Slide 1

Cognitive Complexity in the Sub-Regular Realm

James Rogers
Dept. of Computer Science
Earlham College
jrogers@cs.earlham.edu

http://cs.earlham.edu/~jrogers/slides/UCLA.ho.pdf

Joint work with Jeff Heinz, U. Delaware,
Geoff Pullum and Barbara Scholz, U.Edinburgh,
and a raft of Earlham College undergrads.

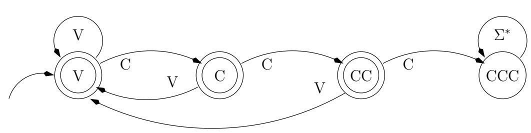
This work completed, in part, while in residence at the Radcliffe Institute for Advanced Study

Slide 2

Yawelmani Yokuts (Kissberth'73)

$$\star\, CCC$$

$$\overline{\Sigma^* CCC \Sigma^*}$$



Contrast: $\star\, C^{2i+1}$

Definition 1 *A* finite-state stringset *is one in which there is an* a priori *bound, independent of the length of the string, on the amount of information that must be inferred in distinguishing strings in the set from those not in the set.*

Regular = Recognizable = Finite-State

Slide 3

Cognitive Complexity from First Principles

What kinds of distinctions does a cognitive mechanism need to be sensitive to in order to classify an event with respect to a pattern?

Reasoning about patterns

- What objects/entities/things are we reasoning about?
- What relationships between them are we reasoning with?

Slide 4

Dual characterizations of complexity classes

Computational classes

- Characterized by abstract computational mechanisms
- Equivalence between mechanisms
- Means to determine structural properties of stringsets

Descriptive classes

- Characterized by the nature of information about the properties of strings that determine membership
- Independent of mechanisms for recognition
- Subsume wide range of types of patterns

Slide 5

Some Assumptions about Linguistic Behaviors

- Perceive/process/generate linear sequence of (sub)events
- Can model as strings—linear sequence of abstract symbols
  - Discrete linear order (initial segment of $\mathbb{N}$).
  - Labeled with alphabet of events

    Partitioned into subsets, each the set of positions at which some event occurs.

Slide 6

Word models

$$\langle \mathcal{D}, \triangleleft, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$$

(+1) $\langle \mathcal{D}, \triangleleft, P_\sigma \rangle_{\sigma \in \Sigma}$ (<) $\langle \mathcal{D}, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$

$\mathcal{D}$	—	Finite
$\triangleleft^+$	—	Linear order on $\mathcal{D}$
$\triangleleft$	—	Successor wrt $\triangleleft^+$
P_σ	—	Subset of $\mathcal{D}$ at which σ occurs
		(P_σ partition $\mathcal{D}$)

$$CCVC = \Big\langle \{0,1,2,3\},\ \{\langle i, i+1\rangle \mid 0 \leq i < 3\},\ \{0,1,3\}_C,\ \{2\}_V \Big\rangle$$

Slide 7

Adjacency—Substrings

$$CVCVCV$$

Definition 2 (k-Factor)
v is a factor *of w if $w = uvx$ for some $u, v \in \Sigma^*$.*

*v is a k-*factor *of w if it is a factor of w and $|v| = k$.*

$$F_k(w) \stackrel{def}{=} \begin{cases} \{v \in \Sigma^k \mid (\exists u, x \in \Sigma^*)[w = uvx]\} & \textit{if } |w| \geq k, \\ \{w\} & \textit{otherwise.} \end{cases}$$

$$F_2(CVCVCV) = \{CV, VC\}$$
$$F_7(CVCVCV) = \{CVCVCV\}$$

Slide 8

Strictly Local Stringsets—SL

Strictly k-Local Definitions
—Grammar is set of permissible k-factors

$$\mathcal{G} \subseteq F_k(\{\rtimes\} \cdot \Sigma^* \cdot \{\ltimes\})$$
$$w \models \mathcal{G} \stackrel{\text{def}}{\iff} F_k(\rtimes \cdot w \cdot \ltimes) \subseteq \mathcal{G}$$
$$L(\mathcal{G}) \stackrel{\text{def}}{=} \{w \mid w \models \mathcal{G}\}$$

Definition 3 (Strictly Local Sets) *A stringset L over Σ is* Strictly Local *iff there is some strictly k-local definition $\mathcal{G}$ over Σ (for some k) such that L is the set of all strings that satisfy $\mathcal{G}$*

Slide 9

SL Hierarchy

Definition 4 (SL)
A stringset is Strictly k-Local *if it is definable with an* SL_k *definition.*

A stringset is Strictly Local *(in SL) if it is* SL_k *for some* k*.*

Theorem 1 (SL-Hierarchy)

$$SL_2 \subsetneq SL_3 \subsetneq \cdots \subsetneq SL_i \subsetneq SL_{i+1} \subsetneq \cdots \subsetneq SL$$

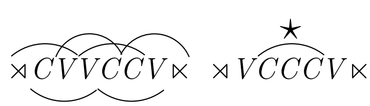
Every Finite stringset is SL_k for some k: Fin $\subseteq$ SL.

There is no k for which SL_k includes all Finite languages.

Slide 10

$\star$ CCC is SL_3

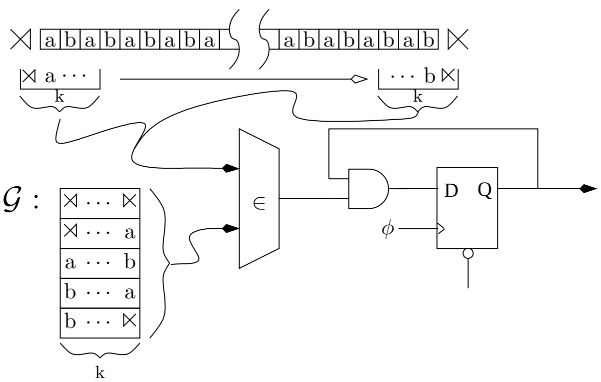
$$\mathcal{G}_{\neg CCC} = F_3(\{\rtimes\} \cdot \Sigma^* \cdot \{\ltimes\}) - \{CCC\}$$



Membership in an SL_k stringset depends only on the individual k-factors which occur in the string.

Slide 11

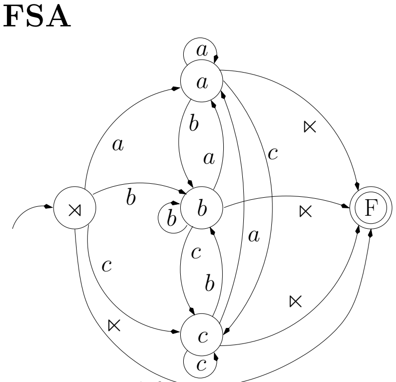
Scanners



Recognizing an SL_k stringset requires only remembering the k most recently encountered symbols.

Slide 12

Scanners as FSA



$$\mathcal{M} \stackrel{\text{def}}{=} \langle Q, \Sigma, q_0, \delta, F \rangle$$

$$Q \stackrel{\text{def}}{=} F_{k-1}(\{\rtimes\} \cdot \Sigma^* \cdot \{\ltimes\} \cup \bigcup_{0 \le i < k-1} [\{\rtimes\} \cdot \Sigma^i]$$

$$q_o \stackrel{\text{def}}{=} \rtimes$$

$$\delta(\sigma \cdot v, \gamma) \stackrel{\text{def}}{=} v \cdot \gamma, \qquad \sigma \in \{\rtimes\} \cup \Sigma,\ \gamma \in \Sigma \cup \{\ltimes\}$$

$$F \stackrel{\text{def}}{=} \{v \cdot \ltimes \mid v \cdot \ltimes \in Q\}$$

Slide 13

Character of Strictly k-Local Sets

Theorem (Suffix Substitution Closure):
A stringset L is strictly k-local iff whenever there is a string x of length $k-1$ and strings w, y, v, and z, such that

$$\begin{array}{ccccccc} w & \cdot & \overbrace{x}^{k-1} & \cdot & y & \in L \\ v & \cdot & x & \cdot & z & \in L \end{array}$$

then it will also be the case that

$$w \cdot x \cdot z \in L$$

E.g.:

$$\begin{array}{ccccc} V \cdot & VC & \cdot & CV & \in \star CCC \\ C \cdot & VC & \cdot & VC & \in \star CCC \\ \hline V \cdot & VC & \cdot & VC & \in \star CCC \end{array}$$

But $\star CCC$ is not SL_2:

$$\begin{array}{ccccc} C \cdot & C & \cdot & VC & \in \star CCC \\ V \cdot & C & \cdot & CV & \in \star CCC \\ \hline C \cdot & C & \cdot & CV & \notin \star CCC \end{array}$$

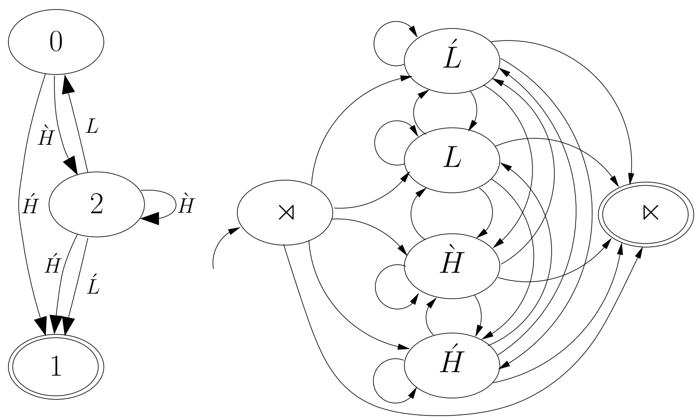
Slide 14

Cognitive interpretation of SL

- Any cognitive mechanism that can distinguish member strings from non-members of an SL_k stringset must be sensitive, at least, to the length k blocks of events that occur in the presentation of the string.
- If the strings are presented as sequences of events in time, then this corresponds to being sensitive, at each point in the string, to the immediately prior sequence of $k-1$ events.
- Any cognitive mechanism that is sensitive *only* to the length k blocks of events in the presentation of a string will be able to recognize *only* SL_k stringsets.

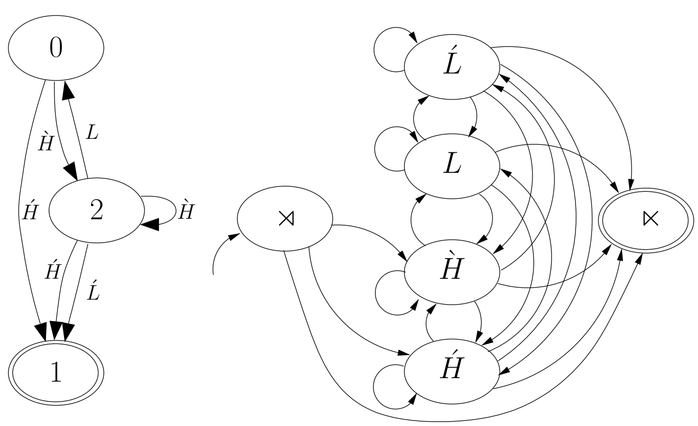
Slide 15

Cambodian



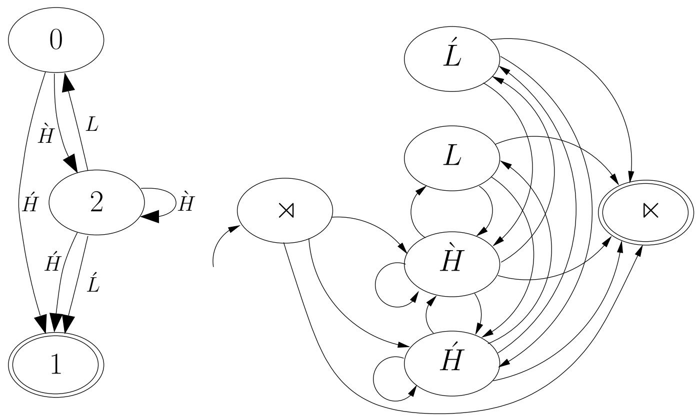
Slide 16

Cambodian—Can’t start with a light syllable



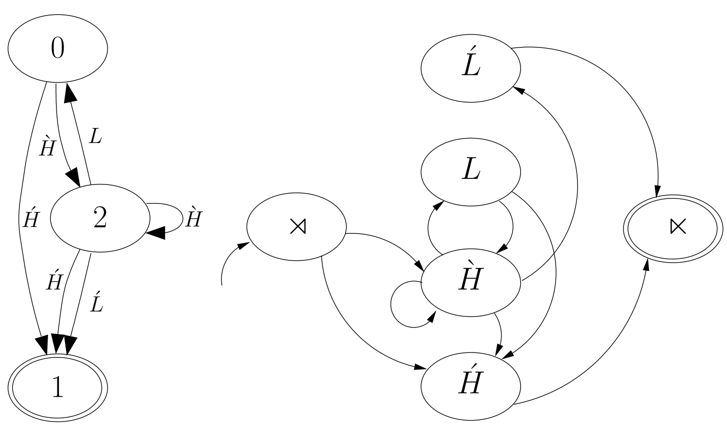
Slide 17

Cambodian—No light follows another light



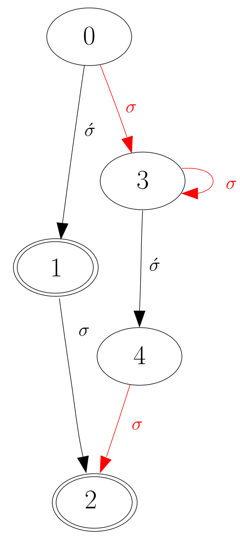
Slide 18

Cambodian—Must end on primary stress



Slide 19

Alawa

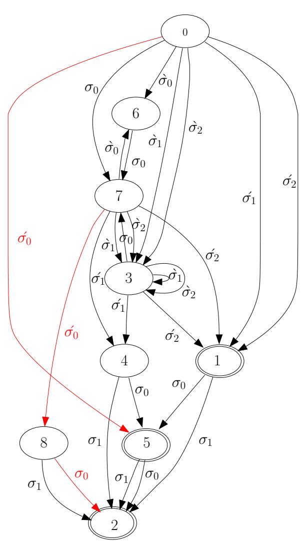


$\rtimes\sigma$	σ	$\acute{\sigma}\sigma\ltimes$
$\rtimes\acute{\sigma}$	σ	$\ltimes$
$\star\ \rtimes\sigma$	σ	$\ltimes$

$$\mathcal{G}_{\text{Alawa}} = \{\ \rtimes\sigma\sigma,\ \rtimes\sigma\acute{\sigma},\ \rtimes\acute{\sigma}\sigma,\ \sigma\sigma\sigma,\ \sigma\sigma\acute{\sigma},\ \sigma\acute{\sigma}\sigma,\ \rtimes\acute{\sigma}\ltimes,\ \acute{\sigma}\sigma\ltimes\ \}$$

Slide 20

Arabic (Bani-Hassan)

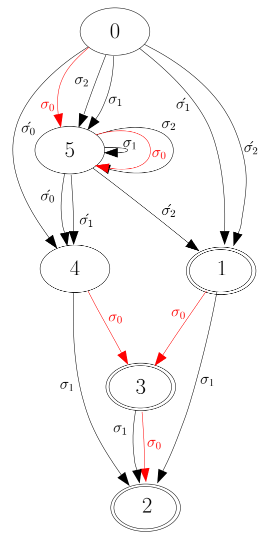


$\rtimes\sigma_0$	$\acute{\sigma_0}$	$\sigma_0\ltimes$
$\rtimes$	$\acute{\sigma_0}$	$\ltimes$
$\star\ \rtimes\sigma_0$	$\acute{\sigma_0}$	$\ltimes$

$$\mathcal{G}_{\text{ArabicBH}} = \{\cdots\} - \{\sigma\acute{\sigma_0}\ltimes \mid \sigma \in \sigma_0, \sigma_1, \sigma_2\}$$

Slide 21

Arabic (Cairene)

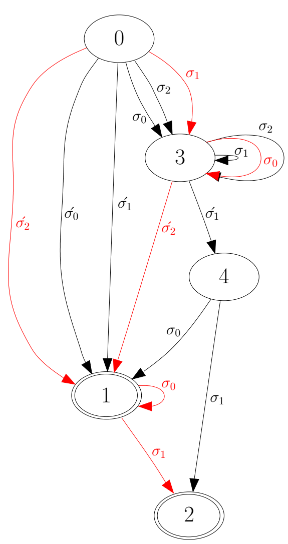


$\rtimes\sigma_0$	$\sigma_0\sigma_0$	$\acute{\sigma}_2\ltimes$
$\rtimes\acute{\sigma}_0$	$\sigma_0\sigma_0$	$\ltimes$
$\star\ \rtimes\sigma_0$	$\sigma_0\sigma_0$	$\ltimes$

$$\mathcal{G}_{\text{ArabicCai}} = \{\cdots\} - \{\sigma\sigma\sigma\ltimes \mid \sigma \in \sigma_0, \sigma_1, \sigma_2\}$$

Slide 22

Arabic (Classical)



$\rtimes\sigma_1$	$\overbrace{\sigma_0\cdots\sigma_0}^{k-1}$	$\acute{\sigma}_2\ltimes$
$\rtimes\acute{\sigma}_2$	$\overbrace{\sigma_0\cdots\sigma_0}^{k-1}$	$\sigma_1\ltimes$
$\star\ \rtimes\sigma_1$	$\overbrace{\sigma_0\cdots\sigma_0}^{k-1}$	$\sigma_1\ltimes$

Slide 23

Strictly Local Stress Patterns

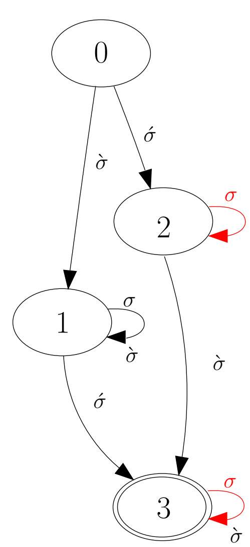
Heinz's Stress Pattern Database (ca. 2007)—109 patterns

9 are SL_2	Abun West, Afrikans, . . . Cambodian,. . . Maranungku
44 are SL_3	Alawa, Arabic (Bani-Hassan),. . .
24 are SL_4	Arabic (Cairene),. . .
3 are SL_5	Asheninca, Bhojpuri, Hindi (Fairbanks)
1 is SL_6	Icua Tupi
28 are not SL	Amele, Bhojpuri (Shukla Tiwari), Arabic Classical, Hindi (Keldar), Yidin,. . .

72% are SL, all $k \leq 6$. 49% are SL_3.

Slide 24

The Problematic Case—Some-$\grave{\sigma}$



$\rtimes\acute{\sigma}$	$\overbrace{\sigma\cdots\sigma}^{k-1}$	$\grave{\sigma}\ltimes$
$\rtimes\acute{\sigma}\grave{\sigma}$	$\overbrace{\sigma\cdots\sigma}^{k-1}$	$\sigma\ltimes$
$\star\ \acute{\sigma}$	$\overbrace{\sigma\cdots\sigma}^{k-1}$	$\sigma\ltimes$

Slide 25

Locally definable stringsets

$$\begin{array}{ccll} f \in F_k(\rtimes \cdot \Sigma^* \cdot \ltimes) & w \models f & \stackrel{\text{def}}{\Longleftrightarrow} & f \in F_k(\rtimes \cdot w \cdot \ltimes) \\ \varphi \wedge \psi & w \models \varphi \wedge \psi & \stackrel{\text{def}}{\Longleftrightarrow} & w \models \varphi \text{ and } w \models \psi \\ \neg\varphi & w \models \neg\varphi & \stackrel{\text{def}}{\Longleftrightarrow} & w \not\models \varphi \end{array}$$

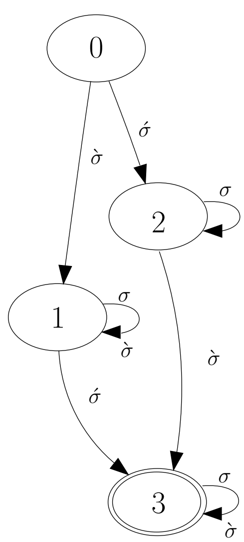
Definition 5 (Locally Testable Sets) *A stringset L over Σ is* Locally Testable *iff (by definition) there is some k-expression φ over Σ (for some k) such that L is the set of all strings that satisfy φ:*

$$L = L(\varphi) \stackrel{def}{=} \{w \in \Sigma^* \mid w \models \varphi\}$$

$$\mathrm{SL}_k \equiv \bigwedge_{f_i \notin \mathcal{G}} [\neg f_i] \subsetneq \mathrm{LT}_k$$

Slide 26

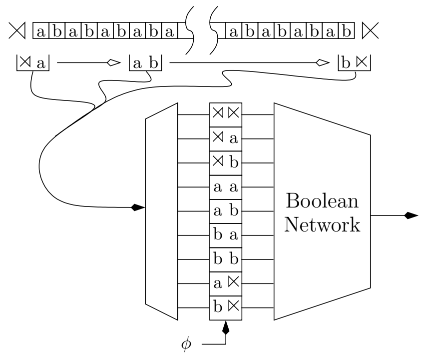
Some-$\grave{\sigma}$



$$\begin{array}{lll} \varphi_{\text{Some-}\grave{\sigma}} = & & \\ & (\rtimes\acute{\sigma} \vee \acute{\sigma}\ltimes) & \text{Starts or ends with } \acute{\sigma} \\ \wedge & & \\ & \grave{\sigma} & \text{Some } \grave{\sigma} \end{array}$$

Slide 27

LT Automata



Membership in an LT_k stringset depends only on the set of k-Factors which occur in the string.

Recognizing an LT_k stringset requires only remembering which k-factors occur in the string.

Slide 28

Character of Locally Testable sets

Theorem 2 (k-Test Invariance) *A stringset L is Locally Testable iff*

there is some k such that, for all strings x and y,

if $\rtimes \cdot x \cdot \ltimes$ and $\rtimes \cdot y \cdot \ltimes$ have exactly the same set of k-factors

then either both x and y are members of L or neither is.

$$w \equiv_k^L v \overset{\text{def}}{\iff} F_k(\rtimes w \ltimes) = F_k(\rtimes v \ltimes).$$

Slide 29

LT Hierarchy

Definition 6 (*LT*)
A stringset is k-Locally Testable *if it is definable with an* LT_k*-expression.*

A stringset is Locally Testable *(in LT) if it is* LT_k *for some* k*.*

Theorem 3 (LT-Hierarchy)

$$LT_2 \subsetneq LT_3 \subsetneq \cdots \subsetneq LT_i \subsetneq LT_{i+1} \subsetneq \cdots \subsetneq LT$$

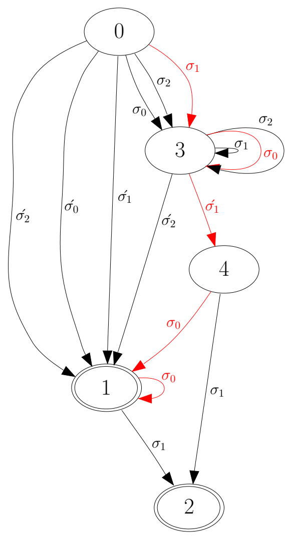
Slide 30

Cognitive interpretation of LT

- Any cognitive mechanism that can distinguish member strings from non-members of an LT_k stringset must be sensitive, at least, to the set of length k blocks of events that occur in the presentation of the string—both those that do occur and those that do not.
- If the strings are presented as sequences of events in time, then this corresponds to being sensitive, at each point in the string, to the length k blocks of events that occur at any prior point.
- Any cognitive mechanism that is sensitive *only* to the set of length k blocks of events in the presentation of a string will be able to recognize *only* LT_k stringsets.

Slide 31

Arabic (Classical)



$$\rtimes\sigma_1 \overbrace{\sigma_0\cdots\sigma_0}^{k-1} \acute{\sigma}_1 \overbrace{\sigma_0\cdots\sigma_0}^{k-1} \ltimes$$

$$\equiv^L_k$$

$$\rtimes\sigma_1 \overbrace{\sigma_0\cdots\sigma_0}^{k-1} \acute{\sigma}_1 \overbrace{\sigma_0\cdots\sigma_0}^{k-1} \acute{\sigma}_1 \overbrace{\sigma_0\cdots\sigma_0}^{k-1} \ltimes$$

Slide 32

FO(+1)

Models: $\langle \mathcal{D}, \triangleleft, P_\sigma \rangle_{\sigma\in\Sigma}$

First-order Quantification (over positions in the strings)

$x \triangleleft y$	$w, [x \mapsto i, y \mapsto j] \models x \triangleleft y$	$\stackrel{\text{def}}{\iff}$	$j = i + 1$
$P_\sigma(x)$	$w, [x \mapsto i] \models P_\sigma(x)$	$\stackrel{\text{def}}{\iff}$	$i \in P_\sigma$
$\varphi \wedge \psi$	$\vdots$		
$\neg\varphi$	$\vdots$		
$(\exists x)[\varphi(x)]$	$w, s \models (\exists x)[\varphi(x)]$	$\stackrel{\text{def}}{\iff}$	$w, s[x \mapsto i] \models \varphi(x)]$ for some $i \in \mathcal{D}$

FO(+1)-Definable Stringsets: $L(\varphi) \stackrel{\text{def}}{=} \{w \mid w \models \varphi\}$.

$$\text{One-}\acute{\sigma} = L((\exists x)[\acute{\sigma}(x) \wedge (\forall y)[\acute{\sigma}(y) \rightarrow x \approx y]\,])$$

Arabic (Classical) is FO(+1)

Slide 33

Character of the FO(+1) Definable Stringsets

Definition 7 (Locally Threshold Testable) *A set L is* Locally Threshold Testable *(LTT) iff there is some k and t such that, for all $w, v \in \Sigma^*$:*

if for all $f \in F_k(\rtimes \cdot w \cdot \ltimes) \cup F_k(\rtimes \cdot v \cdot \ltimes)$
either $|w|_f = |v|_f$ or both $|w|_f \geq t$ and $|v|_f \geq t$,

then $w \in L \iff v \in L$.

Theorem 4 (Thomas) *A set of strings is First-order definable over $\langle \mathcal{D}, \triangleleft, P_\sigma \rangle_{\sigma \in \Sigma}$ iff it is* Locally Threshold Testable.

Membership in an FO(+1) definable stringset depends only on the multiplicity of the k-factors, up to some fixed finite threshold, which occur in the string.

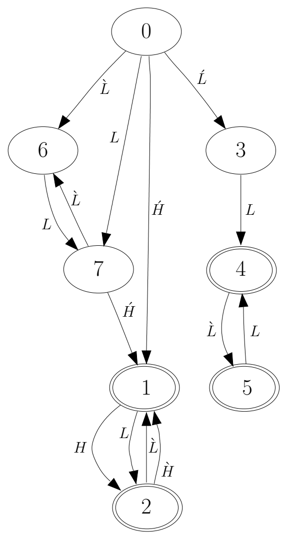
Slide 34

Cognitive interpretation of FO(+1)

- Any cognitive mechanism that can distinguish member strings from non-members of an FO(+1) stringset must be sensitive, at least, to the multiplicity of the length k blocks of events, for some fixed k, that occur in the presentation of the string, distinguishing multiplicities only up to some fixed threshold t.
- If the strings are presented as sequences of events in time, then this corresponds to being able count up to some fixed threshold.
- Any cognitive mechanism that is sensitive *only* to the multiplicity, up to some fixed threshold, (and, in particular, not to the order) of the length k blocks of events in the presentation of a string will be able to recognize *only* FO(+1) stringsets.

Slide 35

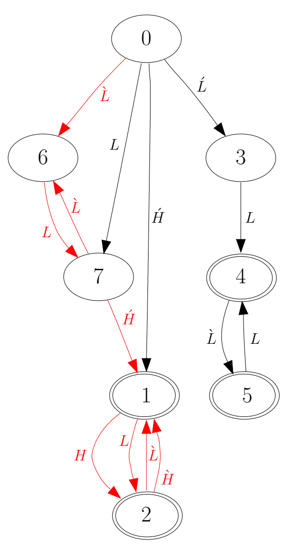
Yidin



- Exactly one $\acute{\sigma}$ (One-$\acute{\sigma}$)
- First H gets primary stress (No-H-before-$\acute{H}$)
- σ and $\grave{\sigma}$ alternate ($(\sigma\grave{\sigma})^*$)
- $\acute{L}$ only if initial (Nothing-before-$\acute{L}$)
- $\acute{L}$ implies no H (No-H-with-$\acute{L}$)
- $\acute{L}$ must be followed by L (L-follows-$\acute{L}$)

Slide 36

Yidin is not FO(+1)



$$\rtimes \overbrace{\grave{L}L\cdots\grave{L}L}^{2kt}\, \acute{H}H\, \overbrace{\grave{L}L\cdots\grave{L}L}^{2kt}\, \grave{H}H\, \overbrace{\grave{L}L\cdots\grave{L}L}^{2kt} \ltimes$$

$$\equiv^{L}_{k,t}$$

$$\star \rtimes \underbrace{\grave{L}L\cdots\grave{L}L}_{2kt}\, \grave{H}H\, \overbrace{\grave{L}L\cdots\grave{L}L}^{2kt}\, \acute{H}H\, \underbrace{\grave{L}L\cdots\grave{L}L}_{2kt} \ltimes$$

- no-H-before-$\acute{H}$ is not FO(+1)
- One-$\acute{\sigma}$ is FO(+1)
- No-H-with-$\acute{L}$ is LT.
- $(\sigma\grave{\sigma})^*$, Nothing-before-$\acute{L}$, and L-follows-$\acute{L}$ are all SL_2.

Slide 37

Long-Distance Dependencies

Sarcee sibilant harmony:

[-anterior] sibilants do not occur after [+anterior] sibilants

a. /si-tʃiz-aʔ/ → **ʃ**í**tʃ**íd**z**àʔ 'my duck'

b. /na-s-ɣatʃ/ → nā**ʃ**ɣát**ʃ** 'I killed them again'

c. cf. ⋆**s**í**tʃ**íd**z**àʔ

$$\overline{\Sigma^* \cdot [+] \cdot \Sigma^* \cdot [\text{-}] \cdot \Sigma^*}$$

Samala (Chumash) sibilant harmony:

[-anterior] sibilants do not occur in the same word as [+anterior] sibilants

[**ʃ**tojonowonowa**ʃ**] 'it stood upright' *[**ʃ**tojonowonowa**s**]

$$\overline{(\Sigma^* \cdot [+] \cdot \Sigma^* \cdot [\text{-}] \cdot \Sigma^*) + (\Sigma^* \cdot [\text{-}] \cdot \Sigma^* \cdot [+] \cdot \Sigma^*)}$$

Slide 38

Complexity of Sibilant Harmony

(Samala and Sarcee)

Symmetric sibilant harmony is LT

$$\neg([+] \wedge [-])$$

Asymmetric sibilant harmony is not FO(+1)

$$\rtimes w\, [-]\, w\, [+]\, w \ltimes$$
$$\equiv^{L}_{k,t}$$
$$\star \rtimes w\, [-]\, w\, [+]\, w\, [-]\, w \ltimes$$

Slide 39

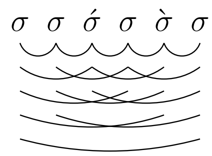
Precedence—Subsequences

Definition 8 (Subsequences)

$$v \sqsubseteq w \overset{def}{\Longleftrightarrow} v = \sigma_1 \cdots \sigma_n \text{ and } w \in \Sigma^* \cdot \sigma_1 \cdot \Sigma^* \cdots \Sigma^* \cdot \sigma_n \cdot \Sigma^*$$

$$P_k(w) \overset{def}{=} \{v \in \Sigma^k \mid v \sqsubseteq w\}$$

$$P_{\leq k}(w) \overset{def}{=} \{v \in \Sigma^{\leq k} \mid v \sqsubseteq w\}$$



$\sigma\sigma$, $\sigma\acute{\sigma}$, $\acute{\sigma}\sigma$, $\sigma\grave{\sigma}$, $\grave{\sigma}\sigma$
$\sigma\acute{\sigma}$, $\sigma\sigma$, $\acute{\sigma}\grave{\sigma}$
$\sigma\sigma$, $\sigma\grave{\sigma}$, $\acute{\sigma}\sigma$
$\sigma\grave{\sigma}$, $\sigma\sigma$
$\sigma\sigma$

$$P_2(\sigma\sigma\acute{\sigma}\sigma\grave{\sigma}\sigma) = \{\sigma\sigma, \sigma\acute{\sigma}, \sigma\grave{\sigma}, \acute{\sigma}\sigma, \acute{\sigma}\grave{\sigma}, \grave{\sigma}\sigma\}$$

$$P_{\leq 2}(\sigma\sigma\acute{\sigma}\sigma\grave{\sigma}\sigma) = \{\varepsilon, \sigma, \acute{\sigma}, \grave{\sigma}, \sigma\sigma, \sigma\acute{\sigma}, \sigma\grave{\sigma}, \acute{\sigma}\sigma, \acute{\sigma}\grave{\sigma}, \grave{\sigma}\sigma\}$$

Slide 40

Strictly Piecewise Stringsets—SP

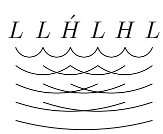
Strictly k-Piecewise Definitions

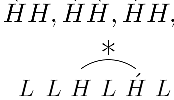
$$\mathcal{G} \subseteq \Sigma^{\leq k}$$

$$w \models \mathcal{G} \overset{\text{def}}{\Longleftrightarrow} P_{\leq k}(w) \subseteq P_{\leq k}(\mathcal{G})$$

$$L(\mathcal{G}) \overset{\text{def}}{=} \{w \in \Sigma^* \mid w \models \mathcal{G}\}$$

$$\mathcal{G}_{\text{No-}H\text{-before-}\acute{H}} = \{HH, H\grave{H}, \grave{H}H, \grave{H}\grave{H}, \acute{H}H, \acute{H}\grave{H}, \ldots\}$$





Membership in an SP_k stringset depends only on the individual ($\leq k$)-subsequences which do and do not occur in the string.

Slide 41

Character of the Strictly k-Piecewise Sets

Theorem 5 *A stringset L is Strictly k-Piecewise Testable iff, for all $w \in \Sigma^*$,*

$$P_{\leq k}(w) \subseteq P_{\leq k}(L) \Rightarrow w \in L$$

Consequences:

Subsequence Closure:	$w\sigma v \in L \Rightarrow wv \in L$
Unit Strings:	$P_1(L) \subseteq L$
Empty String:	$L \neq \emptyset \Rightarrow \varepsilon \in L$

Every naturally occurring stress pattern requires Primary Stress
$\Rightarrow$
No naturally occurring stress pattern is SP.

But SP can forbid multiple primary stress: $\neg\acute{\sigma}\acute{\sigma}$

Slide 42

SP Hierarchy

Definition 9 (SP)
A stringset is Strictly k-Piecewise *if it is definable with an SP_k definition.*

A stringset is Strictly Piecewise *(in SP) if it is SP_k for some k.*

Theorem 6 (SP-Hierarchy)

$$SP_2 \subsetneq SP_3 \subsetneq \cdots \subsetneq SP_i \subsetneq SP_{i+1} \subsetneq \cdots \subsetneq SP$$

SP is incomparable (wrt subset) with the Local Hierarchy

$\mathrm{SP}_2 \not\subseteq \mathrm{FO}(+1)$	No-H-before-$\acute{H} \in \mathrm{SP}_2 - \mathrm{FO}(+1)$
$\mathrm{SL}_2 \not\subseteq \mathrm{SP}$	$(\sigma\grave{\sigma})^* \in \mathrm{SL}_2 - \mathrm{SP}$
$\mathrm{SP}_2 \cap \mathrm{SL}_2 \neq \emptyset$	$A^*B^* \in \mathrm{SP}_2 \cap \mathrm{SL}_2$
$\mathrm{Fin} \not\subseteq \mathrm{SP}$	$\{A\} \in \mathrm{Fin} - \mathrm{SP}$

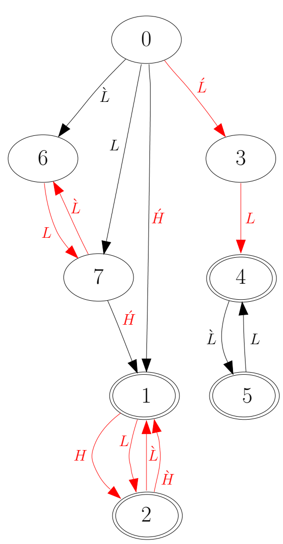
Slide 43

Sarcee Sibilant Harmony is SP$_2$

$$\{\ldots, [-]\,[-],\ [-]\,[+],\ [+]\,[+], \ldots\}$$

Slide 44

Yidin constraints wrt SP



- No-H-before-$\acute{H}$ is SP$_2$: Forbid $H\acute{H}$
- Nothing-before-$\acute{L}$ is SP$_2$: Forbid $\Sigma\acute{L}$
- One-$\acute{\sigma}$ is not SP: $\star\ \sigma\sigma\grave{\sigma}\ \sqsubseteq\ \sigma\acute{\sigma}\sigma\grave{\sigma}$
- $(\sigma\grave{\sigma})^*$ is not SP: $\star\ \sigma\sigma\grave{\sigma}\ \sqsubseteq\ \sigma\grave{\sigma}\sigma\grave{\sigma}$
- L-follows-$\acute{L}$ is not SP: $\star\ \acute{L}\grave{L}\ \sqsubseteq\ \acute{L}L\grave{L}$

Slide 45

Cognitive interpretation of SP

- Any cognitive mechanism that can distinguish member strings from non-members of an SP_k stringset must be sensitive, at least, to the length k (not necessarily consecutive) sequences of events that occur in the presentation of the string.
- If the strings are presented as sequences of events in time, then this corresponds to being sensitive, at each point in the string, to up to $k-1$ events distributed arbitrarily among the prior events.
- Any cognitive mechanism that is sensitive *only* to the length k sequences of events in the presentation of a string will be able to recognize *only* SP_k stringsets.

Slide 46

k-Piecewise Testable Stringsets

PT_k-expressions

$$\begin{array}{ccll} p \in \Sigma^{\leq k} & w \models p & \stackrel{\text{def}}{\iff} & p \sqsubseteq w \\ \varphi \wedge \psi & w \models \varphi \wedge \psi & \stackrel{\text{def}}{\iff} & w \models \varphi \text{ and } w \models \psi \\ \neg\varphi & w \models \neg\varphi & \stackrel{\text{def}}{\iff} & w \not\models \varphi \end{array}$$

k-Piecewise Testable Languages (PT_k):

$$L(\varphi) \stackrel{\text{def}}{=} \{w \in \Sigma^* \mid w \models \varphi\}$$

$$\text{One-}\acute{\sigma} = L(\acute{\sigma} \wedge \neg\acute{\sigma}\acute{\sigma})$$

Membership in an PT_k stringset depends only on the set of ($\leq k$)-subsequences which occur in the string.

SP_k is equivalent to $\bigwedge_{p_i \notin \mathcal{G}} [\neg p_i]$

Slide 47

Character of Piecewise Testable sets

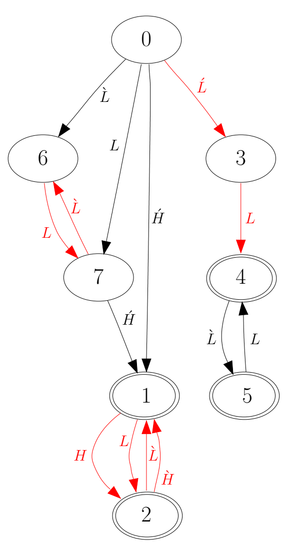
Theorem 7 (k-Subsequence Invariance) *A stringset L is Piecewise Testable iff*

> *there is some k such that, for all strings x and y,*
>
> > *if x and y have exactly the same set of $(\leq k)$-subsequences*
> >
> > *then either both x and y are members of L or neither is.*

$$w \equiv^P_k v \overset{\text{def}}{\iff} P_{\leq k}(w) = P_{\leq k}(v).$$

Slide 48

Yidin constraints wrt PT



- No-H-before-$\acute{H}$ is SP_2:
  Forbid $H\acute{H}$
- Nothing-before-$\acute{L}$ is SP_2:
  Forbid $\Sigma\acute{L}$
- One-$\acute{\sigma}$ is PT_2:
  Require $\acute{\sigma}$, Forbid $\acute{\sigma}\acute{\sigma}$
- $(\sigma\grave{\sigma})^*$ is not PT:
  $\overbrace{\sigma\grave{\sigma}\cdots\sigma\grave{\sigma}}^{2k} \equiv^P_k \overbrace{\sigma\grave{\sigma}\cdots\sigma\grave{\sigma}}^{2k}\grave{\sigma}$
- L-follows-$\acute{L}$ is not PT:
  $\acute{L}L\overbrace{\grave{L}L\cdots\grave{L}L}^{2k} \equiv^P_k \acute{L}L\overbrace{\grave{L}L\cdots\grave{L}L}^{2k}$

Slide 49

PT Hierarchy

Definition 10 (SP)
A stringset is k-Piecewise Testable *if it is definable with an* PT_k *definition.*

A stringset is Piecewise Testable *(in PT) if it is* PT_k *for some* k*.*

Theorem 8 (PT-Hierarchy)

$$PT_2 \subsetneq PT_3 \subsetneq \cdots \subsetneq PT_i \subsetneq PT_{i+1} \subsetneq \cdots \subsetneq PT$$

Slide 50

PT, SP and the Local Hierarchy

$$\begin{aligned} \mathrm{SP}_k &\subsetneq \mathrm{PT}_k & & \\ \mathrm{SP}_{k+1} &\not\subseteq \mathrm{PT}_k & & \\ \mathrm{PT}_2 &\not\subseteq \mathrm{SP} & & \text{One-}\acute{H} \in \mathrm{PT}_2 - \mathrm{SP} \\ \mathrm{PT}_2 &\not\subseteq \mathrm{FO}(+1) & & \text{No-}H\text{-before-}\acute{H} \in \mathrm{PT}_2 - \mathrm{FO}(+1) \\ \mathrm{SL}_2 &\not\subseteq \mathrm{PT} & & (\sigma\grave{\sigma})^* \in \mathrm{SL}_2 - \mathrm{PT} \\ \mathrm{PT}_2 \cap \mathrm{SL}_2 &\neq \emptyset & & A^*B^* \in \mathrm{PT}_2 \cap \mathrm{SL}_2 \end{aligned}$$

$\mathrm{Fin} \subseteq \mathrm{SP}$:

$$\Sigma^* = L(\varepsilon), \quad \emptyset = L(\neg\varepsilon), \quad \{\varepsilon\} = L(\bigwedge_{\sigma\in\Sigma}[\neg\sigma]),$$

$$\{w\} = L(w \wedge \bigwedge_{p\in\Sigma^{|w|+1}}[\neg p])$$

$$\{w_1,\ldots,w_n\} = L(\bigvee_{1\leq i\leq n}[w_i \wedge \bigwedge_{p\in\Sigma^{|w_i|+1}}[\neg p]])$$

Slide 51

Cognitive interpretation of PT

- Any cognitive mechanism that can distinguish member strings from non-members of an PT_k stringset must be sensitive, at least, to the set of length k subsequences of events that occur in the presentation of the string—both those that do occur and those that do not.
- If the strings are presented as sequences of events in time, then this corresponds to being sensitive, at each point in the string, to the set of all length k subsequences of the sequence of prior events.
- Any cognitive mechanism that is sensitive *only* to the set of length k subsequences of events in the presentation of a string will be able to recognize *only* PT_k stringsets.

Slide 52

First-Order(<) definable stringsets

$\langle \mathcal{D}, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$

First-order Quantification over positions in the strings

$$
\begin{array}{lrcl}
x \triangleleft^+ y & w,[x \mapsto i, y \mapsto j] \models x \triangleleft^+ y & \stackrel{\text{def}}{\iff} & i < j \\
P_\sigma(x) & w,[x \mapsto i] \models P_\sigma(x) & \stackrel{\text{def}}{\iff} & i \in P_\sigma \\
\varphi \wedge \psi & \vdots & & \\
\neg\varphi & \vdots & & \\
(\exists x)[\varphi(x)] & w,s \models (\exists x)[\varphi(x)] & \stackrel{\text{def}}{\iff} & w,s[x \mapsto i] \models \varphi(x) \\
 & & & \text{for some } i \in \mathcal{D}
\end{array}
$$

Slide 53

PT, FO(+1) and FO(<)

Theorem 9 *PT* $\subsetneq$ *FO*(<).

$$\sigma_1 \cdots \sigma_n \sqsubseteq w \Leftrightarrow (\exists x_1, \ldots, x_n)[\bigwedge_{1 \leq i < j \leq n} [x_i \triangleleft^+ x_j] \wedge \bigwedge_{1 \leq i \leq n} [P_{\sigma_i}(x_i)]\ ]$$

$$(\sigma\grave{\sigma})^* \subseteq \mathrm{FO}(<) - \mathrm{PT}$$

Theorem 10 *FO*(+1) $\subsetneq$ *FO*(<).

+1 is FO definable from <:

$$x \triangleleft y \equiv x \triangleleft^+ y \wedge \neg(\exists z)[x \triangleleft^+ z \wedge z \triangleleft^+ y]$$

$$\text{No-}H\text{-before-}\acute{H} \subseteq \mathrm{FO}(<) - \mathrm{FO}(+1)$$

Slide 54

Star-Free stringsets

Definition 11 (Star-Free Set) *The class of* Star-Free Sets *(SF) is the smallest class of languages satisfying:*

- *Fin* $\subseteq$ *SF.*
- *If* $L_1, L_2 \in$ *SF then:* $L_1 \cdot L_2 \in$ *SF,*
$L_1 \cup L_2 \in$ *SF,*
$\overline{L_1} \in$ *SF.*

Theorem 11 (McNauthton and Papert) *A set of strings is First-order definable over* $\langle \mathcal{D}, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$ *iff it is* Star-Free.

Slide 55

PT and LT with Order

$$\varphi \bullet \psi \qquad w \models \varphi \bullet \psi \overset{\text{def}}{\Longleftrightarrow} w = w_1 \cdot w_2, \quad w_1 \models \varphi \text{ and } w_2 \models \psi.$$

LTO_k is LT_k plus $\varphi \bullet \psi$

$$\text{No-}H\text{-before-}\acute{H} = L((\neg H) \bullet (\neg \acute{H})) \in \text{LTO}$$

PTO_k is PT_k plus $\varphi \bullet \psi$

Let:

$$\varphi_{A=^i} = A^i \wedge \bigwedge_{p \in \Sigma^{i+1}} [\neg p], \qquad \varphi_{\Sigma^*} = \varepsilon$$

$$L(\varphi_{A=^i}) = \{A^i\} \qquad L(\varphi_{\Sigma^*}) = \Sigma^*$$

Then:

$$\begin{aligned} (\sigma\grave{\sigma})^* \quad = \quad & L(\neg(\varphi_{\grave{\sigma}=^1} \bullet \varphi_{\Sigma^*}) \wedge \neg(\varphi_{\Sigma^*} \bullet \varphi_{\sigma=^1}) \wedge \\ & \neg(\varphi_{\Sigma^*} \bullet \varphi_{\sigma=^2} \bullet \varphi_{\Sigma^*}) \wedge \neg(\varphi_{\Sigma^*} \bullet \varphi_{\grave{\sigma}=^2} \bullet \varphi_{\Sigma^*})) \quad \in \text{PTO} \end{aligned}$$

Slide 56

PTO, LTO and SF

Theorem 12

$$PTO = SF = LTO$$

SF $\subseteq$ PTO, SF $\subseteq$ LTO

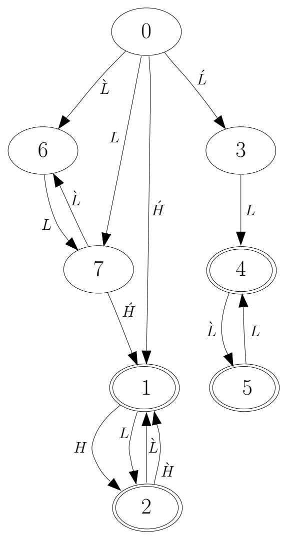
Fin $\subseteq$ PTO, Fin $\subseteq$ LTO and both are closed under concatenation, union and complement.

LTO $\subseteq$ PTO $\subseteq$ SF

Concatenation is FO($<$) definable.

Slide 57

Yidin is FO(<)



- No-H-before-$\acute{H}$ is SP_2: Forbid $H\acute{H}$
- Nothing-before-$\acute{L}$ is SP_2: Forbid $\Sigma\acute{L}$
- One-$\acute{\sigma}$ is PT_2: Require $\acute{\sigma}$, Forbid $\acute{\sigma}\acute{\sigma}$
- $(\sigma\grave{\sigma})^*$ is SL_2: $\{\rtimes\sigma,\ \sigma\grave{\sigma},\ \grave{\sigma}\sigma,\ \grave{\sigma}\ltimes\}$
- L-follows-$\acute{L}$ is SL_2: $\neg\{\acute{L}H,\ \acute{L}\grave{H},\ \acute{L}\acute{H},\ \acute{L}\grave{L},\ \acute{L}\acute{L}\}$

Yidin is $\mathrm{SL}_2 \cap \mathrm{PT}_2$.
Yidin is $\mathrm{LT}_2 \cap \mathrm{SP}_2$.

Slide 58

Character of FO(<) definable sets

Theorem 13 (McNaughton and Papert) *A stringset L is definable by a set of First-Order formulae over strings iff it is recognized by a finite-state automaton that is* non-counting *(that has an* aperiodic *syntactic monoid), that is, iff:*

there exists some $n > 0$ such that

for all strings u, v, w over Σ

if $uv^n w$ occurs in L

then $uv^{n+i}w$, for all $i \geq 1$, occurs in L as well.

E.g.

$$\frac{\{\textbf{people (who were left by people)}^n\ \textbf{left}\} \quad \in L}{\{\textbf{people (who were left by people)}^{n+1}\ \textbf{left}\} \quad \in L}$$

Slide 59

Cognitive interpretation of FO(<)

- Any cognitive mechanism that can distinguish member strings from non-members of an FO(<) stringset must be sensitive, at least, to the sets of length k blocks of events, for some fixed k, that occur in the presentation of the string when it is factored into segments, up to some fixed number, on the basis of those sets with distinct criteria applying to each segment.
- If the strings are presented as sequences of events in time, then this corresponds to being able to count up to some fixed threshold with the counters being reset some fixed number of times based on those counts.
- Any cognitive mechanism that is sensitive *only* to the sets of length k blocks of events in the presentation of a string once it has been factored in this way will be able to recognize *only* FO(<) stringsets.

Slide 60

MSO definable stringsets

$\langle \mathcal{D}, \triangleleft, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$

First-order Quantification (positions)

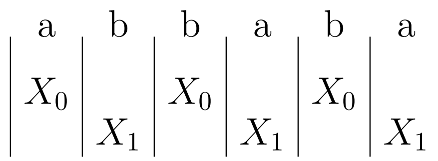
Monadic Second-order Quantification (sets of positions)

$\triangleleft^+$ is MSO-definable from $\triangleleft$.

Slide 61

MSO example

$$
\begin{aligned}
(\exists X_0, X_1)[\quad & (\forall x)[(\exists y)[y \triangleleft x] \vee X_0(x)] \wedge \\
& (\forall x, y)[\neg(X_0(x) \wedge X_1(x))] \wedge \\
& (\forall x, y)[x \triangleleft y \rightarrow (X_0(x) \leftrightarrow X_1(y))] \wedge \\
& (\forall x)[(\exists y)[x \triangleleft y] \vee X_1(x)] \qquad]
\end{aligned}
$$



Slide 62

Theorem 14 (Chomsky Schützenberger) *A set of strings is Regular iff it is a homomorphic image of a Strictly 2-Local set.*

Definition 12 (Nerode Equivalence) *Two strings w and v are* Nerode Equivalent *with respect to a stringset L over Σ (denoted $w \equiv_L v$) iff for all strings u over Σ, $wu \in L \Leftrightarrow vu \in L$.*

Theorem 15 (Myhill-Nerode) *A stringset L is recognizable by a FSA (over strings) iff $\equiv_L$ partitions the set of all strings over Σ into finitely many equivalence classes.*

Theorem 16 (Medvedev, Büchi, Elgot) *A set of strings is MSO-definable over $\langle \mathcal{D}, \triangleleft, \triangleleft^+, P_\sigma \rangle_{\sigma \in \Sigma}$ iff it is regular.*

Theorem 17 *MSO $=$ $\exists$MSO over strings.*

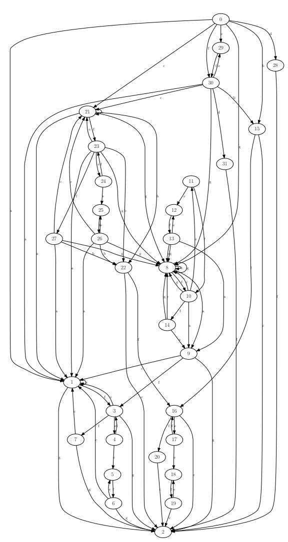
Slide 63

Cognitive interpretation of Finite-state

- Any cognitive mechanism that can distinguish member strings from non-members of a finite-state stringset must be capable of classifying the events in the input into a finite set of abstract categories and are sensitive to the sequence of those categories.
- Subsumes *any* recognition mechanism in which the amount of information inferred or retained is limited by a fixed finite bound.
- Any cognitive mechanism that has a fixed finite bound on the amount of information inferred or retained in processing sequences of events will be able to recognize *only* finite-state stringsets.

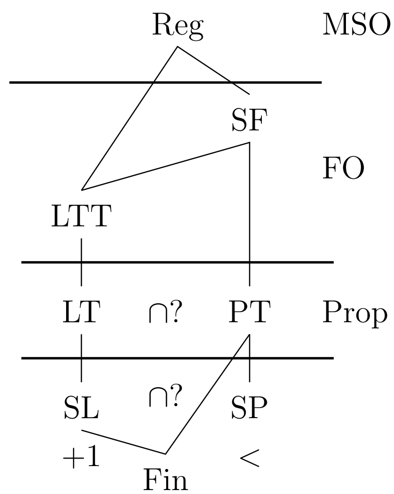
Slide 64

Hindi (Kelkar)



Slide 65

Local and Piecewise Hierarchies



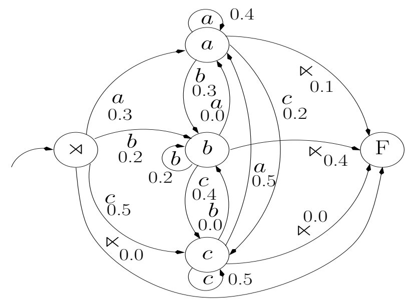
Slide 66

Complexity of some phonological constraints

		MSO (Reg)	Hindi (Kellkar)?		
		FO(<) (SF)	(Yidin)		
FO(+1)	?				
LT				PT	
LT_2	Some-$\acute{\sigma}$, Symmetric SH	$LT_2 \cap PT_2$	Yidin	PT_2	One-$\acute{\sigma}$
SL				SP	
SL_6	72%				
SL_4	Arabic (Cariene)				
SL_3	$\star\, CCC$, Alawa, Arabic (Bani-Hassan), 49%				
SL_2	Cambodian			SP_2	Asymmetric SH, No-H-before-$\acute{H}$, Nothing-before-$\acute{L}$

Slide 67

n-gram Models of Language



$$\Pr_L(\sigma_1 \cdots \sigma_n) = \Pr_L(\sigma_1 \mid \rtimes) \cdot \prod_{1<i\leq n} [\Pr_L(\sigma_i \mid \sigma_i - 1)] \cdot \Pr_L(\ltimes \mid \sigma_n)$$

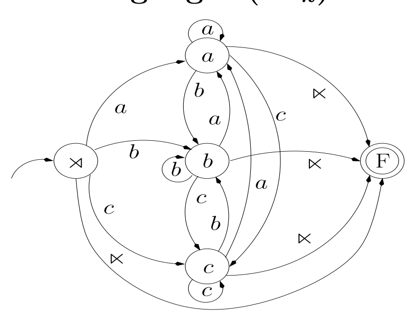
$$F_k(w) \stackrel{\text{def}}{=} \{v \in \Sigma^k \mid w \in \Sigma^* \cdot v \cdot \Sigma^*\}$$

$$F_k^M(w) \stackrel{\text{def}}{=} \{\!\{v \in \Sigma^k \mid w \in \Sigma^* \cdot v \cdot \Sigma^*\}\!\}$$

$$\Pr_L(w) = \prod_{v\cdot\sigma \in F_k^M(\rtimes\cdot w\cdot\ltimes)} [\Pr_L(\sigma \mid v)]$$

Slide 68

Strictly k-Local Languages (SL$_k$)



$$T_{\mathcal{M}} \stackrel{\text{def}}{=} \{v\sigma \in F_k(\rtimes \cdot \Sigma^* \cdot \ltimes) \mid \delta(v, \sigma)\downarrow\}$$

$$L(\mathcal{M}) = \{w \in \Sigma^* \mid F_k(w) \subseteq T_{\mathcal{M}}\}$$

$$L \in \mathrm{SL}_k \stackrel{\text{def}}{\iff} L \text{ is } L(\mathcal{M}) \text{ for some } k\text{-scanner } \mathcal{M}$$

$$L \in \mathrm{SL} \stackrel{\text{def}}{\iff} (\exists k)[L \in \mathrm{SL}_k]$$

Slide 69

Subsequences

v **is a *subsequence* of** w**:**

$$v \sqsubseteq w \overset{\text{def}}{\iff} v = \sigma_1 \cdots \sigma_k \text{ and } w \in \Sigma^* \cdot \sigma_1 \cdot \Sigma^* \cdots \Sigma^* \cdot \sigma_k \cdot \Sigma^*$$

$$P_k(w) \overset{\text{def}}{=} \{v \in \Sigma^k \mid v \sqsubseteq w\} \qquad P_{\leq k}(w) \overset{\text{def}}{=} \bigcup_{0<i\leq k} [P_i(w)]$$

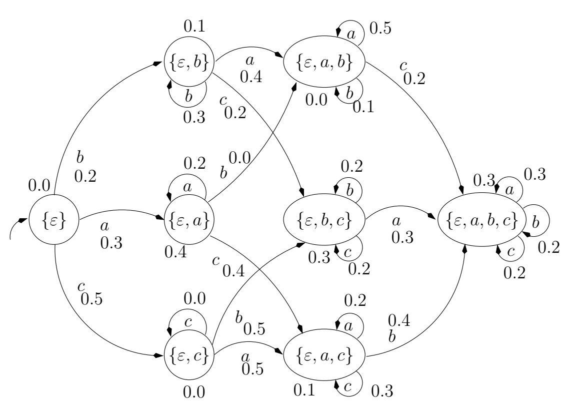
$$P_k^M(w) \overset{\text{def}}{=} \{\!\{v \sqsubseteq w\}\!\}$$

Would like:

$$\Pr_L(w) = \prod_{v\cdot\sigma \in P^M_{\leq k}(w)} [\Pr_L(\sigma \mid v)]$$

Slide 70

Initial Model



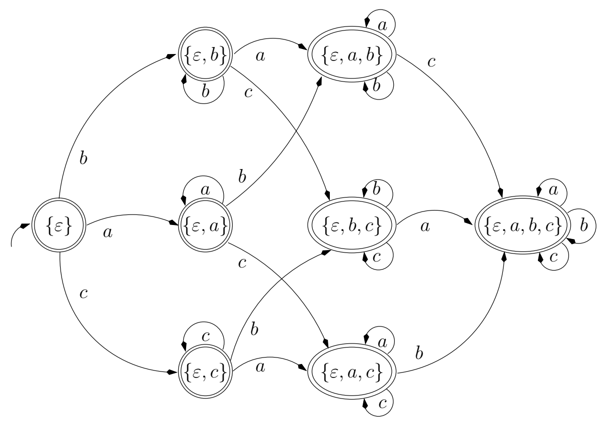
$$Q = \mathcal{P}(P_{\leq k}(\Sigma^*))$$

Let $w = v \cdot \sigma \cdot u,\ q = \hat{\delta}(\{\varepsilon\}, v)$:

$$T(q, \sigma) = \Pr_L(\sigma \mid P_{\leq k}(v) = q)$$

Slide 71

PT-Automata



Slide 72

Piecewise-Testable Languages (PT)

$$\mathrm{SI}(w) \stackrel{\text{def}}{=} \{v \in \Sigma^* \mid w \sqsubseteq v\}$$

L is Piecewise Testable $\stackrel{\text{def}}{\iff}$ L is a finite Boolean combination of principal shuffle ideals.

P_k-expressions

Atoms $v \in P_{\leq k}(\Sigma^*)$

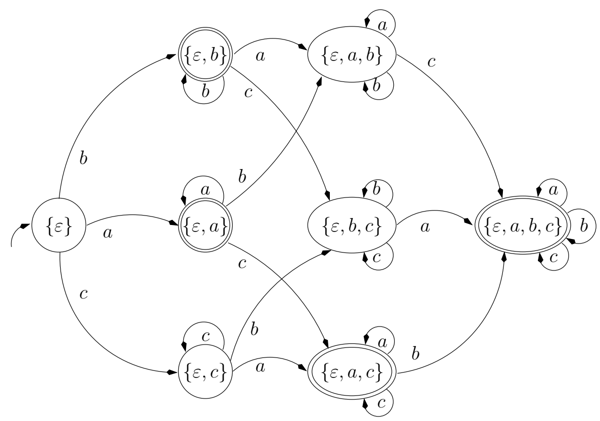
$$w \models v \stackrel{\text{def}}{\iff} w \in \mathrm{SI}(v) \qquad \text{(i.e., } v \sqsubseteq w\text{)}$$

Operators Truth functional connectives

$$L \in \mathrm{PT}_k \Leftrightarrow L = \{w \in \Sigma^* \mid w \models \varphi\} \text{ for some } P_k\text{-expression } \varphi$$

Slide 73

PT-Automata and P_k-expressions



$$F_\varphi = \{q \in \mathcal{P}(P_{\leq k}(\Sigma^*)) \mid (\bigwedge_{s \in q}[s] \wedge \bigwedge_{s \notin q}[\neg s]) \to \varphi\}$$

$$L(\mathcal{M}_\varphi) = \{w \in \Sigma^* \mid w \models \phi\}$$

Slide 74

Strictly Piecewise Testable Languages (SP)

The following are equivalent:

1. $L \in \text{SP}$
2. L is the set of strings satisfying a finite conjunction of negative P_k-literals.
3. $L = \bigcap_{w \in S}[\overline{\text{SI}(w)}],\ S$ finite,
4. $(\exists k)[P_{\leq k}(w) \subseteq P_{\leq k}(L) \Rightarrow w \in L]$,
5. $w \in L$ and $v \sqsubseteq w \Rightarrow v \in L$ (L is *subsequence closed*),
6. $L = \overline{\text{SI}(X)},\ X \subseteq \Sigma^*$ (L is the complement of a shuffle ideal).

Slide 75

DFA representation of SP$_k$ languages

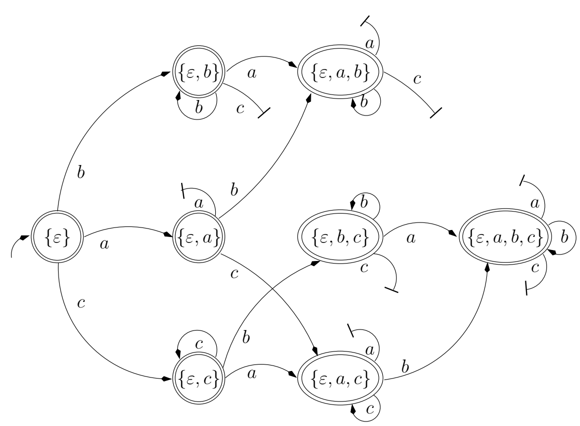
Let $\mathcal{M}$ be a trimmed minimal DFA recognizing an SP$_k$ language. Then:

1. All states of $\mathcal{M}$ are accepting states.
2. If $\delta(q, \sigma)\uparrow$ then there is some $s \in P_{\leq k}(\{w \mid \hat{\delta}(q_0, w) = q\})$ such that for all $q' \in Q$ $s \in P_{\leq k}(\{w \mid \hat{\delta}(q_0, w) = q'\}) \Rightarrow \delta(q, \sigma)\uparrow$

Consequently, for all $q_1, q_2 \in Q$ and $\sigma \in \Sigma$, if $\delta(q_1, \sigma)\uparrow$ and $\hat{\delta}(q_1, w) = q_2$ for some $w \in \Sigma^*$ then $\delta(q_2, \sigma)\uparrow$.
(Missing edges propagate down.)

Slide 76

SP$_k$-automata

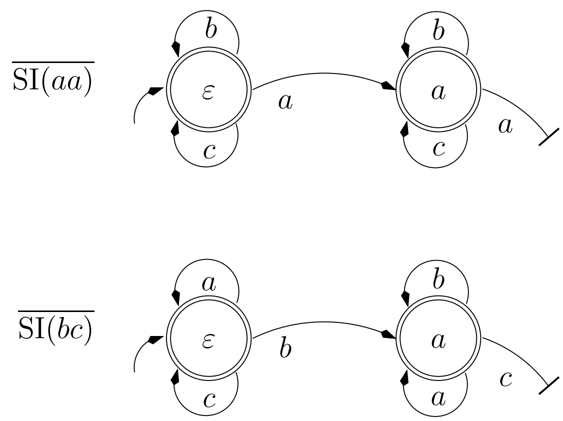


$$Q = \mathcal{P}(P_{\leq k-1}(\Sigma^*))$$

Size of automaton: $\Theta(2^{\mathbf{card}(\Sigma)^k})$

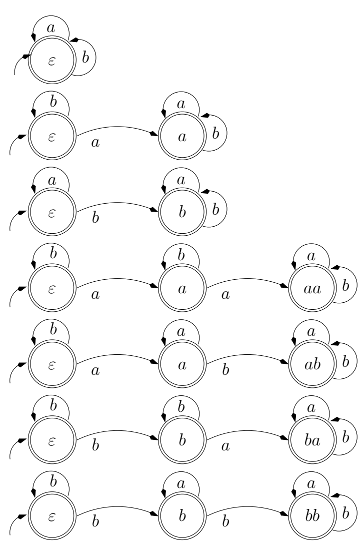
Slide 77

Factored SP$_k$-automata



Slide 78

SP-PDFA



Slide 79

Product PDFAs

Co-emission Probability

$$\mathrm{CT}(\langle \sigma, q_1 \ldots q_n \rangle) = \Pi_{i=1}^{n} T_i(q_i, \sigma)$$

$$\mathrm{CF}(\langle q_1 \ldots q_n \rangle) = \Pi_{i=1}^{n} F_i(q_i)$$

$$Z(\langle q_1 \ldots q_n \rangle) = \mathrm{CF}(\langle q_1 \ldots q_n \rangle) + \sum_{\sigma \in \Sigma} \mathrm{CT}(\langle \sigma, q_1 \ldots q_n \rangle)$$

$$F(\langle q_1 \ldots q_n \rangle) = \frac{\mathrm{CF}(\langle q_1 \ldots q_n \rangle)}{Z(\langle q_1 \ldots q_n \rangle)}$$

$$T(\langle q_1 \ldots q_n \rangle, \sigma) = \frac{\mathrm{CT}(\langle \sigma, q_1 \ldots q_n \rangle)}{Z(\langle q_1 \ldots q_n \rangle)}$$

Slide 80

Product PDFAs—*k*-sets

Positive Co-emission Probability

$$\mathrm{PCT}(\langle \sigma, q_\epsilon \ldots q_u \rangle) = \prod_{\substack{q_w \in \langle q_\epsilon \ldots q_u \rangle \\ q_w = w}} T_w(q_w, \sigma)$$

$$\mathrm{PCF}(\langle q_\epsilon \ldots q_u \rangle) = \prod_{\substack{q_w \in \langle q_\epsilon \ldots q_u \rangle \\ q_w = w}} F_w(q_w)$$

$$Z(\langle q_1 \ldots q_n \rangle) = \mathrm{PCF}(\langle q_1 \ldots q_n \rangle) + \sum_{\sigma \in \Sigma} \mathrm{PCT}(\langle \sigma, q_1 \ldots q_n \rangle)$$

Let $q = \langle \epsilon, \epsilon, b, aa, a, ba, b \rangle$:

$$\begin{aligned} \mathrm{CT}(a, q) &= T_\epsilon(\epsilon, a) \cdot T_a(\epsilon, a) \cdot T_b(b, a) \cdot \\ &\quad T_{aa}(aa, a) \cdot T_{ab}(a, a) \cdot T_{ba}(ba, a) \cdot T_{bb}(b, a) \\ \mathrm{PCT}(a, q) &= T_\epsilon(\epsilon, a) \cdot T_b(b, a) \cdot T_{aa}(aa, a) \cdot T_{ba}(ba, a) \end{aligned}$$

Slide 81

Complexity

Number of automata:

$$\sum_{0 \le i < k} [\mathbf{card}(\Sigma)^i] = \Theta(\mathbf{card}(\Sigma)^{k-1})$$

Number of states:

$$\sum_{0 \le i < k} [(i+1)\,\mathbf{card}(\Sigma)^i] = \Theta(k\,\mathbf{card}(\Sigma)^{k-1})$$

ML estimation $n = \sum_{w \in S}[|w|]$—size of corpus

$$\Theta(n\,\mathbf{card}(\Sigma)^{k-1}) \qquad \text{(v.s. } \Theta(n))$$

$\mathbf{Pr}_L(w)$

$$\Theta(n\,\mathbf{card}(\Sigma)^{k-1}) \qquad \text{(v.s. } \Theta(n))$$

Parameters Only final states matter

$$\mathbf{card}(\Sigma)\Theta(\mathbf{card}(\Sigma)^{k-1}) = \Theta(\mathbf{card}(\Sigma)^k) \qquad (\textit{Same})$$

Slide 82

Remaining issues

- Estimation undercounts
  - counts number of k-sequences that start with first prefix—$\Theta(n)$
  - actual number $\binom{n}{k} \in \Theta(2^n)$.
- Want probability to depend on *multiset* of subsequences
  - infinitely many states
  - but probability of n occurrences is (probability of occurrence)^n
  - same number of parameters/still linear time
- *Not* Regular distribution
  - Not clear that there is a corresponding class of distributions over strings

Slide 83

Summary

SP-Distributions

- Regular distribution
  Model (some) long distance dependencies
- Asymptotic complexity same as SL-distributions (n-gram models)
- SL-distributions can't model long distance dependencies
  SP-distributions can't model local ones
- Both are classes of Regular distributions
  Combination is straightforward

Slide 84

Samala Corpus

- 4800 words drawn from Applegate 2007, generously provided in electronic form by Applegate (p.c).

35 Consonants

	labial	coronal	a.palatal	velar	uvular	glottal
stop	p p^{ʔ} p^{h}	t t^{ʔ} t^{h}		k k^{ʔ} k^{h}	q q^{ʔ} q^{h}	ʔ
affricates		t͡s t͡s^{ʔ} t͡s^{h}	t͡ʃ t͡ʃ^{ʔ} t͡ʃ^{h}			
fricatives		s s^{ʔ} s^{h}	ʃ ʃ^{ʔ} ʃ^{h}	x x^{ʔ}		h
nasal	m	n n^{ʔ}				
lateral		l l^{ʔ}				
approx.	w	y				

6 Vowels

i	ɨ	u
e		o
	a	

(Applegate 1972, 2007)

Slide 85

Samala: results of SP2 estimation

$P(x \mid \{y\} <)$		x			
		t͡ʃ	ʃ	t͡s	s
y	t͡ʃ	0.0313	0.0455	0.	0.0006
	ʃ	0.0353	0.0671	0.	0.0009
	t͡s	0.	0.0009	0.0113	0.0218
	s	0.0002	0.0011	0.0051	0.0335

(Collapsing laryngeal distinctions)

Slide 86

Finnish: Corpus

- 44,040 words from Goldsmith and Riggle (to appear)

19 Consonants

	lab.	lab.dental	cor.	pal.	velar	uvular	glottal
stop	p b		t d	c	k g	q	
fricatives		f v	s		x		h
nasal	m		n				
lateral			l				
rhotic			r				
approx.	w		j				

8 Vowels

-back	+back
i y	u
e oe	o
ae	a

Back vowels and front vowels don't mix (except for [i,e], which are transparent).

Slide 87

Results of SP2 Estimation

$P(b \mid \{c\} <)$		b							
		i	e	y	oe	ae	u	o	a
c	i	0.092	0.08	0.012	0.006	0.026	0.033	0.033	0.099
	e	0.094	0.073	0.014	0.005	0.032	0.035	0.028	0.082
	y	0.092	0.071	0.047	0.03	0.066	0.015	0.017	0.039
	oe	0.097	0.067	0.029	0.014	0.053	0.023	0.026	0.059
	ae	0.095	0.077	0.038	0.015	0.09	0.015	0.015	0.036
	u	0.086	0.07	0.006	0.002	0.007	0.059	0.045	0.12
	o	0.111	0.071	0.005	0.002	0.007	0.047	0.034	0.121
	a	0.099	0.063	0.005	0.002	0.007	0.049	0.035	0.134