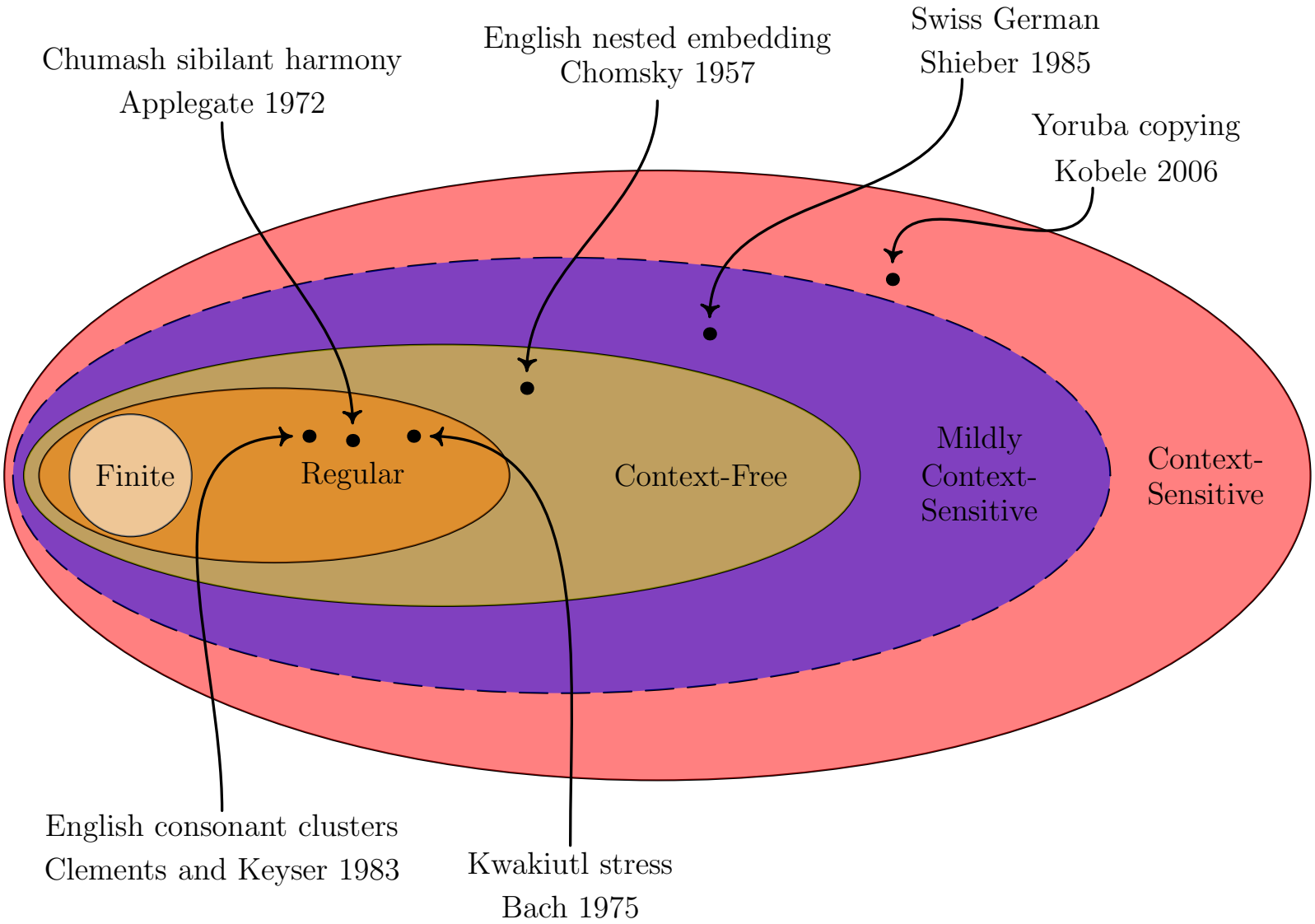
Patterns of stress and rhythm in words: a computational perspective

Jeffrey Heinz
heinz@udel.edu

University of Delaware

University of Connecticut
Decemeber 1, 2011

A famous computational perspective of natural language



Why a computational perspective?

Tension exists between traditional, theoretical approaches to linguistics and computational and mathematical approaches.

> *As many authors have pointed out before, the expressive power of a (formal) language and its place within the so-called Chomsky Hierarchy constitute a fact about what has come to be known as 'weak generativity' (i.e. string-generation), but what the linguist ought to be studying is the generation and conceptualization of structure (i.e., strong generativity).*

Brenchley and Lobina, November 21, 2011, Linguist List Discussion: 22.4650.

Brenchley and Lobina, 11/21/2011 (con't)

In a way, computational linguists are hostage to the fact that strong generativity has so far resisted formalization and that, therefore, their results do not appear to be directly relatable to the careful descriptions and explanations linguists propose; a fortiori, their formulae do not tell us much about the psychological facts of human cognition. In our opinion, then, Chomsky's analysis does not show an 'extremely shallow acquaintance' with computational models, but a principled opposition to them because of what these models assume and attempt to show.

see also Chomsky 1981

Why the computational perspective?

It DOES address STRONG generative capacity.

Indirectly: The weak generative capacity of a language is a property of its strong generative capacity.

Directly:
1. Strong generative capacity (like tree structure) can be encoded into the strings directly (with brackets)
2. The computational regions identified do not only describe classes of *string*sets but also classes of *tree*sets.

Why the computational perspective?

Learnability!

1. The weak generative capacity—the strings—is observable!
2. The strong generative capacity—the tree structures, the derivation trees, the *hidden structure*—is not. To some extent, they must be learned.

Why the computational perspective?

1. The computational perspective can distill necessary properties of natural language,
2. and can identify the contributions such properties can make to learnability.

THIS TALK:

1. Phonological patterns are *subregular*.
2. Locality, formalized as neighborhood-distinctness (a certain subregular property), makes a significant contribution to locality.

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?

Example: Initial Stress

SPE

$$\boxed{\sigma \rightarrow \acute{\sigma} \; / \; \#\_\_\_\_}$$

$$/\sigma\sigma\sigma/ \rightarrow [\acute{\sigma}\sigma\sigma]$$

Example: Initial Stress

Principles and Parameters

Trochaic, Left-to-right, End-Rule-Left

$$/\sigma\sigma\sigma/ \rightarrow (\acute{\sigma}\sigma)\sigma \rightarrow [\acute{\sigma}\sigma\sigma]$$

Example: Initial Stress

Optimality Theory

Trochaic $\gg$ Iambic
Align(Stress,Left) $\gg$ Align(Stress,Right)
BinaryFoot $\gg$ ParseSyllable

$$/\sigma\sigma\sigma/ \rightarrow (\acute{\sigma}\sigma)\sigma \rightarrow [\acute{\sigma}\sigma\sigma]$$

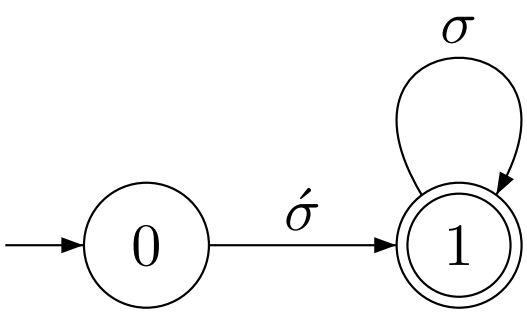
Different grammars, same result

Each of these grammars generates the following *infinite* set of observable strings.

$$\acute{\sigma}$$
$$\acute{\sigma}\sigma$$
$$\acute{\sigma}\sigma\sigma$$
$$\acute{\sigma}\sigma\sigma\sigma$$
$$\ldots$$

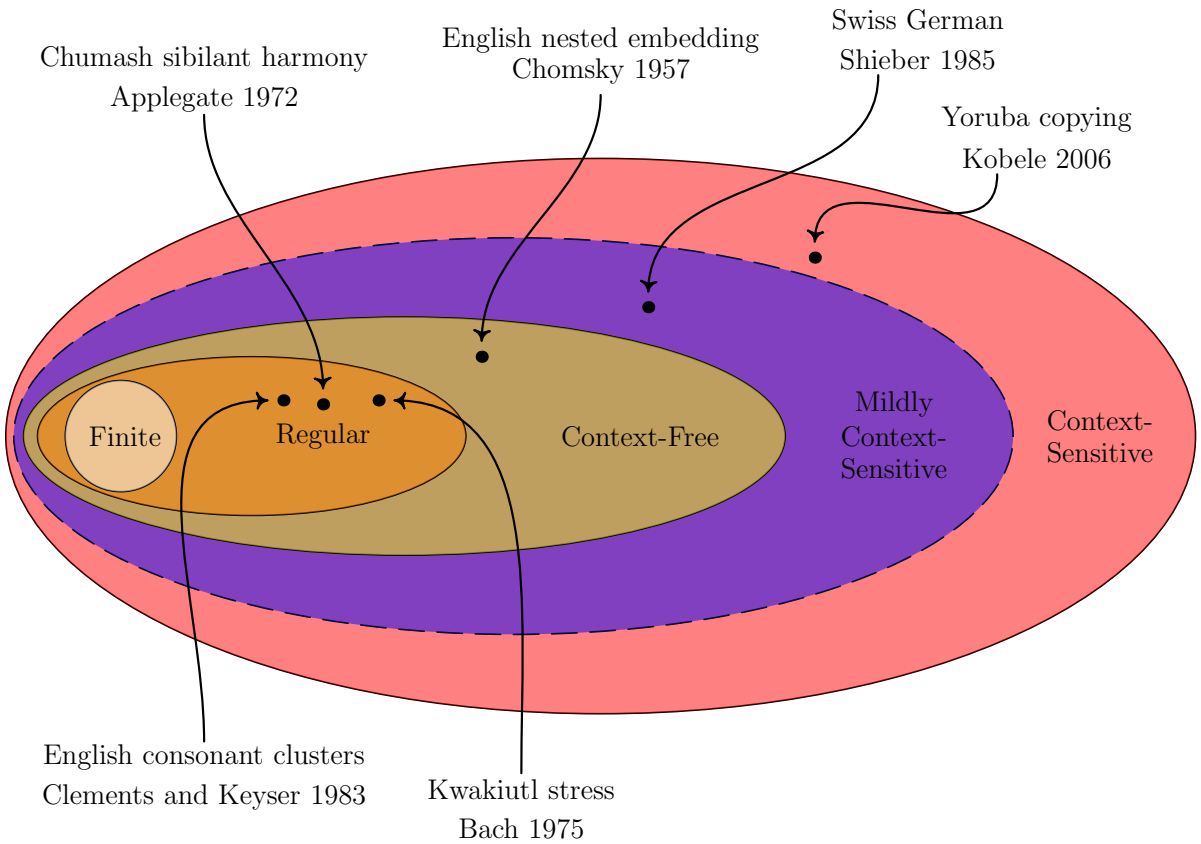
Example: Initial Stress

$$\acute{\sigma}$$
$$\acute{\sigma}\sigma$$
$$\acute{\sigma}\sigma\sigma$$
$$\acute{\sigma}\sigma\sigma\sigma$$
$$\ldots$$



This FSA describes this infinite set too.

Patterns describable with FSA are *regular*.



Hypothesis: "Being regular" is a universal property of phonological patterns.

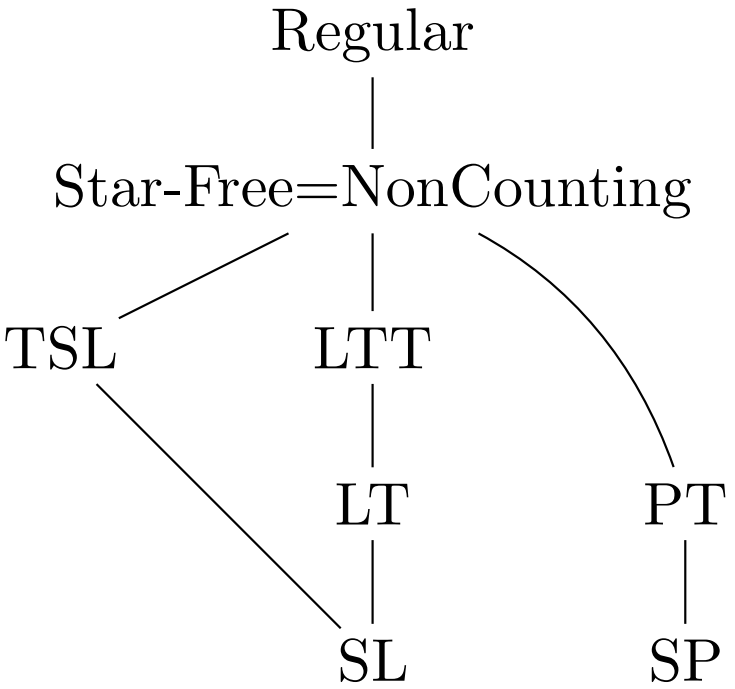
(Johnson 1972, Kaplan and Kay 1994)

Not any regular pattern is phonological.

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 sibilants, ...).
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)

Classifying regular patterns



Proper inclusion relationships among language classes (indicated from top to bottom).

TSL	Tier-based Strictly Local
LTT	Locally Threshold Testable
LT	Locally Testable
SL	Strictly Local
PT	Piecewise Testable
SP	Strictly Piecewise

(McNaughton and Papert 1971, Simons 1975, Rogers et al. 2010, Heinz et al. 2011)

Neighborhood-distinctness

Only some regular patterns are neighborhood-distinct.

1. 107 of the 109 stress patterns (400+ languages represented) are neighborhood-distinct.
2. Many logically possible stress patterns are not (stress every 4th syllable, etc.) are not.

Neighborhood-distinctness is one way to formalize the concept of locality in phonology.

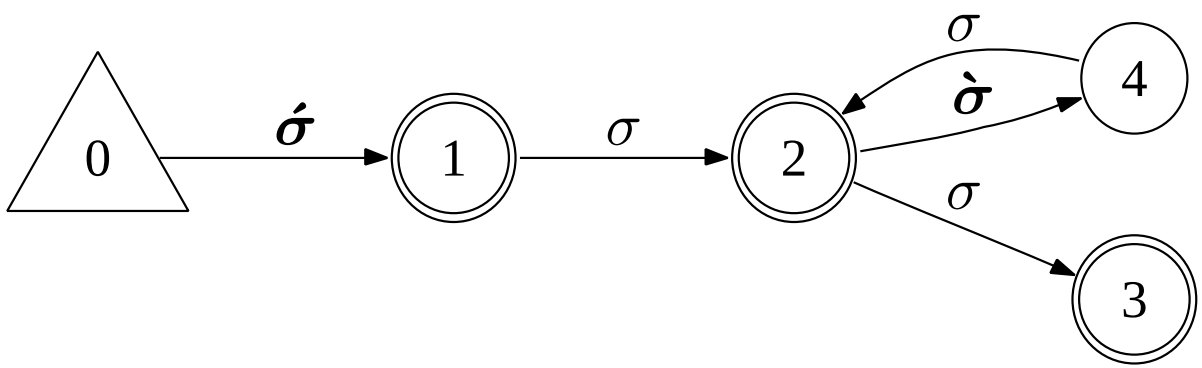
Pintupi Stress (Quantity-Insensitive Binary)

a.	***σ́*** *σ*	páɳa	‘earth’
b.	***σ́*** *σ σ*	tʲúʈaya	‘many’
c.	***σ́*** *σ* ***σ̀*** *σ*	málawàna	‘through from behind’
d.	***σ́*** *σ* ***σ̀*** *σ σ*	púliŋkàlatʲu	‘we (sat) on the hill’
e.	***σ́*** *σ* ***σ̀*** *σ* ***σ̀*** *σ*	tʲámulìmpatʲùŋku	‘our relation’
f.	***σ́*** *σ* ***σ̀*** *σ* ***σ̀*** *σ σ*	ʈílirìŋulàmpatʲu	‘the fire for our benefit flared
g.	***σ́*** *σ* ***σ̀*** *σ* ***σ̀*** *σ* ***σ̀*** *σ*	kúranʲùlulìmpatʲùɻa	‘the first one who is our rela
h.	***σ́*** *σ* ***σ̀*** *σ* ***σ̀*** *σ* ***σ̀*** *σ σ*	yúmaɻìŋkamàratʲùɻaka	‘because of mother-in-law’

- Secondary stress falls on nonfinal odd syllables (counting from left)
- Primary stress falls on the initial syllable

Hayes (1995:62) citing Hansen and Hansen (1969:163)

The Learning Question



Q: How can this finite state acceptor be learned from the finite list of Pintupi words?

A:
- Generalize by writing smaller and smaller descriptions of the observed forms
- guided by some universal property of the target class...

Neighborhoods

The neighborhood of an environment (state) is:

(1) the set of incoming symbols to the state

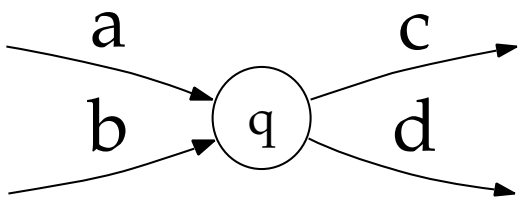
(2) the set of outgoing symbols to the state

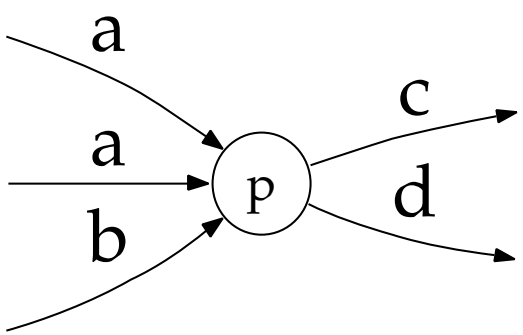
(3) whether it is a final state or not

(4) whether it is a start state or not

Example of Neighborhoods

- States p and q have the same neighborhood.





Neighborhood-distinctness

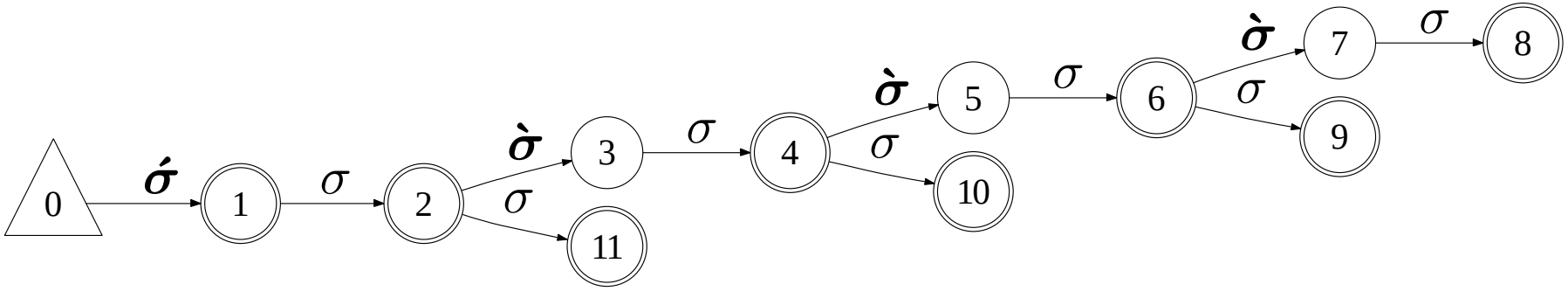
A language (regular set) is *neighborhood-distinct* iff there is an acceptor for the language such that each state has its own unique neighborhood.

Overview of the Neighborhood Learner

- Two stages:
  1. Builds a structured representation of the input list of words
  2. Generalizes by merging states which are redundant:
     i.e. those that have the same local environment—the neighborhood

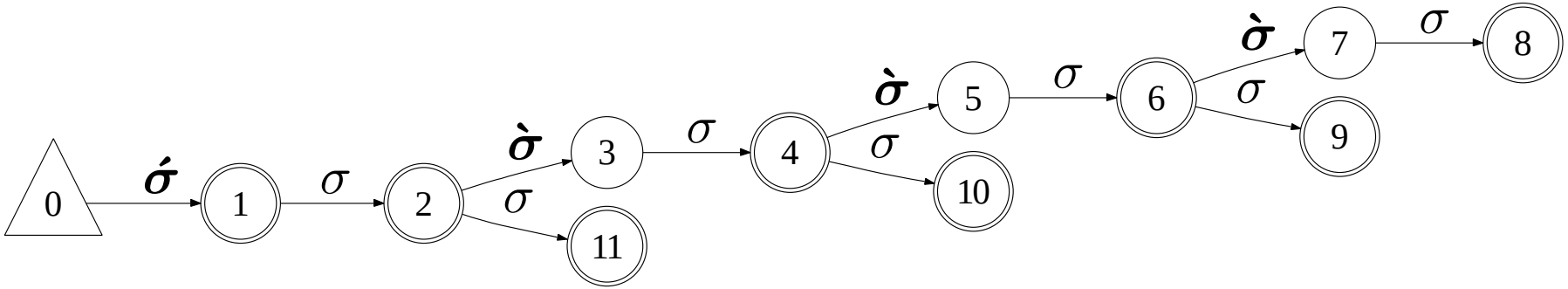
(cf. (Angluin 1982, Muggleton 1990))

The Prefix Tree for Pintupi Stress



- Accepts the words: $\boxed{\acute{\sigma}}$, $\boxed{\acute{\sigma}\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \grave{\sigma}\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \grave{\sigma}\ \sigma}$
- A structured representation of the input (Angluin 1982, Muggleton 1990).
- It accepts only the forms that have been observed.
- Note that environments are repeated in the tree!

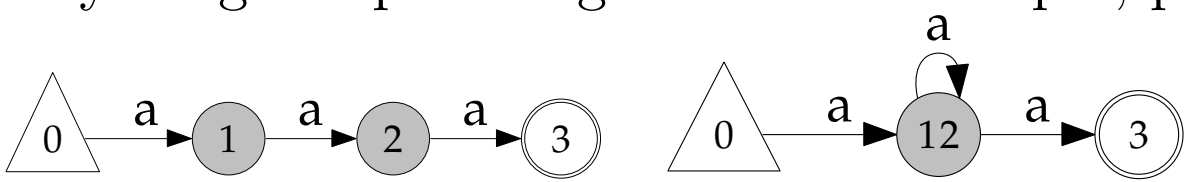
The Prefix Tree for Pintupi Stress



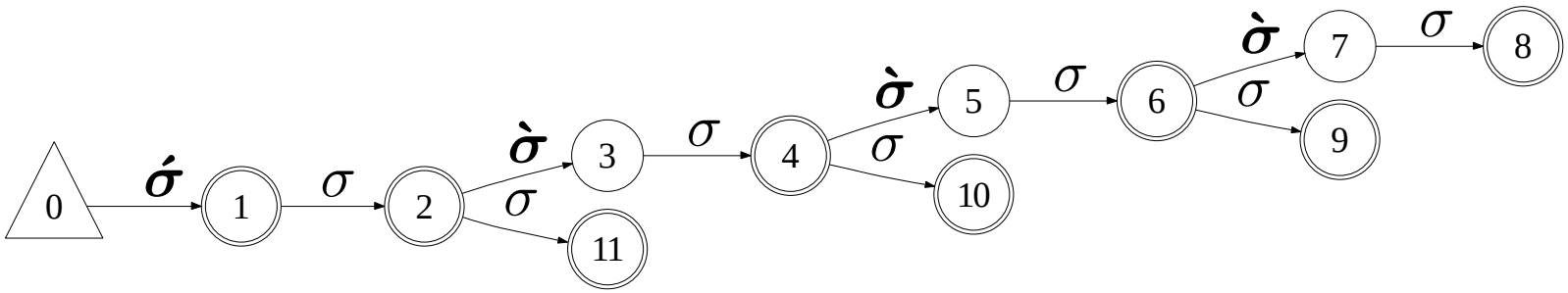
- Accepts the words: **σ́**, **σ́** σ, **σ́** σ σ, **σ́** σ **σ̀** σ, **σ́** σ **σ̀** σ σ, **σ́** σ **σ̀** σ **σ̀** σ, **σ́** σ **σ̀** σ **σ̀** σ σ, **σ́** σ **σ̀** σ **σ̀** σ **σ̀** σ
- A structured representation of the input (Angluin 1982, Muggleton 1990).
- It accepts only the forms that have been observed.
- Note that environments are repeated in the tree!

Generalizing by State-merging

- Eliminate redundant environments by *state-merging.*
- This is a process where two states are identified as equivalent and then *merged* (i.e. combined).
- A key concept behind state merging is that transitions are preserved (Angluin 1982)
- This is one way in which generalizations may occur—because the post-merged machine accepts everything the pre-merged machine accepts, possibly more.



The Learner's State Merging Criteria



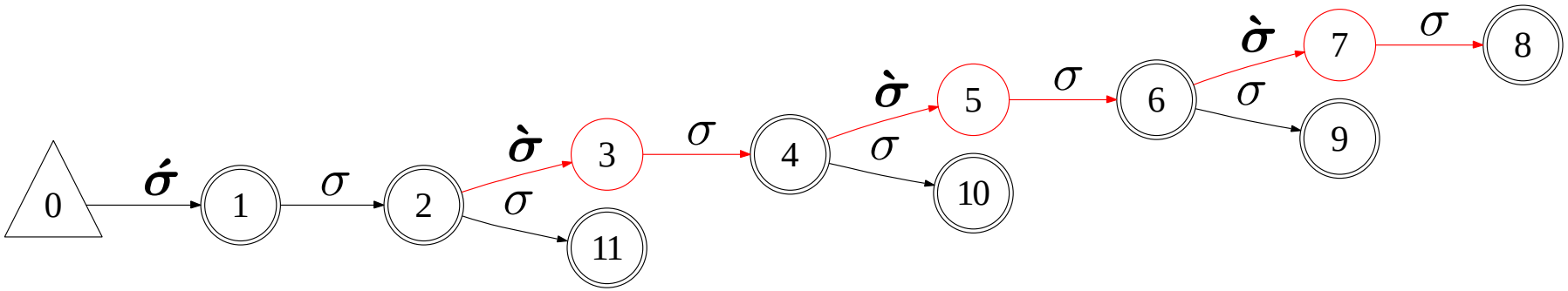
- How does the learner decide whether two states are equivalent in the prefix tree?
- Merge states with the same neighborhood.

Example of Neighborhoods

- States p and q have the same neighborhood.

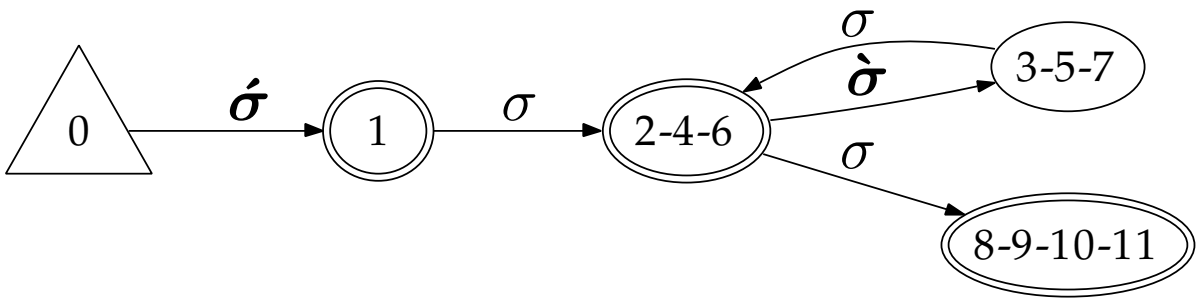
- The learner merges states in the prefix tree with the same neighborhood.

The Prefix Tree for Pintupi Stress



- States 3, 5, and 7 have the same neighborhood.
- So these states are merged.

The Result of Merging Same-Neighborhood States



- The machine above accepts
  $\boxed{\acute{\sigma}}$, $\boxed{\acute{\sigma}\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \sigma}$, $\boxed{\acute{\sigma}\ \sigma\ \grave{\sigma}\ \sigma\ \grave{\sigma}\ \sigma}$, ...
- The learner has acquired the stress pattern of Pintupi, i.e. it has generalized exactly as desired.
- Each state in the acceptor above has a distinct neighborhood.

Summary of the Forward Neighborhood Learner

(1) Builds a prefix tree of the observed words.

(2) Generalize by merging states which have the same neighborhood (local environment).

(3) The acceptor returned by the algorithm is neighborhood-distinct—every state has a distinct neighborhood.

Results

- A more sophisticated version of this learner can learn 100 of the 109 patterns (414 of 423 languages) in the Stress Database (Heinz 2009).
  - 37 of the 39 quantity-insensitive patterns
  - 38 of the 44 quantity-sensitive bounded patterns
  - 25 of the 26 quantity-sensitive unbounded patterns
- The patterns not learned differ only slightly from the target ones.

Examples of unlearned patterns

- Içuã Tupi (not ND) (Abramson 1968): Stress falls on the penult in words with four or fewer syllables and on the antepenult in words with five or more syllables.

  The learned grammar predicts that secondary stress may fall optionally on the penult instead of the antepenult in words five syllables or longer.

Examples of unlearned patterns

- Pirahã (ND) (Everett 1988): Stress falls on the rightmost heaviest/most prominent syllable as long as it occurs in the last three syllables.

  The learner predicts stress falls per the Pirahã pattern but in certain words it may fall optionally on the final syllable.

Examples of unlearned patterns

- Ashéninca (ND) (Payne 1990) has a complicated stress pattern involving foot extrametricality at the right word edge, among other things.

  The learner predicts that words ending with a long vowel followed by three syllables with the high front vowel like attested [má:kiriti] ‘type of bee’ could have two pronunciations: [má:kiriti] and [mà:kiríti]

Comparison to other stress-learning models

- a P&P-based learner (Dresher and Kaye 1990, Gillis et al. 1995)
- a perceptron-based learner (Gupta and Touretzky 1994)
- an OT-based learner (Tesar 1998, Tesar and Smolensky 2000).

Comparison to other stress-learning models

- a P&P-based learner (Dresher and Kaye 1990, Gillis et al. 1995)
  - 10 parameters yield a 216 language typology*
  - 75%-80% are learned given words up to length 4 syllables
- a perceptron-based learner (Gupta and Touretzky 1994)
- an OT-based learner (Tesar 1998, Tesar and Smolensky 2000).

*based on actual patterns, not all actual patterns included

Comparison to other stress-learning models

- a P&P-based learner (Dresher and Kaye 1990, Gillis et al. 1995)
  - 10 parameters yield a 216 language typology*
  - 75%-80% are learned given words up to length 4 syllables
- a perceptron-based learner (Gupta and Touretzky 1994)
  - 19 stress patterns
  - 17 are learned given multiple presentations of all words of length 1 to 7 syllables.
- an OT-based learner (Tesar 1998, Tesar and Smolensky 2000).

*based on actual patterns, not all actual patterns included

Comparison to other stress-learning models

- a P&P-based learner (Dresher and Kaye 1990, Gillis et al. 1995)
  - 10 parameters yield a 216 language typology*
  - 75%-80% are learned given words up to length 4 syllables
- a perceptron-based learner (Gupta and Touretzky 1994)
  - 19 stress patterns
  - 17 are learned given multiple presentations of all words of length 1 to 7 syllables.
- an OT-based learner (Tesar 1998, Tesar and Smolensky 2000).
  - 12 constraints yield a 124 language typology*
  - about 60% learned when given a monostratal initial ranking
  - about 97% learned when given a particular constraint ranking

*based on actual patterns, not all actual patterns included

Learnable Unnatural Patterns

- There are stress patterns that can be learned by neighborhood learning which are not considered natural.
  (1) Leftmost Light otherwise Rightmost.
  (2) A stress pattern requiring both lapses and clashes.
  (3) A stress pattern where all syllables have primary stress.
- If these patterns are harder to learn, do we expect the explanation for those facts to follow from considerations of locality?

Learnable Unnatural Patterns

- There are stress patterns that can be learned by neighborhood learning which are not considered natural.
  (1) Leftmost Light otherwise Rightmost.
  (2) A stress pattern requiring both lapses and clashes.
  (3) A stress pattern where all syllables have primary stress.
- If these patterns are harder to learn, do we expect the explanation for those facts to follow from considerations of locality?

Conclusion

1. Identifying computational constraints on phonological patterns helpd identify stronger and stronger (more restrictive) universal properties of phonological patterns.
2. These properties lead to novel hypotheses regarding how phonological patterns are learned.