Phonology is subregular

Jeffrey Heinz
heinz@udel.edu

University of Delaware

Oct. 9 2010
NECPHON
University of Massachusetts, Amherst

Collaborators: James Rogers (Earlham College)
Cesar Koirala, Darrell Larsen (University of Delaware)

Theories of Phonology

$$F_1 \times F_2 \times \cdots \times F_n = P$$

Theories of Phonology - The Factors

$$F_1 \times F_2 \times \cdots \times F_n = P$$

- The factors are the *individual* generalizations.
- In SPE, these are *rules*.
- In OT, HG, and HS, these are markedness and faithfulness *constraints*.

(Chomsky and Halle 1968, Prince and Smolenksy 1993/2004, Legendre et al. 1990, Pater et al. 2007, McCarthy 2000, 2006 et seq.)

Theories of Phonology - The Interaction

$$F_1 \times F_2 \times \cdots \times F_n = P$$

SPE The output of one rule becomes the input to the next.
(transducer composition)

OT Optimization over ranked constraints.
(transducer lenient composition, or shortest path)

HG Optimization over weighted constraints.
(shortest path, linear programming)

HS Repeated incremental changes w/OT optimization until convergence. *(no computational characterization yet)*

(Johnson 1992, Kaplan and Kay 1994, Frank and Satta 1998, Karttunen 1998, Riggle 2004, Pater et al. 2007, Riggle, submitted)

Theories of Phonology - The Whole Phonology

$$F_1 \times F_2 \times \cdots \times F_n = P$$

- The whole phonology is an *input/output mapping* given by the product of the factors.
- SPE, OT, HG, and HS grammars map underlying forms to surface forms.
- What kind of mapping is this?

Questions for theories of phonology

1. What is the nature of whole phonologies?
2. What is the nature of the individual generalizations?
   - I.e. what is the theory of possible rules?
   - Or what is the theory of Con?
3. How can these things be learned?

What is the nature of whole phonologies and individual generalizations?

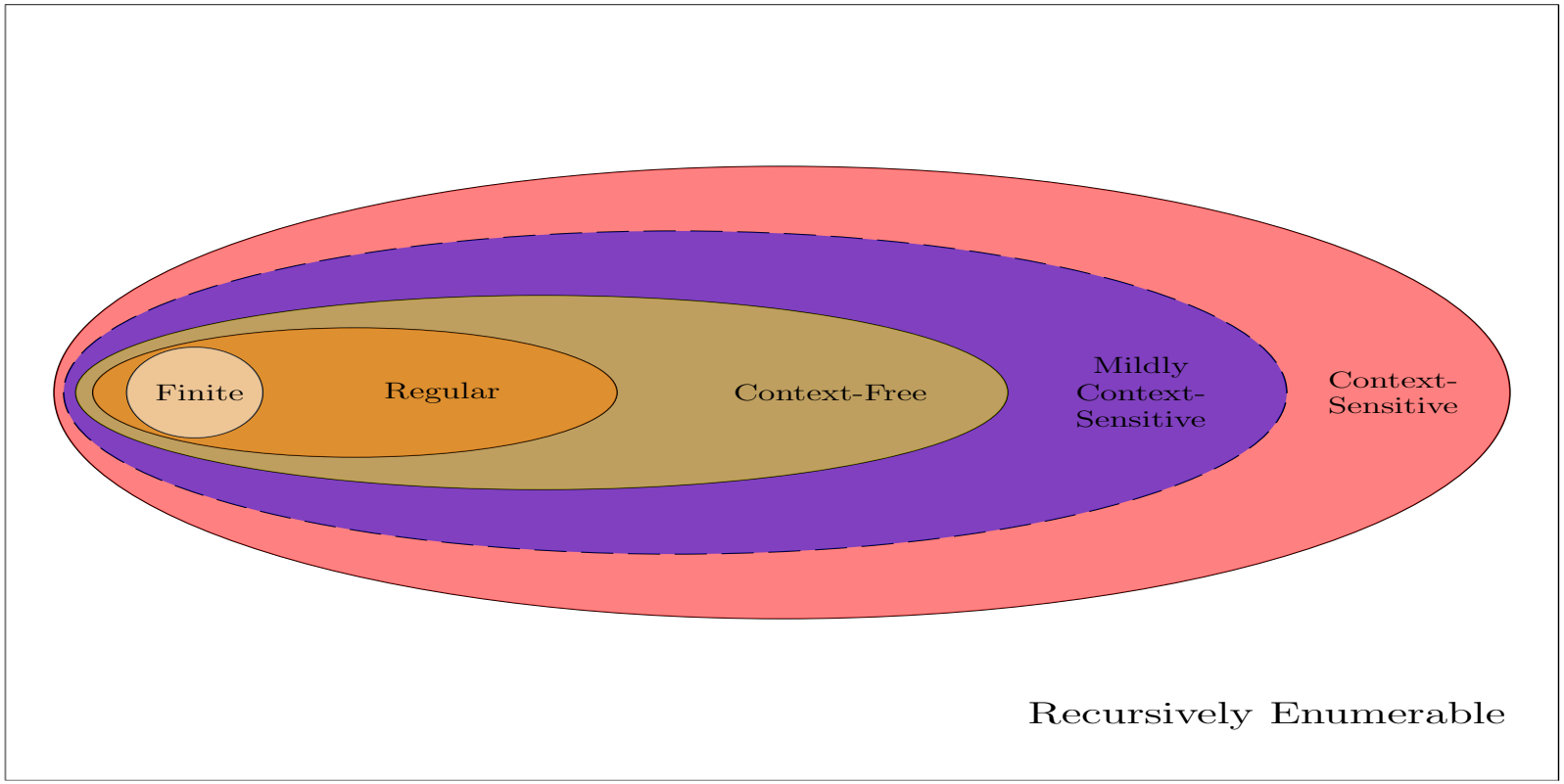


Figure: The Chomsky hierarchy classifies logically possible patterns.

What is the nature of whole phonologies and individual generalizations?

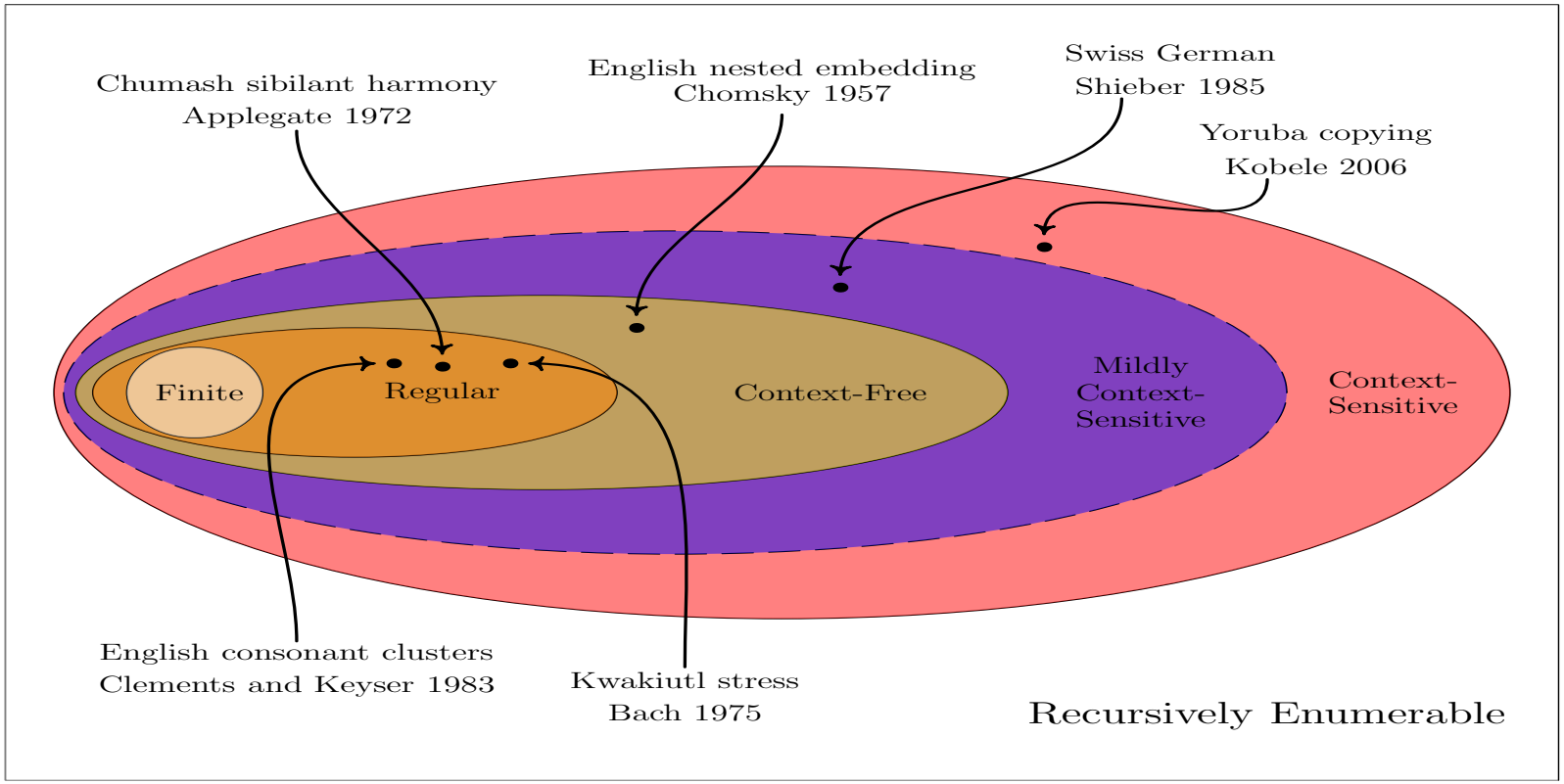


Figure: The Chomsky hierarchy classifies logically possible patterns.

What is the nature of whole phonologies and individual generalizations?

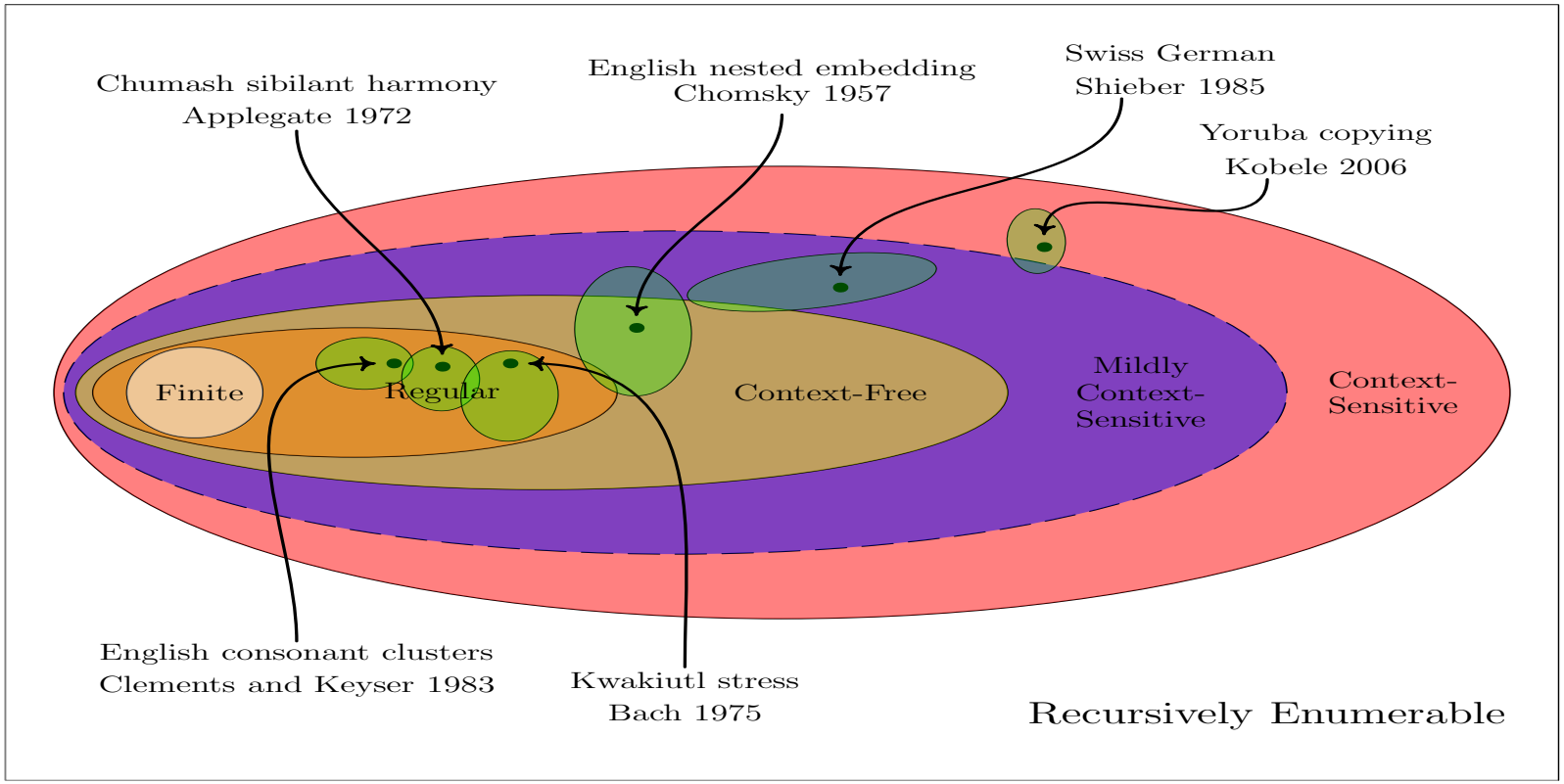


Figure: The Chomsky hierarchy classifies logically possible patterns.

Hypothesis: Phonology is Subregular.

$$F_1 \times F_2 \times \cdots \times F_n = P$$

1. The individual factors and the whole phonologies cannot be *any* regular pattern. Instead they belong to well-defined subregular regions.
2. We ought characterize necessary and sufficient properties of these regions.
3. We ought to aim to prove that these regions are feasibly learnable (under various definitions).
4. We ought to investigate the empirical consequences.

What is at stake if phonology is subregular?

$$F_1 \times F_2 \times \cdots \times F_n = P$$

1. We obtain more precise characterizations of possible phonological patterns.

- We can decide whether some logically possible pattern is a possible phonological one.
- We can *cross-classify* to help understand *why* this is so. For example, we can formulate more precise theories which ground phonology in (articulatory or perceptual) phonetics.

What is at stake if phonology is subregular?

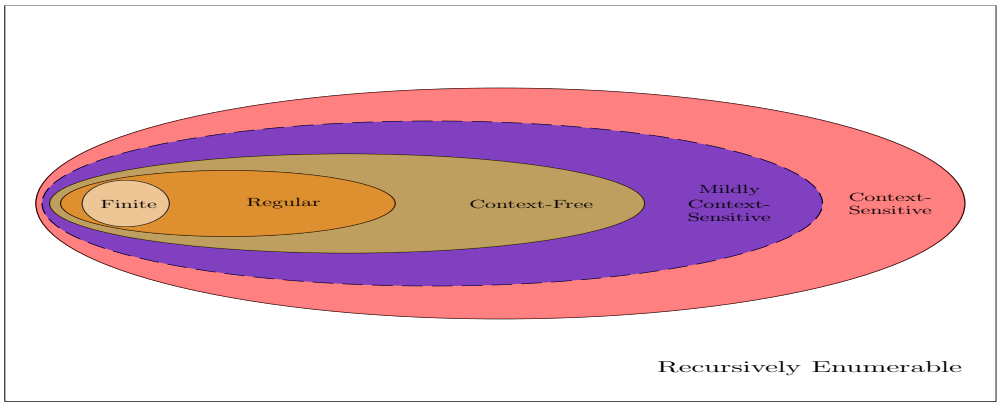
$$F_1 \times F_2 \times \cdots \times F_n = P$$

2. The computational complexity issues may resolve.

- The complexity problems noticed by Barton et al., Eisner and Idsardi stem from the the known fact that the intersection/composition of arbitrarily-many arbitrary regular sets/relations is NP-Hard.
- But if actual phonological patterns belong to more "well-behaved" subregular regions, these issues may disappear.

(Barton et. al 1997, Eisner 1997, Idsardi 2006, Heinz et al. 2007)

What is at stake if phonology is subregular?

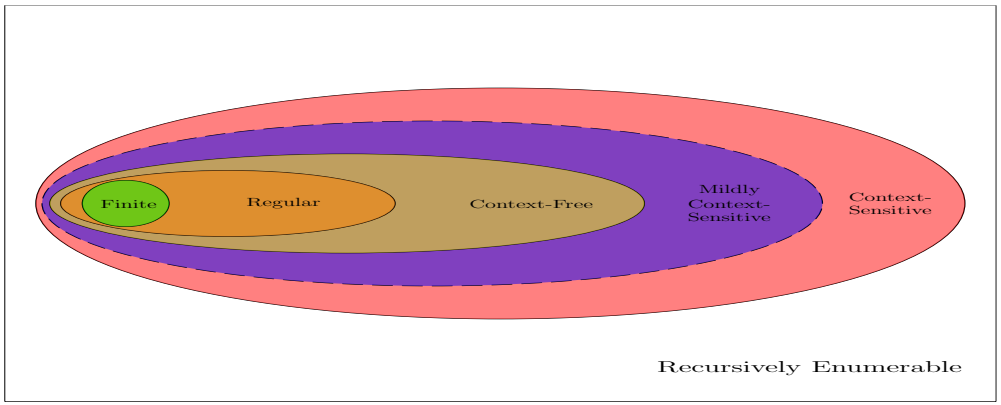


3. The learning problems may become easier to solve.

- No superfinite class of languages is identifiable in the limit from positive data (or with probability $p > 2/3$)
- The finite languages are not PAC-learnable.
- While the class of r.e. languages and stochastic languages is identifiable from positive data from computable classes of texts,
  - these learners are not feasible, and
  - the learning criteria is much weaker than these others
- But many non-superfinite classes of languages are feasibly learnable and include patterns found in natural language (proofs are often constructive)

(Gold 1967, Horning 1969, Angluin 1980, 1982, 1988, Osherson et al. 1984, Wiehagen et. al 1984, Pitt 1985, Valiant 1984, Blum et. al 1989, Garcia et al. 1990, Muggleton 1990, Jain et. al 1999, Kearns and Vazirani 1994, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Chater and Vitanyí 2007, Clark and Eryaud 2007, Heinz 2008, 2010, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

What is at stake if phonology is subregular?

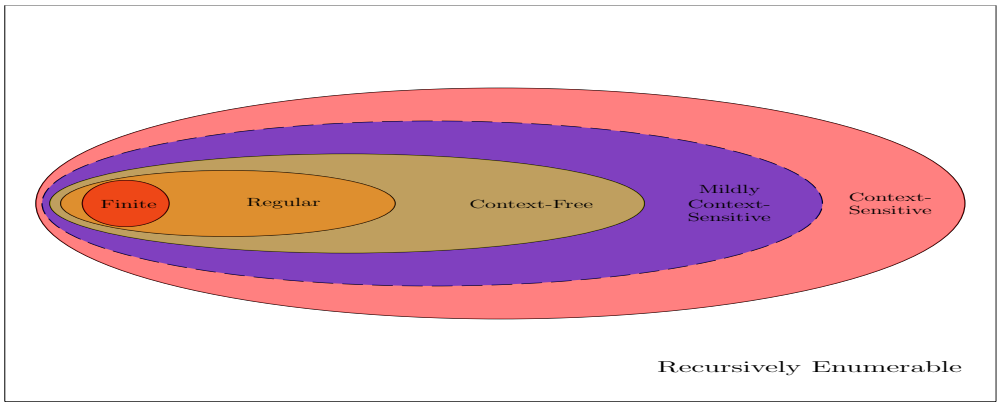


3. The learning problems may become easier to solve.

- No superfinite class of languages is identifiable in the limit from positive data (or with probability $p > 2/3$)
- The finite languages are not PAC-learnable.
- While the class of r.e. languages and stochastic languages is identifiable from positive data from computable classes of texts,
  - these learners are not feasible, and
  - the learning criteria is much weaker than these others
- But many non-superfinite classes of languages are feasibly learnable and include patterns found in natural language (proofs are often constructive)

(Gold 1967, Horning 1969, Angluin 1980, 1982, 1988, Osherson et al. 1984, Wiehagen et. al 1984, Pitt 1985, Valiant 1984, Blum et. al 1989, Garcia et al. 1990, Muggleton 1990, Jain et. al 1999, Kearns and Vazirani 1994, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Chater and Vitanyí 2007, Clark and Eryaud 2007, Heinz 2008, 2010, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

What is at stake if phonology is subregular?

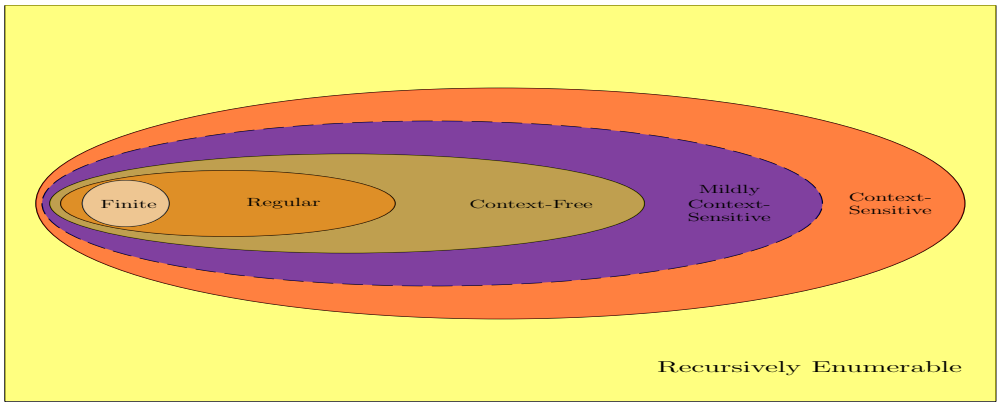


3. The learning problems may become easier to solve.

- No superfinite class of languages is identifiable in the limit from positive data (or with probability $p > 2/3$)
- The finite languages are not PAC-learnable.
- While the class of r.e. languages and stochastic languages is identifiable from positive data from computable classes of texts,
  - these learners are not feasible, and
  - the learning criteria is much weaker than these others
- But many non-superfinite classes of languages are feasibly learnable and include patterns found in natural language (proofs are often constructive)

(Gold 1967, Horning 1969, Angluin 1980, 1982, 1988, Osherson et al. 1984, Wiehagen et. al 1984, Pitt 1985, Valiant 1984, Blum et. al 1989, Garcia et al. 1990, Muggleton 1990, Jain et. al 1999, Kearns and Vazirani 1994, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Chater and Vitanyí 2007, Clark and Eryaud 2007, Heinz 2008, 2010, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

What is at stake if phonology is subregular?

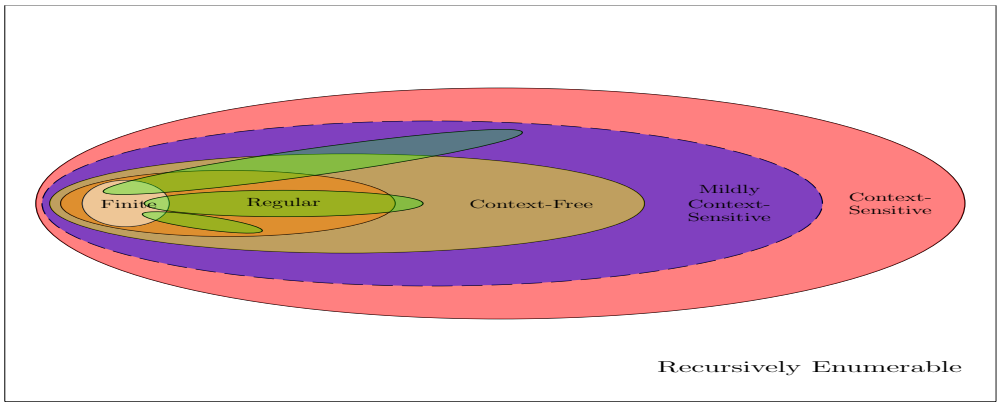


3. The learning problems may become easier to solve.

- No superfinite class of languages is identifiable in the limit from positive data (or with probability $p > 2/3$)
- The finite languages are not PAC-learnable.
- While the class of r.e. languages and stochastic languages is identifiable from positive data from computable classes of texts,
  - these learners are not feasible, and
  - the learning criteria is much weaker than these others
- But many non-superfinite classes of languages are feasibly learnable and include patterns found in natural language (proofs are often constructive)

(Gold 1967, Horning 1969, Angluin 1980, 1982, 1988, Osherson et al. 1984, Wiehagen et. al 1984, Pitt 1985, Valiant 1984, Blum et. al 1989, Garcia et al. 1990, Muggleton 1990, Jain et. al 1999, Kearns and Vazirani 1994, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Chater and Vitanyí 2007, Clark and Eryaud 2007, Heinz 2008, 2010, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

What is at stake if phonology is subregular?

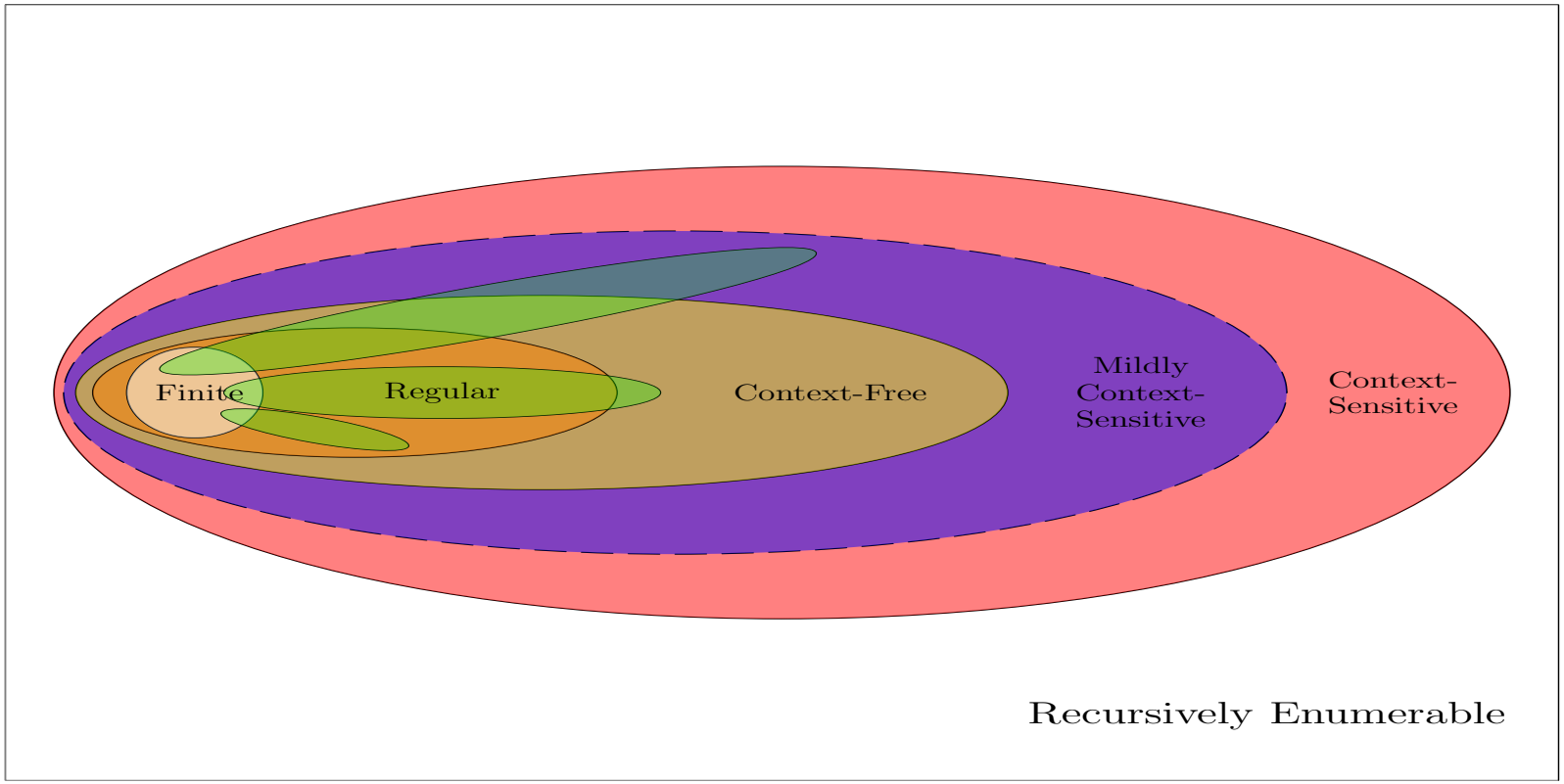


3. The learning problems may become easier to solve.

- No superfinite class of languages is identifiable in the limit from positive data (or with probability $p > 2/3$)
- The finite languages are not PAC-learnable.
- While the class of r.e. languages and stochastic languages is identifiable from positive data from computable classes of texts,
  - these learners are not feasible, and
  - the learning criteria is much weaker than these others
- But many non-superfinite classes of languages are feasibly learnable and include patterns found in natural language (proofs are often constructive)

(Gold 1967, Horning 1969, Angluin 1980, 1982, 1988, Osherson et al. 1984, Wiehagen et. al 1984, Pitt 1985, Valiant 1984, Blum et. al 1989, Garcia et al. 1990, Muggleton 1990, Jain et. al 1999, Kearns and Vazirani 1994, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Chater and Vitanyí 2007, Clark and Eryaud 2007, Heinz 2008, 2010, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

What is at stake if phonology is subregular?



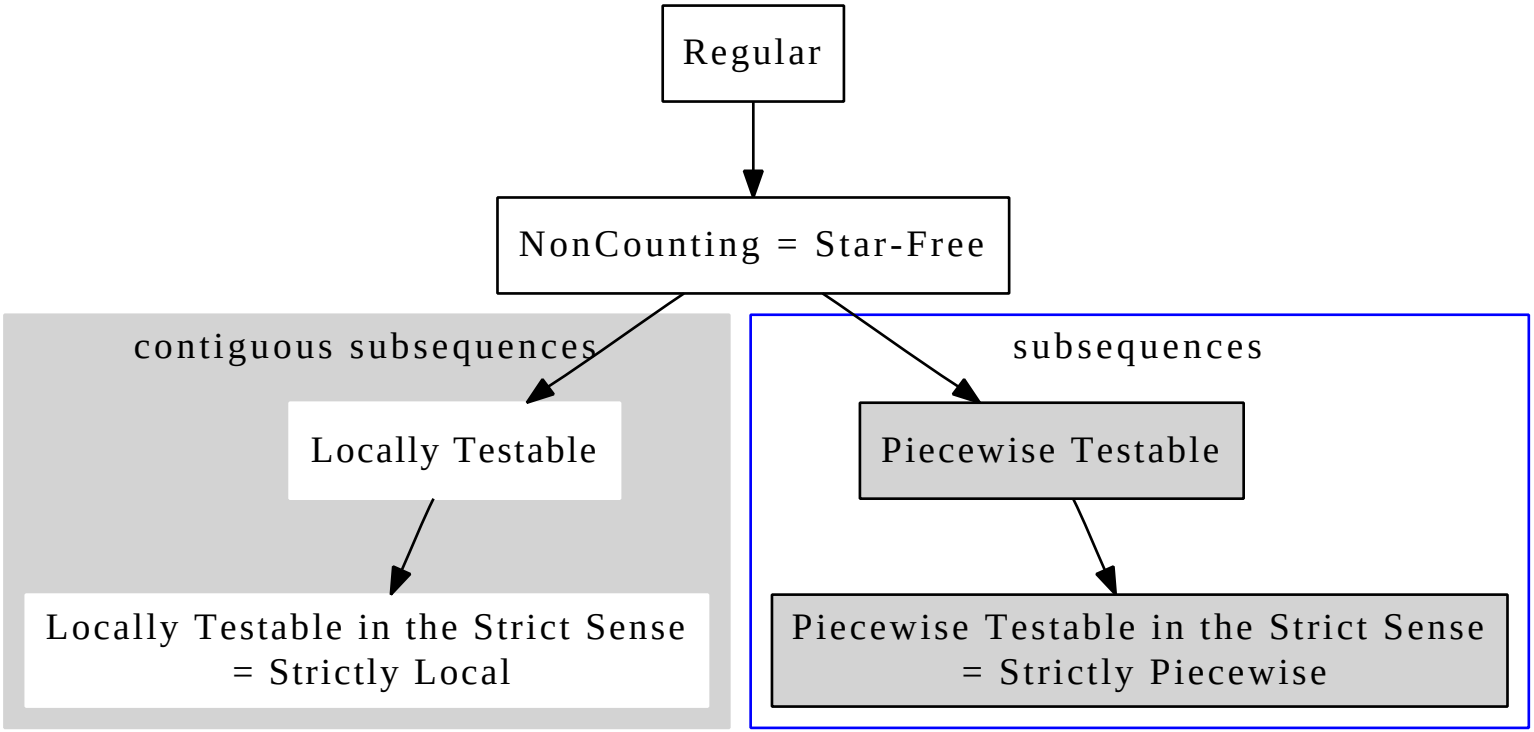
4. The learning solutions can help explain the limits of phonological variation.

Regular Patterns and Markedness Constraints

Phonological Patterns	**Nonphonological Patterns**
Words do not have NT strings.	Words do not have 3 NT strings (but 2 is OK).
Words must have a vowel (or a syllable).	Words must have an even number of vowels (or consonants, or syllables, ...).
If a word has sounds with [F] then they must agree with respect to [F]	If the first and last sounds in a word have [F] then they must agree with respect to [F].
Words have exactly one primary stress.	These six arbitrary words $\{w_1, w_2, w_3, w_4, w_5, w_6\}$ are well-formed.

(Pater 1996, Dixon and Aikhenvald 2002, Baković 2000, Rose and Walker 2004, Liberman and Prince 1977)

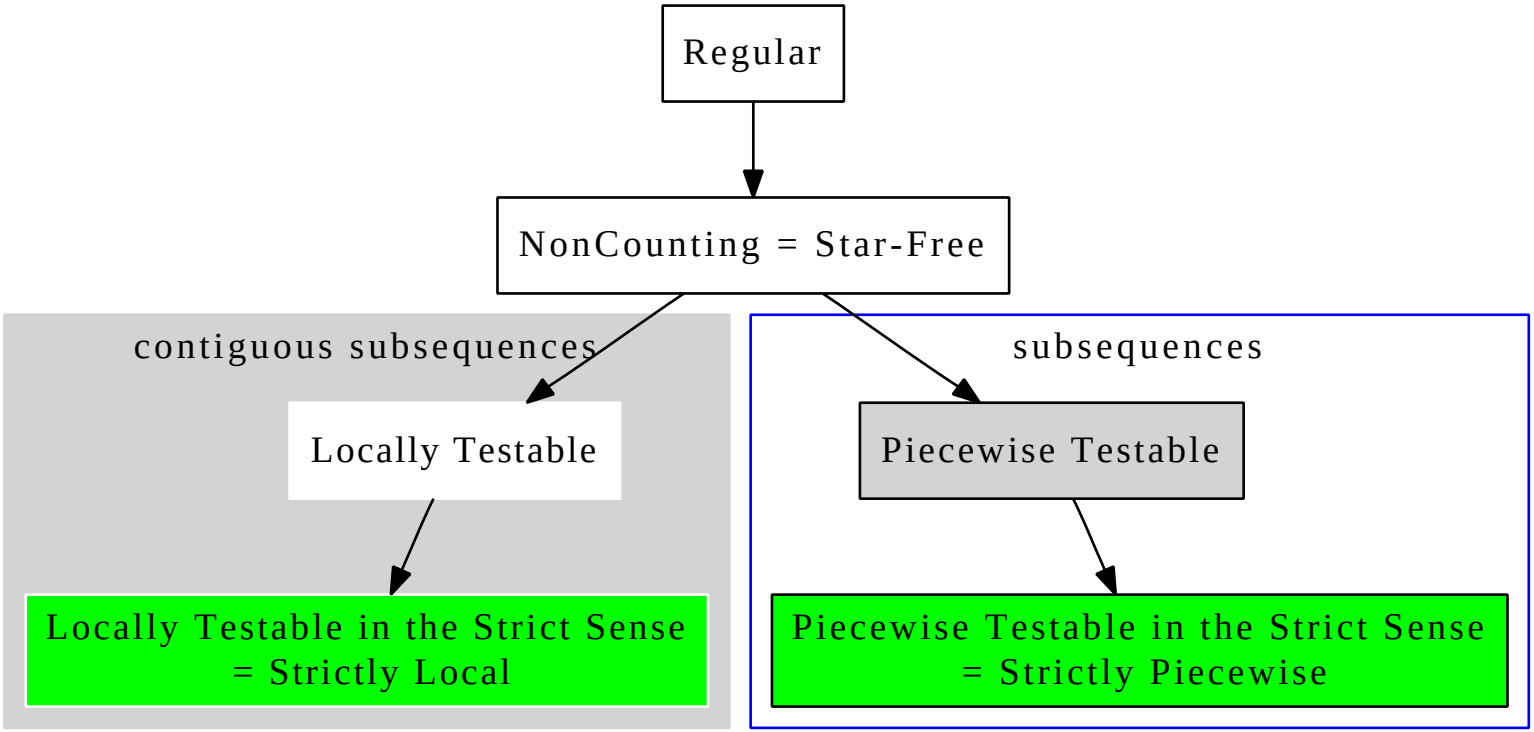
Dual subregular hierarchies (simplified)



- Each class has independent, equivalent characterizations from formal language theory, group theory, logic, and automata theory.

  (McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007, Rogers et. al 2010)

Dual subregular hierarchies (simplified)

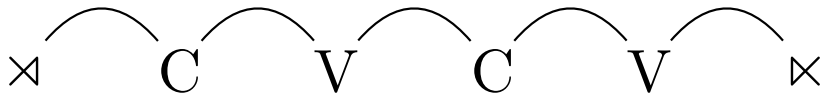


Hypotheses:

- Segmental patterns are largely Strictly Local or Strictly Piecewise.
- Stress patterns are more complex (NonCounting), but have simpler factors.

(McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007,

Strictly k-Local: Adjacency—Substrings



Definition

u is a factor of w iff $w = xuy$ for some $x, y \in \Sigma^*$.

u is a k-factor of w iff u is a factor and $|u| = k$.

$$F_k(w) = \begin{cases} \{v \in \Sigma^k : v \text{ is a k-factor of } w\} \text{ when } |w| \geq k \\ \{w\} \text{ otherwise} \end{cases}$$

Example

1. $F_2(\rtimes CVCV \ltimes) = \{\rtimes C, CV, VC, V \ltimes\}$
2. $F_8(\rtimes CVCV \ltimes) = \{\rtimes CVCV \ltimes\}$

Strictly k-Local Grammars and Languages (simplified)

Definition

A strictly k-local grammar is the set of permissible k-factors.

$$G \subseteq F_k(\{\rtimes\} \cdot \Sigma^* \cdot \{\ltimes\})$$

The strictly k-local language of G is all and only those words whose k-factors belong to G.

$$L(G) = \{w : F_k(\rtimes w \ltimes) \subseteq G\}$$

The strictly k-local languages (SL_k) are those languages that can be described by all such grammars G.

Example

$$\begin{aligned} G &= \{\rtimes C,\ CV,\ VC,\ V\ltimes\} \\ L(G) &= \{\rtimes CV\ltimes,\ \rtimes CVCV\ltimes,\ \rtimes CVCVCV\ltimes, \ldots\} \end{aligned}$$

Examples: Strictly K-Local Markedness Constraints

$$F_1 \times F_2 \times \cdots \times F_n = P$$

1. *a is SL_1.
2. *[F] is SL_1.
3. *NT is SL_2.
4. *$\acute{\sigma} \ltimes$ is SL_2.
5. *CCC is SL_3.

Examples: Stress Patterns

$$F_1 \times F_2 \times \cdots \times F_n = P$$

Edlefsen et. al (2008) classify the 109 patterns in the Stress Pattern Database (Heinz 2007,2009).

9 are SL_2	Abun West, Afrikans, Maranungku, Cambodian, ...
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_k for $k \leq 6$. 49% are SL_3.

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

Learnability: Identification in the limit from positive data of SL_k languages

Example

Consider the SL_2 Language which forbids ba. I.e.
$L = \{\rtimes\} \cdot \Sigma^* \backslash \Sigma^* ba \Sigma^* \cdot \{\ltimes\}$

time	Word w	$F_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\rtimes a, aa, a\ltimes\}$	$\{\rtimes a, aa, a\ltimes\}$	aa^*
1	aab	$\{\rtimes a, aa, ab, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes\}$	$aa^* \cup aa^*b$
2	ϵ	$\{\rtimes\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes, \rtimes\ltimes\}$	$a^* \cup a^*b$
3	$bbbbb$	$\{\rtimes b, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
4	$abbb$	$\{\rtimes a, ab, bb, b\ltimes\}$	$\{\rtimes a, aa, a\ltimes, ab, b\ltimes \rtimes\ltimes, \rtimes b, bb\}$	$\Sigma^* \backslash \Sigma^* ba \Sigma^*$
...				

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.
- 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.

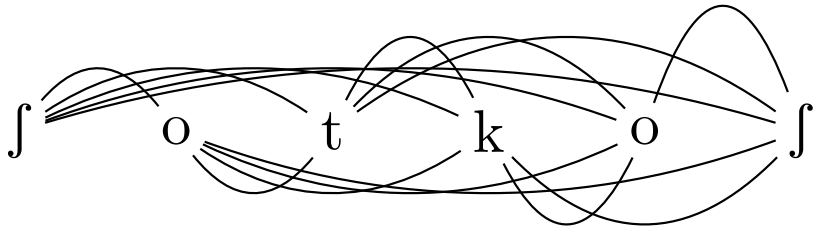
Rogers and Pullum 2007, to appear

What is not SL_k

For any k:

1. Unbounded Stress Patterns (because the primary stress may occur arbitrarily far from a word edge)
2. Long-distance Harmony patterns (because arbitrarily long material may occur between segments)

Strictly Piecewise



Definition

u is a subsequence of w iff $u = a_0 a_1 \cdots a_n$ and $w \in \Sigma^* a_0 \Sigma^* a_1 \Sigma^* \cdots \Sigma^* a_n \Sigma^*$.

u is a k-long subsequence of w iff u is a subsequence of w and $|u| = k$.

$$P_{\leq k}(w) = \{v \in \Sigma^{\leq k} : v \text{ is } (\leq k)\text{-long subsequence of } w\}$$

Example

1. $P_{\leq 2}(\int\text{otko}\int) =$
   $\{\epsilon,\int,\text{o},\text{t},\text{k},\text{o},\text{s},\int\text{o},\int\text{t},\int\text{k},\int\int,\text{ot},\text{ok},\text{oo},\text{o}\int,\text{tk},\text{to},\text{t}\int,\text{ko},\text{k}\int\}$

Strictly k-Piecewise Grammars and Languages

Definition

A strictly k-piecewise grammar is the set of permissible subsequences up to length k.

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

The strictly k-piecewise language of G is all and only those words whose subsequences up to length k belong to G.

$$L(G) = \{w : P_{\leq k}(w) \subseteq G\}$$

The strictly k-local languages (SL_k) are those languages that can be described by all such grammars G.

Example

1. $G = \Sigma^{\leq 2}/\{\mathrm{s}\int\}$ and so $L(G) = \Sigma^*/\Sigma^*\mathrm{s}\Sigma^*\int\Sigma^*$

(Rogers et al. 2010, Heinz 2007, 2010, in press)

Examples: What is and is not SP_k

SP_2 includes

1. Asymmetric consonantal Harmony
   - Sibilant Harmony in Sarcee (Cook 1978a,b, 1984)
   - *s…ʃ
2. Symmetric consonantal Harmony
   - Sibilant Harmony in Navajo (Sapir and Hojier 1967, Fountain 1998)
   - *ʃ…s and *s…ʃ
3. Vowel harmony patterns with transparent vowels
   - Finnish, Korean sound-symbolic harmony, …

For any k, these are not SP_k:

1. Consonantal harmony with blocking (unattested)
   (Hansson 2001, Rose and Walker 2004)
2. Vowel harmony with blocking, i.e. opaque vowels (attested)

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	$\{\epsilon,$ a, aa$\}$	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	$\{\epsilon,$ a, aa, b, ab$\}$	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	$\{\epsilon,$ a, b, aa, ab ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	$\{\epsilon,$ a, b, aa, ab, ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	$\{\epsilon,$ a, aa$\}$	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	$\{\epsilon,$ a, aa, b, ab$\}$	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	$\{\epsilon,$ a, b, aa, ab ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	$\{\epsilon,$ a, b, aa, ab, ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	{ϵ, a, aa}	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	{ϵ, a, aa, b, ab}	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	{ϵ, a, b, aa, ab ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	{ϵ, a, b, aa, ab, ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	{ϵ, a, aa}	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	{ϵ, a, aa, b, ab}	$a^* \cup a^*b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	{ϵ, a, b, aa, ab ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	{ϵ, a, b, aa, ab, ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	$\{\epsilon$, a, aa$\}$	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	$\{\epsilon$, a, aa, b, ab$\}$	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	$\{\epsilon$, a, b, aa, ab ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	$\{\epsilon$, a, b, aa, ab, ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	{ϵ, a, aa}	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	{ϵ, a, aa, b, ab}	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	{ϵ, a, b, aa, ab ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	{ϵ, a, b, aa, ab, ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

Learnability: Identification in the limit from positive data of SP_k

Let $L = \Sigma^* \backslash \Sigma^* b \Sigma^* b \Sigma^*$

time	Word w	$P_2(w)$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\epsilon, a, aa\}$	{ϵ, a, aa}	a^*
1	aab	$\{\epsilon, a, b, aa, ab\}$	{ϵ, a, aa, b, ab}	$a^* \cup a^* b$
2	baa	$\{\epsilon, a, b, aa, ba\}$	{ϵ, a, b, aa, ab ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\epsilon, a, b, ab, ba\}$	{ϵ, a, b, aa, ab, ba}	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

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.
- 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.

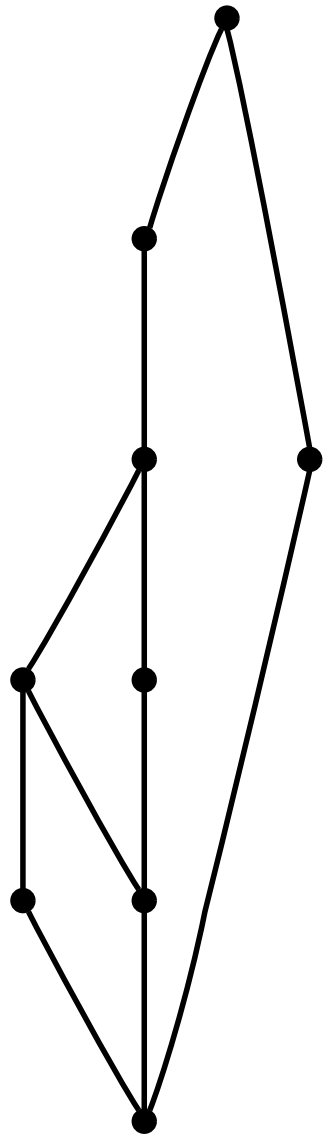
Rogers and Pullum 2007, to appear

Characterizing those learners: Lattice-structured hypothesis spaces

Each node represents a block in the partition of Σ^* given by f (E.g. F_k or P_k).

Each node N also represents a language. The language is all words in all blocks of all nodes dominated by N.

Each node also represents a grammar - a finite description of this potentially infinitely-sized language.



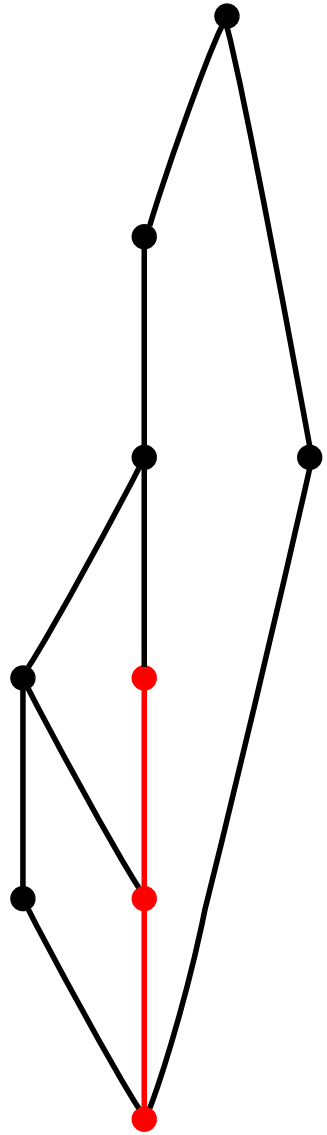
Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

Each node represents a block in the partition of Σ^* given by f (E.g. F_k or P_k).

Each node N also represents a language. The language is all words in all blocks of all nodes dominated by N.

Each node also represents a grammar - a finite description of this potentially infinitely-sized language.



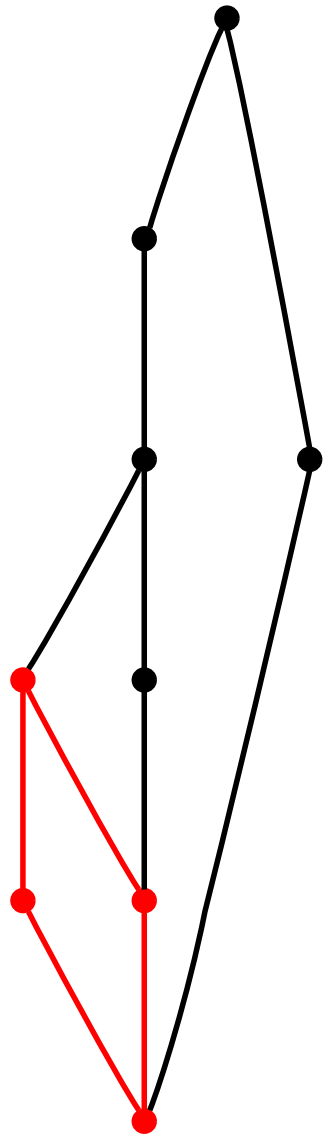
Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

Each node represents a block in the partition of Σ^* given by f (E.g. F_k or P_k).

Each node N also represents a language. The language is all words in all blocks of all nodes dominated by N.

Each node also represents a grammar - a finite description of this potentially infinitely-sized language.



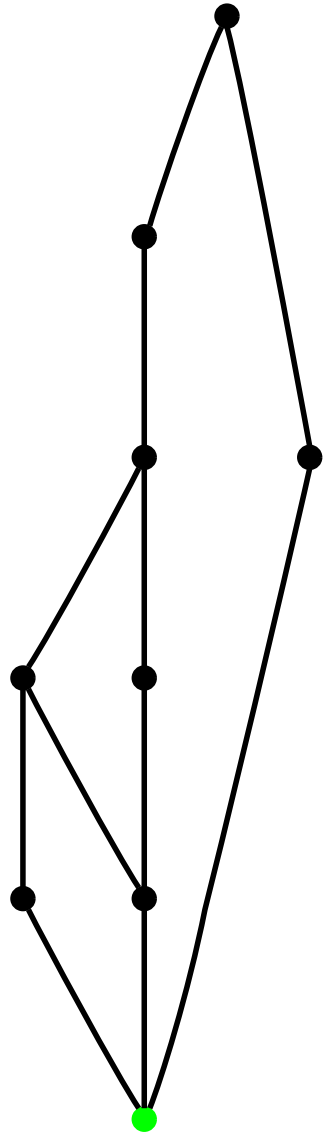
Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

Learners can make inferences in two ways:

1. If a node is part of the language, everything below it is too.
2. If two nodes are part of the language, the least upper bound is too.

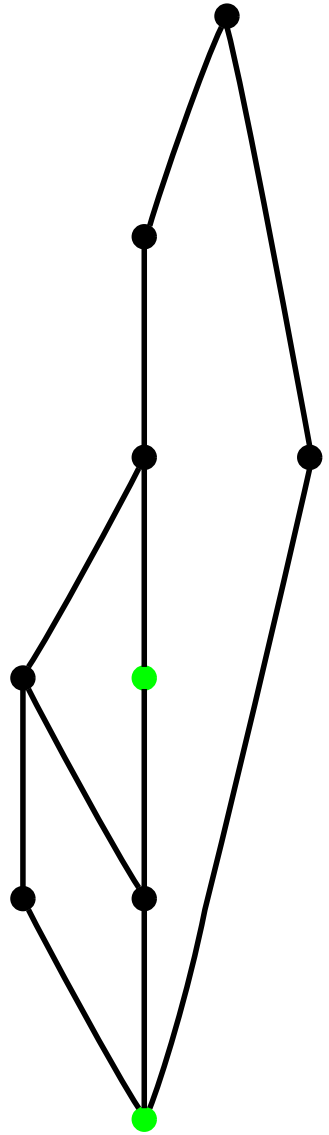
Assume the starting point is the least element in the example.



Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

Suppose the learner observes w_1 and $f(w_1)$ maps to the node shown.

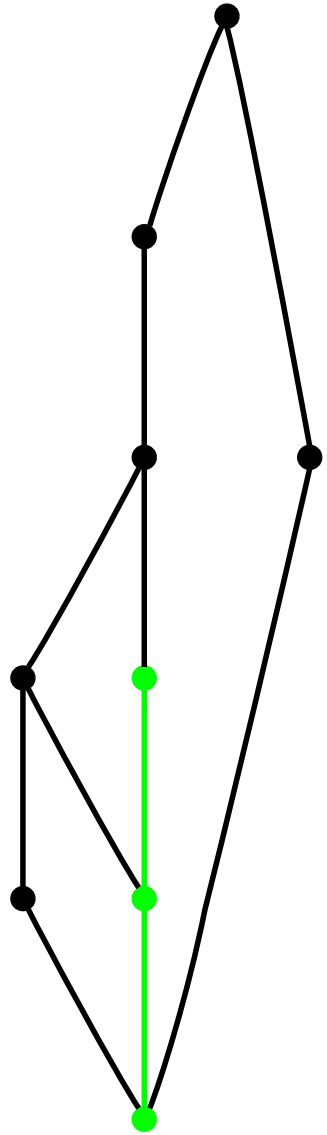


Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

Suppose the learner observes w_1 and $f(w_1)$ maps to the node shown.

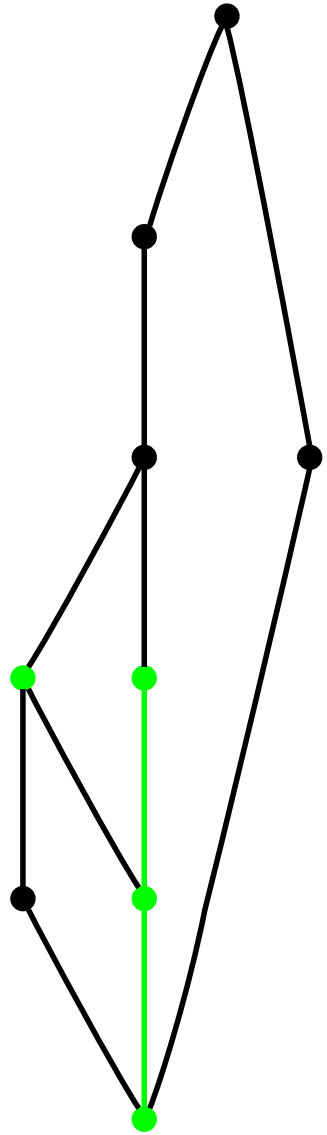
Then the learner can infer everything below that node is also in the language.



Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

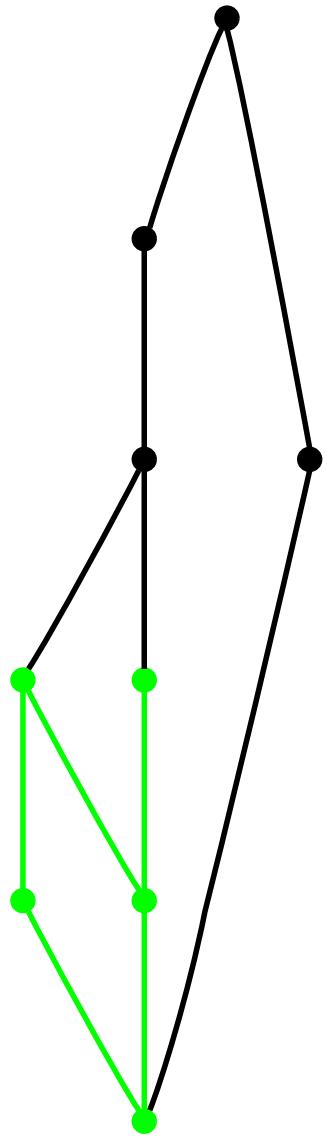
Suppose the learner then observes w_2 and $f(w_2)$ maps to this other node.



Heinz, Kasprzik, and Kötzing, submitted

Characterizing those learners: Lattice-structured hypothesis spaces

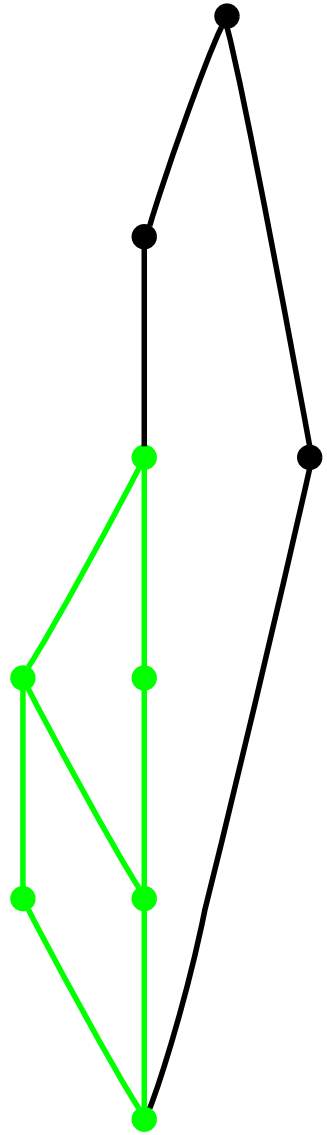
Then the learner can infer all words in blocks below that node are also in the language.



Heinz, Kasprzik, and Kötzing, submitted

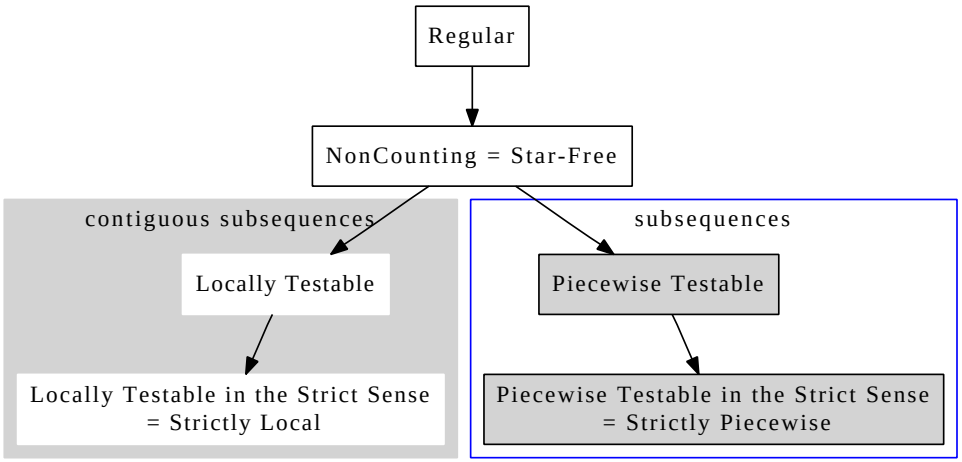
Characterizing those learners: Lattice-structured hypothesis spaces

And the learner can infer words in the least upper bound are also in the language.



Heinz, Kasprzik, and Kötzing, submitted

Locally Testable and Piecewise Testable



- The Locally k-Testable (LT_k) class of languages is the smallest class which closes SL_k under boolean operations.

  (McNaughton and Papert 1971)

- The Piecewise k-Testable (PT_k) class of languages is the smallest class which closes SP_k under boolean operations.

  (Simon 1975, Rogers et al. 2009)

- For fixed k, LT_k and PT_k are identifiable in the limit from positive data, but not feasibly.

  (Garcia and Ruiz 2004, Heinz 2010)

Examples: What is and what is not LT or PT

1. Consonant Harmony (Heinz, in press)
   - Symmetric consonantal harmony patterns are LT_1.
   - Asymmetric consonantal harmony patterns are not LT_k for any k.
2. Stress Patterns:
   - Culminativity is PT_2.
   - The stress pattern of Yidin is the intersection of PT_2 and SL_2 (Rogers, p.c.)
   - Factoring Culminativity out of unbounded stress patterns leaves you with SP_2 patterns (Heinz, in progress)

NonCounting (also known as Star-Free)

Definition

A language L is NonCounting iff there exists some $n > 0$ such that for all strings $u, v, w \in \Sigma^*$, it is the case that if $uv^n w$ belongs to L then $uv^{n+i}w$,for all $i \geq 1$, belongs to L as well.
(McNaughton and Papert 1971)

Example

1. All stress patterns in the Stress Pattern Database (Heinz 2007, 2009) are NonCounting (Edlefsen et al. 2008, Rogers, p.c.).
2. Patterns like "has an even number of vowels, consononants, syllables, etc." are *not* noncounting.

Learning Stochastic Strictly Piecewise Patterns

- Heinz and Rogers (2010) define a family of stochastic languages whose categorical counterpart is the strictly piecewise languages.
- For the $k = 2$ case, the probability of the next symbol in a sequence is determined by a function of the probability of this symbol given each preceding symbol.
- For given k, they prove this family yields a family of well-defined probability distributions with on the order of $|\Sigma|^k$ parameters.
- They show how to find the maximum likelihood estimates of these parameters from a set of positive data.

Samala (Chumash) 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 pˀ pʰ	t tˀ tʰ		k kˀ kʰ	q qˀ qʰ	ʔ
affricates		t͡s t͡sˀ t͡sʰ	t͡ʃ t͡ʃˀ t͡ʃʰ			
fricatives		s sˀ sʰ	ʃ ʃˀ ʃʰ	x xˀ		h
nasal	m	n nˀ				
lateral		l lˀ				
approx.	w	y				

6 Vowels

i	ɨ	u
e		o
	a	

(Applegate 1972, 2007)

Samala: results of SP2 estimation

$P(x \mid \{y\} <)$		x			
		$\widehat{\text{t}\int}$	$\int$	$\widehat{\text{ts}}$	s
y	$\widehat{\text{t}\int}$	0.0313	0.0455	0.	0.0006
	$\int$	0.0353	0.0671	0.	0.0009
	$\widehat{\text{ts}}$	0.	0.0009	0.0113	0.0218
	s	0.0002	0.0011	0.0051	0.0335

(Collapsing laryngeal distinctions)

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).

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

Whither tiers?

Q: Since long-distance patterns are learnable by tier-based n-gram models, do we need SP distributions?
(Goldsmith 1976, Clements 1985, Sagey 1986, Mester 1988,Hayes and Wilson 2008, Goldsmith and Xanthos 2009, Goldsmith and Riggle to appear)

A: The models make different predictions, making it a fruitful area for future research.

tier-based SL (n-gram) models	SP models
Predicts unattested blocking effects in consonantal harmony	Predicts absence of blocking in consonantal harmony
Captures blocking effects in vowel harmony	Unable to capture blocking effects in vowel harmony
Only able to describe patterns with transparent vowels if they are "off" the tier	Able to describe patterns with transparent vowels
Requires independent theory of tiers	Does not require independent theory of tiers
Requires independent theory of similarity	Requires independent theory of similarity

Vowel harmony in sound-symbolic morphemes in Korean (joint work with Darrell Larsen)

	front	front rounded	mid	back	
high	i	ü	ɨ	u	'dark'
mid	e	ö	ə	o	
low	æ		a		'light'

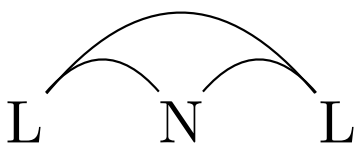
- Vowels [i] and [ɨ] are 'dark' in initial syllables, transparent in noninitial syllables (Kim-Renaud 1976, Cho 1994, inter alia)
- Extracted 4,006 sound-symbolic morphemes from Korea's National Institute of the Korean Language's 'The Great Standard Korean Dictionary' `http://www.hangeul.pe.kr/symbol/words.htm`
- Only unique morphemes of 2 or 3 syllables were selected from reduplicating examples in the corpus for ease of extraction.

Goal of the study

- Compare tier-based SL_2 bigram models to a tier-based SP_2 models.
- These models have the same number of parameters!
- The parameters identify different kinds of phonological relationships.

SL_2

SP_2



Bigram Model (Strictly 2-Local distributions)

- A trained probabilistic bigram model over the vowel tier (Jurafsky & Martin, 2008) fails to make the right distinctions:

Word	Prob(word)
LNL	0.003611
DND	0.006353
LND	0.007325
DNL	0.003132

Bigram Model (Strictly 2-Local distributions)

- A trained probabilistic bigram model over the vowel tier (Jurafsky & Martin, 2008) fails to make the right distinctions:

Word	Prob(word)
LNL	0.003611
DND	0.006353
LND	0.007325
DNL	0.003132

Bigram Model (Strictly 2-Local distributions)

- A trained probabilistic bigram model over the vowel tier (Jurafsky & Martin, 2008) fails to make the right distinctions:

Word	Prob(word)
LNL	0.003611
DND	0.006353
LND	0.007325
DNL	0.003132

Learning Strictly 2-Piecewise Distributions

- A trained probabilistic SP2 learner (Heinz & Rogers 2010) learns the transparency of noninitial N vowels, and to some extent, the behavior of initial-syllable N vowels.

Word	Prob(word)
LNL	0.002893
DND	0.004357
LND	0.000142
DNL	0.000255

Learning Strictly 2-Piecewise Distributions

- A trained probabilistic SP2 learner (Heinz & Rogers 2010) learns the transparency of noninitial N vowels, and to some extent, the behavior of initial-syllable N vowels.

Word	Prob(word)
LNL	0.002893
DND	0.004357
LND	0.000142
DNL	0.000255

Learning Strictly 2-Piecewise Distributions

- A trained probabilistic SP2 learner (Heinz & Rogers 2010) learns the transparency of noninitial N vowels, and to some extent, the behavior of initial-syllable N vowels.

Word	Prob(word)
LNL	0.002893
DND	0.004357
LND	0.000142
DNL	0.000255

Quantitative Comparison

Using the trained SP2 and SL2 probability distributions, we calculated the expected number of each word type.

word type	actual	SP2	SL2
DD	455	502.5	473.4
DL	47	56.5	10.8
DN	637	563.6	237.5
...			

Then we computed the correlation (Spearman's r) between the expected number and the actual number:

	SP2	SL2	# of words
All	0.95	0.55	4006
Disyllabic words	0.97	0.87	3020
Trisyllabic words	0.47	0.31	986

SL2 distributions and SP2 distributions have the same number of parameters!

Local Summary

1. These results are evidence that SP_k constraints are present and active insofar as they extract the right generalization.
2. These results do not mean we don't need SL_k constraints! (or SL-based learners)
3. SP_k patterns don't capture long-distance dissimilation or opaque vowels in vowel harmony patterns, not to mention any kind of local dependency!
4. The view of phonological learning espoused is here is **modular**. Different kinds of patterns have different kinds of learners—both SL-type and SP-type learners are needed.

Conclusion: Future Work

1. Further restrict the SL_k and $\mathrm{SP}_{k'}$ classes with phonological features

   (Hayes and Wilson 2008, Albright 2009, Heinz and Koirala 2010)

2. Learning non-surface true-generalizations.

   (Heinz and Idsardi in prep)

3. Define new subregular classes relevant to phonology.
   - E.g. while Culminativity is PT_2, it's unclear that PT_2 is the natural class of patterns that we are looking for.
   - What subregular class describes blocking patterns? (characterizing tier-based SL_k classes)

4. Develop subregular hierarchies and subregular classes of *regular relations* and classify patterns of alternation

   (cf. Tesar 2009, output-directed maps)

Conclusion: Phonology is Subregular.

$$F_1 \times F_2 \times \cdots \times F_n = P$$

1. We can develop constrained, precise theories of whole phonologies and phonological factors by classifying them with respect to subregular language classes.
2. If factor-interaction is well-defined then we ought to be able to prove conclusions about whole phonologies from characterizations of these factors. E.g. we ought to be able to reduce the computational load.
3. We can profitably investigate the learnability of these classes.

Conclusion: Phonology is Subregular.

$$F_1 \times F_2 \times \cdots \times F_n = P$$

1. We can develop constrained, precise theories of whole phonologies and phonological factors by classifying them with respect to subregular language classes.
2. If factor-interaction is well-defined then we ought to be able to prove conclusions about whole phonologies from characterizations of these factors. E.g. we ought to be able to reduce the computational load.
3. We can profitably investigate the learnability of these classes.

Thank You!