# Modeling morphological productivity with the Minimal Generalization Learner

Adam Albright (albright@mit.edu)

Massachusetts Institute of Technology

July 4, 2012 — DRAEM

# Outline

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence
2. A simple simulation
3. Extensions
   - Phonological rules
   - ‘Impugnment’
   - Non-local and suprasegmental contexts

# Outline

# Goal of the model

- Given: a set of morphologically related forms—e.g., English:

| Present | Past | |
|---|---|---|
| mɪs | mɪst | *'miss(ed)'* |
| prɛs | prɛst | *'press(ed)'* |
| læf | læft | *'laugh(ed)'* |
| hʌg | hʌgd | *'hug(ged)'* |
| rʌb | rʌbd | *'rub(bed)'* |

| Present | Past | |
|---|---|---|
| nid | nidəd | *'need(ed)'* |
| dʒʌmp | dʒʌmpt | *'jump(ed)'* |
| plæn | plænd | *'plan(ned)'* |
| sɪŋ | sæŋ | *'sing/sang'* |
| drɪŋk | dræŋk | *'drink/drank'* |

- Goal: generalize to new items, such as novel verb [splɪŋk]$_{Present}$
  - Expected outputs

    Very likely/acceptable: splɪŋkt
    Somewhat likely/acceptable: splʌŋk, splæŋk
    Quite unlikely/unacceptable: splɔt
  - Not:
    - *splɪŋkd, *splɪŋkəd (misapplication of phonology)
    - *sploʊŋ (valid past, wrong context)
    - *splɪnd (idiosyncratic change: *make*∼*made*)

# Desired information

Some questions

- What string mappings relate the morphological categories? (∅ →d, ɪ→æ, ɪŋk→ɔt, etc.)
- Are some mappings phonological variants of others? (∅ →d, ∅ →t, ∅ →əd)
- Are some mappings restricted (categorically, probabilistically) to specific phonological contexts?
- What is the relative strength of the mappings in various contexts?

# Induction strategy

- Parse pairs to discover changes and contexts—e.g.,
  - mɪs $\sim$ mɪst $=$ $\varnothing \rightarrow$ t / mɪs ___ #
  - dʒʌmp $\sim$ dʒʌmpt $=$ $\varnothing \rightarrow$ t / dʒʌmp ___ #
- Compare contexts to discover broader contexts
  - mɪs, dʒʌmp: /[-voi]___
- Evaluate accuracy of resulting contexts

# Discovering changes

Pinker & Prince (1988), p. 130:

1. "Candidates for rules are hypothesized by comparing base and past tense versions of a word, and factoring apart the changing phonetic portion, which serves as the rule operation, from certain morphologically-relevant phonological components of the stem, which serve to define the class of stems over which the operation can apply." (= the *context*)

- Pinker and Prince (1988), Ling and Marinov (1993), Albright and Hayes (2002)

# A simple parsing strategy

Albright and Hayes (2002, 2003)

- Edge-in alignment: start from both left and right, aligning until there's a mismatch
- Combine leftward and rightward alignments to align as many segments as possible

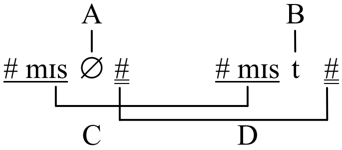
$$\begin{array}{ccc|c} m_1 & ɪ_2 & s_3 & \\ m_1 & ɪ_2 & s_3 & t \end{array} + \begin{array}{cccc|} & m & ɪ & s \\ m & ɪ & s & t \end{array} \mapsto \begin{array}{ccc|c|} m_1 & ɪ_2 & s_3 & \\ m_1 & ɪ_2 & s_3 & t \end{array}$$

- Medial changes: leftward and rightward scans are both able to align some segments; residue left in the middle

$$\begin{array}{c|cc} s_1 & ɪ & ŋ \\ s_1 & æ & ŋ \end{array} + \begin{array}{cc|c} s & ɪ & ŋ_1 \\ s & æ & ŋ_1 \end{array} \mapsto \begin{array}{c|c|c} s_1 & ɪ & ŋ_2 \\ s_1 & æ & ŋ_2 \end{array}$$

# Factoring out changes and contexts

- Mismatched portions = change (A → B)
- Aligned portions = context (C ___ D)




- Resulting rule: ∅ → t / #mɪs ___ #

# Word-specific rules

Result: a set of word-specific rules

| | | | | |
|---|---|---|---|---|
| a. | $\varnothing \rightarrow$ t | / | #mɪs ___ | # |
| b. | $\varnothing \rightarrow$ t | / | #prɛs ___ | # |
| c. | $\varnothing \rightarrow$ t | / | #læf ___ | # |
| d. | $\varnothing \rightarrow$ d | / | #hʌg ___ | # |
| e. | $\varnothing \rightarrow$ d | / | #rʌb ___ | # |
| f. | $\varnothing \rightarrow$ əd | / | #nid ___ | # |
| g. | $\varnothing \rightarrow$ t | / | #dʒʌmp ___ | # |
| h. | $\varnothing \rightarrow$ d | / | #plæn ___ | # |
| i. | ɪ $\rightarrow$ æ | / | s ___ | ŋ# |
| j. | ɪ $\rightarrow$ æ | / | dr ___ | ŋk# |

# Comparing rules

- *miss* and *press* both take $\varnothing \rightarrow t$. Assume that this is because they have some crucial property in common
  - In both cases, the change is word-final
  - In both cases, the segment before the change is an [s]
  - In both cases, the segment before [s] is a non-low lax front V
  - etc...??? (sonorant before the vowel, monosyllabic, etc.)
- Albright and Hayes (2002): Minimal generalization approach
  - Conservative collapsing: new rule keeps *everything* that the pair has in common

# Minimal generalization

- Once again, some alignment scheme is needed. Assume *locality*

| | | | | | |
|---|---|---|---|---|---|
| | m | ɪ | s | ___ | # |
| p | r | ɛ | s | ___ | # |

  Precludes many possible generalizations, such as:
  - End in /s/
  - Vowel is front, lax
  - First segment is labial
- A pragmatic assumption: *myopic generalization*
  - Once a mismatch is encountered and featural generalization is needed, shared similarities farther away from the change location are not likely to be crucial
  - Convert more distant segments to free variables (X, Y)
  - We'll come back to this assumption later

# Minimal generalization

Comparing *miss* and *press*: $\varnothing \rightarrow$ t / ...

| | Residue | Shared Features | Shared Segments | Change Location | Shared Segments | Shared Features | Residue |
|---|---|---|---|---|---|---|---|
| A. | #m | ɪ | s | ___ | # | | |
| B. | #pr | ɛ | s | ___ | # | | |
| C. | X | [+syllabic, −low, −back, −tense, −round] | s | ___ | # | | |

# Minimal generalization

Minimal generalization of features: retain all shared feature values

- Prevents generalization to unseen feature values
- Permits generalization to unseen feature *combinations*
- Example: comparing *b* and *n*

$$\overset{\text{b}}{\begin{bmatrix} -\text{syllabic} \\ +\text{cons} \\ -\text{contin} \\ +\text{voice} \\ -\text{nasal} \\ +\text{labial} \\ -\text{coronal} \\ -\text{velar} \end{bmatrix}} + \overset{\text{n}}{\begin{bmatrix} -\text{syllabic} \\ +\text{cons} \\ -\text{contin} \\ +\text{voice} \\ +\text{nasal} \\ -\text{labial} \\ +\text{coronal} \\ -\text{velar} \end{bmatrix}} \mapsto \overset{\{\text{ b, m, d, n}\}}{\begin{bmatrix} -\text{syllabic} \\ +\text{cons} \\ -\text{contin} \\ +\text{voice} \\ \\ \\ \\ -\text{velar} \end{bmatrix}}$$

# Examples

Comparing *hug*, *plan*, and then *rub*: $\varnothing \rightarrow$ d / ...

| | Residue | Shared Features | Shared Segments | Change Location | Shared Segments | Shared Features | Residue |
|---|---|---|---|---|---|---|---|
| A. | #hʌ | g | | ___ | # | | |
| B. | #plæ | n | | ___ | # | | |
| C. | X | [ −syllabic<br>−continuant<br>−labial<br>−lateral<br>+voice<br>. . . ] | | ___ | # | | |
| And generalizing further with *rub*: | | | | | | | |
| D. | #rʌ | b | | ___ | # | | |
| E. | X | [ −syllabic<br>−continuant<br>−lateral<br>+voice<br>. . . ] | | ___ | # | | |

# Examples

Comparing *sing* and *drink*: ɪ → æ / ...

| | Residue | Shared Features | Shared Segments | Change Location | Shared Segments | Shared Features | Residue |
|---|---|---|---|---|---|---|---|
| A. | # | s | | ___ | ŋ | | # |
| B. | #d | r | | ___ | ŋ | | k# |
| C. | X | [−syllabic, +coronal, +continuant] | | ___ | ŋ | | Y |

# What this model does well

Rapidly discovers most general environment for each pattern

$$\varnothing \rightarrow \text{əd} \;/\; \#\text{want}\ \_\_\_\ \#$$

$$\varnothing \rightarrow \text{əd} \;/\; \text{X} \begin{bmatrix} -\text{syl} \\ +\text{cons} \\ -\text{son} \\ -\text{cont} \\ -\text{strid} \\ +\text{cor} \\ +\text{ant} \end{bmatrix} \_\_\_\ \#$$

$$=\{\text{t, d}\}$$

$$\varnothing \rightarrow \text{əd} \;/\; \#\text{nid}\ \_\_\_\ \#$$

# What this model does well

Rapidly discovers most general environment for each pattern

$\varnothing \rightarrow$ t / #lʊk ___ #

$\varnothing \rightarrow$ t / #stɑp ___ #

$$\varnothing \rightarrow \text{t / X} \begin{bmatrix} -\text{syl} \\ +\text{cons} \\ -\text{son} \\ -\text{cont} \\ -\text{voi} \\ -\text{strid} \end{bmatrix} \_\_\_ \#$$

= {p,t,k}

$\varnothing \rightarrow$ t / #pæs ___ #

$$\varnothing \rightarrow \text{t / X} \begin{bmatrix} -\text{syl} \\ +\text{cons} \\ -\text{son} \\ -\text{voi} \end{bmatrix} \_\_\_ \#$$

= *voiceless obstruents*

# What this model does well

Rapidly discovers most general environment for each pattern

$\varnothing \rightarrow$ d / #sim ___ #

$\varnothing \rightarrow$ d / #juz ___ #

$\varnothing \rightarrow$ d / X $\begin{bmatrix} -\text{syl} \\ +\text{cons} \\ +\text{voi} \end{bmatrix}$ ___ #

= *voiced consonants*

$\varnothing \rightarrow$ d / #ʃoʊ ___ #

$\varnothing \rightarrow$ d / X [+voi] ___ #

# Rapid generalization

- Generalization proceeds rapidly, given sufficiently diverse stems
- The pathways shown here can be found using verbs that are among the 75 most frequent verbs of English (according to CELEX) (of which the majority are actually irregular)

# Rule creation for irregulars

"Irregular" mappings are also compared and generalized

ɪ → æ / #s ___ ŋ#

ɪ → æ / #r ___ ŋ#

ɪ → æ / X [−syl, +cons, +cont, +cor] ___ ŋ#
= θ, ð, s, z, ʃ, ʒ, l, r
*Strength = 2*

ɪ → æ / #dr ___ ŋk#

ɪ → æ / X [−syl, +cons, +cont, +cor] ___ ŋ Y#
= θ, ð, s, z, ʃ, ʒ, l, r
*Strength = 3*

- Provides rules to generalize [splɪŋ]→[splæŋ] in addition to [splɪŋd]

# Outline

# What determines the strength of a rule?

- Goal: suffixed output [splɪŋd] is more probable/acceptable than outputs like [splæŋ], [splʌŋ], etc.
- Irregular patterns (at least in English) tend to cover relatively few forms, which are similar to one another ≈ share a set of phonological properties
- Result: rules that characterize them are more specific, and weaker strength
- These patterns are not productive, except possibly for inputs that fit a very particular phonological shape (≈ prototypical forms)

# What determines the strength of a rule?

Pinker and Prince (1988):

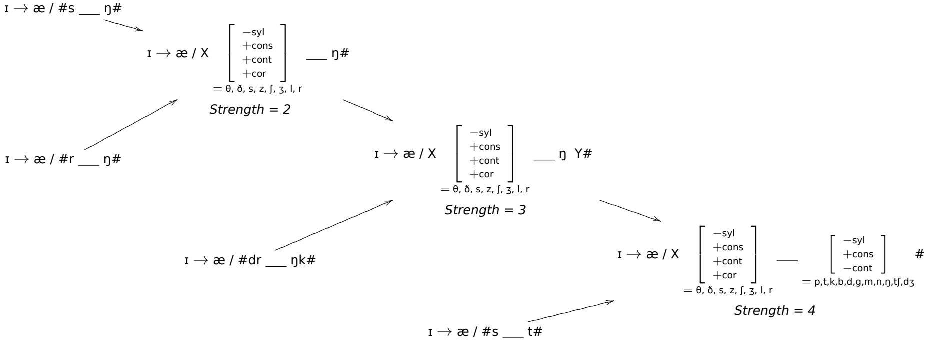
- “Rule candidates increase in strength each time they have been exemplified by an input pair”

| | Rule | Examples covered |
|---|---|---|
| 1. | ∅ → t / #mɪs ___ # | 1 |
| 2. | ∅ → t / #prɛs ___ # | 1 |
| 3. | ∅ → t / X [+syl,−low,−bk,−tns,−rnd] s ___ # | 2 |
| 4. | ∅ → d / #hʌg ___ # | 1 |
| 5. | ∅ → d / #plæn ___ # | 1 |
| 6. | ∅ → d / ___ X [−syl,−cont,−lab,−lat,−del.rel.] | 2 |
| 7. | ɪ → æ / #s ___ ŋ# | 1 |
| 8. | ɪ → æ / #dr ___ ŋk# | 1 |
| 9. | ɪ → æ / X [−syl, +cor, +cont] ___ ŋ Y# | 2 |

- p. 134: “[by various means], some regularities can be enshrined as permanent productive rules whereas others can be discarded or treated differently.”

# Simply counting examples is not enough

- Consider hypothesized rules after exposure to *sing*, *ring*, *drink*, *sit*




- "Prototypical" forms (*sing*, *ring*, *drink*) and "outlier" (*sit*) combine to yield more general rule
- More general rule encompasses more forms, so has higher strength

# Simply counting examples is not enough

- Prediction: ɪ→æ change may apply equally well to any form within this broader context
  - Hypothetical *zick* ∼ *zack* just as well supported as *spling* ∼ *splang*

# Diagnosis

- *Strength*, measured purely in terms of number of examples covered, favors the broadest possible generalizations
- Under this view, “compatible with the data” = “encompasses all the examples”
- Broad generalizations are not always virtuous; sometimes, they incur numerous exceptions.
- A more intuitive notion of “compatible with the data” must incorporate not only generality, but also accuracy

# Rule reliability

Reliability of a hypothesized rule:

$$= \frac{\text{number of items that the rule works for (“}\textit{hits}\text{”)}}{\text{number of items that meet the rule's structural description (CAD) (“}\textit{scope}\text{”)}}$$

- Also known as accuracy, or coverage

| Rule | Hits | Exceptions |
|---|---|---|
| ɪ → æ / X [−syl, +cons, +cont, +cor] ___ ŋ Y# <br> [−syl, +cons, +cont, +cor] = θ, ð, s, z, ʃ, ʒ, l, r | *ring*, *sing*, *drink* <br> = *n* | *wring*, *fling*, ... <br> = *m* |
| ɪ → æ / X [−syl, +cons, +cont, +cor] ___ [−syl, +cons, −cont] Y# <br> [−syl, +cons, +cont, +cor] = θ, ð, s, z, ʃ, ʒ, l, r; [−syl, +cons, −cont] = p,t,k,b,d,g,m,n,ŋ,tʃ,dʒ | *ring*, *sing*, *drink*, *sat* <br> = *n*+1 | *wring*, *fling*, *lick*, *rig*, *ship*, *rid* ... <br> = *m*+lots |

- Island of reliability: context in which a change is especially likely
  - “especially likely” = more likely than in its most general context (Albright 2002)

# The intended effect

- Favoring rules with high reliability should let the model find specific contexts that *uniquely* characterize members of a particular class
- Overly broad contexts are discouraged, because of the cost of adding exceptions
  - Number of positive hits gained (numerator) must exceed number of exceptions added to scope (denominator)
  - For small, “irregular” class, optimal contexts will be narrow
  - For “regular” classes, there may be fewer distinguishing features that uniquely characterize most regular verbs, so no benefit to staying small
  - Model zooms to full generality, incurring some exceptions
- Model should settle on the right level of generality for each class

# How this plays out in practice

- Example: ran the model on the 200 most frequent verbs in CELEX
  - 134 regular (80 *-d*, 30 *-t*, 24 *-əd*); 66 irregular, of various types

# How this plays out in practice

- Example: ran the model on the 200 most frequent verbs in CELEX
  - 134 regular (80 *-d*, 30 *-t*, 24 *-əd*); 66 irregular, of various types
- Most reliable rules for a few of the major classes

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| i: → ε / # [−syl, +cont] ___ d# (= *continuants*) | 3 | 3 | 1.00 | read, lead, feed | (none) | meet |

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| e → o / # [−syl, +voi, +lab] ___ r (= *b,m,v,r*) | 2 | 2 | 1.00 | wear, bear | (none) | break |

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| ∅ → ∅ / # [−syl, +cons, +cont, +cor, +ant] εt ___ # (= *θ,ð,l,s,z*) | 2 | 2 | 1.00 | set, let | (none) | put, cut, get, sit, … |

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| d → t / X [−syl, −son, −voi] εn ___ # (= *p,t,k,f,θ,s,ʃ,tʃ*) | 2 | 3 | 0.67 | send, spend | tend | build |

# How this plays out in practice

- Example: ran the model on the 200 most frequent verbs in CELEX
  - 134 regular (80 -*d*, 30 -*t*, 24 -*əd*); 66 irregular, of various types
- Most reliable rules for a few of the major classes

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| ∅ → t / X [−syl, −son, +cont, −voi] ___ # (= *f,θ,s,ʃ*) | 14 | 14 | 1.00 | pass, produce, wish, laugh, ... | (none) | 52 other [−voi] |
| ∅ → t / X [+syl, −low, +back] k ___ # (= *uː,ʊ,oʊ,ɔ,ʌ,ɚ*) | 4 | 4 | 1.00 | look, work, talk, walk | (none) | 62 other [−voi] |

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| ∅ → əd / X [−syl, +cons] t ___ # (= *consonants*) | 10 | 10 | 1.00 | want, start, inspect, suggest, ... | (none) | 35 other *t,d* |
| ∅ → əd / X [+syl, −high, −back] d ___ # (= *e,ɛ,æ,ɔ,ɪ,aɪ*) | 4 | 4 | 1.00 | provide, decide, add, avoid | (none) | 41 other *t,d* |

# How this plays out in practice

- Example: ran the model on the 200 most frequent verbs in CELEX
  - 134 regular (80 -*d*, 30 -*t*, 24 -*əd*); 66 irregular, of various types
- Most reliable rules for a few of the major classes

| Rule | Hits | Scope | Rel. | Examples | Exceptions | Excludes |
|---|---|---|---|---|---|---|
| $\varnothing \rightarrow$ d / X ɚ ___ # | 11 | 11 | 1.00 | remember, consider, offer, ... | (none) | 69 other [+voi] |
| $\varnothing \rightarrow$ d / X $\begin{bmatrix} +\text{syl} \\ -\text{low} \\ +\text{back} \end{bmatrix}$ v ___ # (= uː,ʊ,oʊ,ɔ,ʌ,ɚ) | 5 | 5 | 1.00 | move, love, serve, prove, remove | (none) | 75 other [+voi] |
| $\varnothing \rightarrow$ d / X $\begin{bmatrix} +\text{syl} \\ -\text{high} \\ -\text{back} \end{bmatrix}$ n ___ # (= e,ɛ,æ,ɔ,ɪ,aɪ) | 5 | 5 | 1.00 | remain, explain, plan, join, contain | (none) | 75 other [+voi] |

# The fate of the more general rules

The more general rules are not nearly so reliable

| Rule | Hits | Scope | Rel. | Examples | Exceptions |
|---|---|---|---|---|---|
| $\varnothing \rightarrow$ d / X [+voi] ___ # | 80 | 134 | .60 | seem, use, try, call, turn, … | do, say, go, know, see, **need**, … |
| $\varnothing \rightarrow$ t / X [−voi] ___ # | 30 | 66 | .46 | look, ask, work, talk, help, … | get, think, take, **want**, put, … |
| $\varnothing \rightarrow$ əd / X $\begin{bmatrix} -\text{syl} \\ +\text{cons} \\ -\text{son} \\ -\text{cont} \\ -\text{strid} \\ +\text{cor} \\ +\text{ant} \end{bmatrix}_{=\{t,\ d\}}$ ___ # | 24 | 45 | .53 | want, need, start, wait, expect, … | get, find, put, sit, stand, … |

# An unintended consequence

- Focusing on reliability has something of the desired effect
  - Finds consists clusters of similar words
  - Identifies cases where it more distant words should be excluded
- However, it takes small-scale generalizations too seriously
- By favoring reliability instead of generality, we end up finding a long list of very accurate, but seemingly quite accidental contexts

# Why are the more general contexts so unreliable?

- Genuine exceptions: almost half of *t*, *d*-final verbs in this sample are irregular
  - Expect this proportion to go down in a larger sample
  - Larger sample will never make the problem disappear completely, however, since some subcontexts are 100% regular
    - E.g., verbs ending in voiceless fricatives, verbs ending in [ɚ], etc.
    - These will always be more reliable than the general context
  - In any system with irregularity, broad generalizations will necessarily involve some exceptions (decreased reliability)
- Inadequate characterization of rule interaction:
  - *t*, *d*-final verbs act as exceptions for more general *-d*, *-t* suffixes
  - More on this below...

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence
2. A simple simulation
3. Extensions
   - Phonological rules
   - ‘Impugnment’
   - Non-local and suprasegmental contexts

# Confidence

- Reliability is defined as a proportion
  - 2/2 = 7/7 = 100/100
- Speaker intuitions: not all ‘perfectly reliable’ rules are equally productive
- 100/100 less likely to be coincidental than 2/2

# A comparison

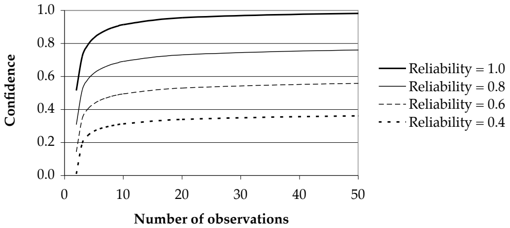
Reliability values from a larger set of English verbs:

- $\varnothing \rightarrow$ əd / $\begin{bmatrix} -\text{high} \\ -\text{low} \\ +\text{tense} \end{bmatrix}$ ___ (regular after *-ate*/*-ote*) 366/366
- $\varnothing \rightarrow$ t / $\begin{bmatrix} -\text{sonorant} \\ +\text{continuant} \end{bmatrix}$ ___ (regular after voiceless frics.) 352/352
- ɪ $\rightarrow$ ʌ / $\begin{bmatrix} +\text{coronal} \\ +\text{voiced} \\ +\text{anterior} \end{bmatrix}$ ___ $\begin{bmatrix} +\text{voiced} \\ +\text{dorsal} \end{bmatrix}$ 4/4 (*dig*, *cling*, *fling*, *sling*)
- $\varnothing \rightarrow \varnothing$ / # $\begin{bmatrix} +\text{coronal} \\ +\text{continuant} \end{bmatrix}$ ɛt ___ 2/2 (*let*, *set*)
- iː $\rightarrow$ ɛ / # $\begin{bmatrix} +\text{sonorant} \\ -\text{nasal} \\ +\text{coronal} \end{bmatrix}$ ___ d 2/2 (*read*, *lead*)

# Confidence

Implementation: lower confidence intervals

- Although 4/4 = 1, still likely that the next example will be an exception
  - Confidence interval: actual proportion might be somewhat lower or higher than observed
  - In this case, we can be 95% sure that the true value is between .875 and 1.125 (if values above 1 made sense here)
- Lower confidence limits:



# The effects of confidence

- Patterns involving small numbers of words may be weak, even if consistent
- Morphological categories that are rare may yield overall less confident rules

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence

2. A simple simulation

3. Extensions
   - Phonological rules
   - ‘Impugnment’
   - Non-local and suprasegmental contexts

# The training data

- Text file of training pairs and test items
  - Training pairs (and optionally, token frequency)
  - Test forms for ‘wug testing’
- Phonological features
  - `.fea` file with feature specifications
  - “ASCII” column (numeric codes): values don’t matter, should just be unique
  - Each segment must be a single character (ASCII, or Unicode)
  - Feature values: 0–1 (binary), 0–n (scalar)
  - Unspecified = -1

# Running the model

- Java archive: MinGenLearner.jar
  - From command line: `java -jar MinGenLearner.jar`
- Output files (tab-delimited text format)
  - `.out` file with list of changes
  - `.sum` file with wug test results
  - `.rules` file with list of rules discovered (if ‘Save rules’ option selected)

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence
2. A simple simulation
3. Extensions
   - Phonological rules
   - 'Impugnment'
   - Non-local and suprasegmental contexts

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence
2. A simple simulation
3. Extensions
   - Phonological rules
   - 'Impugnment'
   - Non-local and suprasegmental contexts

# Insufficient generalization without phonology

- Most general 'regular' rules in the sample simulation

| Rule | Rel. | Conf. |
|---|---|---|
| $\varnothing \rightarrow$ d / X [+voi] ___ # | 80/134 | .568 |
| $\varnothing \rightarrow$ t / X [−syl, 0-2son, 0-1aper] ___ # <br> = *p,k,t,f,θ,s,ʃ,ʧ b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ* | 31/142 | .197 |
| $\varnothing \rightarrow$ əd / X [−syl, 0 son, 0 aper, +COR, +ant] ___ # <br> = *t,d* | 24/45 | .482 |

- Misses fact that t, əd could be derived by phonological rules
  - d $\rightarrow$ t / [-voi] ___
  - $\varnothing \rightarrow$ ə / {t,d} ___ d

# Unifying changes with phonology

Goal:

- Allow model to discover that adding *d* after other contexts would yield illegal sequences such as *pd, *dd
- Posit phonological rules that repair illegal sequences to generate the attested output

# Approach to discovering phonological rules

- Provide model with list of phonotactically (surface) illegal sequences
  - *pd, *kd, *sd, *td, *dd, *bt, *gt, *zt, *tt, *dt, ...
  - Listed in `.ill` file
- 'Doppelgänger' rules
  - When positing a rule A $\rightarrow$ B / C ___ D, try out other changes A$'$ $\rightarrow$B$'$ in the same context C___ D
  - Original: $\varnothing \rightarrow$ t / [-voi] ___ #
  - Doppelgänger $\varnothing \rightarrow$ d / [-voi] ___ #
- Scan outputs of Doppelgänger rules for illegal sequences
- If difference between illegal output and attested output is a possible phonological mapping, posit a rule
  - Change/insertion/deletion of a single segment—e.g., d$\rightarrow$t
- Model options: 'Use Phonology', 'Use Doppelgängers'

# Some limitations

- Can't discover rule ordering
- Can't detect conflicting rules to see which is the productive one
  - kd → kt: lʊkd → lʊkt 'look')
  - kd → d: meɪkd → meɪd 'make'
- Assumes that rules hold in entire string (no NDEB or morphological conditioning)
  - pd → pt: ʌpdeɪtəd → *ʌpteɪtəd 'update'
  - Workaround: add word boundaries and augment list of illegal sequences to restrict to final position (*pd#, *dt#, etc.)
- Also possible to pre-specify mappings by brute force in `.phon` file
  - End-run around limited discovery mechanism

# Results for sample file

Most general 'regular' rules, with phonology enabled

| Rule | Rel. | Conf. |
|---|---|---|
| $\varnothing \rightarrow$ d / X ___ # | 134/200 | .647 |
| $\varnothing \rightarrow$ t / X [−syl, 0-2son, 0-1aper] ___ # <br> = *p,k,t,f,θ,s,ʃ,ʧ b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ* | 79/142 | .528 |
| $\varnothing \rightarrow$ əd / X [−syl, 0 son, 0 aper, +COR, +ant] ___ # <br> = *t,d* | 24/45 | .482 |

- Model discovers more general $\varnothing \rightarrow$d rule
- Reliability of most general $\varnothing \rightarrow$t rule improves, too

# Outline

1. Model overview
   - Minimal generalization
   - Reliability
   - Confidence
2. A simple simulation
3. Extensions
   - Phonological rules
   - 'Impugnment'
   - Non-local and suprasegmental contexts

# Another problem

- Model outputs for [splɪŋ]

| Output | Rule | Rel. | Conf. |
|---|---|---|---|
| splɪŋd | $\varnothing \rightarrow$ d / X $\begin{bmatrix} -\text{syl} \\ \text{1-2son} \\ \text{0-1aper} \end{bmatrix}$ ___ # <br> = *f,m,n,s,v,z,ð,ŋ,ʃ,ʒ,θ* | 50/65 | .729 |
| splɪŋt | $\varnothing \rightarrow$ t / X $\begin{bmatrix} -\text{syl} \\ \text{0-2son} \\ \text{0-1aper} \end{bmatrix}$ ___ # <br> = *p,k,t,f,θ,s,ʃ,ʧ* <br> *b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ* | 79/142 | .528 |
| splæŋ | ɪ $\rightarrow$ æ / X ˈ ___ {m,n,ŋ} | 2/4 | .310 |

- $\varnothing \rightarrow$t rule covers voiceless segments + n (lɚn → lɚnt 'learn')

# Diagnosis

- $\varnothing \rightarrow$ t / X $\begin{bmatrix} -\text{syl} \\ \text{0-2son} \\ \text{0-1aper} \end{bmatrix}$ ___ # covers two distinct sets of words
  = *p,k,t,f,θ,s,ʃ,ʧ*
  *b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ*
  - Regular affixation after voiceless obstruents
  - Irregular forms *learnt*, *burnt*, etc.
- Comparison of [-voi] and *n* should reveal that reliability is very different in these two contexts
  - $\varnothing \rightarrow$ t / $\begin{bmatrix} -\text{syl} \\ \text{0-2son} \\ \text{0-1aper} \\ +\text{voi} \end{bmatrix}$ ___ #: 78/117 = .667
    = *p,k,t,f,θ,s,ʃ,ʧ*
  - $\varnothing \rightarrow$ t / $\begin{bmatrix} -\text{syl} \\ \text{0-2son} \\ \text{0-1aper} \\ \text{-voi} \end{bmatrix}$ ___ #: 1/25 = .04 (*learnt*)
    = *b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ*

# Impugnment

- When generalizing from rule R to superset R′
  - Compare confidence of subset context (R) and context outside subset (R′—R)
  - Be generous: lower confidence limit of subset, upper confidence limit of superset
- Impugnment: if superset confidence is lower, replace score of rule with (upper) confidence of superset region
- Result: $\varnothing \rightarrow$t rule is adjusted from .528 to .090

# Results of impugnment

Outputs for [splɪŋ] with impugnment:

| Output | Rule | Rel. | Conf. |
|---|---|---|---|
| splɪŋd | Ø → d / X [−syl, 1-2son, 0-1aper] ___ # <br> = *f,m,n,s,v,z,ð,ŋ,ʃ,ʒ,θ* | 50/65 | .700 |
| splæŋ | ɪ → æ / X ˈ ___ {m,n,ŋ} | 2/4 | .310 |
| splɪŋt | Ø → t / X [−syl, 0-2son, 0-1aper] ___ # <br> = *p,k,t,f,θ,s,ʃ,ʧ* <br> *b,d,g,v,ð,z,ʒ,ʤ,m,n,ŋ* | 79/142 | .090 |

- Effect is stronger with larger training set

# Outline

# Some limitations of the model

- Input representations are 'flat'
  - No prosodic information (monosyllabic vs. polysyllabic)
  - No morphological information (inflection class, gender, etc.)
- Contexts are strictly local
  - No harmony, other non-local conditioning

# Simulating tiers in a flat representation

Goal:

- Give the model access to prosodic size of the base
- E.g., monosyllabic vs. longer

Approach

- Create dummy segments, specified for just for 'length' feature
  - E.g., M (monosyllabic) vs. P (polysyllabic)
  - Feature: [±polysyllabic]

| | Polysyllabic |
|---|---|
| M | 0 |
| P | 1 |

- Could likewise specify gender/inflection class, nearest vowel, etc.

# Augmenting inputs with dummy segments

Simple case: all changes are suffixal (or prefixal)

- Use 'empty space' on other side of affix to specify suprasegmental context

| | | | |
|---|---|---|---|
| leklebP | leklebiP | fegrivP | fegriviP |
| globlefP | globlefiP | pomM | pomiM |
| nuvM | nuviM | sivodP | sivodiP |
| gorM | goriM | sogM | sogiM |
| fabegP | fabegiP | zibotP | zibotiP |

- Resulting rule format
  - $\varnothing \rightarrow$ i / segmental context ___ prosodic context
- In practice, adding symbols to mark boundary can aid legibility
  - lekleb●P ∼ leklebi○P, sog●M ∼ sogi○M, ...
  - Change: ●→i○ (different input/output symbols ensure that they are not treated as part of the context)

# Discontinuous morphological contexts

- A particularly strong context for ɪ→ʌ (Bybee and Slobin 1982; Bybee and Moder 1983)
  - sC(C)____ ŋ(k)
- Model presented here cannot discover this, due to its locality restriction
  - Alignment outward from change can't skip segments, or find features of multiple segments
  - Also cannot posit rules with literal segment outside featurally underspecified segments (sCC)
- Bybee and Moder (1983): speakers do generalize to novel sC(C)...items more than C(C)

# A better alignment procedure

Albright and Hayes (2006): minimum string edit distance

- Aligns strings by finding optimal segmental correspondence between segments
- Align matching segments with one another
- Mismatches: find phonetically closest correspondent

| | | | | | | |
|---|---|---|---|---|---|---|
| | s | | w | ɪ | m | |
| | \| | | \| | \| | \| | |
| | ʃ | | ɹ | ɪ | ŋ | k |
| ⇒ | [−voi +strid] | | [−syl $4^{+}$son] | ɪ | [+nas] | k? |
| | \| | | \| | \| | \| | |
| | s | p | ɹ | ɪ | ŋ | |
| ⇒ | [−voi +strid] | p | [−syl $4^{+}$son] | ɪ | [+nas] | k? |

# A challenge...

- Discontinuous alignments retain much more information than the 'myopic' alignments of the local MGL
- Explosion in number of rules
- Albright and Hayes (2006) devise a procedure for pruning the grammar to retain the most reliable rules, using the Gradual Learning Algorithm (Boersma 1997)
- Possible to employ more sophisticated selection/weighting techniques (e.g., maximum entropy models)
- Interesting theoretical issue: what kinds of non-local contexts do human learners actually notice?
  - English *v*- initial verbs are all regular, but it appears that no effect of this is seen in wug verbs
- If you are interested in employing a non-local version of the model, let me know

# Links

- Model download: http://www.mit.edu/ albright/mgl/
- These slides: http://www.mit.edu/ albright/mgl/MGL-Tutorial.pdf
- Additional papers, data files, etc.: http://www.linguistics.ucla.edu/people/hayes/learning/

# References I

Albright, A. and B. Hayes (2002).
Modeling English past tense intuitions with minimal generalization.
In *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, pp. 58–69. ACL.

Albright, A. and B. Hayes (2003).
Rules vs. analogy in English past tenses: A computational/experimental study.
*Cognition 90*, 119–161.

Albright, A. and B. Hayes (2006).
Modeling productivity with the Gradual Learning Algorithm: The problem of accidentally exceptionless generalizations.
In G. Fanselow, C. Féry, R. Vogel, and M. Schlesewsky (Eds.), *Gradience in Grammar: Generative Perspectives*, pp. 185–204. Oxford University Press.

# References II

Boersma, P. (1997).
How we learn variation, optionality, and probability.
*Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 21*, 43–58.
http://fon.hum.uva.nl/paul/.

Bybee, J. and C. Moder (1983).
Morphological classes as natural categories.
*Language 59*(2), 251–270.

Bybee, J. and D. Slobin (1982).
Rules and Schemas in the Development and Use of the English Past Tense.
*Language 58*(2), 265–289.

# References III

Ling, C. and M. Marinov (1993).
Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs.
*Cognition 49*, 235–290.

Pinker, S. and A. Prince (1988).
On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition.
*Cognition 28*, 73–193.