\mathsf{ht}(s) < \mathsf{ht}(t)$*, otherwise* $\mathsf{ht}(s) = \mathsf{ht}(t)$.

For each term t on A we will show how to find the sequence $(b_1, \ldots, b_k)$, or (l_1, l_2, m, r_2, r_1), since s may be obtained immediately from it.

The sequences $\mathsf{seq}(t)$ will have the following properties.

- If $|t|$ has at most 4 blocks, denoted in order by the primitive terms $b_1, \ldots, b_k$ in normal form, then $\mathsf{seq}(t) = (\langle\!| b_1 |\!\rangle, \ldots, \langle\!| b_k |\!\rangle)$.

- If $|t|$ has at least 5 but finitely many blocks, $\mathsf{seq}(t) = \langle\!| l_1 |\!\rangle, \langle\!| l_2 |\!\rangle, m, \langle\!| r_2 |\!\rangle, \langle\!| r_1 |\!\rangle)$, where l_1, l_2, r_2, r_1 are letters in $D(A)$ and m is a proper term which is a product of letters.
- If $|t|$ has infinitely many blocks, then l_1 is a single letter iff u has a first block, l_2 is a single letter iff u has a second block, r_1 is a letter iff u has a last block, and r_2 is a letter iff u has a next to last block. Of course, m is a proper term such that $|\sigma(m)|$ has infinitely many blocks.
- The word $|\sigma(m)|$ has no first block if l_1 or l_2 is $\mathbf{1}$; similarly, $|\sigma(m)|$ has no last block if r_1 or r_2 is $\mathbf{1}$. If $l_1 = \mathbf{1}$ then $l_2 = \mathbf{1}$; similarly, if $r_1 = \mathbf{1}$ then $r_2 = \mathbf{1}$.

We will say that terms l, m or l, m, r on $D(A)$ merge if $\sigma(l), \sigma(m)$ or $\sigma(l), \sigma(m), \sigma(r)$ merge.

Notation: If $\langle\!| b |\!\rangle, \langle\!| c |\!\rangle$ are letters in $D(A)$ such that the primitive terms b, c on A merge, then let

$$b \star c := \langle\!| p |\!\rangle,$$

where p is the primitive term in normal form such that $\mathsf{Ax} \vdash b \cdot c = p$. Similarly, if $\langle\!| b |\!\rangle, \langle\!| c |\!\rangle, \langle\!| d |\!\rangle$ are letters in $D(A)$ such that the primitive terms b, c, d merge, then $b \star c \star d$ is $\langle\!| p |\!\rangle$, where p is the primitive term in normal form such that $\mathsf{Ax} \vdash b \cdot c \cdot d = p$.

Remark 66 *The terms $b \star c$, $b \star c \star d$ exist, by the Single block lemma, and can be found in time $O(n^2)$, where n is the size of $b \cdot c$, or $b \cdot c \cdot d$, respectively. Moreover, the size of $b \star c$ and $b \star c \star d$ is at most n.*

We now proceed to the *proof of Theorem 65*. There are five cases, depending on t.

Case 1. If the term t is such that $|t|$ has a single block, let p be the term in normal form such that $\mathsf{Ax} \vdash t = p$. In this case, $\mathsf{seq}(t)$ is a sequence of length one:

$$\mathsf{seq}(t) = (\langle\!| p |\!\rangle).$$

In the remaining cases, we assume $|t|$ has at least two blocks.

Case 2. Assume that $t = t' \cdot t''$. By induction, we assume that for t' and t'' we have found the proper terms s', s'' and the sequences $\mathsf{seq}(t')$, $\mathsf{seq}(t'')$ satisfying the properties of the theorem.

When $\mathsf{seq}(t') = (b_1, \ldots, b_k)$ and $\mathsf{seq}(t'') = (c_1, \ldots, c_j)$, where $k, j \leq 4$, we consider the sequence

$$(b_1, \ldots, b_k, c_1, \ldots c_j).$$

If b_k, c_1 merge, we replace the two letters b_k, c_1 by $b_k \star c_1$. If b_{k-1} exists and b_{k-1}, b_k, c_1 merge, we replace these three letters by $b_{k-1} \star b_k \star c_1$. Last, if c_2 exists and b_k, c_1, c_2 merge, we replace these three letters by $b_k \star c_1 \star c_2$. These three cases are mutually exclusive. Denote the resulting sequence by θ. If θ has length at most 5, we define $\mathsf{seq}(t)$ as this sequence. If the length of θ is more than 5, we define $\mathsf{seq}(t)$ as the 5 element sequence (l_1, l_2, m, r_2, r_1), where l_i is the i-th element of θ, r_1 is the last letter in θ and r_2 is the next to last letter in θ; m is the product of the remaining letters in θ. The fact that the term s determined by the resulting sequence is proper follows, using the First block concatenation corollary, Corollary 30.

There is a similar construction when just one of $\mathsf{seq}(t')$, $\mathsf{seq}(t'')$ has length less than 5.

Now, assume we have

$$\begin{aligned}\mathsf{seq}(t') &= (l'_1, l'_2, m', r'_2, r'_1)\\ \mathsf{seq}(t'') &= (l''_1, l''_2, m'', r''_2, r''_1)\end{aligned}$$

satisfying the hypotheses, such that $\mathsf{Ax} \vdash t' = \sigma(s')$ and $\mathsf{Ax} \vdash t'' = \sigma(s'')$.

We would like to define the sequence $\mathsf{seq}(t) = (l_1, l_2, m, r_2, r_1)$ by:

$$\begin{aligned}l_1 &= l'_1\\ l_2 &= l'_2\\ m &= m'_1 \cdots m'_{k'} \cdot r'_2 \cdot r'_1 \cdot l''_1 \cdot l''_2 \cdot m''_1 \cdots m''_{k''}\\ r_2 &= r''_2\\ r_1 &= r''_1,\end{aligned}$$

where $m' = m'_1 \cdots m'_{k'}$, $m'' = m''_1 \cdots m''_{k''}$ and none of the terms m'_i and m''_j can be written as the product of two terms. However, m need not be proper, due to the possible merging in the product $r'_2 \cdot r'_1 \cdot l''_1 \cdot l''_2$. Thus, if r'_1, l''_1 merge, we replace $r'_1 \cdot l''_1$ by $r'_1 \star l''_1$; similarly, if r'_2, r'_1, l''_1 merge, we replace their product by $r'_2 \star r'_1 \star l''_1$; if r'_1, l''_1, l''_2 merge, we replace their product by $r'_1 \star l''_1 \star l''_2$. We also must eliminate redundant occurrences of $\mathbf{1}$. The fact that the term s determined by the resulting sequence is proper again follows from the First block concatenation corollary, Corollary 30.

For example, if

$$\begin{aligned}\mathsf{seq}(t') &= (\llparenthesis a^\eta \rrparenthesis)\\ \mathsf{seq}(t'') &= (\llparenthesis a^\eta \rrparenthesis, \llparenthesis b^\eta \rrparenthesis)\end{aligned}$$

then, since $t' \cdot t''$ has two blocks, we define $\mathsf{seq}(t)$ as

$$\mathsf{seq}(t' \cdot t'') = (\llparenthesis a^\eta \rrparenthesis, \llparenthesis b^\eta \rrparenthesis).$$

As a second example, suppose

$$\begin{aligned}\mathsf{seq}(t') &= (\llparenthesis c \rrparenthesis, \llparenthesis d \rrparenthesis,\ \llparenthesis a^\omega \rrparenthesis^\omega, \llparenthesis b^\eta \rrparenthesis, \llparenthesis b \rrparenthesis)\\ \mathsf{seq}(t'') &= (\llparenthesis b^\eta \rrparenthesis, \llparenthesis a^{\omega^{op}} \rrparenthesis, \llparenthesis b^{\omega^{op}} \rrparenthesis).\end{aligned}$$

After merging the last two letters of $\mathsf{seq}(t')$ with the first letter of $\mathsf{seq}(t'')$, we see that the corresponding sequence $\mathsf{seq}(t)$ is

$$\mathsf{seq}(t' \cdot t'') = (\llparenthesis c \rrparenthesis, \llparenthesis d \rrparenthesis,\ (\llparenthesis a^{\omega} \rrparenthesis^{\omega} \cdot \llparenthesis b^{\eta} \rrparenthesis)\ ,\ \llparenthesis a^{\omega^{op}} \rrparenthesis, \llparenthesis b^{\omega^{op}} \rrparenthesis),$$

since there is no more merging.

Note that the size of s is not greater than the size of $s' \cdot s''$.

Case 3. Suppose that $t = [\![t_1, \ldots, t_n]\!]^{\eta}$ and assume $|t|$ has more than one block. Then, if s_i are proper terms such that $\mathsf{Ax} \vdash t_i = \sigma(s_i)$, we know from Proposition 33 that $[\![s_1, \ldots, s_n]\!]^{\eta}$ is proper. Thus, we let

$$\mathsf{seq}(t) = (\mathbf{1}, \mathbf{1},\ [\![s_1, \ldots, s_n]\!]^{\eta},\ \mathbf{1}, \mathbf{1}).$$

In this case, the size of s is at most the size of $[\![s_1, \ldots, s_n]\!]^{\eta}$.

Case 4. Next, we assume $t = (t')^{\omega}$. By induction, we have either for some k, $1 \leq k \leq 4$,

$$\mathsf{seq}(t') = (b_1, \ldots, b_k),$$

or

$$\mathsf{seq}(t') = (l'_1, l'_2, m', r'_2, r'_1).$$

In the first case, either $k = 1$ or $2 \leq k \leq 4$. When $k = 1$, since we know $|t|$ has more than one block, the first two blocks of $|t|$ are isomorphic to $|\sigma(b_1)|$. Thus, we define

$$\mathsf{seq}(t) = (b_1, b_1, b_1^{\omega}, \mathbf{1}, \mathbf{1}).$$

If $k = 2$ and b_2, b_1 do not merge, we define $\mathsf{seq}(t) = (b_1, b_2, (b_1 \cdot b_2)^{\omega}, \mathbf{1}, \mathbf{1})$. However, if b_2 and b_1 merge, we let

$$\mathsf{seq}(t) = (b_1, b_2 \star b_1, (b_2 \star b_1)^{\omega}, \mathbf{1}, \mathbf{1}).$$

It is not possible that b_1, b_2, b_1 or b_2, b_1, b_2 merge, or $|t|$ will have only one block.

If $k > 2$, we first consider whether $b_3 \cdots b_k \cdot b_1 \cdot b_2$ has any consecutive triple, or pair of letters that merge; if so, we replace that triple or pair by the corresponding 'star' product. Let c denote the resulting term. We define

$$\mathsf{seq}(t) = (b_1, b_2, c^{\omega}, \mathbf{1}, \mathbf{1}).$$

The fact that $\mathsf{Ax} \vdash t = \sigma(s)$ in this case follows from Axiom (4).

In the case that $\mathsf{seq}(t') = (l'_1, l'_2, m', r'_2, r'_1)$, $|t|$ has a first and second block iff $|t'|$ does, and they are the same. We would like to define $\mathsf{seq}(t)$ by:

$$\mathsf{seq}(t) = (l_1', l_2',\ (m_1' \cdots m_k' \cdot r_2' \cdot r_1' \cdot l_1' \cdot l_2')^{\omega}\ ,\ \mathbf{1}, \mathbf{1}),$$

where $m' = m_1' \cdots m_k'$ and none of the m_i' can be written as the product of two terms, since $|t|$ has no last or next to last block. However, the term $r_2' \cdot r_1' \cdot l_1' \cdot l_2'$ may not be proper. We replace any consecutive pairs or triples that merge by their "star" product, and continue as above. The same axiom shows that $\mathsf{Ax} \vdash t = \sigma(s)$. In either case, it follows using the First and Second block concatenation corollaries, Corollaries 30, 31 and the General block concatenation corollary, Corollary 32, that the resulting sequence determines a proper term s.

If s' is the proper term for t', the height of s' is **at most** one less than the height of t', and the size of s is at most the size of $(s')^{\omega}$, plus, perhaps, the sizes of the *letters in* $D(A)$ denoting the first and second blocks of $|t|$ - thus an upper bound on the size of s is *at most 3 times the size of* $(s')^{\omega}$. Note also that no new omega term of size larger than $\mathsf{size}(t)$ is created.

Case 5. The case that $t = (t')^{\omega^{op}}$ is similar, so that if s is the proper term for t, the size of s is at most the size of $(s')^{\omega^{op}}$, plus, perhaps, the sizes of the *letters in* $D(A)$ denoting the last and next to last blocks of $|t|$ - thus an upper bound on the size of s is *at most 3 times the size of* $(s')^{\omega^{op}}$. No new omega-op term of size larger than $\mathsf{size}(t)$ is created.

This completes the proof. $\square$

We note an important property of this construction.

Lemma 67 *Suppose that t is a term on A and s is the proper term on $D(A)$ constructed in Theorem 65. Then the size of any letter that occurs in s and the size of any subterm of s of the form u^{ω} or $u^{\omega^{op}}$ is at most the size of t.*

Proof by induction on t. When $|t|$ has a single block, s is a single letter $\langle\!\langle p \rangle\!\rangle$, where $\mathsf{size}(p) \leq \mathsf{size}(t)$.

If $t = t_1 \cdot t_2$, or $t = [\![t_1, \ldots, t_n]\!]^{\eta}$, the statement follows by induction and Cases 2 and 3 of the proof of Theorem 65.

If t is $(t')^{\omega}$, the largest s can be is $l_1 \cdot l_2 \cdot m^{\omega}$, where l_1, l_2 and m are, respectively, letters and a product of terms that occur in the proper term for t'. It is clear by induction that the size of any letter in s is at most $\mathsf{size}(t') < \mathsf{size}(t)$, and by Case 4, $\mathsf{size}(m) < \mathsf{size}(t)$.

The case that $t = (t')^{\omega^{op}}$ is similar, and is omitted. $\square$

Corollary 68 *Suppose that t is a term on A and s is the proper term on $D(A)$ obtained in Theorem 65. If k size expanding rules (Cases 4 or 5) were applied in order to obtain s, then*

$$\mathsf{size}(s) \leq (2k+1)\mathsf{size}(t). \quad \square$$

Corollary 69 *Suppose that $t = t_0$ is a term on A of height $h > 0$, and size m. Let*

t_1 be the proper term on $D(A)$ with $\mathsf{Ax} \vdash t_0 = \sigma(t_1)$. Similarly, if $i-1 \geq 0$ and t_{i-1} has positive height, let t_i be the proper term on $D^i(A)$ with $\mathsf{Ax} \vdash t_{i-1} = \sigma(t_i)$. If $|t_0|$ has one block, then t_1 is a letter and $\mathsf{size}(t_1) \leq \mathsf{size}(t_0)$. Suppose t_i has been defined for $i \leq h'$, where $h' \leq h$. Then,

$$\mathsf{size}(t_{h'}) \leq (2m+1) \cdot \mathsf{size}(t_0).$$

Proof. We can apply the size expanding rules for t^ω and $t^{\omega^{op}}$ at most m times, one for each occurrence of a subterm of the form u^ω or $u^{\omega^{op}}$, and each time, we might increase the size of the resulting term by the size of two letters, each of size at most $\mathsf{size}(t_0)$, by Lemma 67. Thus, if we apply the expanding rules k_1-times to obtain t_1, $\mathsf{size}(t_1) \leq 2k_1\mathsf{size}(t_0) + \mathsf{size}(t_0)$, and if $v = u^\omega$ or $v = u^{\omega^{op}}$ is a subterm of t_1, then $\mathsf{size}(v) \leq \mathsf{size}(t_0)$, by Cases 4 and 5 of Theorem 65. If we apply the expanding rules k_2-times to obtain t_2,

$$\begin{aligned}\mathsf{size}(t_2) &\leq 2k_2\mathsf{size}(t_0) + \mathsf{size}(t_1)\\ &\leq 2(k_2+k_1)\mathsf{size}(t_0) + \mathsf{size}(t_0).\end{aligned}$$

Similarly, if we apply the expanding rules k_i-times to obtain t_i,

$$\mathsf{size}(t_{h'}) \leq 2(k_{h'} + \ldots + k_1)\mathsf{size}(t_0) + \mathsf{size}(t_0),$$

and $\sum_{i=1}^{h'} k_i \leq m$. □

As an example, suppose that t_0 is $\llbracket (a^\omega)^\omega, (b^{\omega^{op}})^{\omega^{op}} \rrbracket^\eta$. Then

$$\begin{aligned}t_1 &= \llbracket \langle\!| a^\omega |\!\rangle \cdot \langle\!| a^\omega |\!\rangle \cdot \langle\!| a^\omega |\!\rangle^\omega, \langle\!| b^{\omega^{op}} |\!\rangle^{\omega^{op}} \cdot \langle\!| b^{\omega^{op}} |\!\rangle \cdot \langle\!| b^{\omega^{op}} |\!\rangle \rrbracket^\eta\\ t_2 &= \llbracket \langle\!| c^\omega |\!\rangle, \langle\!| d^{\omega^{op}} |\!\rangle \rrbracket^\eta, \qquad \text{where } c = \langle\!| a^\omega |\!\rangle,\ d = \langle\!| b^{\omega^{op}} |\!\rangle\\ t_3 &= \langle\!| t_2 |\!\rangle\\ &= \langle\!| \llbracket \langle\!| c^\omega |\!\rangle, \langle\!| d^{\omega^{op}} |\!\rangle \rrbracket^\eta |\!\rangle.\end{aligned}$$

So $\mathsf{size}(t_1) < 3\mathsf{size}(t_0)$, and $\mathsf{size}(t_2) = \mathsf{size}(t_3) = \mathsf{size}(t_0)$!

As yet another consequence of Theorem 64, we easily obtain the following fact, which will be used below.

Corollary 70 *Each block of a regular word on the set A is isomorphic to a word denoted by some primitive term on A.*

Proof. Since each regular word is isomorphic to $|\sigma(t)|$, for some proper term t on $D(A)$, we use induction on the structure of proper terms. The basis step is: t is a single letter, so that $\sigma(t)$ is primitive and $|\sigma(t)|$ has just one block.

- Assume that $t = t_1 \cdots t_n$. Since the terms $t_1, t_2, \ldots, t_n$ are proper, no two adjacent pairs or triples merge, it follows that each block of $|\sigma(t)|$ is a block of some word $|\sigma(t_i)|$, and thus, is named by a primitive term.

- Now suppose that $t = (t')^{\omega}$. Since t is proper, the blocks of $|\sigma(t)|$ are copies of those of $|\sigma(t')|$, which, by induction, are named by primitive terms.
- The case $t = (t')^{\omega^{op}}$ is similar.
- Finally, if $t = [\![t_1, \ldots, t_n]\!]^{\eta}$, since $|\sigma(t)|$ has at least two blocks, the blocks of $|\sigma(t)|$ are those of the words $|\sigma(t_i)|$, which are named by primitive terms, by induction.

This completes the proof. □

Example 71 *The word whose underlying order is ω whose labeling is*

$$abaaba^3b\cdots ba^nb\cdots$$

has one block, but it is not denoted by a primitive term.

Proposition 72 *Suppose that u is a word on A. Then u is regular iff $\widehat{u}$ exists and is regular.*

Proof. First, suppose that u is regular. Then, each block of u is a primitive word, by Corollary 70, so $\widehat{u}$ exists. We have proved that there is a proper term t with $u = |\sigma(t)|$ and t denotes $\widehat{u}$, so $\widehat{u}$ is regular.

Conversely, if $\widehat{u}$ exists and is regular, $\mathbf{alph}(\widehat{u})$ is finite. Let $\mathbf{alph}(\widehat{u}) = \{\langle\!\langle v_1\rangle\!\rangle, \ldots, \langle\!\langle v_k\rangle\!\rangle\}$, where $v_1, \ldots, v_k$ are primitive words on A. We have

$$u \cong \widehat{u}(\langle\!\langle v_1\rangle\!\rangle/v_1, \ldots, \langle\!\langle v_k\rangle\!\rangle/v_k),$$

so that by Proposition 40, since $\widehat{u}$ is regular, so is u. □

Corollary 73 *If u, v are regular words on A, then*

$$u \cong v \iff \widehat{u} \cong \widehat{v}.$$

Proof. By Proposition 62 and Proposition 72. □

Proper terms and the condensation map interact nicely.

Proposition 74 *Let t be a proper term on $D(A)$. Then*

$$\widehat{|\sigma(t)|} = \begin{cases} \widehat{|\sigma(t_1)|}\ \cdots\ \widehat{|\sigma(t_n)|} & \text{if } t = t_1 \cdots t_n \\ (\ \widehat{|\sigma(t_1)|}\ )^{\omega} & \text{if } t = (t_1)^{\omega} \\ (\ \widehat{|\sigma(t_1)|}\ )^{\omega^{op}} & \text{if } t = (t_1)^{\omega^{op}} \\ [\![\ \widehat{|t_1|}, \ldots, \widehat{|t_n|}\ ]\!]^{\eta} & \text{if } t = [\![t_1, \ldots, t_n]\!]^{\eta}. \end{cases}$$

Proof. We prove only the first case. When t is $t_1 \cdots t_n$, $n \geq 2$, then since each term t_i is proper and since no two or three consecutive terms merge, the blocks of $|\sigma(t)|$ are the blocks in each word $|\sigma(t_i)|$, by the General block concatenation corollary, Corollary 32. Thus,

$$\widehat{\sigma(t)} = \widehat{\sigma(t_1)} \cdots \widehat{\sigma(t_n)}\,.$$

since no two or three consecutive terms merge. Thus,

$$\widehat{|\sigma(t)|} = \widehat{|\sigma(t_1)|} \cdots \widehat{|\sigma(t_n)|}\,. \quad \square$$

Corollary 75 *If the term t on $D(A)$ is proper, then the words $|t|$ and $\widehat{|\sigma(t)|}$ are the same:*

$$|t| = \widehat{|\sigma(t)|}\,. \tag{19}$$

Proof. We use induction on the structure of t to prove equation (19).

When t is $\langle\!|s|\!\rangle$, where $s \in D(A)$, then $\sigma(t) = s$. Thus, $\widehat{|s|} = t = |t|$.

If $t = t_1 \cdots t_n$, $n \geq 2$, then

$$\sigma(t) = \sigma(t_1) \cdots \sigma(t_n).$$

Since each term t_i is proper and each block of $\sigma(t)$ is a block of some $\sigma(t_i)$,

$$\begin{aligned} |\,\widehat{\sigma(t)}\,| &= \widehat{|\sigma(t_1)|} \cdots \widehat{|\sigma(t_n)|} \\ &= |t_1| \cdots |t_n|, \quad \text{by induction} \\ &= |t|. \end{aligned}$$

In the cases that $t = t_1^{\omega}$ or $t = t_1^{\omega^{op}}$, the argument is the same, using the General block concatenation corollary, since if t_1, t_1 do not merge, then t_1, t_1, t_1 do not merge.

Last, if $t = [\![t_1, \ldots, t_k]\!]^{\eta}$, and each of $t_1, \ldots, t_k$ is proper, and the above assumption holds, then by Proposition 33, the blocks of $|\sigma(t)|$ are those of $|\sigma(t_i)|$, showing again that

$$|t| = \widehat{|\sigma(t)|}\,.$$

This concludes the proof. $\square$

10 The completeness theorem

We can now prove the Completeness theorem.

Theorem 76 (Completeness theorem) *For terms s, t on the set A,*

$$|s| \cong |t| \iff \mathsf{Ax} \vdash s = t.$$

Proof. Since we have already dealt with the soundness of the axioms, Proposition 54, we prove only completeness. Assume that $|s| \cong |t|$. We will prove that $\mathsf{Ax} \vdash s = t$ by induction on $h = \max\{\mathsf{ht}(s), \mathsf{ht}(t)\}$.

If $h = 0$, then both s, t are finite terms, and by the associativity axiom, Axiom (3), $\mathsf{Ax} \vdash s_1 = t_1$.

If $h > 0$, then neither term is finite. Thus, by Theorem 64, there are proper terms s_1, t_1 on $D(A)$ such that $\mathsf{ht}(s_1) < \mathsf{ht}(s)$, $\mathsf{ht}(t_1) < \mathsf{ht}(t)$, and

$$\mathsf{Ax} \vdash s = \sigma(s_1), \quad \text{and} \tag{20}$$
$$\mathsf{Ax} \vdash t = \sigma(t_1), \tag{21}$$

so that, by Corollary 75, and Proposition 62,

$$|s_1| \cong |t_1|.$$

Since the maximum of $\mathsf{ht}(s_1), \mathsf{ht}(t_1)$ is strictly less than h,

$$\mathsf{Ax} \vdash s_1 = t_1,$$

by the induction hypothesis. Thus,

$$\mathsf{Ax} \vdash \sigma(s_1) = \sigma(t_1), \tag{22}$$

by Proposition 55. Lastly, by (20), (21), and (22), it follows that $\mathsf{Ax} \vdash s = t$. $\square$

We end this section with a structural characterization of the regular words. Recall that $\widehat{u}$ is defined for the word u when each block of u is denoted by a primitive term. Also, note that if u is a single letter, so is $\widehat{u}$.

Suppose that u is a word on A. For the purposes of the next theorem, define the finite or infinite sequence of words $u(0), u(1), \ldots$ as follows.

$$u(k) := \begin{cases} u & k = 0 \\ \widehat{u(k-1)} & \text{if } k > 0 \text{ and } \widehat{u(k-1)} \text{ is defined.} \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{23}$$

Theorem 77 *A word u is regular iff there is a positive integer k such that $u(j)$ is defined for all $0 \leq j \leq k$ and $u(k)$ is a single letter.*

Proof. Assume u is regular. If u consists of a single block, or has only finitely many blocks, $u(1) = \widehat{u}$ is a finite product of letters in $D(A)$, and $u(2)$ is a single letter. Otherwise, since u is regular, there is a proper term t on $D(A)$ such that u is denoted by $\sigma(t)$, and by Corollary 75, $v = \widehat{u}$ is regular and is denoted by t, and the height of t is one less than the height of $\sigma(t)$. Thus, by induction, for some positive k, $v(k)$ is a single letter. But $v(k) = u(k + 1)$.

Conversely, using Proposition 72, if $u(k+1) = \widehat{u(k)}$ is regular, so is $u(k)$. $\square$

In the next Corollary, we identify each linear order with a word on a one-letter alphabet, so that $L(k)$ is meaningful, for linear orders L. Recall the definition of the class M in Remark 39.

Corollary 78 *A nonempty linear order L belongs to M iff $L(k)$ is a single letter, for some nonnegative integer k.* $\square$

11 The decision algorithm

Consider the following problem. *Given two terms s, t on A, is $|s| \cong |t|$?*

The version of this problem for scattered terms was raised by Courcelle [Cour78], and the general question was posed in [Heil80]. We recall that Thomas [Th86] showed that this problem is decidable, using Rabin tree automata. However, his method did not provide an explicit upper bound on the complexity of this problem. We show that the method used to prove the completeness theorem using the condensation in Definition 61 gives a polynomial upper bound.

Theorem 79 *There is an $O(n^5)$-algorithm to decide, given terms s, t on A whether $|s| \cong |t|$,where n is $\mathsf{size}(s) + \mathsf{size}(t)$.*

Proof sketch.

Here is pseudo-code for the algorithm. The input is a pair s, t of terms on A.

```
$h \leftarrow \max(\mathsf{ht}(s), \mathsf{ht}(t))$;
while ($h > 0$)
{
 if the height of one term is 0,
   return false.
 else
   { find proper terms $s'$, $t'$ using Theorem 9.6
   such that $\mathsf{Ax} \vdash s = \sigma(s')$ and $\mathsf{Ax} \vdash t = \sigma(t')$;
    $s \leftarrow s'$;
    $t \leftarrow t'$;
    $h \leftarrow \max(\mathsf{ht}(s), \mathsf{ht}(t))$;
   }
}
determine whether $|s| \cong |t|$.
```

The algorithm proceeds by induction on $h = \max(\mathsf{ht}(s),\ \mathsf{ht}(t))$. Note that h may be computed in linear time.

We will show that the question of whether $|s| \cong |t|$ can be decided in $O(n)$ steps if $h = 0$, and in $O(h \cdot n^4)$ steps otherwise. Since $h < n$, this fact is sufficient to prove the theorem.

If h is zero, then $|s| \cong |t|$ iff the terms, considered as finite trees, have the same frontier. This may be decided in linear time.

If $h > 0$, but the height of one term is 0, the terms are not equivalent. Otherwise, we repeat the following procedure at most h times until we are considering two terms of height zero.

Let $s_0 = s$, $t_0 = t$. Find proper terms s_i, t_i on $D^i(A)$, $i = 1, \ldots, h'$, such that $|s_{i-1}| \cong |t_{i-1}|$ iff $|s_i| \cong |t_i|$, by Theorem 65. When the maximum height of s_i, t_i is zero we stop.

The size of each s_i, t_i is at most $3n^2$, by Corollary 69. Further, there is a fixed constant C such that the time needed to obtain s_i, t_i from s_{i-1}, t_{i-1}, respectively, is at most Cn^4, by Theorem 65. Since $h < n$, it follows that the algorithm is $O(n^5)$. □

12 Adding 1 and reverse

In order to enrich the above operations with a constant for the empty word and the reverse operation, we add the following axioms Rev.

$$(x^r)^r = x \tag{24}$$
$$(x \cdot y)^r = y^r \cdot x^r \tag{25}$$
$$(x^\omega)^r = (x^r)^{\omega^{op}} \tag{26}$$
$$([\![x_1, \ldots, x_n]\!]^\eta)^r = [\![x_1^r, \ldots, x_n^r]\!]^\eta \tag{27}$$
$$\mathbf{1} \cdot x = x \;=\; x \cdot \mathbf{1} \tag{28}$$
$$(\mathbf{1})^\omega = (\mathbf{1})^{\omega^{op}} \;=\; \mathbf{1} \tag{29}$$
$$\mathbf{1}^r = \mathbf{1} \tag{30}$$
$$[\![\mathbf{1}]\!]^\eta = \mathbf{1} \tag{31}$$
$$[\![\mathbf{1}, x_1, \ldots, x_n]\!]^\eta = [\![x_1, \ldots, x_n]\!]^\eta \tag{32}$$

A term built from letters in the set A using the regular operations, the constant $\mathbf{1}$ and the reverse operation ^{r} are **extended terms**.

Remark 80 *If v is a block of the word u, then the reverse of v, v^r, is a block of u^r - i.e., the blocks of u^r are the reversals of blocks of u.*

As a corollary of Theorem 76, we may obtain this extension.

Theorem 81 *Let s, t be two extended terms on A. Then*

$$|s| \cong |t| \iff \mathsf{Ax} \cup \mathsf{Rev} \vdash s = t.$$

Proof. It is clear that any provable equation is valid. As for completeness, let t be an extended term on A. If $|t| = \mathbf{1}$, we use induction on the number of operation symbols in t to show $\mathsf{Ax} \cup \mathsf{Rev} \vdash t = \mathbf{1}$. Otherwise, we prove that there is a term t' having no occurrences of $\mathbf{1}$ in which the reverse operation is applied only to letters in A such that $\mathsf{Ax} \cup \mathsf{Rev} \vdash t = t'$. Now assume that t, s are extended terms with $|s| \cong |t|$. Obtain the terms s', t' as just described, which we may regard as terms over the larger alphabet $A \cup A^r$, where A^r has a distinct letter a^r for each letter $a \in A$. Since $|s'| \cong |t'|$, by Theorem 76, we have $\mathsf{Ax} \vdash s' = t'$. Thus, $\mathsf{Ax} \cup \mathsf{Rev} \vdash s = t$. □

13 Finite axiomatizability

No finite subset of Ax is complete.

Theorem 82 *For any finite subset E of the axioms enumerated in Section 7, and even the axioms involving the reverse operation ^{r} and the neutral element $\mathbf{1}$ in Section 12, there is some prime number p and an algebra M such that each equation in E is* *true in M, but the power identity $(x^p)^\omega = x^\omega$ fails in M.*

Thus, by the Compactness Theorem, we obtain this fact.

Corollary 83 *There is no finite axiomatization of the variety consisting of the models of* Ax.

Proof of Theorem 82. Let $M = \mathbb{N} \cup \{\mathbf{1}, \top, \bot\}$, the disjoint union of the nonnegative integers with a three element set. Let p be a prime. Define the operations $x \cdot y$ and x^ω on M as follows.

$$x \cdot y := \begin{cases} x + y & \text{if } x, y \in \mathbb{N} \\ x & \text{if } y = \mathbf{1} \\ y & \text{if } x = \mathbf{1} \\ \top & \text{if exactly one of } x, y \text{ is } \top \text{ and the other is in } \mathbb{N} \cup \{\mathbf{1}\} \\ \bot & \text{otherwise.} \end{cases}$$

$$x^\omega := \begin{cases} \mathbf{1} & \text{if } x = \mathbf{1} \\ \top & \text{if } x \in \mathbb{N} \text{ and } p \text{ divides } x \\ \bot & \text{otherwise.} \end{cases}$$

$$x^{\omega^{op}} := x^\omega$$

$$x^r := x.$$

Lastly, define

$$[\![x_1, \ldots, x_n]\!]^\eta := \begin{cases} \mathbf{1} & \text{if all } x_i = \mathbf{1} \\ \bot & \text{otherwise.} \end{cases}$$

By [BlEs03a], everything follows once we show that the shuffle axioms hold. There are always two cases: all arguments are $\mathbf{1}$ or not. If not, both sides of each shuffle axiom is $\bot$; otherwise, both sides are $\mathbf{1}$.

Another way to show that Ax is not equivalent to a finite set of axioms is to show that no finite subset of the shuffle axioms will suffice.

Theorem 84 *For any finite subset F of the shuffle axioms on the set A, there is a model of F and all scattered axioms which is not a model of* Ax.

Proof. Indeed, let the model be the word algebra AW with a new definition of the shuffle operations. For a given positive integer q, define the operation $[x_1, \ldots, x_n]^\eta$ on the word algebra AW by:

$$[x_1, \ldots, x_n]^\eta := \begin{cases} [\![x_1, \ldots, x_n]\!]^\eta & \text{if size of } \{x_1, \ldots, x_n\} \text{ is at most } q \\ x_1 & \text{otherwise.} \end{cases}$$

Then on sets of size at most q, the shuffle axioms hold with $[\,\cdot\,]^\eta$ in place of $[\![\cdot]\!]^\eta$. But, on larger sets, just about every shuffle axiom fails. $\square$

14 Free algebras

Let W be the variety of all algebras $(X, \cdot, {}^\omega, {}^{\omega^{op}}, {}^\eta)$ generated by the word algebras AW. Thus, W is the collection of all models of Ax. Recall that AR is the subalgebra of AW consisting of the regular words on A. In the case that A is a singleton, a word in AW is just a linear order, and $AR = \mathsf{M}$ consists of the regular linear orders, those linear orders generated from a singleton by the regular operations. See Remark 39.

As an immediate corollary of the Completeness Theorem, we obtain the following concrete description of the free algebras in W.

Corollary 85 *For any set A, the algebra AR is freely generated by A in the variety W. In detail, for any algebra $(X, \cdot, {}^\omega, {}^{\omega^{op}}, {}^\eta)$ in W, and any function $f : A \to X$ there is a unique homomorphism $f^\# : AR \to X$ extending f.* $\square$

Proposition 86 *Suppose that $A_n = \{a_1, \ldots, a_n\}$, for $n \geq 1$. Then there is an injective morphism $\varphi : A_nR \to A_1R$.*

Proof. Let φ be the unique morphism determined by the map

$$a_n \mapsto b^{\omega^{op}} b^\omega b^n b^{\omega^{op}} b^\omega,$$

where we write b instead of a_1. We need show only that φ is injective. But we can recover the word u on A_n from the word $\varphi(u)$ by noting that each finite block in $\varphi(u)$ is bracketed by blocks isomorphic to $\mathbb{Z}$, i.e., to $b^{\omega^{op}} b^\omega$, and every block of $\varphi(u)$ is either finite, or is isomorphic to $\mathbb{Z}$. There are no three consecutive $\mathbb{Z}$-blocks - every two consecutive $\mathbb{Z}$-blocks are followed by a finite block. $\square$

Corollary 87 *Let W' be the variety generated by linear orders enriched by the regular operations. Then $\mathsf{W}' = \mathsf{W}$. Thus, the equational theory of W' is decidable in polynomial time.* $\square$

Proof. Since $\mathsf{W}' \subseteq \mathsf{W}$, any equation valid in W is also valid in W'. But since all finitely generated free algebras in W are subalgebras of an algebra in W', any equation valid in W' is also valid in W. The two equational theories are the same. Hence the complexity result in Theorem 79 applies to W'. $\square$

Recall Remark 39.

Corollary 88 *The equational theory of W is the same as that of M.* $\square$

Remark 89 *The first order theory of $<$ for linear orderings is decidable, as first proved in [Ehr59]. Another proof is given in [Ro82]. Thomas [Th86] shows that the equational theory of the regular operations on linear orders is also decidable. Corollary*

87 gives a polynomial upper bound to the complexity of this theory. Also, it follows from the same Corollary that the equational theory of linear orders enriched with the regular operations is not finitely based.

15 Other models

Aside from the word algebras AW and their subalgebras, we note two other classes of models of Ax.

Labeled Partial Orders. In place of labeled linear orders, we may use labeled partial orders, with the same operations, since the substitution of partial orders for letters as in Definition 5 makes sense.

Language algebras. We extend the definitions of the operations on words on A to nonempty sets of words on A, as follows. For $U, V, U_i \subseteq AW$,

$$\begin{aligned} U \cdot V &:= \{uv : u \in U, v \in V\} \\ U^\omega &:= \{u_1 \cdot u_2 \cdots : u_i \in U\} \\ U^{\omega^{op}} &:= \{\cdots u_2 \cdot u_1 : u_i \in U\} \end{aligned}$$

The shuffle operation is more complicated to describe. Let $[\![U_1, \ldots, U_k]\!]^\eta$ be the set of all words obtained by substituting a word in U_i for an occurrence of the letter a_i in the word ρ_k, for $i = 1, \ldots, k$, where different words in U_i may be substituted for distinct occurrences of a_i. (This is in essence what happens in the definition of U^ω and $U^{\omega^{op}}$: it is the set of all words obtained by substituting a word in U for an occurrence of the letter a_1 in the word a_1^ω, and $a_1^{\omega^{op}}$, respectively.)

Proposition 90 *The variety generated by all word algebras AW, as A ranges over all sets, is the same as the variety of algebras generated by the labeled partial orders is the same as the variety generated by the language algebras.*

Proof. It is clear that each algebra of labeled partial orders and each language algebra satisfies the axioms. Also, the free algebras in W belong to both classes. □

16 Open question

Briefly, the question is this: *Is the iteration theory of regular words the free iteration theory on an associative binary operation?* (See [BlEs93] for more than you want to know about iteration theories.)

Let R_A be the algebraic theory whose morphisms $1 \to p$ are the regular words on the set $A \cup \{x_1, \ldots, x_p\}$, and where composition is defined via substitution (see Definition 4 above). Morphisms $n \to p$ are n-tuples of morphisms $1 \to p$.

If $t : 1 \to p$ and $s_i : 1 \to q$, for $i \in [p]$, then

$$t \cdot \langle s_1, \ldots, s_p \rangle := t(x_1/s_1, \ldots, x_p/s_p).$$

In this theory, the coproduct injections $1 \to 2$ are the words x_1 and x_2, respectively.

Let $\overline{\sigma} : 1 \to 2$ in R_A be the word x_1x_2. Note that for any two words $u, v : 1 \to p$, the word uv is $\overline{\sigma} \cdot \langle u, v \rangle$. Thus the equation

$$\begin{aligned}\overline{\sigma} \cdot (\overline{\sigma} \oplus \mathbf{1}_1) &= (x_1x_2)x_3 \\ &= \overline{\sigma} \cdot (\mathbf{1}_1 \oplus \overline{\sigma}) \\ &= x_1(x_2x_3)\end{aligned}$$

holds in R_A, since the product of words is associative.

It can be shown [BE05] that R_A is in fact an iteration theory.

In this theory, each of the regular operations is an instance of the fixed point operation: when $u, u_1, \ldots, u_n$ are words on A,

$$\begin{aligned}u^{\omega} &= (u \cdot x_1)^{\dagger} \\ u^{\omega^{op}} &:= (x_1 \cdot u)^{\dagger} \\ [\![u_1, \ldots, u_n]\!]^{\eta} &= (x_1 \cdot u_1 \cdot x_1 \cdots x_1 \cdot u_n \cdot x_1)^{\dagger}.\end{aligned}$$

Now let Σ_A be the signature containing one function symbol σ of rank 2, and for each letter $a \in A$, a function symbol a of rank 0. Let T_{Σ_A} be the free iteration theory generated by Σ_A. A morphism $1 \to p$ in T_{Σ_A} is a regular, full binary tree whose leaves are labeled with letters in the set $A \cup \{x_1, \ldots, x_p\}$; each interior node has two successors and is labeled σ. Let $\varphi : T_{\Sigma_A} \to R_A$ be the unique iteration theory morphism determined by

$$\begin{aligned}&\varphi(\sigma) := x_1x_2 \\ &\varphi(a) := a, \quad a \in A.\end{aligned}$$

Let $\sim$ be the congruence on T_{Σ_A} determined by φ: $f \sim g \iff \varphi(f) = \varphi(g)$.

The open problem is this. *Is the congruence $\sim$ the* ***smallest*** *iteration theory congruence $\cong$ on T_{Σ_A} such that*

$$\sigma \cdot (\sigma \oplus \mathbf{1}_1) \cong \sigma \cdot (\mathbf{1}_1 \oplus \sigma). \tag{33}$$

The equation (33) means that σ is interpreted as an associative operation.

Acknowledgement

The authors would like to thank two anonymous referees and Walter Taylor, University of Colorado, for suggesting several significant improvements to what we thought was our final draft.

References

[BedCar98] N. Bedon and O. Carton. An Eilenberg theorem for words on countable ordinals. in *Latin'98: Theoretical Informatics* (C. L. Lucchesi and A. V. Moura, eds.), LNCS 1380, Springer-Verlag, 1998, 53–64.

[BlCho01] S.L. Bloom and C. Choffrut. Long words: the theory of concatenation and ω-power. *Theoretical Computer Science*, 259(2001), 533–548.

[BlEs03] S.L. Bloom and Z. Ésik. Deciding whether the frontier of a regular tree is scattered. *Fundamenta Informatica*, 55(2003), 1–21.

[BlEs03a] S.L. Bloom and Z. Ésik. Axiomatizing omega and omega-op powers on words. *Theoretical Informatics and Applications*, **38** (2004), 3–17.

[BE05] S.L. Bloom and Z. Ésik. Some 2-theories of words. To appear.

[BlEs93] S.L. Bloom and Z. Ésik. *Iteration Theories.* Springer, 1993.

[BruyCar] V. Bruyère and O. Carton. Automata on linear orderings. Proceedings *Mathematical Foundations of Computer Science*, 2001, LNCS 2136, Springer, 2001, 236–247.

[BruyCar2] V. Bruyère and O. Carton. Hierarchy among automata on linear orderings. In: *Foundation of Information Technology in the Era of Network and Mobile Computing, Proc. TCS 2002*, Kluwer Academic Publishers, 2002, 107–118.

[Car03] O. Carton. Unambiguous automata on bi-infinite words. In MFCS'2003 (B. Rovan and P. Vojtá, eds.), vol. 2747 of Lect. Notes in Comput. Sci., pp. 308-317, 2003.

[Chou78] Y. Choueka. Finite automata, definable sets and regular expressions over ω^n-tapes. J. Comp. Sys. Sci., 17:81–97, 1978.

[Cour78] B. Courcelle. Frontiers of infinite trees. *RAIRO Informatique théorique/Theoretical Computer Science*, 12(1978), 319–337.

[Ehr59] A. Ehrenfeucht. Decidability of the theory of the linear ordering relation. *Notices Amer. Math. Soc.* 6(1959), 556-38.

[GTW02] E. Grädel, W. Thomas, and T. Wilke. Automata, Logics and Infinite Games. A Guide to Current Research. LNCS 2500 Tutorial, Springer, 2002. 1–385.

[Heil80] S. Heilbrunner. An algorithm for the solution of fixed-point equations for infinite words. *Theoretical Informatics and Applications*, 14(1980), 131–141.

[Ro82] J.B. Rosenstein. *Linear Orderings*. Academic Press, New York, 1982.

[Th86] W. Thomas. On frontiers of regular trees. *Theoretical Informatics and Applications*, vol. 20, 1986, 371–381.