What Does "Simplicity" Mean in Grammar Learning?

Jon Rawski & Jeff Heinz

Department of Linguistics
Institute for Advanced Computational Science
Stony Brook University

Phonotactic Knowledge (Halle 1978)

Speakers' knowledge of possible and impossible sequences:

ptak, thole, hlad, plast, sram, mgla, vlas, flitch, dnom, rtut

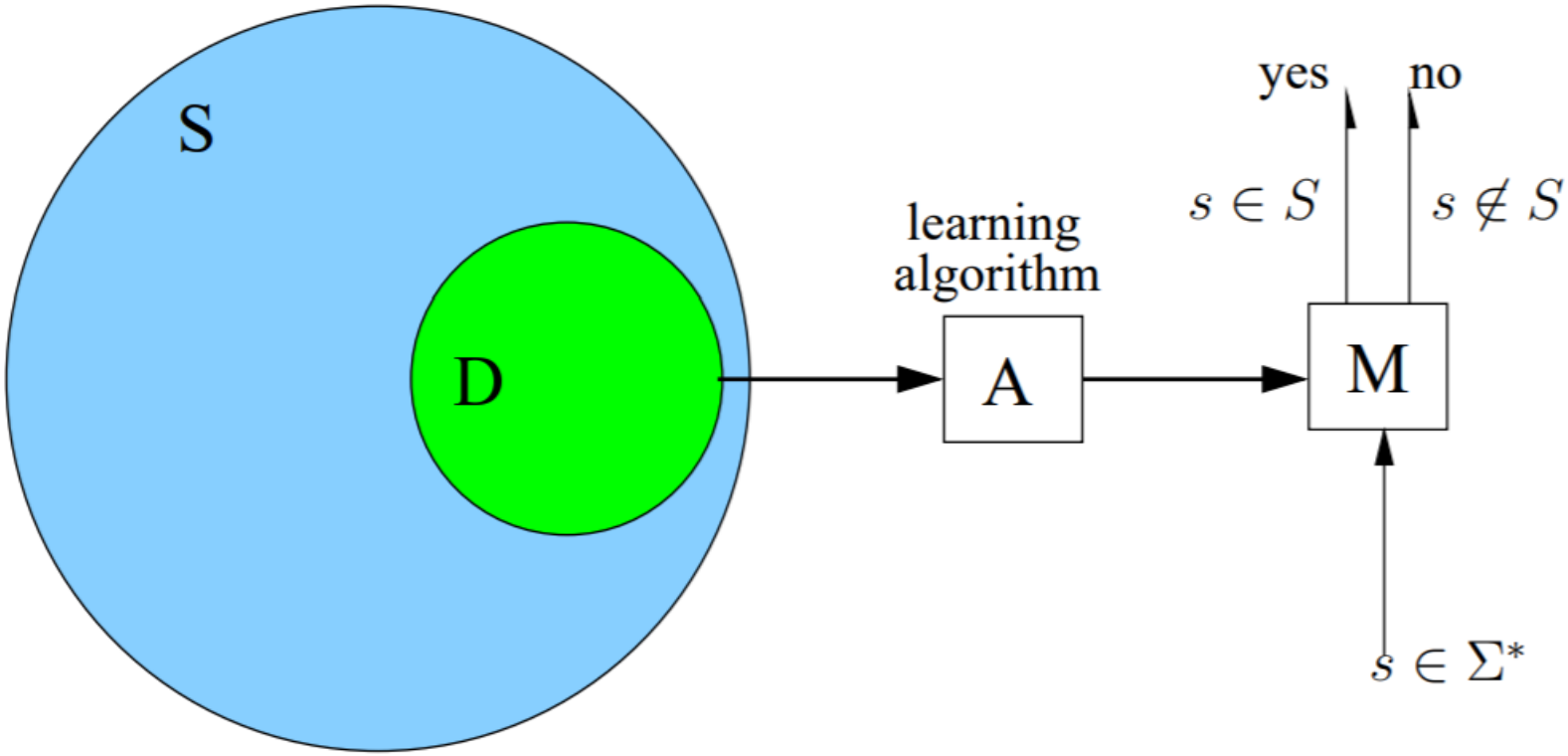
Infants show early sensitivity to phonotactic patterns (5-9mo)
(Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993,Jusczyk, Luce, & Charles-Luce, 1994, Sundara & Breiss 2020)

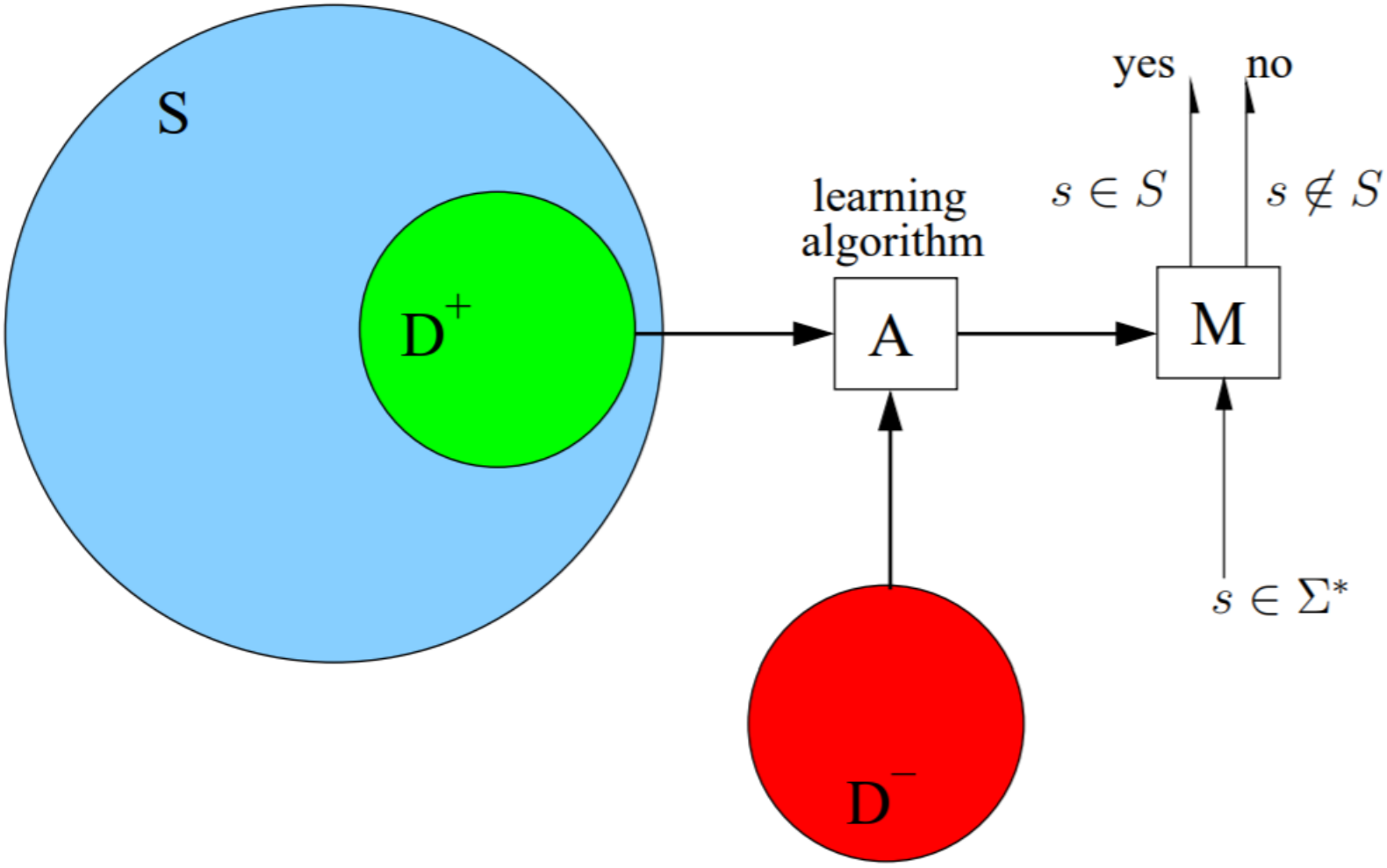
Phonotactic knowledge robust in speech, sign, and pro-tactile sign
Sandler & Lillo-Martin 2006, Edwards 2014

Some Assumptions about Phonological Behavior

- Perceive/process/generate linear sequence of (sub)events
- Can model as strings—linear sequence of abstract symbols
  - Discrete linear order (initial segment of $\mathbb{N}$).
  - Labeled with alphabet of events
  - Partitioned into subsets, each the set of positions at which some event occurs.

Function	Description	Linguistic Correlate
$f : \Sigma^* \to \{0, 1\}$	Binary classification	(well-formedness)
$f : \Sigma^* \to [0, 1]$	strings to real values	(gradient well-formedness)
$f : \Sigma^* \to \Delta^*$	strings to strings	(single-valued transformation)
$f : \Sigma^* \to \wp(\Delta^*)$	strings to stringsets	(multi-valued transformation)


S
D
learning
algorithm
A
M
yes
no
$s \in S$
$s \notin S$
$s \in \Sigma^*$


S
D^+
learning
algorithm
A
M
yes
no
$s \in S$
$s \notin S$
$s \in \Sigma^*$
D^-

"It's really more intelligent to be able to simplify things than to complicate them. Even if some people think it makes you look stupid."

Eugenia Cheng, 2015

"We should somehow specify additional constraints on the generating process. But which constraints are plausible? Which reflect the "natural" concept of simplicity?"

Schmidhuber 2002

"The real problem is that of developing a hypothesis about initial structure that is sufficiently restrictive to account for acquisition of language, yet not so restrictive as to be inconsistent with the known diversity of language."

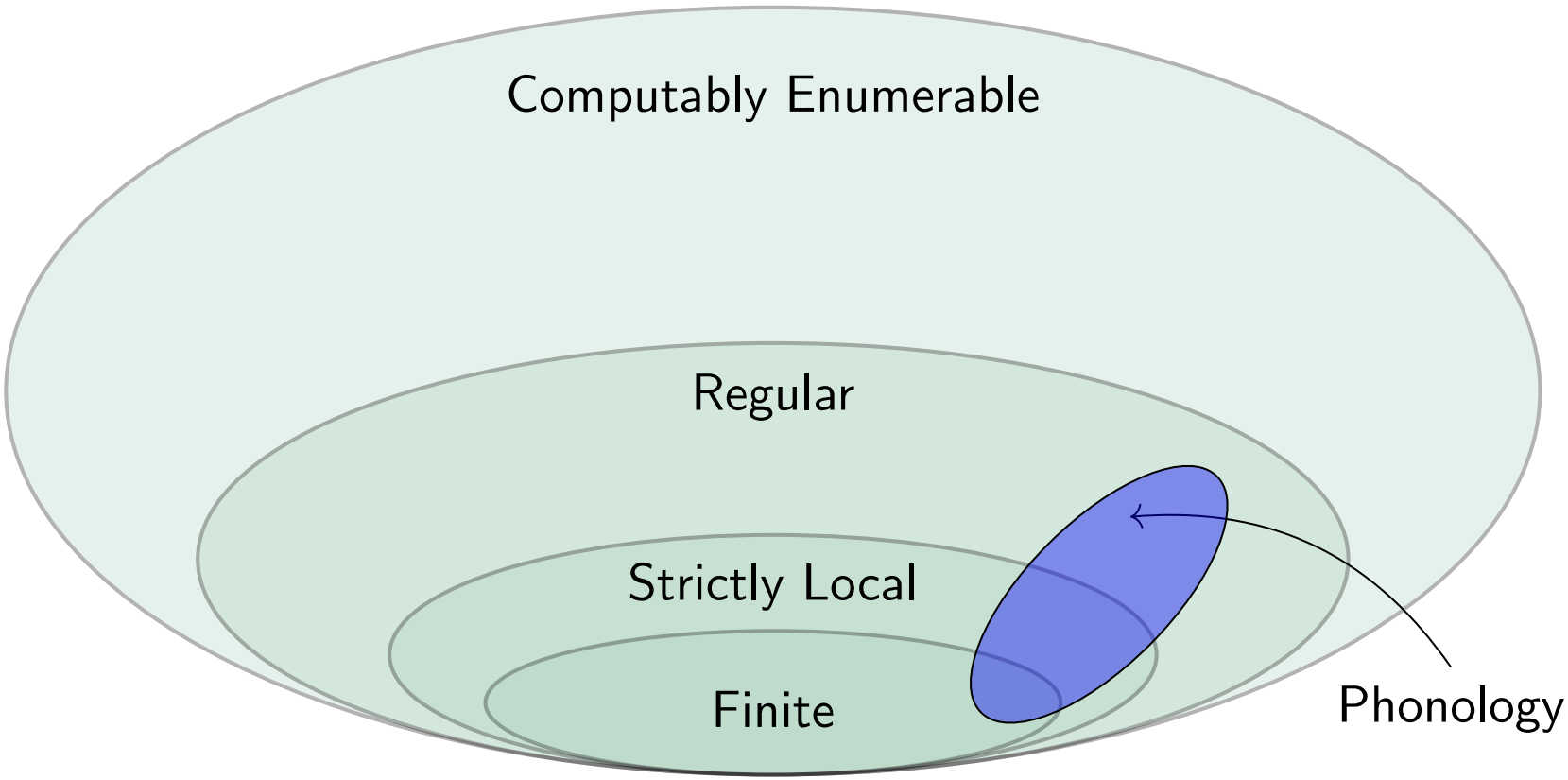
Chomsky 1965

Three notions of simplicity

Three Notions of Simplicity

- Descriptive Simplicity
- Bayesian Simplicity
- Interacting Grammar Compactness and Descriptive Complexity

Descriptive Pattern Complexity



Defining Property of Strict Locality

Substring Substitution Closure

There is some number $k > 1$ s.t. if a language L contains strings uxv and $u'xv'$, where x has length $k-1$, L contains uxv'

Theorem: A language is Strictly k-Local if it satisfies k-SSC

$(ab)^+$ is SL-2

	x		
ab	a	b	$\in L$
$abab$	a	bab	$\in L$
ab	a	bab	$\in L$

Even-A is not SL

	x		
a	$a \cdots a$	a	$\in L$
	$a \cdots a$		$\in L$
a	$a \cdots a$		$\notin L$

German Intervocalic *s* voicing is SL

- In GERMAN, [s] is not allowed in-between two vowels:
  - (1) fa[z]er 'fiber'
  - (2) rei[z]en 'to.travel'

ok r e **ɪ z e** n

* r e **ɪ s e** n

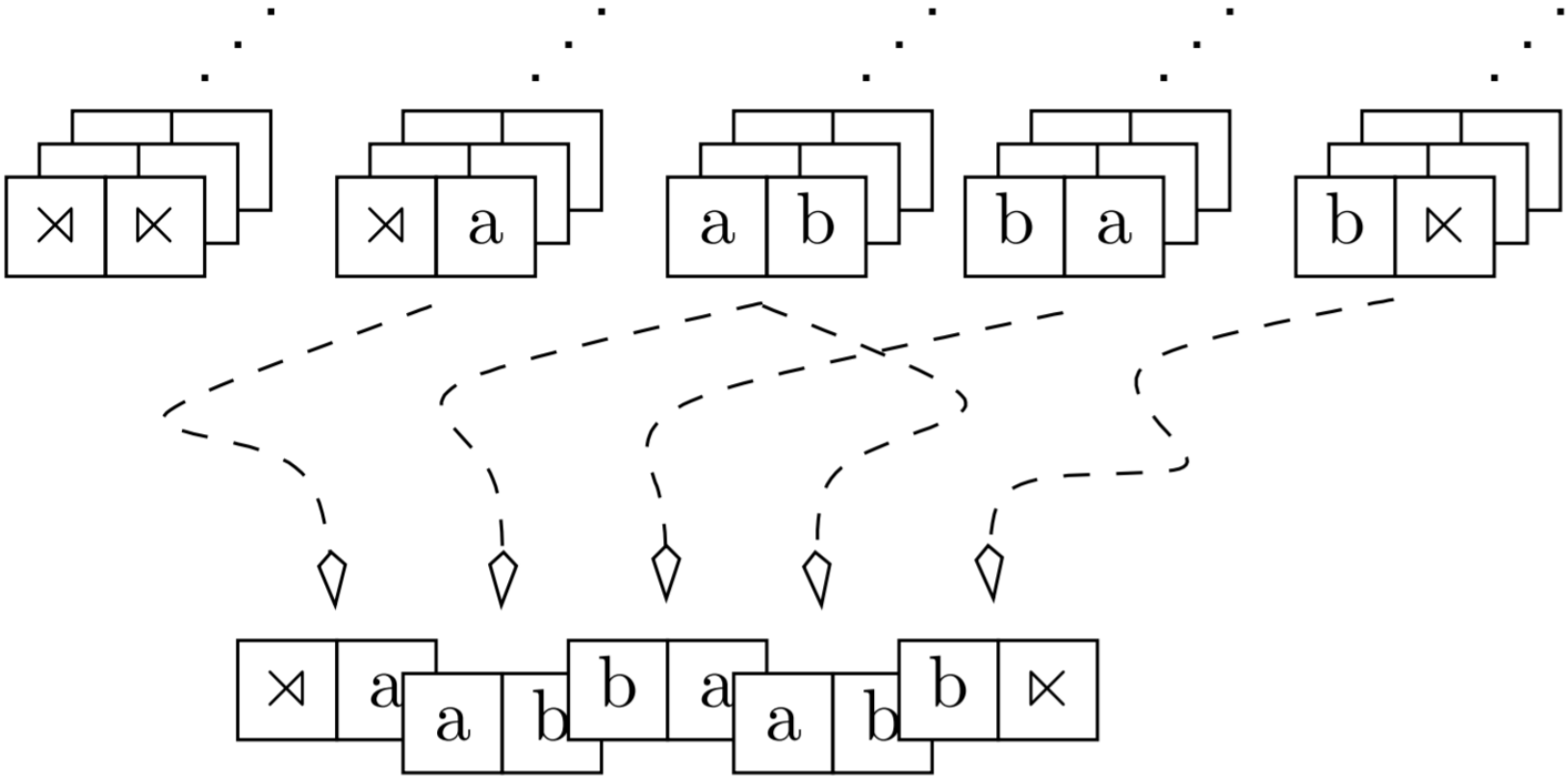
Aari long distance sibilant harmony is not SL

- In Aari, all sibilants agree in anteriority.
  - (3) baʔse 'he brought'
  - (4) ʒaʔʃit 'I arrived'

ok **ʒ** a ʔ **ʃ** i t

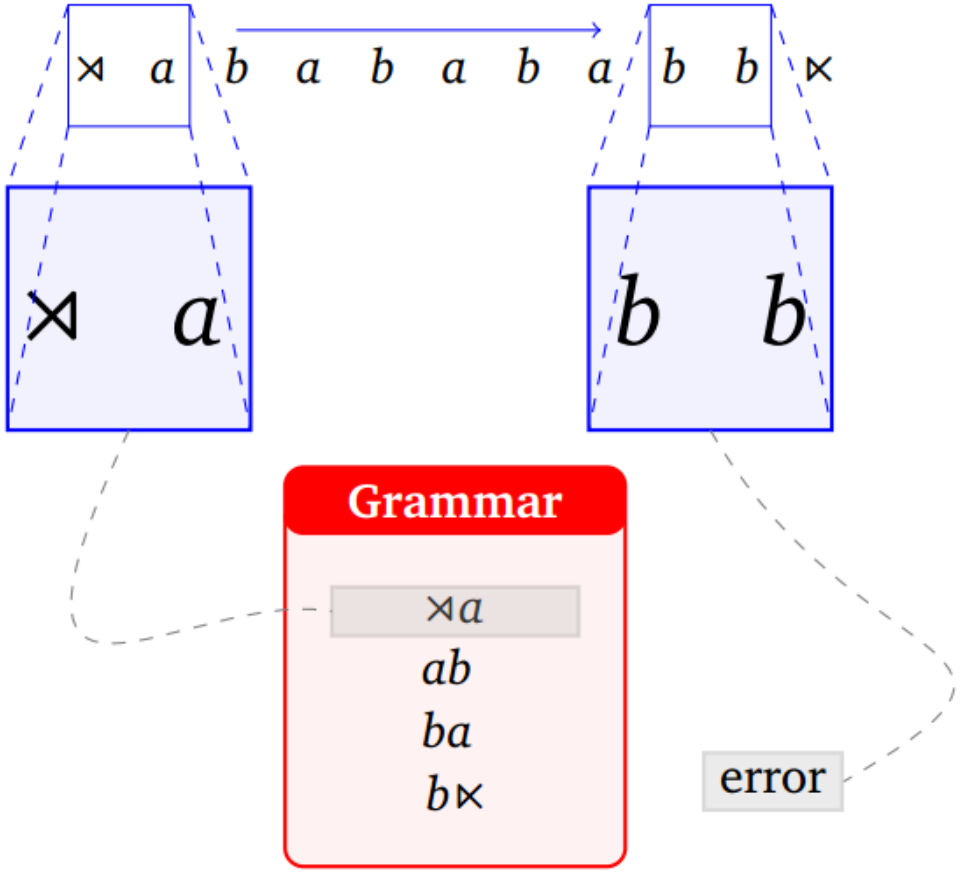
* **ʒ** a ʔ **s** i t

SL Grammaticality of *abab*



Pics courtesy of Heinz and Rogers 2014 ESSLLI course.

Strictly Local Grammars



(Garcia et al. 1991, Heinz 2010), p.c. Thomas Graf

Cognitive interpretation of SL (Rogers et al 2013)

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

Regular Languages & Finite-State Automata

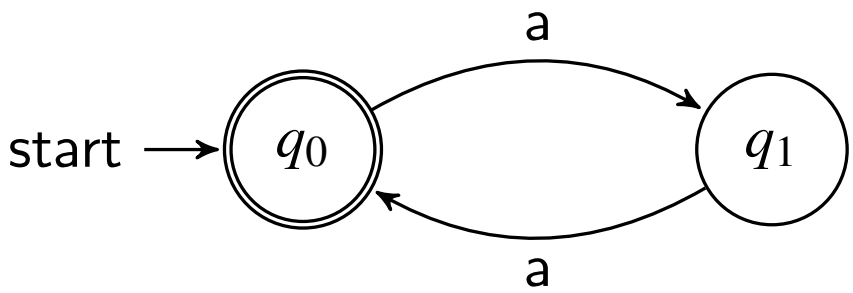
Closed under union, intersection, complement, reversal, set-difference, homomorphism, Kleene concatenation, and others

First characterization of neural nets (Kleene 1956)

Many, many equivalent models (See Pin 2020)

Length of generalization is not preserved across models

Even-A: {aa, aaaa, aaaaaa,...}



Not regular: $a^n b^n$
(but generated by a finite-state machine with a counter on top)

Cognitive Interpretation of Regular (Rogers et al 2013)

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

Learning Distinctions via Subregular Properties

Linear Time, Space-efficient online learning algorithms from positive data (Lambert, Rawski, Heinz in review)

Many finite neural network models practically function as subregular automata (Nelson et al 2020, Merrill et al 2020)

Laboratory learning: Learners biased towards subregular distinctions, even in the presence of feasible alternatives (Lai 2015, McMullin & Hansson 2019, Finley 2009, Avcu et al 2020)

A Bayesian Perspective

Bayes' Theorem organizes unknowns relating hypotheses and data
Popper: simplicity = low prior probability

- discourages generalization, rewards including irrelevant info

Jeffreys: simplicity = high prior probability
Bayesian organizing has played a simplicity role in phonological learning:

> *Given two alternative descriptions of a particular body of data, the description containing fewer such symbols will be regarded as simpler, and will, therefore, be preferred over the other*

(Halle, 1964)
Berwick (1985): Simplest hypothesis is the one requiring fewest training examples (prefix code)

Minimum Description Length (Rissanen, Solomonoff, etc)

$L(h)$: length of the shortest encoding of hypothesis h

probability of h: $P(h) \approx 2^{-L(h)}$.

joint probability: $P(h,d) \approx 2^{-L(h,d)}$

conditional probability of data given h:
$P(d \mid h) = P(h,d)/P(h) \approx 2^{-L(d|h)}$, where $L(d \mid h) = L(h,d) - L(h)$

Substituting into Bayes' theorem:

$$P(h \mid d) \approx \frac{2^{-L(d|h)}}{P(d)} 2^{-L(h)}$$

Maximal probability occurs when $L(d \mid h) + L(h)$ is minimised — when the encoding of the hypothesis and of the data in terms of the hypothesis has **minimal description length**

Chater & Vitanyi 2007

"the learner postulates the underlying structure in the linguistic input that provides the simplest, that is, briefest, description of that linguistic input."

- linguistic data can come from a nonstationary (but still computable) distribution, not just fixed-probability
- Makes "calculations that are known to be uncomputable"
- can learn any computably enumerable distribution from positive evidence

Piantdosi (2020), Piantadosi et al (in review)

- Chater & Vitanyi, but with a resource prior on top (monotonically decreasing in runtime)
- Can learn several formal languages given a Church encoding (alphabet primitives and logical operations)
- "Fix your favorite programming language, and consider hypotheses to be any program you can write in that language." - Steve (p.c.)

Important Assumption: Encoding is dependent on primitives.

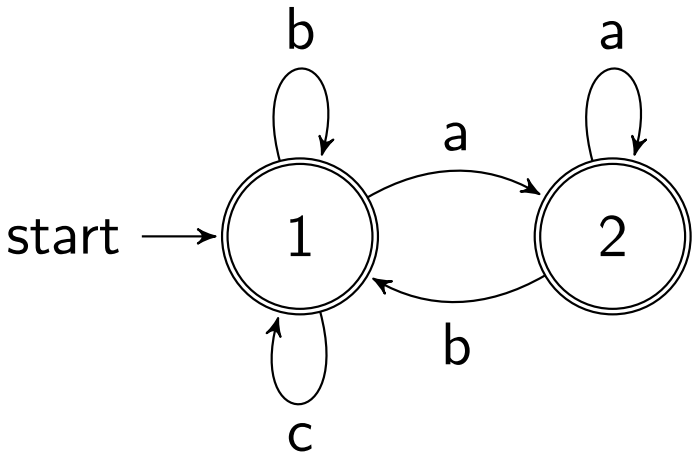
Li & Vitanyi: Kolmogorov complexity only depend up to an additive constant on the encoding language

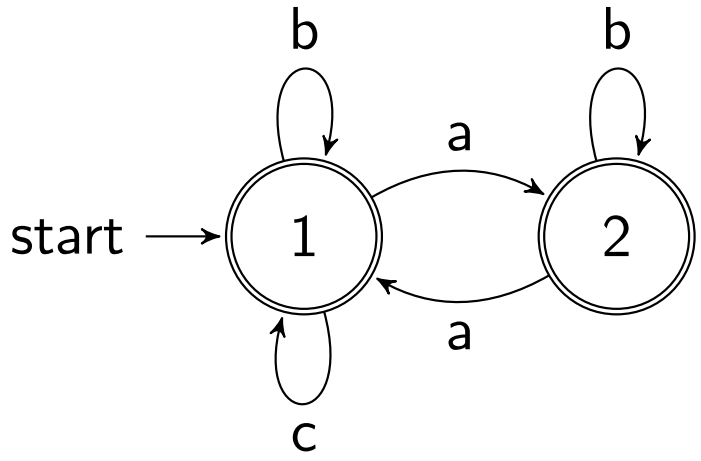
If in encoding 1, $x < y$, no guarantee that $x < y$ in encoding 2

Question: How do we get guarantees on targets, convergence, etc?

Consider two grammars:

1. $G_1 = \{$*ac$\}$
2. $G_2 = \{$*c given an even number of a's in the left context.$\}$





G_1 corresponds to an SL pattern, G_2 corresponds to a regular one. A language that is recognized by both automata:

$$L_{12} = L(G_1) \cap L(G_2) = \{c^*, abc, aaabc, aaaaabc, \dots\}$$

Subregular complexity predicts that a learner given a subset of L_{12} as an input will infer G_1, as it is the simplest grammar generating the language. What does MDL predict?

A Cautionary Tale: Morita & Koda 2019

1. analyse gibbon data (regular language) via probabilistic context-free grammars. PCFGs do not improve fit
2. Use Bayes and invoke compactness of the analysis as a fundamental advantage of this approach

The gibbon data is regular. additional hierarchical structure given by PCFG analysis is **unnecessary**

About compactness:

1. context-free descriptions of regular languages can exponentially improve the size of the recognition machine
2. This succinctness comes from adding more structure (a push-down store) to the recognizer (a finite automata)

A smaller grammar pays the price of additional computational resources.

Conclusions and discussion Questions

Notions of simplicity in grammar learning problematically interact

Questions:

1. What evidence do we have that phonology necessarily draws from a computable hypothesis space?
2. is there some encoding scheme where MDL (or any Bayesian) generalizations match subregular predictions?
3. how can we get reliable guarantees across Bayesian encoding schemes?
4. How does MDL account for typological gaps?

References I

Phonology: local assimilations

Assimilation and word-final devoicing in Russian

- *Anticipatory obstruent voicing assimilation:*
  ot dveri '*from the door*' → o[ddv]eri
  iz korobki '*out of the box*' → i[sk]oro[pk]i
- *Obstruent word final devoicing:*
  moroz '*frost*' → moro[s]
  morozy '*frosts*' → moro[z]y

- $\Sigma = \{\text{b, s, z, …, k}\}$ $G = \langle \text{*td, *zk, *bk, …, *z}\ltimes\rangle$ $n = 2$
- mozg '*brain*' → mo[sk]

  $^{*}\rtimes\text{mozg}\ltimes$ $^{*}\rtimes\text{mosg}\ltimes$ $^{*}\rtimes\text{mozk}\ltimes$ $^{ok}\rtimes\text{mosk}\ltimes$

Phonology: local assimilations

Assimilation and word-final devoicing in Russian

- *Anticipatory obstruent voicing assimilation:*
  ot dveri '*from the door*' → o[ddv]eri
  iz korobki '*out of the box*' → i[sk]oro[pk]i
- *Obstruent word final devoicing:*
  moroz '*frost*' → moro[s]
  morozy '*frosts*' → moro[z]y

- $\Sigma = \{\text{b, s, z, …, k}\}$ $G = \langle$*td, *zk, *bk, …, *z⋉$\rangle$ $n = 2$
- mozg '*brain*' → mo[sk]

  *⋊mozg⋉ *⋊mosg⋉ *⋊mozk⋉ ^{ok}⋊mosk⋉

Phonology: local assimilations

Assimilation and word-final devoicing in Russian

- *Anticipatory obstruent voicing assimilation:*
  ot dveri '*from the door*' → o[ddv]eri
  iz korobki '*out of the box*' → i[sk]oro[pk]i
- *Obstruent word final devoicing:*
  moroz '*frost*' → moro[s]
  morozy '*frosts*' → moro[z]y

- $\Sigma = \{\text{b, s, z, …, k}\}$ $G = \langle$*td, *zk, *bk, …, *z⋉$\rangle$ $n = 2$
- mozg '*brain*' → mo[sk]

  *⋊mozg⋉ *⋊mosg⋉ *⋊mozk⋉ ^{ok}⋊mosk⋉

Morphotactics: prefixes and suffixes

English affixes

- *Prefix un-:*
  unlock, unhash
- *Suffix -able:*
  lockable, hashable

- $\Sigma = \{\text{un}, \text{able}, \text{hash}, \ldots, \text{lock}\} \quad n = 2$
  $G = \langle *\rtimes\text{able}, *\text{un}\ltimes, *\text{ableun}, \ldots, *\text{lockun}, *\text{ablehash}\rangle$

- $^{ok}\rtimes\text{lock}\ltimes \qquad *\rtimes\text{able-lock}\ltimes \qquad ^{ok}\rtimes\text{un-hash-able}\ltimes$

Morphotactics: prefixes and suffixes

English affixes

- *Prefix un-:*
  unlock, unhash
- *Suffix -able:*
  lockable, hashable

- $\Sigma = \{\text{un}, \text{able}, \text{hash}, \ldots, \text{lock}\} \quad n = 2$
  $G = \langle *\rtimes\text{able}, *\text{un}\ltimes, *\text{ableun}, \ldots, *\text{lockun}, *\text{ablehash}\rangle$
- $^{ok}\rtimes\text{lock}\ltimes$ $\quad *\rtimes\text{able-lock}\ltimes$ $\quad ^{ok}\rtimes\text{un-hash-able}\ltimes$

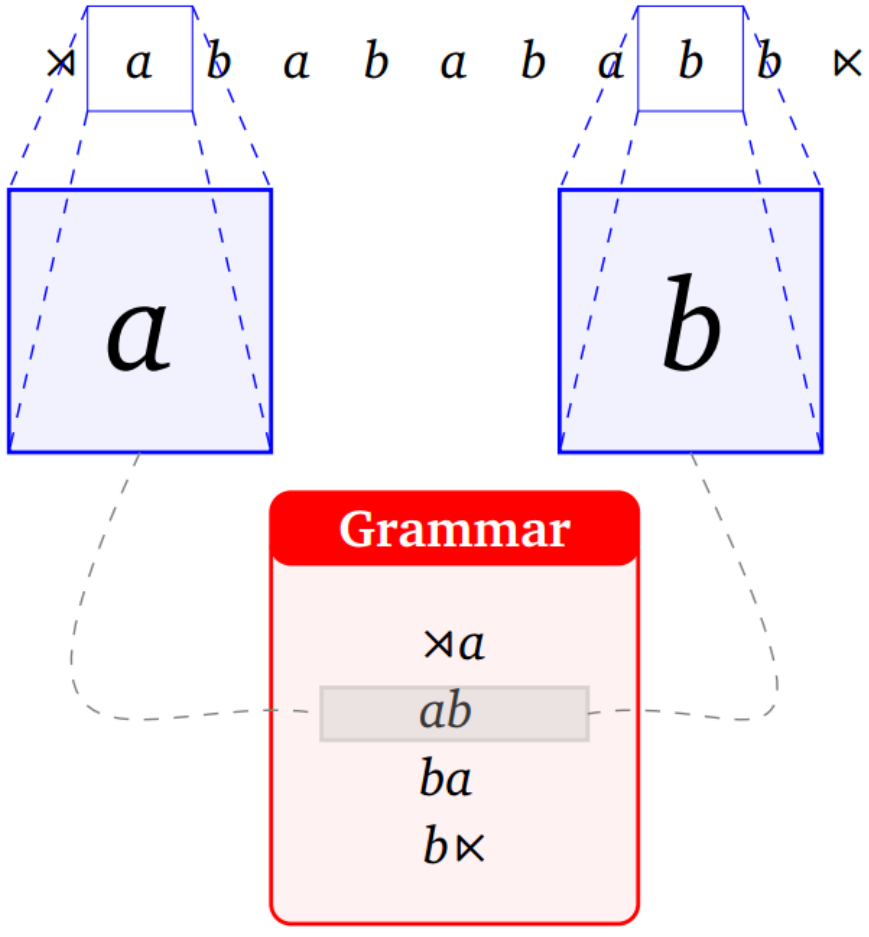
Morphotactics: prefixes and suffixes

English affixes

- *Prefix un-:*
  unlock, unhash
- *Suffix -able:*
  lockable, hashable

- $\Sigma = \{\text{un}, \text{able}, \text{hash}, \ldots, \text{lock}\} \quad n = 2$
  $G = \langle *\rtimes\text{able}, *\text{un}\ltimes, *\text{ableun}, \ldots, *\text{lockun}, *\text{ablehash}\rangle$
- $^{ok}\rtimes\text{lock}\ltimes$ $\quad *\rtimes\text{able-lock}\ltimes$ $\quad ^{ok}\rtimes\text{un-hash-able}\ltimes$

Strictly Piecewise Grammars



- 1 window of size k
- Closed under subsequence

(Heinz 2010), p.c. Thomas Graf