Some classes of sets of structures definable without quantifiers

James Rogers
Earlham College
jrogers@cs.earlham.edu

Dakotah Lambert
Earlham College
djlambe11@earlham.edu

Abstract

We derive abstract characterizations of the Strictly Piecewise Local (SPL) and Piecewise Locally Testable (PLT) stringsets. These generalize both the Strictly Local/Locally Testable stringsets (SL and LT) and Strictly Piecewise/Piecewise Testable stringsets (SP and PT) in that SPL constraints can be stated in terms of both adjacency and precedence.

We do this in a fully abstract setting which applies to any class of purely relational models that label the points in their domain with some finite labeling alphabet. This includes, for example, labeled trees and graphs. The actual structure of the class of intended models only shows up in interpreting the abstract characterizations of the definable sets in terms of the structure of the models themselves.

1 Introduction

The ultimate goal of this paper is a characterization of the Piecewise Locally Testable and Strictly Piecewise Local classes of sets of strings. But in getting there we employ a very general technique that, with the exception of a single step (the definition of *realizability*) applies to any class of relational structures and yields:

- a quantifier-free logic that is propositional in the sense of "can be interpreted via truth tables as a canonical, but uninteresting, class of models" (but not PC, as in Propositional Calculus),
- an algebraic setting that, modulo the definition of realizability, provides an abstract characterization of the sets of structures definable in that logic, which may or may not be all that useful in itself, but which is strong enough to support more natural characterizations.

The reason that definability with respect to these quantifier-free logics is interesting is that it identifies the sets of structures that are definable purely in terms of the explicit components of the structures themselves, without any auxiliary mechanisms such as distinguishing points in terms independent of their labels (by assigning variables to them, for instance, or associating them with states) or refinements of the label alphabet (by adding features, for instance). This gives a near minimal notion of definability and a class of constraints that can be checked without inferring any information beyond what is explicitly present in the structure itself.

1.1 Overview of the paper

In Section 2 we introduce the Piecewise Local Hierarchy and provide some motivation for exploring its propositional levels. In Section 3 we introduce relational models and their local factors. These are ordinary mathematical models over a purely relational signature which include unary relations that we can interpret as labeling the points in the domain. Beyond that, while the actual structural properties of the models are given by a definition of what counts as an intended model, those properties are inconsequential for nearly all of what follows. As an example we define a class of word models, models for strings that include relations for both successor and precedence.

In Section 4 we introduce a propositional logic based on local factors as atoms and define the class of *Locally Definable* sets of structures as those definable in that logic and the class of *Strictly Locally Definable* sets of structures as those definable by conjunctions of negative literals of that logic. This notion of locality extends that of McNaughton and Papert (1971) to adjacency with respect to any of the non-unary relations of the signature.

In Section 5 we consider sets of factors as models of the logic rather than the relational structures themselves. The function taking a structure to the set of its factors maps models of the first sort (the structures themselves) to models of the second (sets of factors). The advantage of this move is that the space of sets of factors is a finite Boolean Algebra. In this setting it is easy to prove that a set of sets of factors is Strictly Local if and only if (iff) it is a principal ideal in that space. The cost of this move is that not all sets of factors are the image of one of the intended relational structures. Those that are we refer to as being *realizable*.

We then return, in Section 6, to the properties of the Locally and Strictly Locally Definable sets of structures and develop abstract characterizations of these classes. Up until this point, everything we have done applies to any class of relational structures, regardless of the actual structural properties of the intended models (strings, for example, or trees). The characterization of the Local sets is valid for all classes of relational structures, but the last step of the characterization of the Strictly Locally Definable sets depends critically on the notion of realizability. The section ends by completing the characterization for models of strings.

These characterizations can be hard to apply in their fully abstract form. In Section 7 we fix the notion of realizability for models of strings over a signature that includes both successor and precedence, and derive closure properties that are generalizations of the well-known characterizations of the Strictly Local (successor only) and Strictly Piecewise (precedence only) sets of strings.

In Section 8 we consider the learnability of the definable sets of structures. Following that we give both an example of a phonotactic constraint not that is SPL definable and one that separates SPL from both SF and the Tier-based Strictly Local stringset (defined there). We then close with some concluding remarks.

2 The Piecewise-Local Hierarchy

The Piecewise-Local Hierachy (Figure 1) organizes the Local and Piecewise classes of stringsets, introduced in McNaughton and Papert (1971) and extended by Brzozowski and Simon (1973), Simon (1975), Straubing (1985), Thérien and Weiss (1985), Beauquier and Pin (1991) and others, on the basis of model-theoretic definability with respect to word models (see Example 1) along two dimensions: signature (successor alone, less-than alone, or both) and strength of the logical machinery, from the propositional logic discussed below (Section 4) to Monadic Second-Order.

The characterization of Regular stringsets by MSO definability is due to Medvedev (1964), Büchi (1960) and Elgot (1961). This work established the relationship between model-theory of ordered structures and computational structures that spawned the study of Descriptive and Structural Complexity, Finite Model-Theory and other areas of Graph Theory, Abstract Algebra, Theorem Proving and Discrete Math. The characterization of the Star-Free stringsets (SF—definable by regular expressions with complement but not Kleene-closure) by FO($<$), First-Order definability with less-than (or both less-than and successor, since successor is FO definable from less-than) is due to McNaughton and Papert (1971), which spawned the work of Brzozowski, Simon, Beauquier and Pin cited above. Thomas (1978) established the characterization of the Locally Threshold Testable (LTT) stringsets by FO definability with successor alone, FO($+1$).

Our exploration of the Piecewise-Local hierarchy was motivated by Heinz's exploration of learnability of phonotactic stress patterns (Heinz, 2007). Our research group at Earlham College, over the course of several years, constructed computational tools to classify the patterns in the StressTyp2 (Goedemans et al., 2015) collection of stress patterns that have automata-theoretic semantics, about two-thirds of the 750 lects in the collection, covering a broad range of human languages. These fall into 106 distinct patterns.

Initially, we identified the 82 that are Strictly-Local. In exploring the remainder, we started working with constraints expressed in the propositional logic introduced here in Section 4. Constraints definable using just successor are Locally Testable (LT); Strictly Local (SL) constraints are those that are definable by conjunctions of negative literals. Some constraints, the requirement that every word assigns primary stress to some syllable (obligatoriness) or the requirement that primary stress either falls on a heavy syllable or on the final syllable, while not SL, are clearly the complement of SL constraints (co-SL), disjunctions of positive literals, which share the explicit nature of SL constraints.

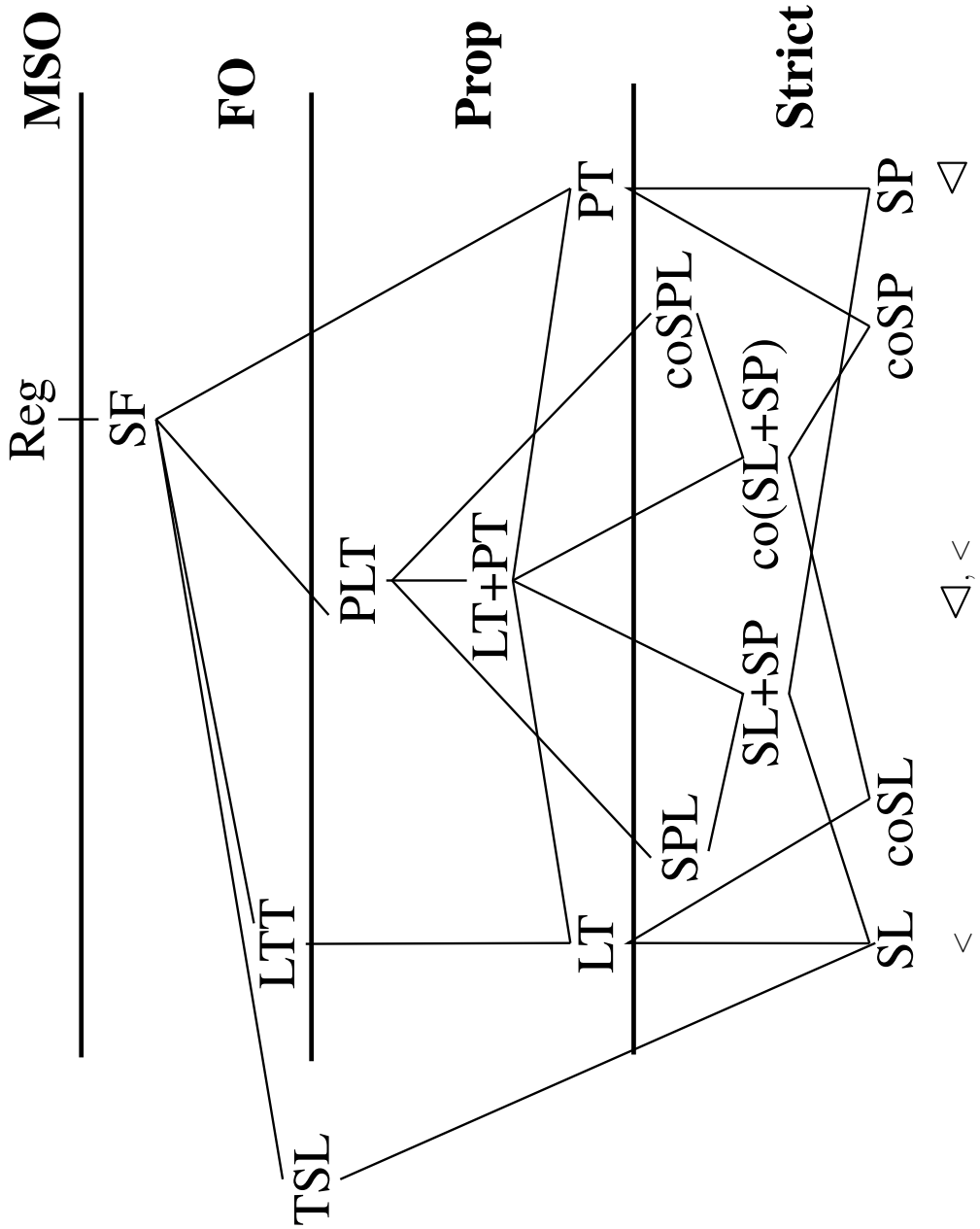

Figure 1: The Piecewise Local Hierarchy

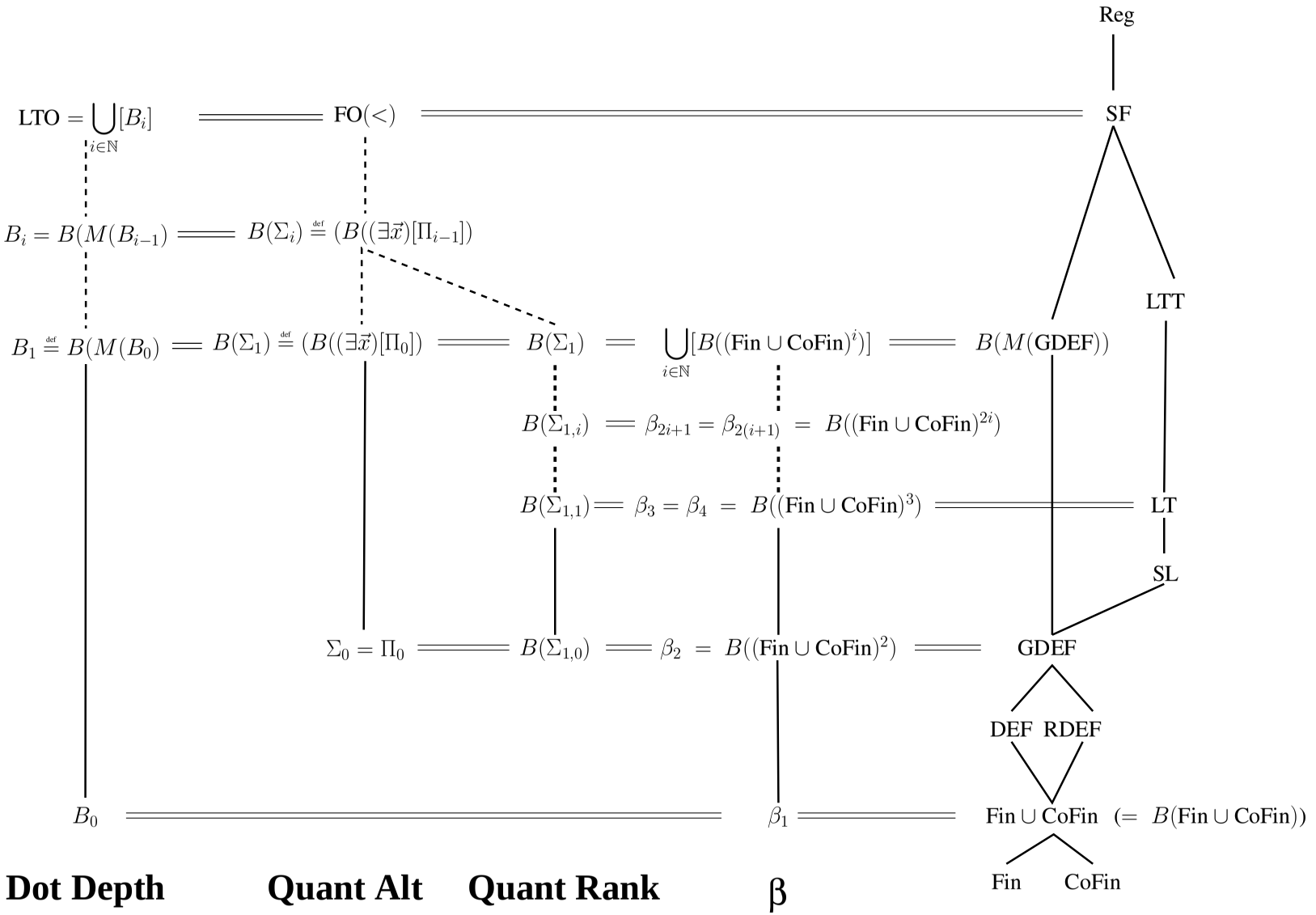

Figure 2: Some Other Local Hierarchies

Some constraints, the requirement that if primary stress falls on a heavy syllable it must be the first heavy syllable, for example, are not even definable in FO(+1). These are examples of long-distance phonotactics which are amenable to being defined in terms of subsequences (sequences of symbols that occur in order but not necessarily adjacently). These are definable using less-than without the aid of successor; they are Piecewise Testable (PT) constraints. The Piecewise Testable stringsets were introduced by Simon (1975) and are the analog of the LT stringsets based on subsequences rather than substrings. Some, the requirement that primary stress does not fall on more than one syllable (culmanitivity), for example, can be expressed as conjunctions of negative piecewise literals, in SP (Rogers et al., 2010). Obligatoriness, since the only factor involved is a single symbol, can be expressed as the complement of an SP stringset, it is co-SP, as well as co-SL.

Conjunctions of SL, co-SL and SP constraints (SL + co-SL + SP) cover 98 of the 106 patterns in the database. Of the remaining eight, two are properly regular, involving covert alternation. The rest are all of the form: if stress falls on a final syllable that is heavy, then a syllable of some other type (an unstressed heavy, for example) does not occur. While LT they are not expressible as SL or SP constraints or their complement. They are, on the other hand expressible as a negative literal that uses both adjacency (for the identification of the final syllable) and less-than (for the long-distance aspect). This, in addition to the obvious theoretical interest, is what led us to explore the Piecewise Locally (PLT) Testable and Strictly Piecewise Local (SPL) stringsets. Except for those two properly regular patterns, all of the automata-theoretic patterns in StressTyp2 are definable in SPL + co-SL.

2.1 Some Other Sub-Regular Hierarchies

An alternative way of partitioning the Star-Free stringsets is via the dot-depth and β hierarchies (Figure 2). The former is due to Schützenberger (1965) and Brzozowski and Knast (1978). McNaughton and Papert (1971) had already established that the Star-Free stringsets are equal to closure of the Locally Testable stringsets under mixed concatenation and Boolean operations (Locally Testable with Order, LTO). Brzozowski and Knast (1978) establishes an infinite hierarchy, building from the class of finite and co-finite stringsets via concatenation closure (M in the figure) followed by Boolean closure (B), alternately, which partitions LTO. Brzozowski and Simon (1973) refines the dot-depth 1 level into an infinite hierarchy built on concatenation of successively more factors, closed under Boolean operations. The second level (β_2) is equivalent to the class of Generalized Definite stringsets (GDEF, stringsets determined by their initial and final substrings); The third, equivalently fourth, level is equivalent to the LT stringsets.

Thomas (1982), using a somewhat different notion of word model, characterizes these with respect to FO definability. The dot-depth hierarchy (albeit starting at GDEF for Σ_0/Π_0) corresponds to the standard notion of quantifier alternation. The β hierarchy corresponds to Σ_1 stratified by quantifier rank.

The citations given here are a very sparse (and idiosyncratic) sample of the incredibly broad and deep body of work over the last 60 years that has its foundations in those initial results, a testament to the fundamental nature of the results. Perhaps the best route into the theories of word models is the books by McNaughton and Papert and by Straubing (1994).

3 Some definitions

3.1 Relational Models

To be precise about terminology, a relational signature, $\mathbb{R}$, is a ranked alphabet of relation symbols $\{\mathbb{R}^i \mid i \in \mathbb{N}\}$, where the symbols in $\mathbb{R}^i$ represent i-ary relations. Let $\mathbb{R}^*$ be the union of the symbols in $\mathbb{R}^i$ for all i. We assume that $\mathbb{R}$ is finite.

An $\mathbb{R}$-structure is a tuple $\mathcal{A} = \langle A, R_1^{\mathcal{A}}, R_2^{\mathcal{A}}, \ldots\rangle$ where A is the domain and the $R_i^{\mathcal{A}}$ are interpretations of the appropriate arity of symbols chosen from $\mathbb{R}^*$.

Example 1 (Word Models)**.** *Let w be a string over the alphabet Σ. Let $|w|$ be the length of w. A* Word Model *for w is a structure:*

$$\mathcal{M}_\Sigma^{\lhd,<}(w) \stackrel{def}{=} \langle \mathcal{D}^w, \lhd^w, <^w, \rtimes^w, \ltimes^w, P_\sigma^w\rangle_{\sigma\in\Sigma}$$

where:

- $\mathcal{D}^w$*—is isomorphic to an initial segment $\langle 0, 1, \ldots, |w|+1\rangle$ of $\mathbb{N}$ (the Natural numbers).*
- $\lhd^w$*—is the successor relation on $\mathcal{D}^w$.*
- $<^w$*—is the proper precedence relation on $\mathcal{D}^w$.*

$\rtimes^w$—*is the singleton set containing the minimum position in* D^w.

$\ltimes^w$—*is the singleton containing the maximum position in* D^w.

P^w_σ—*is the set of positions in* w *at which the symbol* σ *occurs.*

The sets $\rtimes^w$, $\ltimes^w$ *and* P^w_σ, *for each* $\sigma \in \Sigma$ *partition* D^w *(they are pairwise disjoint and their union is* D^w*).*

Let $\mathcal{M}^{\lhd,<}_\Sigma$ *denote the class of all word models over* Σ.

This definition of a word model differs in certain respects from definitions that may be familiar from prior work. In particular, the endmarkers $\rtimes, \ltimes$ are explicit in the structure and mark points that are adjoined to the ends of the set of positions in the string. Thus, if $w = \langle \sigma_1, \sigma_2, \ldots, \sigma_{|w|} \rangle$, then $\mathbf{card}(D^w) = |w| + 2$ and $i \in P^w_\sigma$ iff $\sigma_i = \sigma$. Also this type of word model includes both the successor and precedence relations. When we look at specific classes in the Piecewise Local sub-regular hierarchy we have, heretofore, employed reducts of this signature including one or the other of the ordering relations, but not both. Here we can obtain the same restrictions by varying parameters restricting their usefulness.

It is important to note that a structure is not necessarily a word model simply because it shares the same signature as these word models. In particular, the interpretations of '$\mathcal{D}^w$', '$\lhd^w$', '$<^w$', '$\rtimes^w$' and '$\ltimes^w$' are not arbitrary, but required to satisfy the axioms of finite discrete linear orders under the usual interpretation of the symbols. We refer to these as the *structural* relations—they form the 'bones' of the intended class of structures—and we refer to those structures that satisfy the axioms as the *intended models*. These notions generalize to other classes of labeled relational structures which exhibit particular structural properties. Word models corresponding to distinct strings differ only in the size of the domain and in the interpretations of the P_σ.[1]

In the core of this paper we temporarily turn to sets of factors (defined in Section 3.3) as models, which characterize the freely generated structures over the given signature, and encapsulate the theory of the intended structures in the notion of "realizability" (Section 5.3) which picks out the sets of factors that actually correspond to a particular well-formed structure. The core results are valid independent of the definition of realizability, which puts the focus squarely on the defining power of the quantifier-free logic, abstracting away from properties that distinguish a class of intended models from another with the same signature.

Henceforth, when we refer $\mathbb{R}$-structures, we will mean the class of intended structures, however that definition may be restricted. When we discuss the freely generated structures over the signature $\mathbb{R}$, including those that may not be well-formed, we will explicitly say so.

3.2 Homomorphisms and Embeddings

Suppose $\mathcal{A}$ and $\mathcal{B}$ are $\mathbb{R}$-structures. The following definitions are from (Hodges, 1993).

A *homomorphism* from $\mathcal{A}$ to $\mathcal{B}$ is a (total) function $h : A \to B$ such that:

$$R \in \mathbb{R}^i \text{ and } \vec{a} \in R^{\mathcal{A}} \Rightarrow h(\vec{a}) \in R^{\mathcal{B}}.$$

Note that this only requires that the images of the $\vec{a}$ that are in the interpretation of R in $\mathcal{A}$ are included in the interpretation of R in $\mathcal{B}$. It says nothing about other $h(\vec{a})$ that might also be included in $R^{\mathcal{B}}$.

An *embedding* (or strong homomorphism) from $\mathcal{A}$ to $\mathcal{B}$ is a (total) function $h : A \to B$ such that h is a homomorphism that is is strengthened to

$$R \in R^i \text{ and } \vec{a} \in R^{\mathcal{A}} \Leftrightarrow h(\vec{a}) \in R^{\mathcal{B}}$$

We note the difference because "homomorphism" is often taken in the stronger sense, but we necessarily need the weak sense. Otherwise if the image of A in B includes any tuple $h(\vec{a})$ in the domain of the interpretation of a relation R in $\mathcal{B}$, then $\mathcal{A}$ must include the pre-image of $R^{\mathcal{B}}(h(\vec{a}))$ in $R^{\mathcal{A}}$ as well. In this way the interpretation of R in $\mathcal{B}$ would restrict the structure of $\mathcal{A}$.

3.3 Neighborhoods and factors

The next few definitions are based on those in Libkin (2004). The first two are ubiquitous in the Theory of Finite Models.

Definition 1. Let $\mathcal{A}$ be a relational structure as above and $a \in A$. The (domain of the) r-*Ball*

[1] Although in this case, the interpretations of the P_σ is not entirely free, either, in that we require those interpretations, along with those of the end markers, to partition the domain. Relaxing that actually leads to a more flexible notion of string-like structures that are useful in many applications.

around a in $\mathcal{A}$ (denoted $B_r^{\mathcal{A}}$) is defined inductively as follows:

$$\begin{array}{rcl} B_0^{\mathcal{A}}(a) & = & \{a\} \\ B_{i+1}^{\mathcal{A}}(a) & = & B_i^{\mathcal{A}}(a) \cup \\ & & \{a' \mid (\exists R \in \mathbb{R}, \\ & & \quad \vec{a} \in R^{\mathcal{A}}, a'' \in B_i^{\mathcal{A}}(a)) \\ & & \quad [a', a'' \text{ both occur in } \vec{a}]\} \end{array}$$

The members of $B_i^{\mathcal{A}}(a)$ are all the members of A that are within distance i of a in the Gaifmann graph of $\mathcal{A}$.

Definition 2 (Connectivity)**.** Let $\mathcal{A}$ be an $\mathbb{R}$ structure.

$\mathcal{A}$ is *k-connected* iff for all $a \in A$, $B_{k-1}^{\mathcal{A}}(a) = A$.

Note that every k-connected domain is also j-connected for all $j > k$.

Definition 3 (k-Factor)**.** Let $\mathcal{A}$ and $\mathcal{B}$ be $\mathbb{R}$ structures.

$\mathcal{A}$ is a *k-factor* of $\mathcal{B}$ iff[2]

$$\begin{array}{ll} 1: & \mathbf{card}(A) \leq k \\ 2: & A \text{ is } \mathbf{card}(A)\text{-connected} \\ 3: & \exists h : A \to B, \text{ a homomorphism} \end{array}$$

N.B. In this definition the set of k-factors of a structure includes all of its j-factors for $j \leq k$.

In general there will be many such h. Each one picks out an occurrence of the factor $\mathcal{A}$ in $\mathcal{B}$.

Let $\mathrm{F}_k(\mathcal{B})$ be the set of isomorphism classes of the set of all k-factors of $\mathcal{B}$. We will ignore the difference between an isomorphism class and a canonical representative of that class, so we can consider this to be a set of structures over an anonymous domain of cardinality k.

Proposition 1. *If* ***card***(B) *is finite then there are only finitely many distinct (up to isomorphism) k-factors of $\mathcal{B}$.*

Lemma 1. *If $\mathbb{R}$ is relational and finite then the cardinality of*

$$F_k(\mathbb{R}) \stackrel{def}{=} \bigcup_{\mathcal{B}} [F_k(\mathcal{B})],\ \mathcal{B} \textit{ an } \mathbb{R}\textit{-structure}.$$

is finite.

Proof Sketch. If $\mathbb{R}$ is relational and finite then there are, up to isomorphism, only finitely many $\mathbb{R}$-structures of cardinality k. This is Exercise 6 (Pg. 10) of Hodges, an easy exercise. ■

We extend F_k to sets of structures in the standard way, as the union of the sets of k-factors of the structures in the set.

3.4 Supposing $k < n$

Suppose $\mathcal{A}$ is an $\mathbb{R}$-structure, $\mathbf{card}(A) = k$ and $R \in \mathbb{R}^n \subseteq \mathbb{R}$, as above, and $k < n$. Then, by the pigeon-hole-principle,

$$(\forall \vec{a} \in R^{\mathcal{A}})[(\exists a \in A)[\\ a \text{ occurs in at least two places in } \vec{a}]].$$

Let's say that an n-ary relation is anti-reflexive if no individual occurs more than once in any of its tuples. If all $R \in \mathbb{R}$ are anti-reflexive then, for all $R \in R^{n>k}$, $R^{\mathcal{A}} = \emptyset$.

This is not deep. It just says that in the anti-reflexive case (which will be common) k-factors have nothing to say about relations of arity greater than k.

3.5 Aspects of partial orders

The following is taken, primarily, from MacLane and Birkhoff (1967, 1970). A *partial order* is a set equipped with a partial ordering relation $\sqsubseteq$ that is reflexive, transitive and antisymmetric. If $\sqsubseteq$ is not antisymmetric, then it defines a *quasiorder*.[3]

A *lattice* is a partial order that is closed with respect to two binary operators: a greatest lower bound (meet, $\wedge$) and least upper bound (join, $\vee$). Meets and joins are idempotent, associative and commutative and satisfy the absorption law ($x \wedge (x \vee y) = x = x \vee (x \wedge y)$). If they distribute over each other, then the lattice is distributive.

All finite lattices have a unique minimum element ($\bot$) and a unique maximum element ($\top$). If a lattice has a maximum and minimum element and every element has a complement with respect to these ($x \wedge \overline{x} = \bot$) and ($x \vee \overline{x} = \top$) then it is a complemented lattice. If it is complemented and distributive it is a *Boolean* lattice, equivalently, Boolean algebra. If the lattice is Boolean then every element x has a unique complement $\overline{x}$.

[2] The term "factor" comes from the well known interpretation of strings in a monoid over concatenation, where the definition is immediate. We adopt this fully general definition in order to extend it to arbitrary relational structures, independent of their specific properties.

[3] In our usage, the relationship between quasiorders and partial orders is analogous to that between preorders and total orders.

An element a of a Boolean algebra is an atom iff $a > \bot$ and there is no b such that $a > b > \bot$.

If a set S is partially ordered by $\sqsubseteq$ and I is a non-empty subset of S that is downward closed ($x \in I$ and $y \sqsubseteq x$ implies $y \in I$) and each pair of elements in I has at least one upper bound in I, then I is an *ideal*. If the ideal includes a unique maximum element a then it is the *principal ideal* generated by a, which we will denote $\mathcal{I}(a)$. S and $\sqsubseteq$ will always be clear from the context.

(Principal) filters are defined dually: upward closed and with lower bounds. We will denote the principal filter generated by a as $\mathcal{F}(a)$.

4 Local and Strictly Local Sets of Structures

Definition 4. Let $\mathbb{R}$ be a relational signature and $\mathcal{G}$ be a subset of $\mathrm{F}_k(\mathbb{R})$.

Let $L(\mathcal{G}) \stackrel{\text{def}}{=} \{\mathcal{A} \mid \mathrm{F}_k(\mathcal{A}) \subseteq \mathcal{G}\}$.

Then $L(\mathcal{G})$ is a *strictly local set* of $\mathbb{R}$-structures.

A set S of $\mathbb{R}$-structures is a *locally definable set* of $\mathbb{R}$-structures iff it is a Boolean combination of strictly local sets.[4]

4.1 Local Logics

4.1.1 Well-Formed Formulae

Let $\mathbf{wff}_k(\mathbb{R})$ be the set of Boolean formulae in which the atomic formulae are the factors in $\mathrm{F}_k(\mathbb{R})$. Usually we can be ambiguous about k, letting it be determined by the formula itself.

4.1.2 Satisfaction with respect to $\mathbb{R}$-structures

Each $\mathbb{R}$-structure provides a valuation of the formulae in $\mathbf{wff}_k(\mathbb{R})$ based on its set of factors: if $f \in \mathrm{F}_k(\mathbb{R})$ and $\mathcal{A}$ is an $\mathbb{R}$-structure then

$$\mathcal{A} \models f \stackrel{\text{def}}{\iff} f \in \mathrm{F}_k(\mathcal{A}).$$

Let Φ be a set of $\mathbf{wff}(\mathbb{R})$ formulae and $\mathcal{A}$ a $\mathbb{R}$-structure. Then

$$\mathcal{A} \models \Phi \stackrel{\text{def}}{\iff} (\forall \varphi \in \Phi)[\mathcal{A} \models \varphi]$$

and the *models* of Φ is the set

$$\mathbf{Mod}(\Phi) \stackrel{\text{def}}{=} \{\mathcal{A}, \text{a } \mathbb{R}\text{-structure} \mid \mathcal{A} \models \Phi\}.$$

Φ is *consistent* iff $\mathbf{Mod}(\Phi) \neq \emptyset$.

Let Φ and Ψ be sets of $\mathbf{wff}(\mathbb{R})$ formulae. Φ *entails* (logically implies) Ψ ($\Phi \models \Psi$) iff, by definition, $\mathbf{Mod}(\Phi) \subseteq \mathbf{Mod}(\Psi)$ (i.e., for all $\mathbb{R}$-structures $\mathcal{A}$, $\mathcal{A} \models \Phi \Rightarrow \mathcal{A} \models \Psi$). Φ and Ψ are logically equivalent ($\Phi \equiv \Psi$) iff, by definition, $\Phi \models \Psi$ and $\Psi \models \Phi$.

4.2 Local and Strictly Local Definitions

Let $L = L(\mathcal{G})$ for some $\mathcal{G} \subseteq \mathrm{F}_k(\mathbb{R})$ be a k-strictly local set of $\mathbb{R}$-structures. $\mathcal{G}$ is the set of permitted factors; the structures in $L(\mathcal{G})$ may not include any factors but these. Let $\overline{\mathcal{G}} = \mathrm{F}_k(\mathbb{R}) - \mathcal{G}$, the set of forbidden factors of L. Since $\mathrm{F}_k(\mathbb{R})$ is finite, $\overline{\mathcal{G}}$ is as well. Then L includes all and only those structures that do not include any of the factors in $\overline{\mathcal{G}}$. Formally:

$$L = \mathbf{Mod}(\bigwedge_{f \in \overline{\mathcal{G}}} [\neg f]).$$

Lemma 2. *A set of $\mathbb{R}$-structures is strictly k-local iff it is the set of models of a conjunction of negative literals of $\boldsymbol{wff}_k(\mathbb{R})$.*

As usual, we interpret sets of formulae conjunctively, thus a set of $\mathbb{R}$-structures is strictly k-local iff it is $\boldsymbol{Mod}(\Phi)$ where $\Phi \subseteq \{\neg f \mid f \in F_k(\mathbb{R})\}$.

Lemma 3. *Since a set of structures is local iff it is a Boolean combination of Strictly Locally Definable structures, a set of $\mathbb{R}$-structures is k-Locally Definable if it is $\boldsymbol{Mod}(\Phi)$ for any $\Phi \subseteq \boldsymbol{wff}_k(\mathbb{R})$.*

5 Definable sets of subsets of $\mathrm{F}_k(\mathbb{R})$

Consider the space of subsets of $\mathrm{F}_k(\mathbb{R})$, partially ordered by subset (this is the *powerset algebra* of $\mathrm{F}_k(\mathbb{R})$). It is a Boolean algebra in which $\top$ is $\mathrm{F}_k(\mathbb{R})$, $\bot$ is $\emptyset$ and the atoms are the singleton sets of individual factors in $\mathrm{F}_k(\mathbb{R})$. We will refer to this space of subsets as $\mathbb{B}_k(\mathbb{R})$. Since $\mathrm{F}_k(\mathbb{R})$ is finite, $\mathbb{B}_k(\mathbb{R})$ is as well.

Note that F_k maps $\mathbb{R}$-structures to elements of $\mathbb{B}_k(\mathbb{R})$; it is many-one and generally not onto. While we have restricted our attention to $\mathbb{R}$-structures that are well-formed, those well-formedness properties show up in $\mathbb{B}_k(\mathbb{R})$ only in the structure of the sets of factors. $\mathbb{B}_k(\mathbb{R})$ is the freely generated powerset of the set of k-factors that occur in any well-formed $\mathbb{R}$-structure; those subsets may or may not be in the range of F_k.

[4] Note that Locally Definable sets of strings form the classes that are usually referred to as Locally or Piecewise Testable. In McNaughton and Papert (1971), these are specified by sets of permitted initial and final strings of length k, usually $k-1$ in later work, along with sets of permitted internal strings of length k. In this particular model-theoretic setting the endmarkers obviate the need for three sets of permitted factors, moreover "Testable" is more or less implied and we have, for the most part, replaced it with "Definable". On the other hand, we have not been completely consistent in doing so. This inconsistency should not prove to be overly confusing.

Let $\mathcal{A} \sqsubseteq_k \mathcal{B} \stackrel{\text{def}}{\iff} \mathrm{F}_k(\mathcal{A}) \subseteq \mathrm{F}_k(\mathcal{B})$. This induces a quasiorder on $\mathbb{R}$-structures, in which two $\mathbb{R}$-structures $\mathcal{A}$ and $\mathcal{B}$ are equivalent with respect to $\sqsubseteq_k$ iff they are logically equivalent with respect to $\mathbf{wff}_k(\mathbb{R})$.

N.B. We denote the order relation of the powerset algebra of $\mathbb{B}_k(\mathbb{R})$ by '$\subseteq$' and from this point on reserve '$\sqsubseteq$' for the quasiorder it induces in the space of $\mathbb{R}$-structures.

We are ultimately interested in the properties of the definable sets in that space of $\mathbb{R}$-structures, but will derive them from the properties of the definable subsets of $\mathbb{B}_k(\mathbb{R})$. One of the advantages of $\mathbb{B}_k(\mathbb{R})$ is that it is finite, while the set of $\mathbb{R}$-structures is infinite. More importantly, it has a simple and regular structure that is independent of the details of the properties of well-formed $\mathbb{R}$-structures.

5.1 Satisfaction with respect to Sets of k-factors

To that end, extend '$\models$' to sets of k-factors in the natural way: $S \subseteq \mathrm{F}_k(\mathbb{R})$ satisfies $f \in \mathrm{F}_k(\mathbb{R})$ iff $f \in S$, with the semantics of the Boolean connectives being defined in the usual way. In order to distinguish definable sets of sets of factors from definable sets of $\mathbb{R}$-structures, we will refer to the sets of sets of k-factors that satisfy a given $\varphi \in \mathbf{wff}_k(\mathbb{R})$ as $\mathbf{Mod}^*(\varphi) = \{S \in \mathbb{B}_k(\mathbb{R}) \mid S \models \varphi\}$.

The semantics of the logical connectives '$\wedge$', '$\vee$' and '$\neg$' correspond directly to the order-theoretic operations '$\wedge$', '$\vee$' and '$\bar{\ }$'. This is, of course, no coincidence.

5.2 Strictly Local Sets of k-factors

Following Lemma 2, a subset of $\mathbb{B}_k(\mathbb{R})$ is strictly local iff it is $\mathbf{Mod}^*(\bigwedge_{f\in\Phi}[\neg f])$, for some $\Phi \subseteq \mathrm{F}_k(\mathbb{R})$.

Note that if $f \in \mathrm{F}_k(\mathbb{R})$ then $\mathbf{Mod}^*(f)$ is the principal filter $\mathcal{F}(f)$ in $\mathbb{B}_k(\mathbb{R})$. Thus:

Lemma 4. *A subset of $\mathbb{B}_k(\mathbb{R})$ is strictly local iff it is the intersection of the complements of a (finite) set of principal filters in $\mathbb{B}_k(\mathbb{R})$.*

Let $\mathbf{S}$ be a strictly local subset of $\mathbb{B}_k(\mathbb{R})$. Since filters are upward-closed, their complements are downward-closed, as is $\mathbf{S}$, the intersection of their complements. The elements of $\mathbf{S}$ are necessarily subsets of $\overline{\Phi}$ (i.e., $\mathrm{F}_k(\mathbb{R}) - \Phi$) and $\overline{\Phi} \in \mathbf{S}$. Thus $\mathbf{S} = \mathcal{I}(\overline{\Phi})$, the principal ideal generated by $\overline{\Phi}$.

Lemma 5. *If $\mathbf{S}$ is a strictly local subset of $\mathbb{B}_k(\mathbb{R})$ then $\mathbf{S}$ is a principal ideal in $\mathbb{B}_k(\mathbb{R})$.*

Let $\mathbf{S}$ be any principal ideal in $\mathbb{B}_k(\mathbb{R})$. Since ideals are downward closed, complements of ideals are necessarily upward closed. Let $\Upsilon(\mathbf{S})$ be the set of minimal elements in $\overline{\mathbf{S}}$. Since $\mathbb{B}_k(\mathbb{R})$ is finite, such minimal elements exist. Since it is a Boolean algebra, each of those elements generates a principal filter in $\mathbb{B}_k(\mathbb{R})$. Then $\overline{\mathbf{S}} = \bigcup_{\upsilon\in\Upsilon(S)}[\mathcal{F}(\upsilon)]$. Thus, $\mathbf{S} = \bigcap_{\upsilon\in\Upsilon(S)}[\overline{\mathcal{F}(\upsilon)}]$. Since $\mathbb{B}_k(\mathbb{R})$ is finite, so is Υ, thus $\mathbf{S}$ is strictly local.

Theorem 1. *A subset of $\mathbb{B}_k(\mathbb{R})$ is Strictly Locally Definable iff it is a principal ideal.*

5.3 Realizability

So properties of the strictly local subsets of $\mathbb{B}_k(\mathbb{R})$ are, as promised, extremely simple. What we need now is an abstract characterization of the sets of strictly local sets of $\mathbb{R}$-structures based on these properties.

Some caution is required here, since F_k, as a map between the space of structures and the space of sets of factors, is not onto. The fact that an arbitrary set of factors is a subset of the set of factors of an $\mathbb{R}$-structure $\mathcal{A}$ of the intended type does not necessarily mean that it is the set of factors of a well-formed $\mathbb{R}$-structure—for the word models of Example 1 the factors will need to include both '$\rtimes$' and '$\ltimes$', at least. That type of requirement is not, in general, Strictly-Locally Definable. We have incorporated the properties of the intended models implicitly by considering only well-formed structures in our space of structures. The complexity of defining what it means to be well-formed is a meta-logical issue.

We do know that if L is a k-strictly local set of $\mathbb{R}$-structures then $\mathrm{F}_k(L)$ is a subset of a principal ideal in $\mathbb{B}_k(\mathbb{R})$. Moreover, every $\mathbb{R}$-structure that maps into that ideal is in L. But not every element of that ideal is the image of a well-formed $\mathbb{R}$-structure. Those that are, we refer to as *realizable*.

Definition 5. A subset S of $\mathrm{F}_k(\mathbb{R})$ is *realizable* iff there is some set of well-formed $\mathbb{R}$-structures $\mathbf{A}$ such that $\mathrm{F}_k(\mathbf{A}) = S$.

Every strictly local set of $\mathbb{R}$-structures is the preimage, under F_k, of the set of realizable elements in a principal ideal in $\mathbb{B}_k(\mathbb{R})$.

6 Structure of the Definable Sets of $\mathbb{R}$-structures

Recall that $\sqsubseteq_k$ is the quasiordering of $\mathbb{R}$-structures that corresponds to $\subseteq$ in $\mathbb{B}_k(\mathbb{R})$.

6.1 A Closure Property of Strictly k-Locally Definable Sets

Since F_k maps every strictly k-local set of $\mathbb{R}$-structures into an downward closed set in $\mathbb{B}_k(\mathbb{R})$ if $\mathcal{A} \in L$, a strictly k-local set of $\mathbb{R}$-structures, and $\mathcal{B} \sqsubseteq_k \mathcal{A}$ then $\mathcal{B} \in L$ as well. So k-strictly local sets are all downward closed under $\sqsubseteq_k$.

But we know much more about $\mathrm{F}_k(L)$ than it is downward closed. It is, in fact, a subset of a principal ideal that is generated by some set of k-factors, in particular the $\mathcal{G}$ of Definition 4, and that every realizable subset of $\mathcal{G}$ is the image of some structure in L. So, k-strictly local sets will be closed under any operation that does not increase the set of k-factors of its operands and which preserves realizability.

Lemma 6. *If $\oplus$ is an operation on $\mathbb{R}$-structures such that the set of k-factors of the result is a subset of the union of the sets of k-factors of the operands and which preserves realizability, then every strictly k-local set of $\mathbb{R}$-structures is closed under $\oplus$.*

We will refer to such operations as being *conservative*.

This is a closure condition on strictly k-local sets but not a characterization. The other direction of the characterization depends on the theory of the well-formed $\mathbb{R}$-structures, i.e., on the notion of realizability.

6.2 Characterization of the Local and Strictly Local Sets of Structures

6.2.1 Local Sets

Since $\sqsubseteq_k$ also corresponds to entailment with respect to $\mathbf{wff}_k(\mathbb{R})$, two $\mathbb{R}$-structures are equivalent with respect to $\sqsubseteq_k$ iff they are logically equivalent with respect to $\mathbf{wff}_k(\mathbb{R})$. Thus sets of k-local $\mathbb{R}$-structures cannot break the equivalence classes with respect to $\sqsubseteq_k$.

Even stronger, every such equivalence class is determined by the set of factors of the structures in the class.

Lemma 7. *Let $\equiv_k$ denote equivalence with respect to $\sqsubseteq_k$ and $[\mathcal{A}]_k \stackrel{def}{=} \{\mathcal{B} \mid \mathcal{A} \equiv_k \mathcal{B}\}$. Then $[\mathcal{A}]_k = \boldsymbol{Mod}(\bigwedge_{f \in F_k(\mathcal{A})}[f] \wedge \bigwedge_{f \in F_k(\mathbb{R}) - F_k(\mathcal{A})}[\neg f])$*

Theorem 2. *A set of $\mathbb{R}$-structures L is k-local iff whenever $\mathcal{B} \equiv_k \mathcal{A}$ then either both $\mathcal{A}, \mathcal{B} \in L$ or both $\mathcal{A}, \mathcal{B} \notin L$.*

This is a completely general characterization. Every k-local set of $\mathbb{R}$-structures, regardless of the theory of the structures, is the union of a set of equivalence classes with respect to $\equiv_k$.

6.2.2 Strictly Local Sets

Note that, in the space of $\mathbb{R}$-models, the inverse of $\sqsubseteq_k$ is conservative, that is, if L is k-Strictly Piecewise Locally Definable (SPL_k), $w \in L$ and $v \sqsubseteq_k w$ then $v \in L$. By definition it does not increase the set of k-factors, and v is trivially realizable. This is very close to a characterization of SPL_k, but not quite fully general.

For $f \in \mathrm{F}_k(\mathbb{R})$, with mild abuse of notation, let $\mathcal{F}_k^{\sqsubseteq}(f) \stackrel{\text{def}}{=} \{\mathcal{A} \in \mathbb{R} \mid f \in \mathrm{F}_k(\mathcal{A})\}$. This is the set of $\mathbb{R}$-models, upper-closed with respect to $\sqsubseteq$, that is generated by f. Similarly, let $\mathcal{F}_k^{\sqsubseteq}(S)$, for $S \subseteq \mathrm{F}_k(\mathbb{R})$ be the union of $\mathcal{F}_k^{\sqsubseteq}(f)$ for $f \in S$.

Lemma 8. *Each of the following is a consequence of the preceding statements:*

1. *$L \in SPL_k$.*
2. *$L = \bigcap_{f \in S} \left[\overline{\mathcal{F}_k^{\sqsubseteq}(f)}\right]$, $S \subseteq F_k(\mathbb{R})$, finite.*
3. *$w \in L$ and $v \sqsubseteq_k w \Rightarrow v \in L$. ($L$ is downward closed with respect to $\sqsubseteq_k$.)*
4. *$L = \overline{\mathcal{F}_k^{\sqsubseteq}(S)}$, for some $S \subseteq F_k(\mathbb{R})$.*

Proof. Each step is nearly immediate. By Lemma 2, $L \in \mathrm{SPL}_k \Leftrightarrow L = \mathbf{Mod}(\bigwedge_{f \in \overline{\mathcal{G}}}[\neg f])$, where $\mathcal{G}$ is finite, and each of the $f \in \mathcal{G}$ generates an upper-closed set $\mathcal{F}_k^{\sqsubseteq}(f)$. Since these are upper-closed, their complements are downward closed with respect to $\sqsubseteq_k$, as is their intersection.

To see that 3 implies 4, the complement of L is upper-closed with respect to $\sqsubseteq$. Then S, the set of minimal points in $\overline{L}$ witnesses statement 4. That such a set of minimal points exists is a consequence of the fact that there are no infinite properly descending sequences with respect to $\sqsubseteq$, which itself is a consequence of the finiteness of $\mathbb{B}_k(\mathbb{R})$. ∎

The only difference between statements 4 and 2 is the requirement that S be finite. This is where the theory of the well-formed structures comes in. For word models, it is a consequence of Higman's Lemma (Higman, 1952) which says that there are

no infinite sequences of strings that are pairwise unrelated by $\sqsubseteq$. For certain classes of tree models, it is a consequence of Kruskal's Tree Theorem (Kruskal, 1960), which is similar.

Theorem 3 (Characterization of Strictly Local Sets of Word Models)**.** *A set of word models is SPL_k iff it is downward closed with respect to $\sqsubseteq_k$.*

7 Strictly Piecewise Local Stringsets

SPL is the class of stringsets corresponding to the class of strictly local word models of Example 1. Since these models are linear, factors can be resolved into blocks of positions connected by '$\lhd$' which are, themselves, connected by '$<$'. In the terminology of the Piecewise Local hierarchy, these are subsequences of substrings. Rather than a single parameter to indicate the size of a factor we use j to denote the maximum number of substrings and k to denote the maximum size of the substrings: $\text{SPL}_{j,k}$.

Note that $\text{SPL}_{1,k}$ coincides with the well known class of SL_k stringsets, which are all and only those strictly definable in the reduct of our word models that eliminates the precedence relation. And $\text{SPL}_{j,1}$ coincides with the SP_j stringsets which are all and only those strictly definable in the reduct of our word models that eliminates successor and the end markers.[5]

In what follows we use $\text{F}_{j,k}$ and $\sqsubseteq_{j,k}$ rather than the less precise F_{jk+j-1} and $\sqsubseteq_{jk+j-1}$.

We know, already that $\text{SPL}_{j,k}$ sets are closed under $\sqsubseteq_{j,k}$, and that $\mathcal{M}_\Sigma^{\lhd,<}(v) \sqsubseteq_{j,k} \mathcal{M}_\Sigma^{\lhd,<}(w)$ iff $\text{F}_{j,k}(\mathcal{M}_\Sigma^{\lhd,<}(v)) \subseteq \text{F}_{j,k}(\mathcal{M}_\Sigma^{\lhd,<}(w))$ (modulo realizability), and that, more generally, they are closed under every operation that is conservative in the sense of Lemma 6. What we need is a natural operation on strings that is conservative. That depends on realizability.

7.1 Realizability of sets of $\text{F}_{j,k}$ factors

Definition 6 (Minimally Realizable)**.** A set of factors $S \subseteq \text{F}_{j,k}(\mathcal{M}_\Sigma^{\lhd,<})$ is *minimally realizable* iff there is a sequence of subsets of S: $q_0 \subsetneq q_1 \subsetneq \cdots \subsetneq q_n \subsetneq q_{n+1}$ such that:[6]

$$
\begin{aligned}
q_0 &= \{\rtimes\} \\
q_{i+1} &= \text{F}_{j,k}(w_i \cdot \sigma_{i+1}), \\
&\quad \text{for some } w_i \in \{\rtimes\}\Sigma^*,\ \sigma_{i+1} \in \Sigma \\
&\quad \text{such that } \text{F}_{j,k}(w_i) = q_i \\
q_{n+1} &= \text{F}_{j,k}(w_n \cdot \ltimes), \\
&\quad \text{for some } w_n \in \{\rtimes\}\Sigma^* \\
&\quad \text{such that } \text{F}_{j,k}(w_n) = q_n \\
q_{n+1} &= S.
\end{aligned}
$$

In this case the S is the set of $\text{F}_{j,k}$-factors of $\mathcal{M}_\Sigma^{\lhd,<}(w)$, where $w = \sigma_1\sigma_2\cdots\sigma_n$, and w is a *minimal witness* that such a well-formed word model exists.

Note that every word model that is equivalent to w with respect to $\sqsubseteq_{j,k}$ and also a witness of the realizability of S but only those that have the same length as w are minimal witnesses.

Every $w \in \Sigma^*$ is a witness of the realizability of $\text{F}_{j,k}(\mathcal{M}_\Sigma^{\lhd,<}(w))$. If $|w| \leq |v|$ for every $v \in [w]_{j,k}$ then it is a minimal witness.

Proposition 2. *A subset of $F_{j,k}(\mathcal{M}_\Sigma^{\lhd,<})$ is realizable iff it is the union of a finite set of minimally realizable subsets of $F_{j,k}(\mathcal{M}_\Sigma^{\lhd,<})$.*

7.2 Some closure properties of $\text{SPL}_{j,k}$ sets

Using the characterization of Theorem 3 to prove non-definability in $\text{SPL}_{j,k}$ can be cumbersome. The following closure conditions, extensions of the characterizations in earlier work on SL_k and SP_j, may be somewhat easier to apply.

Theorem 4 (Generalized Suffix-Substitution Closure)**.** *Suppose L is $SPL_{j,k}$.*
Then if

- *$u_1 \cdot x \cdot v_1 \in L$ and $u_2 \cdot x \cdot v_2 \in L$, where $|x| = k-1$,*
- *and either $F_{j-1,k}(u_1) \subseteq F_{j-1,k}(u_2)$ or $F_{j-1,k}(v_2) \subseteq F_{j-1,k}(v_1)$*

then $u_1 \cdot x \cdot v_2 \in L$.

Proof. From Lemma 6 we know that if substitution of suffixes under these conditions is conservative then $\text{SPL}_{j,k}$ is closed with respect to it. To see that it does not increase the set of $\text{F}_{j,k}(\mathcal{M}_\Sigma^{\lhd,<})$ note, to begin with, that every $\text{F}_{1,k}$ factor in $u_1 \cdot x \cdot v_2$ is also in either $u_1 \cdot x$ or in $x \cdot v_2$, thus in $\text{F}_{1,k}(u_1 \cdot x \cdot v_1) \cup \text{F}_{1,k}(u_2 \cdot x \cdot v_2)$. Suppose $f_1 \cdot f_2$ is a (j,k)-factor of $u_1 \cdot x \cdot v_2$, and that f_1 is a (i,k)-factor of u_1 and f_2 a $(j-i,k)$-factor of $x \cdot v_2$ for some $i > 1$ (otherwise it is necessarily in either $u_1 \cdot x \cdot v_1$ or $u_2 \cdot x \cdot v_2$). Since, by Definition 3,

[5] Since the contiguous blocks of a $(j,1)$-factor are all single symbols the presence of '$\lhd$' is inconsequential for the definable sets. Since $\rtimes < x < \ltimes$ for all positions x in the string, their presence is inconsequential as well.

[6] We denote these subsets as q_i to suggest the connection to a finite state automaton, but we have no need to actually construct such an automaton.

$\mathrm{F}_{j-1,k}(w)$ includes $\mathrm{F}_{j-i,k}(w)$ for all strings w and $i > 1$, $f_1 \in \mathrm{F}_{j-1,k}(u_1)$, $f_2 \in \mathrm{F}_{j-1,k}(v_1)$ and $f_1 \cdot f_2 \in \mathrm{F}_{j,k}(u_1 \cdot x \cdot v_1)$.

To see that realizability is maintained, note that $u_1 \cdot x \cdot v_2$ is a minimal witness in which the initial segment (up through x) of the sequence of subsets of factors is from the minimal witness for $u_1 x \cdot v_1$ and the final segment (from x on) is from the minimal witness for $u_2 \cdot x v_2$. ■

Theorem 5 (Generalized Subsequence Closure)**.** *Suppose* $w \in L \in SPL_{j,k}$.
Then if

- $w = u_1 x_1 v x_2 u_2$, *where either* $|x_1| = |x_2| = k-1$ *or* $u_1 = \varepsilon$ *and* $|x_1| < k-1$ *or* $u_2 = \varepsilon$ *and* $|x_2| < k-1$
- *and* $F_{1,k}(x_1 x_2) \subseteq F_{1,k}(x_1 v x_2)$

then $u_1 x_1 x_2 u_2 \in L$.

Proof. First of all, note that whenever f is in $\mathrm{F}_{1,k}(u_1 x_1 x_2 u_2)$ then either $f \in \mathrm{F}_{1,k}(u_1 x_1)$ or $f \in \mathrm{F}_{1,k}(x_1 x_2)$ or $f \in \mathrm{F}_{1,k}(x_2 u_2)$. In each case f is also in $\in \mathrm{F}_{1,k}(u_1 x_1 v x_2 u_2)$. Thus the blocks of k consecutive factors in $u_1 x_1 x_2 u_2$ occur in the same order in $u_1 x_1 v x_2 u_2$. Consequently $\mathrm{F}_{j,k}(u_1 x_1 x_2 u_2) \subseteq \mathrm{F}_{j,k}(u_1 x_1 v x_2 u_2)$.

That this preserves realizability follows from the same reasoning as for Generalized Suffix Substitution Closure. ■

8 Learnability

Strictly Local, Strictly Piecewise and Strictly Piecewise Local Stringsets were studied in a somewhat different form in Heinz (2007), where they were shown to be learnable in the limit from positive data in the sense of Gold (1967). In Heinz (2010b) he generalizes the learning algorithm to a broad class of stringsets on the based on the notion of string extension.

Let A be a class of objects (factors, for example). A *string extension function* is a total function f, mapping Σ^* to finite subsets of A. Each finite subset of A can be interpreted as a grammar G by letting $L(G) = \{w \in \Sigma^* \mid f(w) \subseteq G\}$. Each string extension function f determines a class of stringsets $\mathcal{L}_f$, the class of all stringsets licensed by subsets of A in the range of f.

Clearly $\mathrm{F}_{j,k}$ for word models is a string extension function, with A being the set of all factors of word models, and $\mathcal{L}_{\mathrm{F}_{j,k}}$ is the class of Strictly Local, Strictly Piecewise or Strictly Piecewise Local stringsets, depending on j and k. If we take A to be the powerset of the set of all factors of word models then $f(w) \stackrel{\text{def}}{=} \{\mathrm{F}_{j,k}(w)\}$ is one, as well, and $\mathcal{L}_f$ is the class of Locally, Piecewise and Piecewise Locally Definable sets.

A *text* for a stringset L is an enumeration of $L \cup \{\#\}$ in arbitrary order, possibly with repeats. If t is a text, then $t[i]$ denotes the initial segment $t(0) \ldots t(i)$. The learning function ϕ for a string extension function f maps initial segments of a text to finite subsets of A:

$$\phi_f(t[i]) = \begin{cases} \emptyset & \text{if } i = -1 \\ \phi_f(t[i-1]) & \text{if } t(i) = \# \\ \phi_f(t[i-1]) \cup f(t(i)) & \text{otherwise.} \end{cases}$$

For sets of word models, this provides a practical learning algorithm.

More generally, F_k for arbitrary $\mathbb{R}$ structures is an extension function for that class of structures. The issue in those cases is where the the enumeration of members of the set comes from. For non-phonotactic linguistic applications, it essentially requires an annotated sample. If the sample is less than fully characteristic, the learned grammar will undergenerate. On the other hand, in all cases it is useful even if the set is non-PLT. It will learn a set of constraints that define the minimal PLT approximation of that set. For an example of the usefulness of these constraints see Rogers and Lambert (2017, to appear).

9 Some Examples from Phonology

In Section 2 we discussed the automata-theoretic patterns in the StressTyp2 database. The six that are Star-Free and require something more than SL+co-SL+SP can each be shown to include just one additional LT constraint of the form if stress falls on a final syllable that is heavy, then a syllable of some other type (an unstressed heavy, for example) does not occur. Formally these constraints can be expressed as $\acute{\mathrm{H}} \ltimes \rightarrow \neg X$, which is logically equivalent to $\neg(X \wedge \acute{\mathrm{H}} \ltimes)$. Since if this fails the X must precede the ultimate $\acute{\mathrm{H}}$, we can capture this in $\mathrm{SPL}_{2,2}$ with the constraint$\neg(X < \acute{\mathrm{H}} \ltimes)$.

Since both SL and SP constraints are expressible as SPL constraints all of these stress patterns, other than the two lects of Arabic, are definable in SPL + co-SL. This is significant from a cognitive perspective because in order to check constraints of these forms a mechanism needs only to attend

to factors that actually are present, in isolation, in the input string. (See Rogers et al. (2012) for more on this notion of cognitive complexity.)

9.1 Separating PLT from SF and TSL

In their simplest form (Heinz et al., 2011), Tier-based Strictly Local (TSL) constraints are based on a subset of the input alphabet (the *tier* alphabet) along with strictly local constraints in terms of that alphabet. Operationally, the input string is subjected to an alphabetic homomorphism which erases all symbols except for those in the tier alphabet and the remaining string is checked against the SL constraint. The TSL stringsets are all Star-free, properly include the SL stringsets but are incomparable with the LT, PT and SP stringsets, although the intersection of TSL and SP includes long distance phonotactic patterns derived from asymmetric assimilation processes (Heinz, 2010a).

The canonical separation between TSL and these classes is long distance phonotactic dissimulation patterns. As an example of the application of the closure conditions in Section 7.2, we can show that these patterns are not SPL or even PLT.

Latin liquid dissimulation (LLD): every pair of 'l's is separated by at least one 'r' and every pair of 'r's is separated by at least one 'l':

$$\begin{aligned}(\forall x, y)[\quad & (x < y \wedge l(x) \wedge l(y)) \\ & \rightarrow (\exists z)[x < z \wedge z < y \wedge r(z)]\ ] \\ \wedge \\ (\forall x, y)[\quad & (x < y \wedge r(x) \wedge r(y)) \\ & \rightarrow (\exists z)[x < z \wedge z < y \wedge l(z)]\ ]\end{aligned}$$

This definition demonstrates that LLD is SF. It is also TSL based on the tier alphabet $\{l, r\}$ and the constraint $\neg(rr) \wedge \neg(ll)$.

We can demonstrate that it is not $\mathrm{SPL}_{j,k}$ for any j and k using either Generalized Suffix Substitution Closure (GSSC) or Generalized Subsequence Closure (GSSeqC),

9.1.1 Using GSSC

Let

$w_1 = \rtimes(s^{jk}ls^{jk}r)^{jk} \cdot s^{jk} \cdot ls^{jk}r(s^{jk}ls^{jk}r)\ltimes$

and

$w_2 = \rtimes(s^{jk}ls^{jk}r)^{jk}s^{jk}l \cdot s^{jk} \cdot r(s^{jk}ls^{jk}r)\ltimes.$

Both $w_1,\ w_2 \in L$, but

$\rtimes(s^{jk}ls^{jk}r)^{jk} \cdot s^{jk} \cdot r(s^{jk}ls^{jk}r)\ltimes \notin L.$

Therefore, LLD is not $\mathrm{SPL}_{j,k}$ for any j and k.

9.1.2 Using GSSeqC

Let $w_3 \in L$ be a similar string, divided into $u_1x_1vx_2u_2$ as follows:

$w_3 =$
$\rtimes(s^{jk}ls^{jk}r)^{jk} \cdot s^{k-1} \cdot s^{k}ls^{k} \cdot s^{k-1} \cdot r(s^{jk}ls^{jk}r)\ltimes$

Then $|x_1| = |x_2| = k - 1$ and $\mathrm{F}_{1,k}(x_1x_2) \subseteq \mathrm{F}_{1,k}(x_1vx_2)$, but

$w_4 = u_1x_1x_2u_2$
$= \rtimes(s^{jk}ls^{jk}r)^{jk} \cdot s^{k-1} \cdot s^{k-1} \cdot r(s^{jk}ls^{jk}r)\ltimes$

is not in L.

9.1.3 Using $\equiv_{(j,k)}$.

It is not hard to see that $[\mathcal{W}_3]_{(j,k)} = [\mathcal{W}_4]_{(j,k)}$ (equivalently $\mathrm{F}_{j,k}(\mathcal{W}_3) = \mathrm{F}_{j,k}(\mathcal{W}_4)$), where $\mathcal{W}_3$ and $\mathcal{W}_4$ are word models of w_3 and w_4 equivalently. But w_3 satisfies LLD, while w_4 does not.

10 Conclusion

We have explored the model theory of a type of propositional logic based on factors (connected fragments) of structures defined as labeled purely relational models and given characterizations of the Locally and Strictly Locally Definable sets of these structures. Using those tools, we have derived a characterization of the SPL and PLT definable stringsets, which completes the characterization of the propositional levels of the main sequence of the Piecewise Local hierarchy (See Figure 1).

SPL extends SL and SP by adding, on the one hand, precedence constraints and, on the other, adjacency constraints. The interplay of constraints of these types motivated the original definition of TSL and continues to motivate extensions of the class. But TSL remains incomparable with the sub-Star-Free part of the hierarchy. Ultimately, we hope to find a class of structures that will allow us to incorporate TSL in a natural way.

More importantly, we expect that these model-theoretic tools, when applied to trees and other types of labeled graphs will provide insight into local accounts of autosegmental structures (Jardine, 2017) and other multi-tiered structures as well as model-theoretic accounts of syntactic constraints (e.g. Rogers (1998); Graf (2018)).

Acknowledgments

The authors are indebted to Jeff Heinz, Larry Moss and the anonymous referees for detailed and extremely helpful comments.

References

D. Beauquier and Jean-Eric Pin. 1991. Languages and scanners. *Theoretical Computer Science*, 84:3–21.

J.A. Brzozowski and R. Knast. 1978. The dot-depth hierarchy of star-free languages is infinite. *Journal of Computer and System Sciences*, 16(1):37 – 55.

J.A. Brzozowski and I. Simon. 1973. Characterization of locally testable events. *Discrete Math*, 4:243–271.

J. R. Büchi. 1960. Weak second-order arithmetic and finite automata. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 6:66–92.

Calvin C. Elgot. 1961. Decision problems of finite automata design and related arithmetics. *Transactions of the American Mathematical Society*, 98:21–51.

R. W. Goedemans, Jeffrey Heinz, and Harry van der Hulst. 2015. `http://st2.ullet.net/files/files/st2-v1-archive-0415.tar.gz`. Retrieved 24 Jun 2015.

E.M. Gold. 1967. Language identification in the limit. *Information and Control*, 10:447–474.

Thomas Graf. 2018. Why movement comes for free once you have adjunction. In *Proceedings of CLS 53*, pages 117–137.

Jeffrey Heinz. 2007. *The Inductive Learning of Phonotactic Patterns*. Ph.D. thesis, University of California, Los Angeles.

Jeffrey Heinz. 2010a. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623–661.

Jeffrey Heinz. 2010b. String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 897–906, Uppsala, Sweden. Association for Computational Linguistics.

Jeffrey Heinz, Chetan Rawal, and Herbert G. Tanner. 2011. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 58–64, Portland, Oregon, USA. Association for Computational Linguistics.

Graham Higman. 1952. Ordering by divisibility in abstract algebras. *Proceedings of the London Mathematical Society*, s3-2(1):326–336.

Wilfird Hodges. 1993. *Model Theory*. Cambridge University Press, Cambridge, UK.

Adam Jardine. 2017. The local nature of tone-association patterns. *Phonology*, 34:385–405.

J. B. Kruskal. 1960. Well-quasi-ordering, the tree theorem, and Vazsonyi's conjecture. *Transactions of the American Mathematical Society*, 95:210–225.

Leonid Libkin. 2004. *Elements of Finite Model Theory*. Texts in Theoretical Computer Science. Springer, Berlin and New York.

Saunders MacLane and Garrett Birkhoff. 1967, 1970. *Algebra*. Macmillan, New York.

Robert McNaughton and Seymour Papert. 1971. *Counter-Free Automata*. MIT Press.

Yu. T. Medvedev. 1964. On the class of events representable in a finite automaton. In Edward F. Moore, editor, *Sequential Machines; Selected Papers*, pages 215–227. Addison-Wesley. Originally published in Russian in Avtomaty, 1956, 385–401.

James Rogers. 1998. *A Descriptive Approach to Language-Theoretic Complexity*. CSLI Publications, Stanford, CA.

James Rogers, Jeff Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert, and Sean Wibel. 2012. Cognitive and sub-regular complexity. In Glyn Morrill and Mark-Jan Nederhof, editors, *Formal Grammar 2012*, volume 8036 of *Lecture Notes in Computer Science*, pages 90–108. Springer.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2010. On languages piecewise testable in the strict sense. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *The Mathematics of Language: 10th and 11th Biennial Conference, MOL 10, Los Angeles, CA, USA, July 28-30, 2007, and MOL 11, Bielefeld, Germany, August 20-21, 2009, Revised Selected Papers*, pages 255–265. Springer Berlin Heidelberg, Berlin, Heidelberg.

James Rogers and Dakotah Lambert. 2017. Extracting forbidden factors from regular stringsets. In *Proceedings of the 15th Meeting on the Mathematics of Language*, pages 36–46. Association for Computational Linguistics.

James Rogers and Dakotah Lambert. to appear. Extracting subregular constraints from regular stringsets. In press.

M.P. Schützenberger. 1965. On finite monoids having only trivial subgroups. *Information and Control*, 8(2):190 – 194.

Imre Simon. 1975. Piecewise testable events. In *Automata Theory and Formal Languages: 2nd Grammatical Inference conference*, pages 214–222, Berlin. Springer-Verlag.

Howard Straubing. 1985. Finite semigroup varieties of the form $\boldsymbol{v}^*\boldsymbol{d}$. *Journal of Pure and Applied Algebra*, 36:53–94.

Howard Straubing. 1994. *Finte Automata, Formal Logic, and Circuit Complexity*. "Birkhäuser".

Wolfgang Thomas. 1978. The theory of successor with and extra predicate. *Mathematische Annalen*, 237:121–232.

Wolfgang Thomas. 1982. Classifying regular events in symbolic logic. *Journal of Computer and Systems Sciences*, 25:360–376.

Denis Thérien and Alex Weiss. 1985. Graph congruences and wreath products. *Journal of Pure and Applied Algebra*, 36:205 – 215.