Phonological Patterns and Phonological Learners

Jeffrey Heinz
heinz@udel.edu

University of Delaware

Cornell University
Grammar Induction Workshop
May 15, 2010

Collaborators

James Rogers (Earlham College)

Bill Idsardi (University of Maryland)

Cesar Koirala, Regine Lai, Darrell Larsen, Dan Blanchard, Tim O'Neill, Jane Chandlee, Robert Wilder, Evan Bradley (University of Delaware)

How can something learn?

1. How do people generalize beyond their experience?
2. How can any*thing* that computes generalize beyond its experience?

- Linguistics / Language Acquisition
- Computer Science
- Psychology
- Artificial Intelligence
- Philosophy
- Natural Language Processing
- ...

Phonological Patterns and Phonological Learners

1. Different phonological patterns are learned by different learning mechanisms
2. Illustrate with a learner for long-distance patterns (Strictly k-Piecewise languages and distributions)

Phonological Patterns and Phonological Learners

1. Different phonological patterns are learned by different learning mechanisms
2. Illustrate with a learner for long-distance patterns (Strictly k-Piecewise languages and distributions)

Hypothesis: Phonological learning is modular. There is more than one highly specialized learning mechanism for learning phonology.

Phonological Patterns and Phonological Learners

1. Different phonological patterns are learned by different learning mechanisms
2. Illustrate with a learner for long-distance patterns (Strictly k-Piecewise languages and distributions)

Hypothesis: Phonological learning is modular. There is more than one highly specialized learning mechanism for learning phonology.

The debate isn't likely to be settled soon. All the empirical evidence isn't in yet nor have all models been fully compared.

Phonotactics - Knowledge of word well-formedness

ptak thole hlad plast sram mgla vlas flitch dnom rtut

Halle, M. 1978. In *Linguistic Theory and Pyschological Reality*. MIT Press.

Phonotactics - Knowledge of word well-formedness

possible English words	impossible English words
thole	ptak
plast	hlad
flitch	sram
	mgla
	vlas
	dnom
	rtut

1. Question: How do English speakers know which of these words belong to different columns?

Phonotactics - Knowledge of word well-formedness

possible English words	impossible English words
thole	ptak
plast	hlad
flitch	sram
	mgla
	vlas
	dnom
	rtut

1. Question: How do English speakers know which of these words belong to different columns?

Phonotactics - Knowledge of word well-formedness Chumash Version

ʃtoyonowonowaʃ

stoyonowonowaʃ

stoyonowonowas

ʃtoyonowonowas

pisotonosikiwat

pisotonoʃikiwat

Phonotactics - Knowledge of word well-formedness
Chumash Version

possible Chumash words	impossible Chumash words
ʃtoyonowonowaʃ	stoyonowonowaʃ
stoyonowonowas	ʃtoyonowonowas
pisotonosikiwat	pisotonoʃikiwat

1. Question: How do Chumash speakers know which of these words belong to different columns?
2. By the way, *ʃtoyonowonowaʃ* means ‘it stood upright’

(Applegate 1972)

Phonotactics - Knowledge of word well-formedness Chumash Version

possible Chumash words	impossible Chumash words
ʃtoyonowonowaʃ	stoyonowonowaʃ
stoyonowonowas	ʃtoyonowonowas
pisotonosikiwat	pisotonoʃikiwat

1. Question: How do Chumash speakers know which of these words belong to different columns?
2. By the way, *ʃtoyonowonowaʃ* means 'it stood upright'

(Applegate 1972)

Limits on the variation of segmental phonotactics

1. **Local sound patterns**; e.g. consonant clusters
   - tendencies: sonority sequencing, n-long clusters can be resoloved into two $(n-1)$-long clusters, ...
     (Greenberg 1978, Clements and Keyser 1983, ... Albright today)
2. **Long-distance sound patterns**; e.g. consonantal and vowel harmony
   - Similar segments are involved in long-distance patterns
   - Consonantal harmony patterns do not exhibit blocking: e.g. *s...ʃ unless [z] intervenes.
     (Hansson 2001, Rose and Walker 2004)
   - No harmony pattern applies only to the first and last sounds.
3. Logically possible but unattested segmental phonotactic patterns:
   - The nth sound after x must be y.
   - Words must contain an even number of sounds of type x.
   - ...

Formal Language Theory

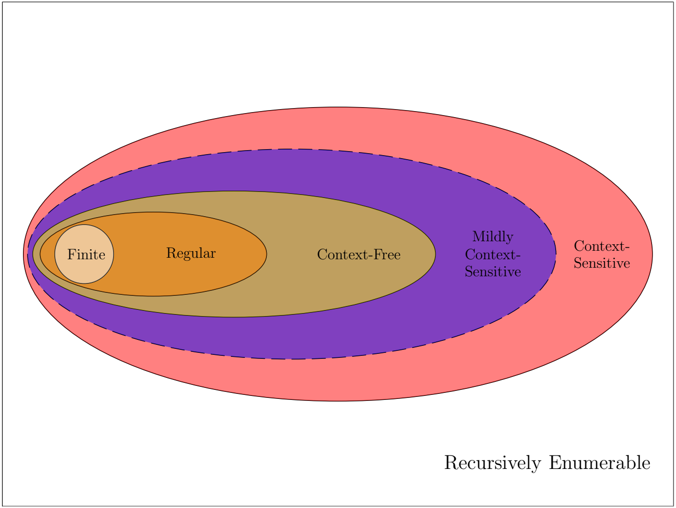


Figure: The Chomsky hierarchy classifies logically possible patterns.

(Chomsky 1956, 1959, Harrison 1978)

Formal Language Theory

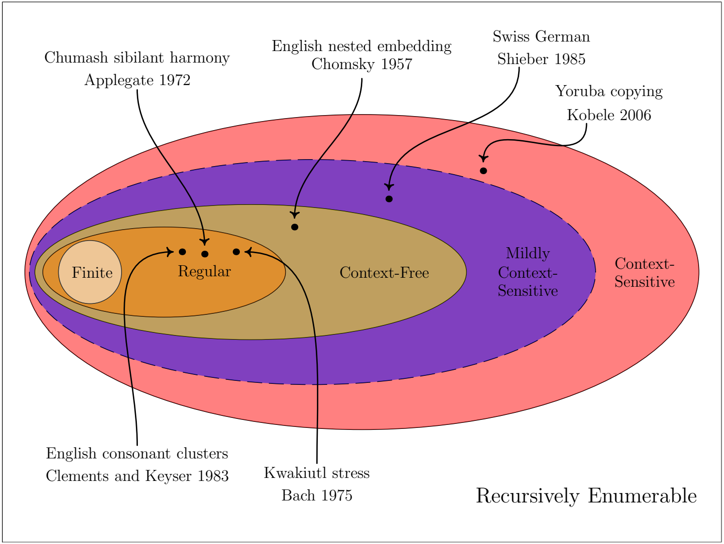


Figure: Natural language patterns in the Chomsky hierarchy.

Formal Language Theory

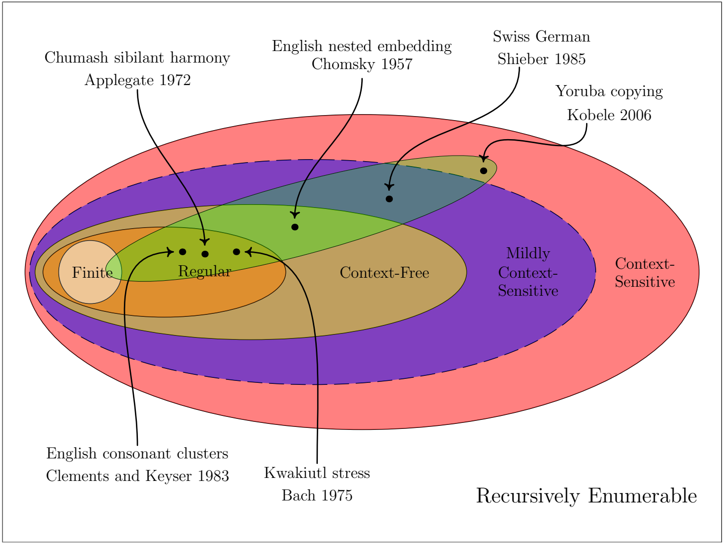


Figure: Possible theories of natural language.

Formal Language Theory

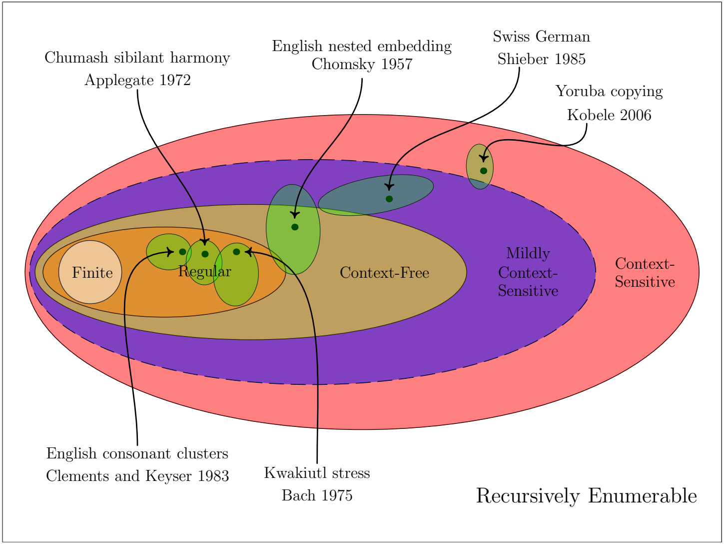


Figure: Possible theories of natural language.

Formal Learning Theories

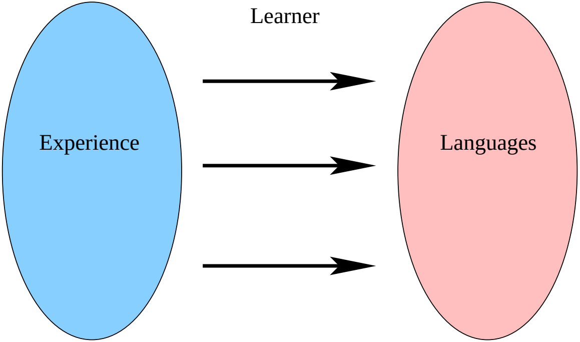


Figure: Learners are functions ϕ from experience to languages.

(Gold 1967, Horning 1969, Angluin 1980, Osherson et al. 1984, Angluin 1988, Anthony and Biggs 1991, Kearns and Vazirani 1994, Vapnik 1994, 1998, Jain et al. 1999, Niyogi 2006, de la Higuera 2010)

The Experience

1. It is a sequence.
2. It is finite.

$$\begin{array}{ll} w_0 & \\ w_1 & \\ w_2 & \big\downarrow \text{ time} \\ \ldots & \\ w_n & \end{array}$$

Types of Experience

1. Positive evidence
2. Positive and negative evidence
3. Noisy evidence
4. Queried Evidence

$$\begin{array}{c} w_0 \in L \\ w_1 \in L \\ w_2 \in L \\ \dots \\ w_n \in L \end{array} \qquad \Big\downarrow \text{time}$$

Types of Experience

1. Positive evidence
2. Positive and negative evidence
3. Noisy evidence
4. Queried Evidence

$$\begin{array}{c} w_0 \in L \\ w_1 \notin L \\ w_2 \notin L \\ \ldots \\ w_n \in L \end{array} \qquad \Big\downarrow \text{time}$$

Types of Experience

1. Positive evidence
2. Positive and negative evidence
3. Noisy evidence
4. Queried Evidence

$$
\begin{array}{c}
w_0 \in L \\
w_1 \notin L \\
w_2 \in L \text{ (but in fact } w_2 \notin L\text{)} \\
\ldots \\
w_n \in L
\end{array}
\qquad \downarrow \text{time}
$$

Types of Experience

1. Positive evidence
2. Positive and negative evidence
3. Noisy evidence
4. Queried Evidence

$w_0 \in L$

$w_1 \notin L$

$w_2 \in L$ (because learner specifically asked about w_2)

$\ldots$

$w_n \in L$

$\downarrow$ time

The Languages

1. They can be sets of words or distributions over words.
2. They are computable.

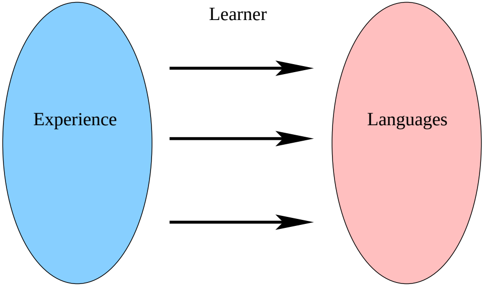


Figure: Learners are functions ϕ from experience to languages.

The Languages

1. They can be sets of words or distributions over words.
2. They are computable.

   I.e. they are describable with grammars.

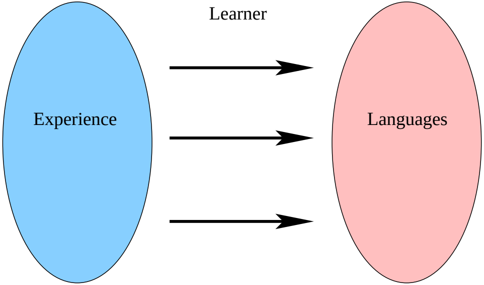


Figure: Learners are functions ϕ from experience to languages.

The Languages

1. They can be sets of words or distributions over words.
2. They are computable.

   I.e. they are describable with grammars.

   I.e they are r.e. languages.

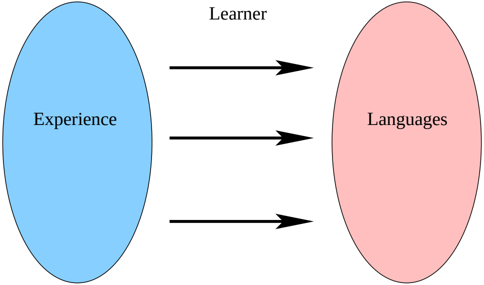


Figure: Learners are functions ϕ from experience to languages.

The Languages

1. They can be sets of words or distributions over words.
2. They are computable.

   I.e. they are describable with grammars.

   I.e they are r.e. languages.

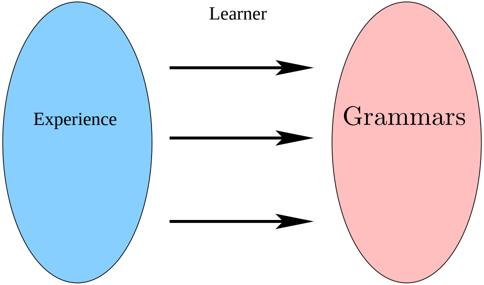


Figure: Learners are functions ϕ from experience to grammars.

Learning Criteria

1. What does it mean to learn a language?
2. What kind of experience is required for success?
3. What counts as success?

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$
w_2	$\phi(\langle w_0, w_1, w_2 \rangle) = G_2$

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$
w_2	$\phi(\langle w_0, w_1, w_2 \rangle) = G_2$
$\ldots$	

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$
w_2	$\phi(\langle w_0, w_1, w_2 \rangle) = G_2$
$\dots$	
w_n	$\phi(\langle w_0, w_1, w_2, \dots, w_n \rangle) = G_n$

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$
w_2	$\phi(\langle w_0, w_1, w_2 \rangle) = G_2$
$\ldots$	
w_n	$\phi(\langle w_0, w_1, w_2, \ldots, w_n \rangle) = G_n$
$\ldots$	

$\downarrow$ time

What does it mean to learn a language?

1. Convergence.
2. Imagine an infinite sequence. Is there some point n after which the learner's hypothesis doesn't change (much)?

datum	Learner's Hypothesis
w_0	$\phi(\langle w_0 \rangle) = G_0$
w_1	$\phi(\langle w_0, w_1 \rangle) = G_1$
w_2	$\phi(\langle w_0, w_1, w_2 \rangle) = G_2$
...	
w_n	$\phi(\langle w_0, w_1, w_2, \ldots, w_n \rangle) = G_n$
...	
w_m	$\phi(\langle w_0, w_1, w_2, \ldots, w_m \rangle) = G_m$

$\downarrow$ time

Does $G_m \simeq G_n$?

What kind of experience is required for success?

Types of Experience

1. Positive-only or positive and negative evidence.
2. Noisless or noisy evidence.
3. Queries allowed or not?

Which infinite sequences require convergence?

1. only complete ones? I.e. where every piece of information occurs at some finite point
2. only computable ones? I.e. the infinite sequence itself is describable by some grammar

What kind of experience is required for success?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

What kind of experience is required for success?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

1. Identification in the limit from positive data (Gold 1967)

What kind of experience is required for success?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

2. Identification in the limit from positive and negative data

(Gold 1967)

What kind of experience is required for success?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

3. Identification in the limit from positive data from r.e. texts
   (Gold 1967)
4. Learning context-free and r.e. distributions
   (Horning 1969, Angluin 1988)

What kind of experience is required for success?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

5. Probably Approximately Correct learning
   (Valiant 1984, Anthony and Biggs 1991, Kearns and Vazirani 1994

What counts as success?

We are interested in learners of *classes of languages* and not just a single language.

Why?

What counts as success?

We are interested in learners of *classes of languages* and not just a single language.

Why?

Because every language can be learned by a constant function!

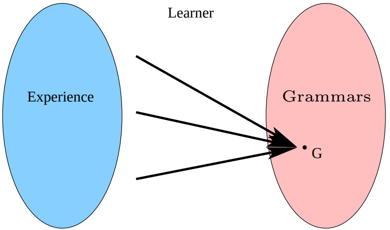


Figure: Learners are functions ϕ from experience to grammars.

Formal Learning Theory

Learning requires a structured hypothesis space, which excludes at least some finite-list hypotheses.

Gleitman 1990, p. 12:

> *'The trouble is that an observer who notices everything can learn nothing for there is no end of categories known and constructable to describe a situation [emphasis in original].'*

Formal Learning Theory

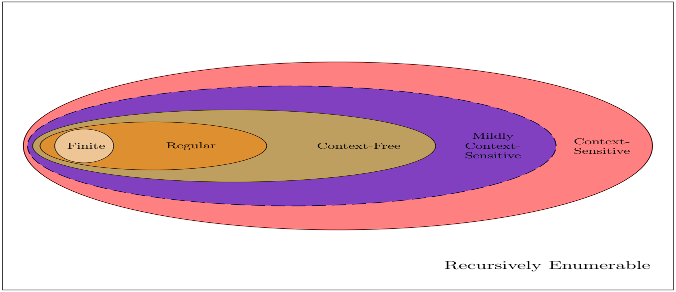
Learning requires a structured hypothesis space, which excludes at least some finite-list hypotheses.

Gleitman 1990, p. 12:

> *'The trouble is that an observer who notices everything can learn nothing for there is no end of categories known and constructable to describe a situation [emphasis in original].'*

Results of Formal Learning Theories: Do feasible learners exist?

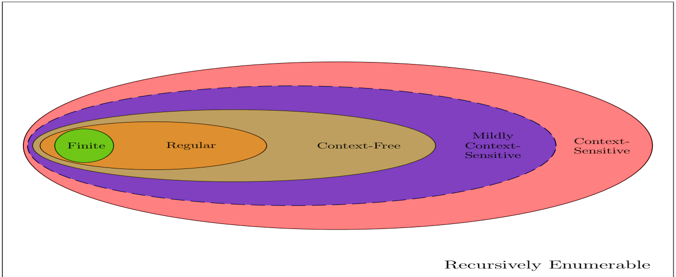
Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence



Results of Formal Learning Theories: Do feasible learners exist?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

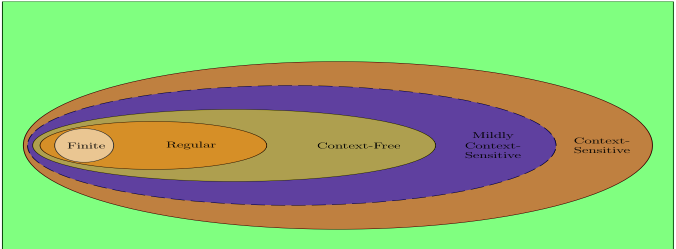
1. Identification in the limit from positive data (Gold 1967)



Results of Formal Learning Theories: Do feasible learners exist?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

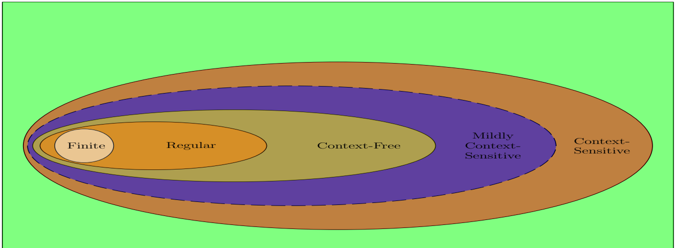
2. Identification in the limit from positive and negative data (Gold 1967)



Results of Formal Learning Theories: Do feasible learners exist?

Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

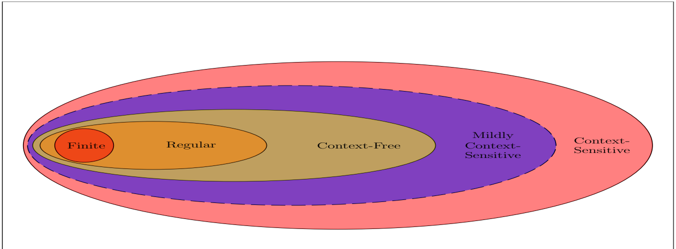
3. Identification in the limit from positive data from r.e. texts (Gold 1967)
4. Learning context-free and r.e. distributions (Horning 1969, Angluin 1988) (See Clark and Thollard 2004 and other refs in Clark's earlier talk today.)



Results of Formal Learning Theories: Do feasible learners exist?

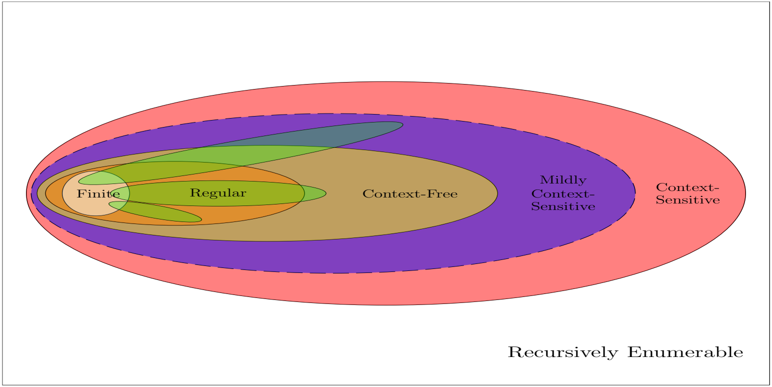
Makes learning easier	Makes learning harder
positive and negative evidence	positive evidence only
noiseless evidence	noisy evidence
queries permitted	queries not permitted
approximate convergence	exact convergence
complete infinite sequences	any infinite sequence
computable infinite sequences	any infinite sequence

5. Probably Approximately Correct learning
   (Valiant 1984, Anthony and Biggs 1991, Kearns and Vazirani 1994



Formal Learning Theory: Positive Results

Many classes which cross-cut the Chomsky hierarchy and exclude some finite languages are feasibly learnable in the senses discussed (and others).



(Angluin 1980, 1982, Garcia et al. 1990, Muggleton 1990, Denis et al. 2002, Fernau 2003, Yokomori 2003, Clark and Thollard 2004, Oates et al. 2006, Niyogi 2006, Clark and Eryaud 2007, Heinz 2008, to appear, Yoshinaka 2008, Case et al. 2009, de la Higuera 2010)

Summary

1. Natural language patterns are not arbitrary: there are limits to the variation.
2. Structured, restricted hypothesis spaces, which crucially exclude some finite languages, can be feasibly learned.
3. The positive learning results are proven results, and the proofs are often constructive.

What is the space of possible phonolgical patterns?

Wilson (earlier today): What is the space of possible constraints?

1. I am not claiming the following learners are the full story.
2. I am claiming that they are good approximations to the full story and that the full story will incorporate their key elements.
3. The role of phonological features, prosody, similarity, sonority, and phonetic factors more generally is ongoing and fully compatible with the present proposals. (Wilson 2006, Hayes and Wilson 2008, Moreton 2008, Albright 2009, and their talks at this event)

Local sound patterns

Distinctions are made on the basis of contiguous subsequences.

possible English words	impossible English words
thole	ptak
plast	hlad
flitch	sram
	mgla
	vlas
	dnom
	rtut

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip ✓

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip ✓

ptip

Local sound patterns and formal language theory

1. The formal languages which make distinctions on the basis of k-long contiguous subsequences are called Strictly k-Local (McNaughton and Papert 1971, Rogers and Pullum 2007)
2. They are subregular and exclude some finite languages.
3. If every k-long contiguous subsequence is licensed by the grammar, the word belongs to the language.

stip ✓

ptip $\times$

Long-distance sound patterns

Distinctions are made on the basis of potentially discontiguous subsequences.

possible Chumash words	impossible Chumash words
shtoyonowonowash	stoyonowonowaʃ
stoyonowonowas	ʃtoyonowonowas
pisotonosikiwat	pisotonoʃikiwat

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos   sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos   sotoʃ

Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

sotos ✓ sotoʃ

Long-distance sound patterns and formal language theory

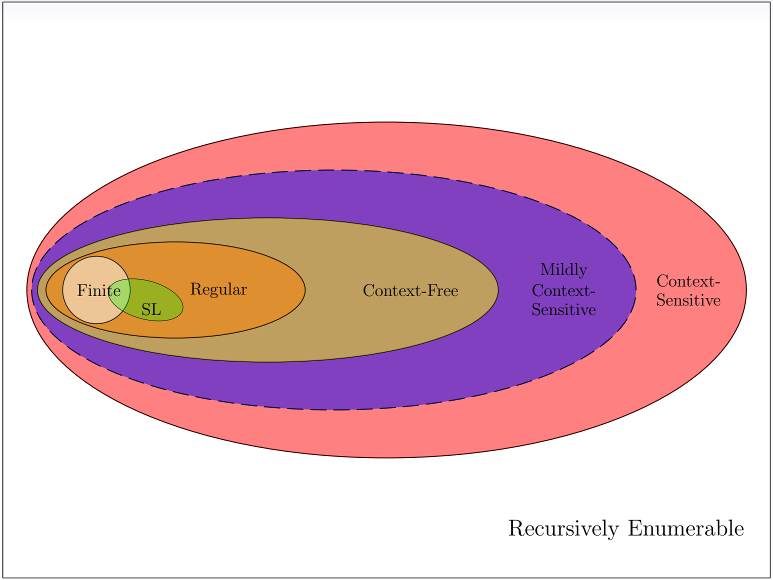
1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

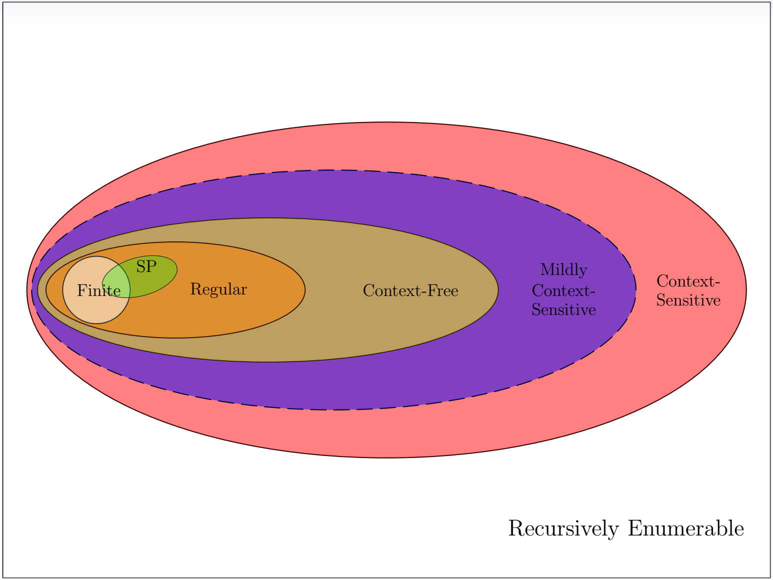
sotos ✓ sotoʃ

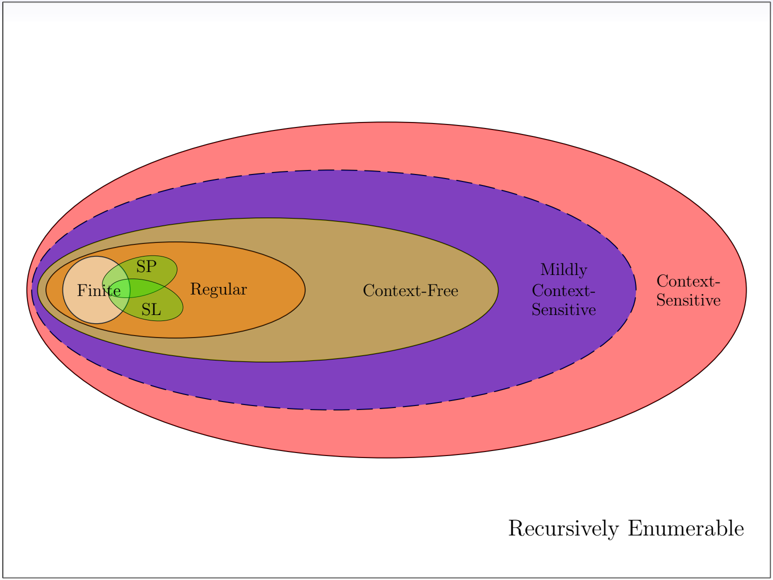
Long-distance sound patterns and formal language theory

1. The formal languages and distributions which make distinctions on the basis of k-long (potentially discontiguous) subsequences are called Strictly k-Piecewise (Heinz 2007, Rogers et al. 2009, Heinz to appear, Heinz and Rogers to appear).
2. They are subregular and exclude some finite languages.
3. Consonantal harmony patterns with blocking are not Strictly Piecewise for any k.
4. Harmony patterns which apply only to the first and last sounds are not Strictly Piecewise for any k.
5. Strictly k-Piecewise models underlie models of reading comprehension (Schoonbaert and Grainger2004, Grainger and Whitney2004)
6. If every k-long subsequence is licensed by the grammar, the word belongs to the language.

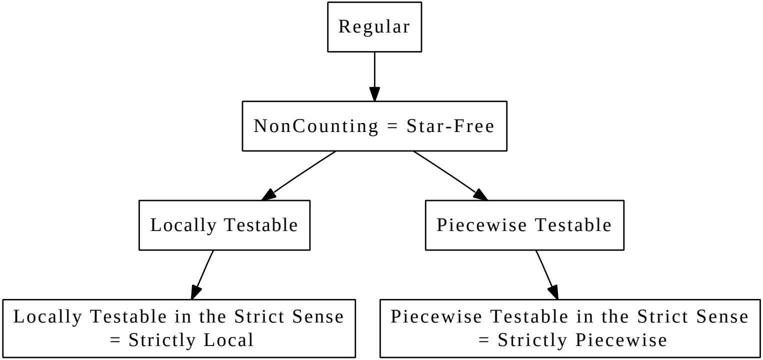
sotos ✓ sotoʃ ×


Finite
SL
Regular
Context-Free
Mildly
Context-
Sensitive
Context-
Sensitive
Recursively Enumerable


SP
Finite
Regular
Context-Free
Mildly
Context-
Sensitive
Context-
Sensitive
Recursively Enumerable

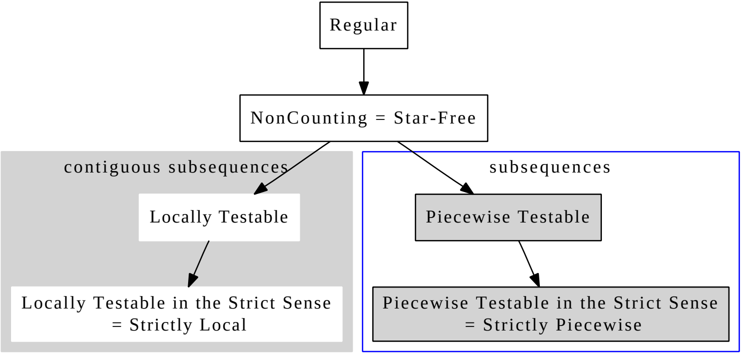

SP
Finite
Regular
SL
Context-Free
Mildly
Context-
Sensitive
Context-
Sensitive
Recursively Enumerable

Background - Subregular Hierarchies



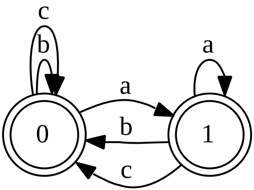
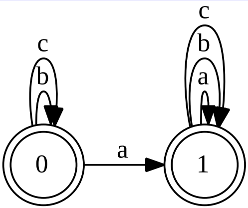
(McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007, Rogers et. al 2009, Heinz and Rogers to appear)

Background - Subregular Hierarchies



(McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007, Rogers et. al 2009, Heinz and Rogers to appear)

Strictly Local and Strictly Piecewise Models

Strictly 2-Local	Strictly 2-Piecewise
Contiguous subsequences	Subsequences (discontiguous OK)
Successor (+1)	Less than (<)
`.*ab.*`	`.*a.*b.*`
Immediate Predecessor	Predecessor
	
0 = have not just seen an [a]	0 = have never seen an [a]
1 = have just seen an [a]	1 = have seen an [a] earlier

Similar but different functions

Strictly k-Local The function SL_k picks out the k-long contiguous subsequences.

$$\mathrm{SL}_2(\mathrm{stip}) = \{st, ti, ip\}$$

Strictly k-Piecewise The function SP_k picks out the k-long (potentially discontiguous) subsequences.

$$\mathrm{SP}_2(\mathrm{stip}) = \{st, si, sp, ti, tp, ip\}$$

Similar but different

Strictly k-Local Grammars are subsets of k-long sequences. Languages are all words w such that $\mathrm{SL}_k(w) \subseteq G$.

$$\text{stip} \in L(G)$$
$$\text{iff}$$
$$SL_2(\text{stip}) \in G$$

Strictly k-Piecewise Grammars are subsets of k-long sequences. Languages are all words w such that $\mathrm{SP}_k(w) \subseteq G$.

$$\text{stip} \in L(G)$$
$$\text{iff}$$
$$\mathrm{SP}_2(\text{stip}) \in G$$

Learning is also similar but different.

1. Stricly k-Local languages are identifiable in the limit from positive data (Garcia et al. 1990).
2. **Keep track of the observed k-long contiguous subsequences.**

time	word w	$SL_2(w)$	Grammar G	$L(G)$
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{aa\}$	$\{\mathbf{aa}\}$	aaa^*
1	aab	$\{aa,\ ab\}$	$\{aa,\ \mathbf{ab}\}$	$aaa^* \cup aaa^*b$
2	ba	$\{ba\}$	$\{aa,\ ab,\ \mathbf{ba}\}$	$\Sigma^*/\Sigma^*bb\Sigma^*$
...				

The Strictly 2-Local learner learns *bb

Learning long-distance sound patterns

1. Stricly k-Piecewise languages are identifiable in the limit from positive data (Heinz 2007, to appear).
2. **Keep track of the observed k-long subsequences.**

i	$t(i)$	$SP_2(t(i))$	Grammar G	Language of G
-1			$\emptyset$	$\emptyset$
0	$aaaa$	$\{\lambda, a, aa\}$	$\{\boldsymbol{\lambda}$**, a, aa**$\}$	a^*
1	aab	$\{\lambda, a, b, aa, ab\}$	$\{\lambda$, a, aa, **b, ab**$\}$	$a^* \cup a^*b$
2	baa	$\{\lambda, a, b, aa, ba\}$	$\{\lambda$, a, b, aa, ab, **ba**$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
3	aba	$\{\lambda, a, b, ab, ba\}$	$\{\lambda$, a, b, aa, ab, ba$\}$	$\Sigma^* \backslash (\Sigma^* b \Sigma^* b \Sigma^*)$
...				

The learner ϕ_{SP_2} learns *b...b

What about distributional learning?

1. Stricly k-Local distributions can be efficiently estimated (Jurafsky & Martin 2008) (they are n-gram models)
2. Strictly k-Piecewise distributions can be efficiently estimated (Heinz and Rogers to appear)

Regular Languages and Distributions

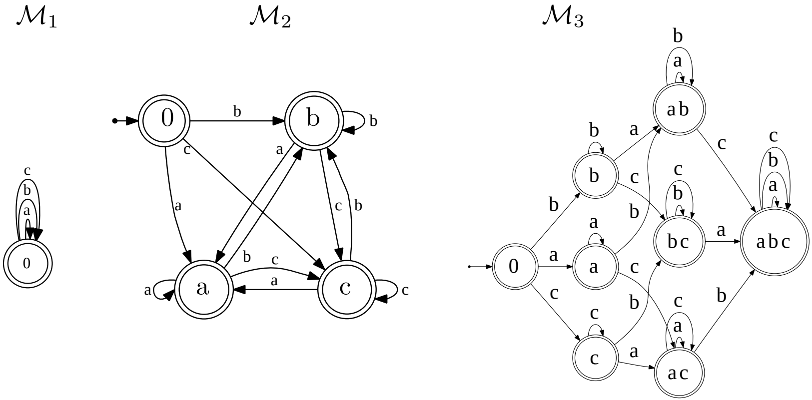
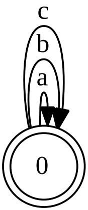


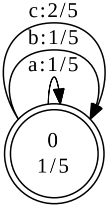
Figure: $\Sigma = \{a, b, c\}$. Each FSA is deterministic and accepts Σ^*. Each DFA represents a family of distributions. A particular distribution is given by assigning probabilities to the transitions.

Background - ML Estimatation of Subregular Distributions (structure is known)

$\mathcal{M}$



$\mathcal{M}'$



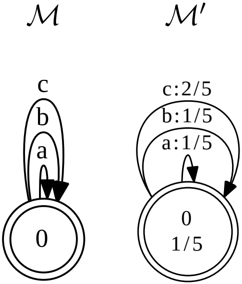
$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

Theorem (1)

Let $\mathcal{M}$ and $\mathcal{M}'$ be DFAs with the same structure and let $\mathcal{D}_{\mathcal{M}'}$ generate a sample S. Then ***the maximum-likelihood estimate*** *(MLE) of S with respect to $\mathcal{M}$ guarantees that $\mathcal{D}_{\mathcal{M}}$ approaches $\mathcal{D}_{\mathcal{M}'}$ as the size of S goes to infinity.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Background - ML Estimatation of Subregular Distributions (structure is known)



$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

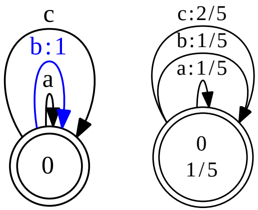
Theorem (2)

For a sample S and deterministic finite-state acceptor $\mathcal{M}$, ***counting the parse of S through $\mathcal{M}$ and normalizing at each state*** *optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Background - ML Estimatation of Subregular Distributions (structure is known)

$\mathcal{M}$ $\mathcal{M}'$



$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

$\downarrow$

S = {bc}

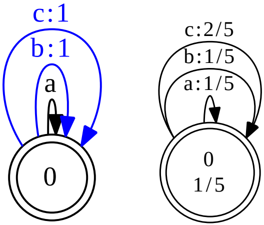
Theorem (2)

*For a sample S and deterministic finite-state acceptor $\mathcal{M}$, **counting the parse of S through $\mathcal{M}$ and normalizing at each state** optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Background - ML Estimatation of Subregular Distributions (structure is known)

$\mathcal{M}$ $\mathcal{M}'$



$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

$\downarrow$

S = {bc}

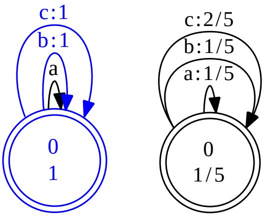
Theorem (2)

For a sample S and deterministic finite-state acceptor $\mathcal{M}$, ***counting the parse of S through $\mathcal{M}$ and normalizing at each state*** *optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Background - ML Estimatation of Subregular Distributions (structure is known)

$\mathcal{M}$ $\mathcal{M}'$



$\downarrow$

S = {bc}

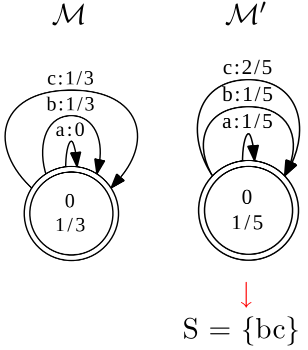
$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

Theorem (2)

For a sample S and deterministic finite-state acceptor $\mathcal{M}$, ***counting the parse of S through $\mathcal{M}$ and normalizing at each state*** *optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Background - ML Estimatation of Subregular Distributions (structure is known)



$\mathcal{M}$ represents a family of distributions with 4 parameters. $\mathcal{M}'$ represents a particular distribution in this family.

Theorem (2)

For a sample S and deterministic finite-state acceptor $\mathcal{M}$, ***counting the parse of*** *S* ***through*** *$\mathcal{M}$* ***and normalizing at each state*** *optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Bigram models (Strictly 2-Local Distributions)

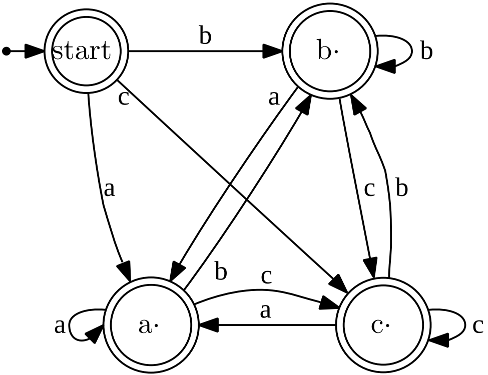


Figure: The structure of a bigram model. The 16 parameters of this model are given by associating probabilities to each transition and to “ending” at each state.

Bigram models (Strictly 2-Local Distributions)

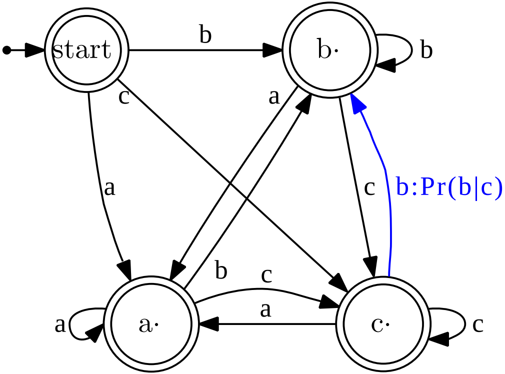


Figure: The structure of a bigram model. The 16 parameters of this model are given by associating probabilities to each transition and to "ending" at each state.

Regular Languages and Distributions

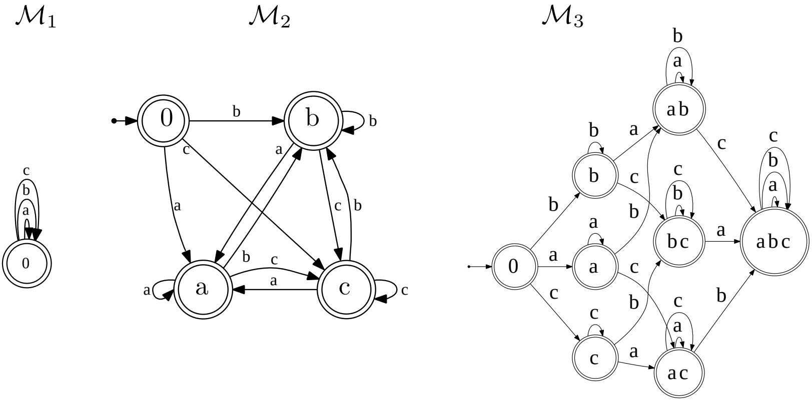
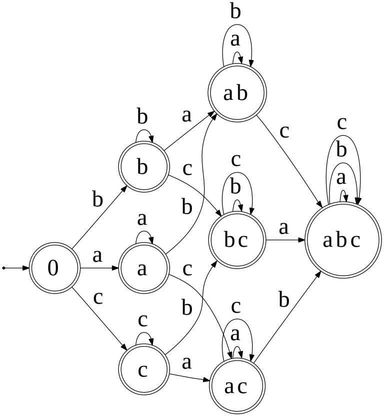


Figure: $\Sigma = \{a, b, c\}$. Each FSA is deterministic and accepts Σ^*. Each DFA represents a family of distributions. A particular distribution is given by assigning probabilities to the transitions. What do the states distinguish?

Strictly 2-Piecewise Distributions: The Problem



Equation 1

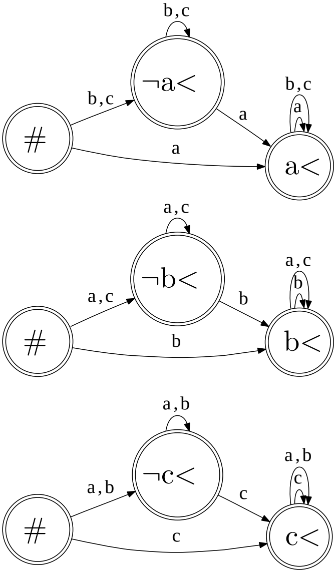
Piecewise Assumption

$$w = a_1 a_2 \dots a_n$$

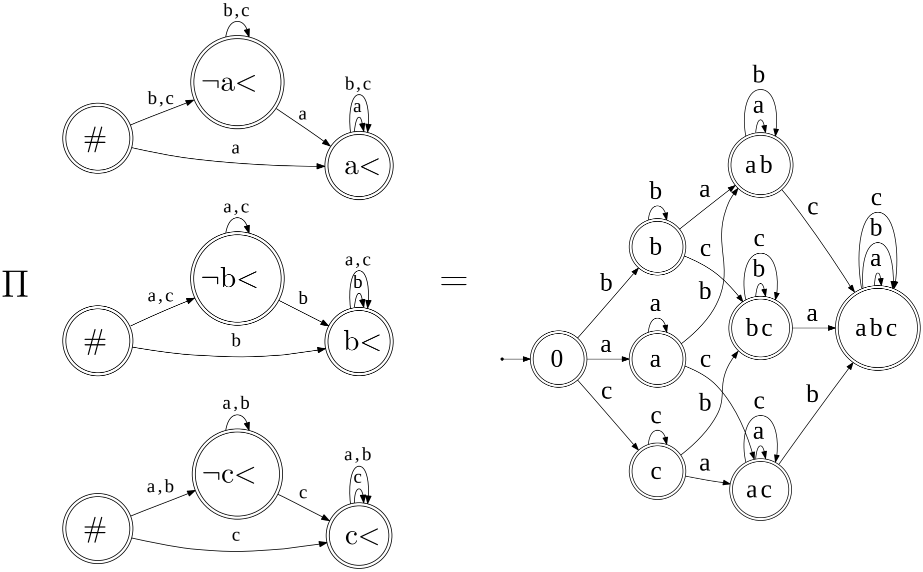
$$\begin{aligned} Pr(w) = \; & Pr(a_1 \mid \#) \\ & \times Pr(a_2 \mid a_1 <) \\ & \times \dots \\ & \times Pr(a_n \mid a_1, \dots, a_{n-1} <) \\ & \times Pr(\# \mid a_1, \dots a_n <) \end{aligned}$$

- What is $Pr(a \mid S <)$?
  There are $2^{|\Sigma|}$ distinct sets S which suggests there are too many(!) independent parameters in the model.
- Fails to capture intuition regarding *ʃtoyonowonowas*:
  $Pr(\text{s} \mid \text{ʃ,t,o,y,w,n,a} <)$ is **not** independent of $Pr(\text{s} \mid \text{ʃ} <)$.

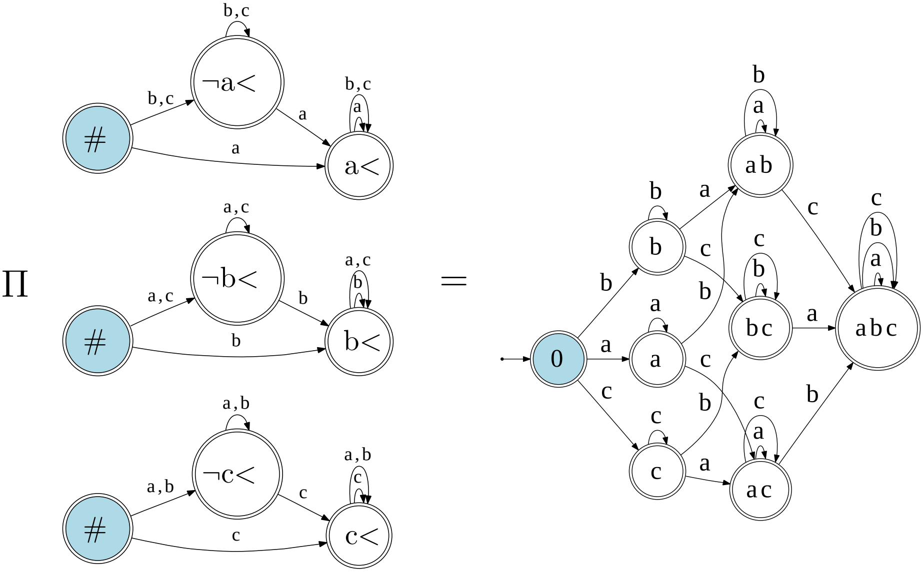
Factors of Strictly 2-Piecewise Distributions



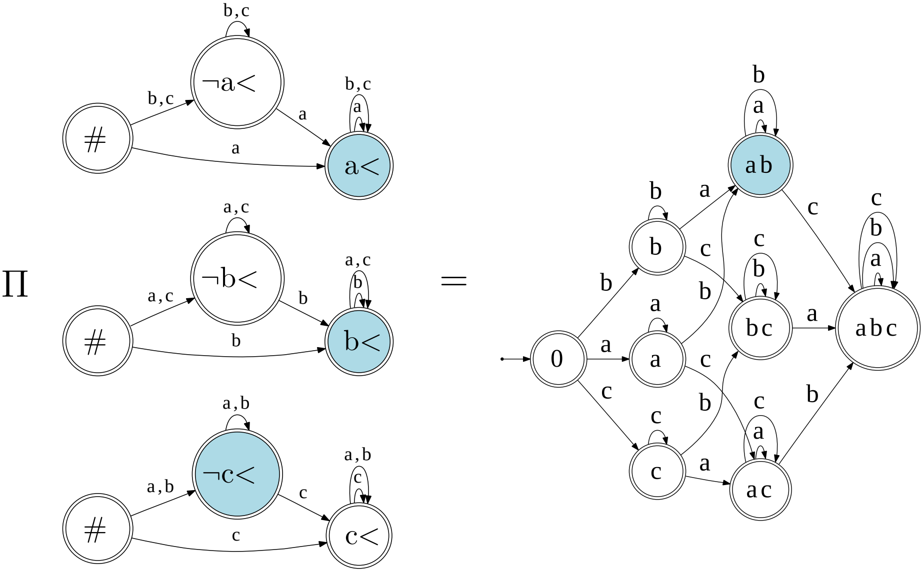
Factors of Strictly 2-Piecewise Distributions



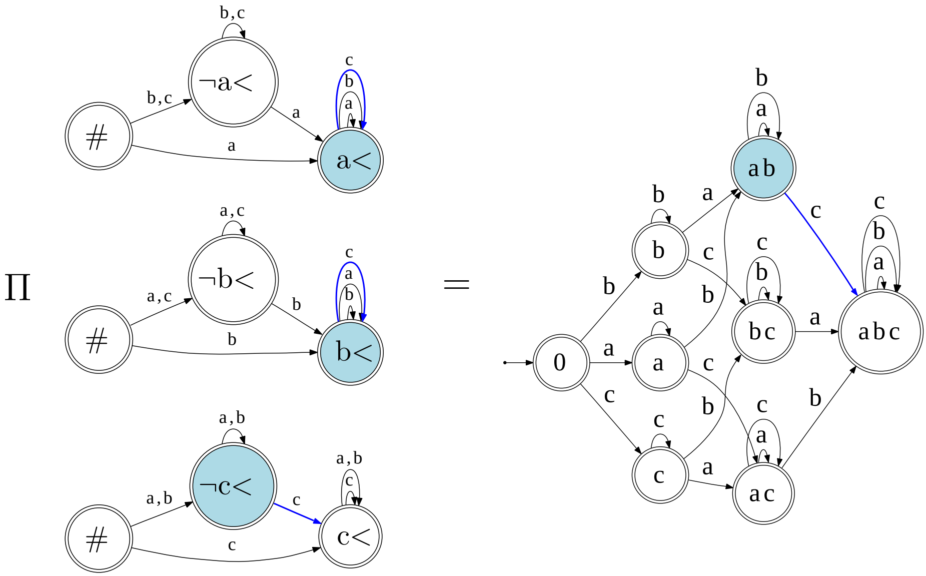
Factors of Strictly 2-Piecewise Distributions



Factors of Strictly 2-Piecewise Distributions



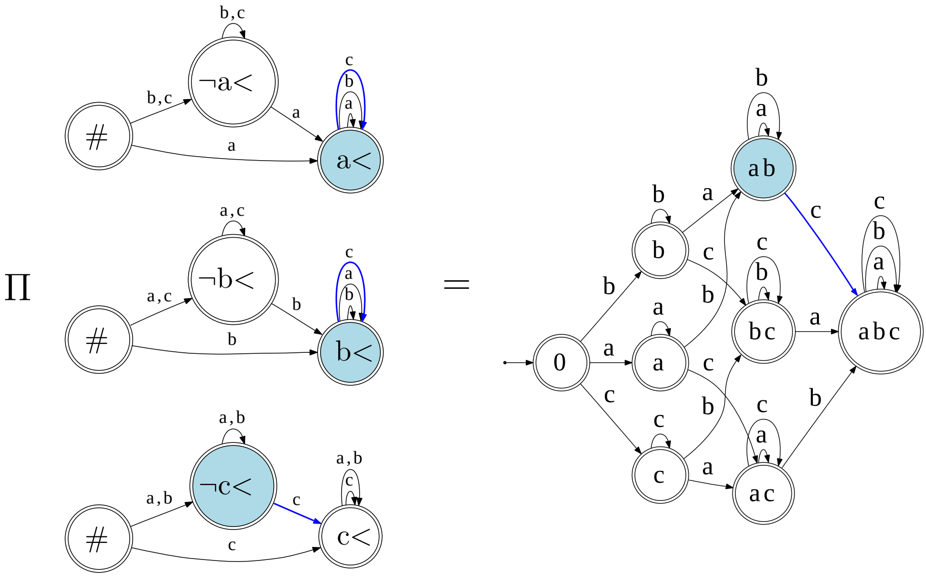
Factors of Strictly 2-Piecewise Distributions



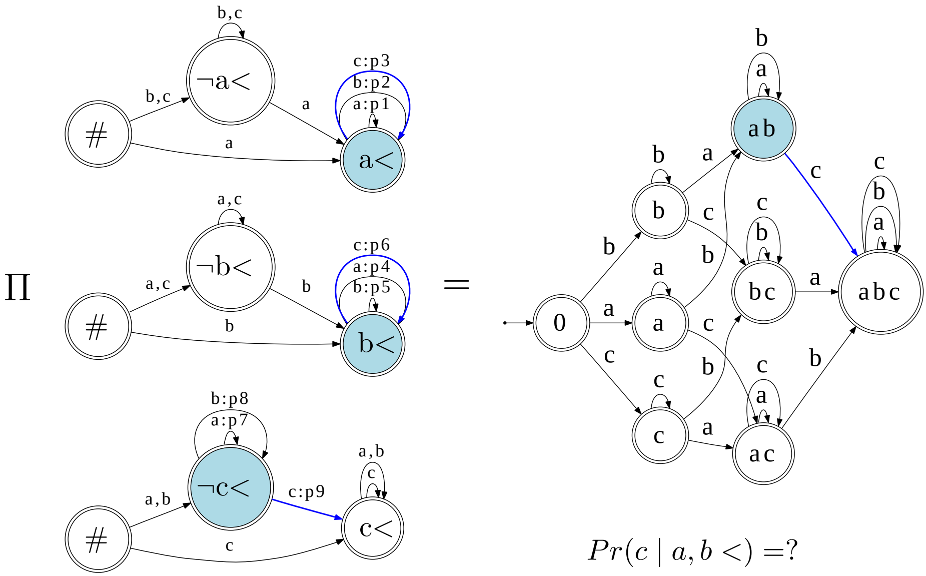
Strictly 2-Piecewise Distributions: Probabilities

How are the probabilities determined?

Strictly 2-Piecewise Distributions: Probabilities

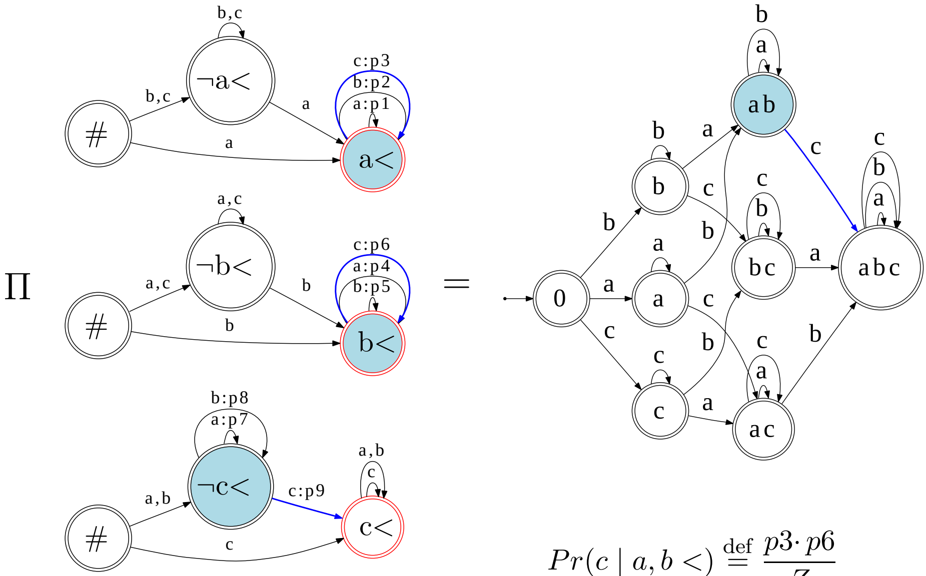


Strictly 2-Piecewise Distributions: Probabilities



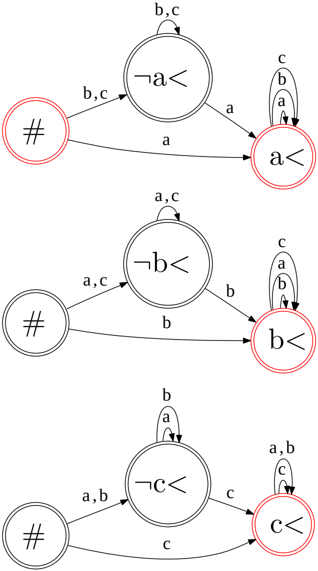
$$Pr(c \mid a, b <) = ?$$

Strictly 2-Piecewise Distributions: Probabilities



$$Pr(c \mid a, b <) \stackrel{\text{def}}{=} \frac{p3 \cdot p6}{Z}$$

Strictly 2-Piecewise Distributions: Theorem



Equation 2
(normalized co-emission product)

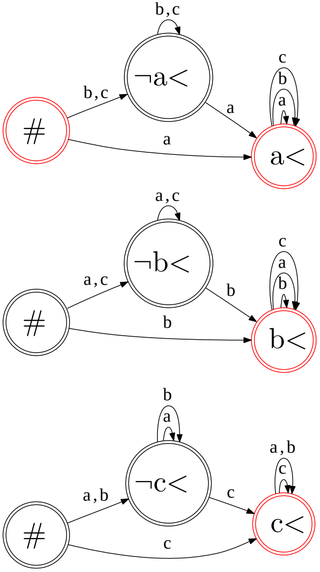
$$Pr(a \mid S <) \stackrel{\text{def}}{=}$$

$$\frac{\prod_{s \in S} Pr(a \mid s <)}{Z = \sum_{a' \in \Sigma \cup \{\#\}} \prod_{s \in S} Pr(a' \mid s)}$$

Theorem (Heinz and Rogers)

Equations (1) and (2) guarantee a well-formed probability distribution over all logically possible words. The distribution has $(|\Sigma| + 1)^2$ *parameters (but distinguishes* $2^{|\Sigma|}$ *states).*

ML Estimation of Factorable Distributions



$$\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \mathcal{M}_n$$

Estimate the factors, not their product!

Theorem (Heinz and Rogers)

The maximum likelihood estimate of a data sample drawn from a Strictly k-Piecewise distribution is obtained by finding the MLE estimates of the sample with respect to the PDFAs which factor the distribution.

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)

Chumash: Results of SP2 ML estimation

$P(x \mid y <)$		x			
		s	t͡s	ʃ	t͡ʃ
y	s	0.0325	0.0051	0.0013	0.0002
	t͡s	0.0212	0.0114	0.0008	0.
	ʃ	0.0011	0.	0.067	0.0359
	t͡ʃ	0.0006	0.	0.0458	0.0314

(Collapsing laryngeal distinctions)

It follows that, according to the model,
$Pr(\text{ʃ}toyonowonowa\text{ʃ}) \gg Pr(stoyonowonowa\text{ʃ})$.

Local Summary

1. Like the regions in the Chomsky hierarchy, the Strictly Local and Strictly Piecewise classes have multiple, independent, converging characterizarions from formal language theory, automata theory, and logic.
2. The possible grammars and languages (distributions?) form a lattice structure (Kasprzik and Kötzing, to appear).
3. They are incomparable.
4. Consequently, Strictly Local learners cannot learn Strictly Piecewise patterns and vice versa.
5. Strictly Piecewise learners cannot learn:
   - blocking patterns, e.g. *s. . . ʃ unless [z] intervenes.
   - harmony patterns which apply only to the first and last sounds.

Competing theories within phonology

1. The main alternative is the tier-based model.
   (Goldsmith 1976, Clements 1985, Sagey 1986, Mester 1988,Hayes and Wilson 2008, Goldsmith and Xanthos 2009, Goldsmith and Riggle to appear)

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

Learning unattested patterns
First and Last sound agreement

Words that start with [s] cannot end with [ʃ].

✓	×
sabika	sotoʃ
stotaʃikop	sibaʃ
pabafri	sitiʃ
...	...

Learning unattested patterns
First and last sound agreement

Words that start with [s] cannot end with [ʃ].

The function FL makes distinctions on the basis of the first and last sounds in words.

$$\mathrm{FL}(\text{sabika}) = \{\text{sa}\}$$
$$\mathrm{FL}(\text{stotaʃikop}) = \{\text{sp}\}$$
$$\mathrm{FL}(\text{pabafri}) = \{\text{pi}\}$$

Learning unattested patterns
First and last sound agreement

Words that start with [s] cannot end with [ʃ].

The function FL makes distinctions on the basis of the first and last sounds in words.

$$\text{FL(sabika)} = \{\text{sa}\}$$
$$\text{FL(stotaʃikop)} = \{\text{sp}\}$$
$$\text{FL(pabafri)} = \{\text{pi}\}$$

1. The class of such languages is identifiable in the limit from positive data.
2. The class of languages and grammars form a lattice structure.
3. The class of such distributions is efficiently estimable from positive data.

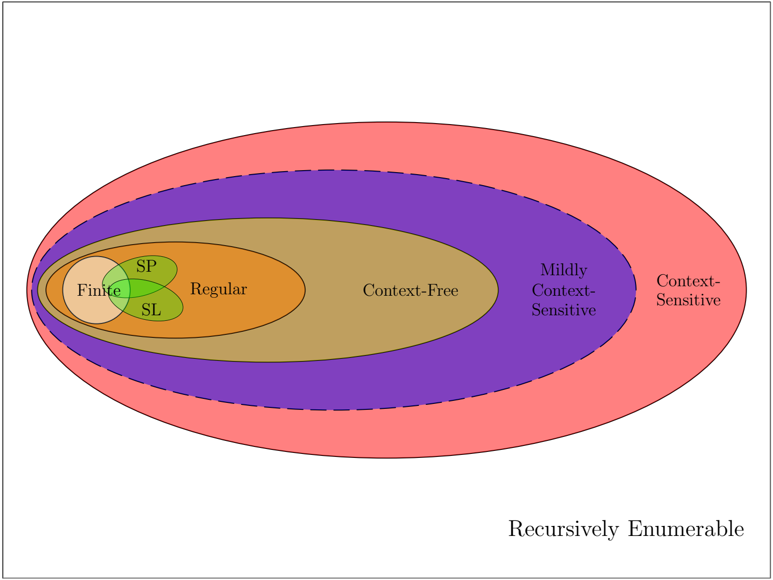
Do phonologies make First-Last distinctions?

1. To my knowledge, no such phonotactic has ever been proposed, nor is any morpho-phonological alternation conditioned by such a phonotactic.
2. Can people learn such patterns if robustly present in the data?

Do phonologies make First-Last distinctions?

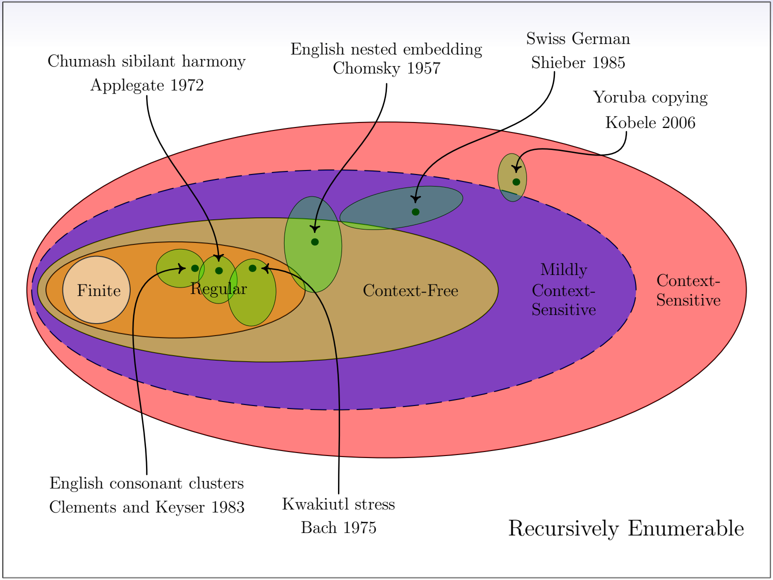
1. To my knowledge, no such phonotactic has ever been proposed, nor is any morpho-phonological alternation conditioned by such a phonotactic.
2. Can people learn such patterns if robustly present in the data?

Domain-specific vs. domain-general?



Conclusion

1. Linguistic patterns are not arbitrary.
2. Only structured classes of patterns can be learned.
3. Distinct, feasible learning models for distinct phonological patterns exist.
4. These help explain the character of the typology.
5. A single, feasible learning model for these distinct phonological patterns will likely have to attribute the character of the typology to something else.
6. Artificial language learning experiments can help.


Chumash sibilant harmony
Applegate 1972
English nested embedding
Chomsky 1957
Swiss German
Shieber 1985
Yoruba copying
Kobele 2006
Finite
Regular
Context-Free
Mildly Context-Sensitive
Context-Sensitive
English consonant clusters
Clements and Keyser 1983
Kwakiutl stress
Bach 1975
Recursively Enumerable

Thank you

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).

Finnish: Results of SP2 estimation

$P(x \mid b <)$		x							
		u	o	a	y	oe	ae	i	e
b	u	0.056	0.040	0.118	0.006	0.002	0.007	0.084	0.072
	o	0.046	0.033	0.120	0.005	0.002	0.007	0.110	0.067
	a	0.045	0.031	0.130	0.005	0.002	0.007	0.095	0.060
	y	0.015	0.016	0.038	0.044	0.026	0.066	0.091	0.072
	oe	0.023	0.027	0.058	0.030	0.014	0.053	0.095	0.067
	ae	0.014	0.014	0.034	0.036	0.015	0.086	0.091	0.073
	i	0.030	0.031	0.097	0.011	0.006	0.0240	0.088	0.080
	e	0.031	0.026	0.077	0.014	0.005	0.031	0.089	0.071

	F	G
a	+	-
b	+	+
c	-	+

Table: An example of a feature system with $\Sigma = \{a, b, c\}$ and two features F and G.

Feature-based generalizations

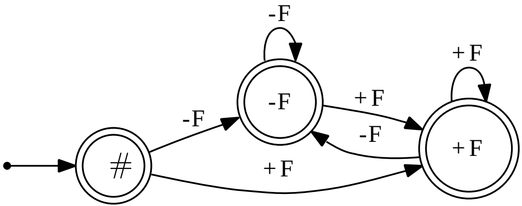


Figure: $\mathcal{M}_F$ represents a SL_2 distribution with respect to feature F.

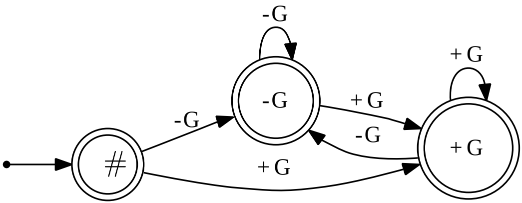


Figure: $\mathcal{M}_G$ represents a SL_2 distribution with respect to feature G.

Feature-based generalization

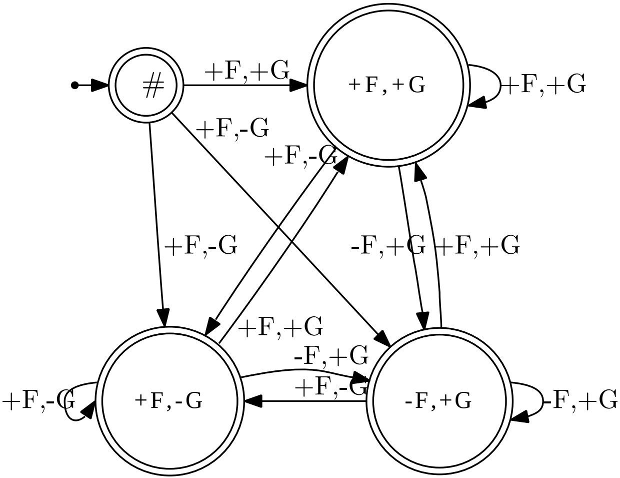


Figure: The structure of the feature product of $\mathcal{M}_F$ and $\mathcal{M}_G$.