Class 4: Inductive Learning by Minimal Generalization

(1) Today

- Some somewhat-old work on how to learn alternations inductively
- Problems involving generalizations of different sizes and overlap: can maxent help?

(2) Readings

- Albright and Hayes (2002) (changed)
- Software for this paper, in user-friendly version, is available if you want to try it: course website

(3) What follows

The next part of this handout is a modified version of a handout for a talk given seven years ago at the Workshop on Morphological and Phonological Learning, ACL 2002, Philadelphia.

(4) Overall Goal

- This is about phonotactics, complementing last time’s work on alternations.
- A shared theme is experimentation with low-UG models:
  ➢ Can intensive scrutiny of the data yield accurate grammars using less UG?

(5) Specific goals

- Develop a system that apprehends the regularities in morphological paradigms, and uses them to generate novel forms.
- Goal is to model people; i.e. an adequate system should mimic human judgments and behavior.
- For example, when given a wug test (Berko 1958):
  ➢ “John like to plim; yesterday he ___."
  the model should give the same answers as are given by native speakers of English.
- Modeling people implies a number of criteria of adequacy.

(6) We’re not the first

- The creation of similar models (Rumelhart-McClelland 1986, Seidenberg, Plunkett) was a striking achievement of the connectionists, and launched the famous “past tense debate.”
• Work by Mark Johnson (1984), which I wish we had read...

CRITERIA OF ADEQUACY

(7) **Generate Complete Output Forms**

rather than just grouping the outputs into (possibly arbitrary) categories such as “regular,” “irregular,” “vowel change.”

(8) **Make Multiple Guesses for Each Word**

in cases where people feel this is appropriate

> spling: splinged, splung, splang

(9) **Rate Each Output on a Scale**

• Human judgments are characteristically gradient (Class 1)

  ➢ Human ratings for *plim*, from our own Wug test:

<table>
<thead>
<tr>
<th>Output</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>plimmed</td>
<td>6.1</td>
</tr>
<tr>
<td>plum</td>
<td>4.2</td>
</tr>
<tr>
<td>plam</td>
<td>3.6</td>
</tr>
</tbody>
</table>

• Since people can rate forms on a numerical scale, the model should be able to as well.

WHAT A MODEL MUST DO TO SATISFY THESE CRITERIA

(10) **Locate Detailed Generalizations**

• Example: here are all the *i* → *ʌ* verbs of English (one dialect only; you may differ):

  ➢ fling-flung, cling-clung, sting-stung, wring-wring, sling-slung, string-tring, swing-swung, spring-sprung
  ➢ slink-slunk, shrink-shrunk, stink-stunk
  ➢ spin-spun, win-won
  ➢ dig-dug, stick-stuck

• There is a specific phonological context that strongly favors *i* → *ʌ*, namely / ___ ñ

• Experimental work (Bybee and Moder 1983, Prasada and Pinker 1993) shows that human speakers have a stronger preference for *i* → *ʌ* for wug verb stems that match this context.

• Hence this context must be learned by the model.

(11) **Locate Detailed Generalizations II: Regulars**

• All verbs in English ending in voiceless fricatives ([f, θ, s, ñ]) are regular (e.g. *laughed, missed, wished*).
• Our experiments show that human speakers have a stronger preference for the regular outcome when the wug verb matches the /voiceless fricative/ context.
• Hence the model must be able to learn this context.

(12) Defn. island of reliability

• An island of reliability is an environment where a particular change applies with greater-than-average consistency.
  ➢ /voiceless η is an island of reliability for I → A.
  ➢ /fricative/ is an island of reliability for Ø → -ed.

(13) Locate Broad Generalizations

• Sometimes the model must derive outputs for which no close analogues are present in the training data.
• Example: in Pinker’s (1999) “Handel out-Bached Bach,” [aʊtaɪbæt] must be derived, even though there may be no stems in the training data ending in the (non-English) sound [x].
• This can be done only if the model discovers broad generalizations (using ordinary data) that will encompass the unusual novel forms.

DESCRIPTION OF THE MODEL

(14) Training Data

• Pairs of morphologically related forms, e.g. verb stems + past tenses

  \[
  \begin{align*}
  ([\text{mis}]_{\text{pres.}}, [\text{must}]_{\text{past}}) & \quad \text{‘miss(ed)’} \\
  ([\text{pres}]_{\text{pres.}}, [\text{prest}]_{\text{past}}) & \quad \text{‘press(ed)’} \\
  ([\text{kəf}]_{\text{pres.}}, [\text{kəeft}]_{\text{past}}) & \quad \text{‘laugh(ed)’} \\
  ([\text{hæg}]_{\text{pres.}}, [\text{hægd}]_{\text{past}}) & \quad \text{‘hug(ed)’} \\
  ([\text{ræb}]_{\text{pres.}}, [\text{ræbd}]_{\text{past}}) & \quad \text{‘rub(ed)’} \\
  ([\text{nid}]_{\text{pres.}}, [\text{nidɔd}]_{\text{past}}) & \quad \text{‘need(ed)’} \\
  ([\text{dʒæmp}]_{\text{pres.}}, [\text{dʒæmp}]_{\text{past}}) & \quad \text{‘jump(ed)’} \\
  ([\text{plæn}]_{\text{pres.}}, [\text{plænd}]_{\text{past}}) & \quad \text{‘plan(ed)’}
  \end{align*}
  \]

• Goal is to create a grammar that generates the second form from the first.

(15) Situating the task

• We conjecture that children start out memorizing present-past pairs, then use that database to produce a grammar, upon which they can synthesize.
This gets us what seems to be right about the “U-shaped curve” (Marcus et al. 1992)

(16) Overall Strategy (Pinker and Prince 1988: 130-136)

- Parse each input pair into a changing portion and a context, yielding word-specific rules.
- Compare rules with one another to construct more general rules.
- Iterate.

(17) Parsing Pairs into Changing Portion and Context

- Assuming rule format $A \rightarrow B / C ___ D$, maximize $C, D$.  
- For miss/missed:

  $\begin{align*}
  A & \quad B \\
  \quad | & \quad |
  \# m s \emptyset \# & \quad \# m s t \# \\
  \hline
  C & \quad D
  \end{align*}

  yields $\emptyset \rightarrow t / \# m s ___ \#$

- This has intriguing complications in ambiguous cases, e.g.

  $pita \sim p-um-ita, muma \sim m-um-uma$ (prefix? infix?)

  These will generally will be fixed by our preference for generality (below)

(18) Generalizing by Comparing Word-Specific Rules

$\emptyset \rightarrow t / m \; i \; s \; \#$  \hspace{1cm} (from miss-missed)
$+ \emptyset \rightarrow t / pr \; \varepsilon \; s \; \#$  \hspace{1cm} (from press-pressed)

$= \emptyset \rightarrow t / X \; s \; \#$

(19) Formula for Rule Generalization

$A \rightarrow B / \quad C_1 \quad _\quad D_1$ \hspace{1cm} word-specific rule
$+ A \rightarrow B / \quad C_2 \quad _\quad D_2$ \hspace{1cm} word-specific rule

$= A \rightarrow B / \quad X \; C'_{\text{feat}} C' \; _\quad D' \; D'_{\text{feat}} \; Y$ \hspace{1cm} generalized rule

---


2 Details: where more than one parse is available, prefer suffixation over prefixation, and prefixation over infixation: hence $(ta, tata)$ yields $\emptyset \rightarrow ta / \# ta ___ \#$; $(tapa, tatapa)$ yields $\emptyset \rightarrow ta / \# ___ tapa$.
• Going leftward from the change location,
  ➢ Locate the maximal shared segmental string (C');
  ➢ Then, if the material in the two words is not yet exhausted, form a feature matrix containing all features shared by the next adjacent segments (C'_{feat}).
  ➢ Then, if the material in the two words is still not exhausted, form a free variable (X).

• Repeat going rightward from the change location, to find D', D'_{feat}, and Y as necessary.

(20) Example

\[
\begin{align*}
\emptyset & \rightarrow t / \# m_{1} s_{1} C_{1} \quad \# D_{1} \\
A & \quad B \\
\emptyset & \rightarrow t / \# pr_{2} e_{2} C_{2} \quad \# D_{2} \\
A & \quad B \\
\end{align*}
\]

\[
\begin{align*}
\emptyset & \rightarrow t / X_{1} s_{1} C'_{feat} \quad C' \quad \# D' \\
A & \quad B \\
\end{align*}
\]

(21) General Philosophy

• Form the tightest rule that covers both of original rules; hence the name minimal generalization.

(22) Traffic Control

• Grammar is constructed incrementally by considering one input pair at a time.
• For each input pair, a word-specific rule is formed ((17)), which is then compared with all existing rules, generalizing wherever possible.\(^3\)

(23) Virtues of Minimal Generalization

• Minimal generalization yields rules for every change, so that the resulting grammar can generate multiple outputs for the same input.

---

\(^3\) We believe, but have not proven, that no additional rules are discovered by comparing generalized rules against generalized rules.
• Minimal generalization discovers detailed generalizations. In particular, as applied to English it discovers
  - the /__ŋ context for ₁ → λ
  - the voiceless-fricative context for regulars
• With sufficient iteration (usually, just a few dozen pairs), minimal generalization also discovers highly general rules, by generalizing over a diverse set of cases.
  - With phonology (see below), the system discovers the standard, very simple English past tense rule ∅ → d / # X __#.

EVALUATING RULES AND OUTPUTS

(24) Gradient Well-Formedness
• Goal: assign gradient well-formedness scores to each output.
• Method: evaluate the reliability of rules, then evaluate outputs on the basis of the rules that derive them.

(25) Reliability of Rules
• How well does a rule perform in the existing lexicon? To determine this:
  - Let scope be the number of forms in the training data that meet the structural description of the rule (for A → B / C__D, these are the forms that contain CAD).
  - Let hits be the number of forms that a rule derives correctly
  - The reliability of a rule is hits/scope.

(26) Why should be trust a rule? I
• Pinker and Prince (1989) suggest scope is all that matters.
• This can’t work: we find that tiny rules compete well with huge ones, if they are accurate enough: spling:
  - splung average rating 5.45
  - splinged average rating 4.36

(27) Why should be trust a rule? II
• Pure accuracy is another candidate.
• Here is a rule that is perfect:
  ₁ → λ / [ [–voice] l __ ŋ ]
• It works for cling, fling, and sling, 3/3.
• Yet it is not much stronger than the regular past rule (spling, above)
(28) Adjusting for the Quantity of Evidence

- Intuition: reliability based on high scope (for example, 990 correct predictions out of 1000) is better than reliability based on low scope (for example, 5 out of 5).
- Implementation (Mikheev 1997): adjust reliability using lower confidence limit statistics.\(^4\)
  - The amount of the adjustment is a parameter (\(\alpha\)), which ranges from \(.5 < \alpha < 1\); the higher the value of \(\alpha\), the more drastic the adjustment.
- Adjusted reliability is termed confidence.

(29) Deriving Outputs for a Novel Form

- Use all the applicable rules in the grammar to generate a set of outputs.
- Each output gets a well-formedness score, which is defined as the confidence score of the best rule that derives it. Scale is 0-1.
- We propose such scores as a model for human well-formedness intuitions. Thus, for \(plim\) ((9) above):

<table>
<thead>
<tr>
<th>Humans (1-7 scale)</th>
<th>Model (0-1 scale)</th>
<th>Rule Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>(plimmed) 6.1</td>
<td>.97</td>
<td>(\emptyset \rightarrow d / X) [(+\text{voice}) [-contin]] #</td>
</tr>
<tr>
<td>(plum) 4.2</td>
<td>.41</td>
<td>(1 \rightarrow \Lambda / X) [(-\text{syllabic}) [+\text{voice}] [-\text{syllabic}] [+nasal] ]</td>
</tr>
<tr>
<td>(plam) 3.6</td>
<td>.19</td>
<td>(1 \rightarrow \varepsilon / X) [(-\text{syllabic}) [+sonorant] [-\text{nasal}] [-\text{syllabic}] [+nasal] ]</td>
</tr>
</tbody>
</table>

(30) Qualms, 7 years later

- This is an algorithm made up for the purpose; there ought to be an algorithm that is reliable on principled grounds…

---

\(\text{Following Mikheev, we use the following formula to calculate lower confidence limits: first, a particular reliability value (\(\hat{p}\)) is smoothed to avoid zeros in the numerator or denominator, yielding an adjusted value } \(\hat{p}^{*}\):\)

\[
\hat{p}^{*} = \frac{\text{Hits} + 0.5}{\text{Scope} + 1.0}
\]

This adjusted reliability value is then used to estimate the true variance of the sample:

\[
\text{estimate of true variance} = \sqrt{\frac{\hat{p}^{*}(1 - \hat{p}^{*})}{n}}
\]

Finally, this variance is used to calculate the lower confidence limit (\(\pi_{L}\)), at the confidence level \(\alpha\):

\[
\pi_{L} = \hat{p}^{*} - z(1-\alpha/2) \times \sqrt{\frac{\hat{p}^{*}(1 - \hat{p}^{*})}{n}}
\]

(The value \(z\) for confidence level \(\alpha\) is found by look-up table.)
DISCOVERING PHONOLOGY

(31) The Traditional Generative Model

- Morphological rules concatenate morphemes in their underlying forms, creating phonological underlying representations.
  - for *jumped*: /dʒæmp/ + /d/
- These are submitted to the phonology, which derives surface representations.
  - /dʒæmp+d/ → [dʒæmp], by Progressive Voicing Assimilation
- Result: by making use of the phonological regularities, the morphology of the language is simplified and generalized: a single [-d] suffixation rule now suffices.
- How can this system be learned by a model like ours?

(32) Approach

- We assume that before human language learners take on morphology they have a fairly good idea of the phonotactics of their language (i.e. what is phonotactically legal/illegal).
  - Experimental support for this view: work by Jusczyk and colleagues with 8-10 month old infants (see Jusczyk et al., 1993; Friederici and Wessels, 1993)
  - Also, last time, using a quick application of the phonotactic algorithm, we discovered the voicing-agreement constraint. (Don’t know about the alveolar cluster constraint…)
- Moreover, the wrong guesses of preliminary rules can be used to discover phonology.

(33) Example

- Example: generalizing over ([hæg], [hægd]), ([ræb], [ræbd]), ([juz], [juzd]), we get

\[ \emptyset \rightarrow d / X^{\text{sonorant}} + \text{voice} \quad \_\# = \text{“attach [d] after any voiced obstruent”} \]

- Applied to *need* [nid], this derives the useful error *[nidd]*.

- Given *[nidd], ✓[nɪdð]*, and prior knowledge that *[dd]* is illegal, the system posits phonology:

  
  /nɪd+d/

  \[ \emptyset \rightarrow \_ / d \_ _ d \]

  [nɪdð] output

  Schwa Epenthesis: \[ \emptyset \rightarrow \_ / d \_ _ d \]
Proceeding similarly, the system is able to learn the “Linguistics 101” English past tense rule: suffixation of /-d/ across the board, followed by phonological rules of epenthesis and devoicing.

(34) A Further Challenge

- Minimal generalization is characteristically conservative, and often fails to generate the informative errors needed to learn phonology.
- We generate these errors by forming “doppelgängers”—constraint that attach alternative allomorphs in the same context.
- We don’t really use underlying forms, but this is our “poor man’s underlying form”.
- For underlying forms of stems, see Adam Albright’s work, http://web.mit.edu/albright/www/.

(35) General Prediction

- The base form of affix (used for attachment) must be one of the allomorphs present in the paradigm; hence no abstract segments, etc.
- For defense of this view see Albright (2002).

THE DISTRIBUTIONAL ENCROACHMENT PROBLEM

(36) The Core of the Minimal Generalization Approach

- Learn the distribution of allomorphs by generalizing over the contexts in which they occur.
- But some broad generalizations are quite misleading.

(37) Example: burnt-class Verbs in English

Question: “Where is /-t/ used in forming past tenses?”

- Answer I: after voiceless obstruents
  
  - [mɒs]-[mɒst] ‘miss(ed)’
  - [lɛf]-[lɛft] ‘laugh(ed)’
  - [dʒʌmp]-[dʒʌmp] ‘jump(ed)’

- Answer II: assuming a (perfectly workable) phonological rule
  
  \[
  \begin{array}{c}
  t \rightarrow d/
  \hline
  \text{−syllabic} \\
  \text{−sonorant} \\
  \text{+voice}
  \end{array}
  \]

  we can cover voiced obstruent examples like
  
  - [hʌɡ]-[hʌgd] ‘hug(ged)’
  - [rʌb]-[rʌbd] ‘rub(bed)’
  - [juz]-[juzd] ‘use(d)’
Now the answer is: “after any obstruent.”

• Answer III: Suppose the learning set includes at least one of the following dialectal irregular forms, where [-t] occurs after a sonorant:

(\([b\circ n]\)_{\text{pres.}}, \([b\circ nt]_{\text{past}}\)) ‘burn(t)’
(\([l\circ n]\)_{\text{pres.}}, \([l\circ nt]_{\text{past}}\)) ‘learn(t)’
(\([d\ell w]\)_{\text{pres.}}, \([d\ell wt]_{\text{past}}\)) ‘dwell(t)’
(\([s\ell p]\)_{\text{pres.}}, \([s\ell lt]_{\text{past}}\)) ‘spell(t)’
(\([s\ell m]\)_{\text{pres.}}, \([s\ell ml]_{\text{past}}\)) ‘smell(t)’

Then there will be further generalization, and the answer becomes “after any consonant.”

• This is not a good idea! burnt etc. are irregular forms, and should not be determining a high-level generalization—especially because the confidence score for this generalization would be rather high (.7).

(38) The Problem Stated More Generally

• Occasionally, an affix has multiple allomorphs, and there are a few irregular forms in which one allomorph “encroaches” on the context of another.

  ➢ In burnt, [-t] encroaches on [-d]’s territory.

• Distribution encroachment shows that one should pay attention to the internal homogeneity of generalizations.

(39) Our Solution (in outline)

• Force all rules to outperform the rules that cover a subset of their cases; if a rule fails to outperform its subsets, it incurs a penalty.

• This penalizes the overly-general rule \(\emptyset \rightarrow t / \# X [\text{consonant}] \_\#\); the penalty is enough that this rule is never used in deriving output forms.

(40) Looking ahead

This looks like a “credit assignment” problem that maxent might be able to solve for us.

TESTING THE MODEL

(41) Training

• Training corpus: 4253 verbs = all verbs of frequency \(\geq 10\) in the English portion of the CELEX database (Burnage 1991)

• We trained the model to predict the past tense form from the present stem.
(42) **Corpus testing**

- When you have just a few constraints, overlearning is probably not a peril, but it definitely is here, so we took a standard precaution:
  - Divide training data randomly into ten parts.
  - Predict past tenses for the verbs of each tenth based on the remaining nine tenths.
- Results:
  - For virtually every verb, the first choice of our model was the regular past tense.
  - Past suffix took the phonologically correct form: [-t], [-d], or [-ød], depending on the last segment of the stem.
- This mimics a general preference English speakers have for regular pasts.
- When humans speakers output irregular pasts for existing verbs, this is best attributed to their having memorized them (see Pinker 1999).

(43) **Generalization Beyond the Training Data**

- Examining the inflection of novel forms is the best way to compare a model with human performance, because it forces both humans and model to create new forms productively (Ling and Marinov 1993).

(44) **Some Simple Examples**

- Because it learns general rules, the model assigns correct past tenses to unusual words of a type not occurring in the training data.
  - e.g. Prasada and Pinker’s (1993) forms *ploamph* and *smairg* were assigned the correct pasts *plomft* and *smergd*.
- This extends to sounds that don’t occur in English: *out-Bach* is derived correctly as *[autbax]*.

(45) **Modeling Native Speaker Judgments in a Wug Test (Albright and Hayes 2003)*

- Stimulus:
  
  “The chance to *rive* would be very exciting. My friend Sam _______ once, and he loved it.”

- Tasks:
  - Fill in the blank.
  - Rate different possibilities on a numerical scale.

<table>
<thead>
<tr>
<th><em>rifed</em></th>
<th>worst</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>best</th>
</tr>
</thead>
</table>

* Linked from course web page
• 41 subjects volunteered forms; of these, 21 also provided ratings.

(46) **Verbs Tested**

• Four kinds, classified according to the model’s predictions:

  I. should sound especially good as regular, but not as irregular
     Example: *blaf* [ends in a voiceless-fricative; cf. (11)]

  II. should sound especially good as (some kind of) irregular, but not as regular
     Example: *sling* [falls in the /___ŋ island for i → Λ; cf. (10)]

  III. should sound good both as regular and as some kind of irregular
     Example: *bize* [fricative stems typically regular, at → o frequent before coronals]

  IV. should not sound especially good either as regular or as any kind of irregular
     Example: *gude*

(47) **Results: Mean Ratings**

Fig. 1: Effect of Islands of Reliability (IOR) on Irregulars and Regulars

(a) IOR Effect on Ratings (adjusted)  
(b) IOR Effect on Production Probabilities

• See Albright and Hayes (2003, *Cognition*) for full details.

(48) **Discussion**

As our model predicts, English speakers

• Have gradient intuitions.
• Show a strong general preference for regulars.
• Give relatively higher scores to irregulars when they fall within an island of reliability for an irregular change, e.g. i → Λ/___ŋ (columns II/III higher than I/IV)
• Give relatively higher scores to regulars when they fall within an island of reliability for the regular change, e.g. $\emptyset \rightarrow -ed / X \left[ \text{voiceless fricative} \right] \#$ (columns I/III higher than II/IV)
• Do not in general (Albright and Hayes 2003) produce responses supported by one single model ($gezz - gozz, zay - zed$). That would not be minimal.

(49) Word-by-Word Correlations

• Ratings Data ($n = 41$)
  
  regulars $r = .745, p < .0001$
  irregulars $r = .570, p < .0001$

• Volunteered Data (% volunteered, $n = 41$)
  
  regulars $r = .695, p < .0001$
  irregulars $r = .333, p < .05$

(50) The Level of Detail in Human Linguistic Knowledge

• Applying our model to other languages, we have consistently found that it locates generalizations that were missed in earlier paper-and-pencil analyses.
• To some extent, we have also been able to show that these generalizations are internalized by human speakers.
  
  ➢ Italian conjugation classes are partially predictable from the phonological form of the stem (Albright, 2002, Language)
  ➢ Spanish diphthongization is partially predictable from segmental context of the changing vowel (Albright, Andrade, and Hayes 2001)
  ➢ The location of subject marking in Lakhota (infix vs. prefix) is partially predictable from the phonological form of the stem (Albright, 2000)
  ➢ It is partially predictable (postdictable) which stems underwent the “honor” analogy of Latin (Albright, 2002)

SOME WAYS THE MODEL COULD BE IMPROVED

(51) Phonological Representations and Rules

➢ Representations are from Chomsky and Halle (1968) (sequences of feature matrices).
➢ Rules follow the very simple schema $A \rightarrow B / X C_{\text{feat}} C \_\_ D_{\text{feat}} Y$.

• Phonology is richer than this, and in a number of areas, generalization will not be possible until the model incorporates more elaborate rules and representations.
• Both of the areas to be mentioned got addressed in Hayes and Wilson (2008), but not yet in this learner.

---

(52) Example 1: Nonlocal Rules

- The concept of “closest vowel” is needed for e.g. Hungarian vowel harmony:

  \[ k\text{ö} \, n^\text{ö}-n \, Ak \rightarrow k\text{ön}^\text{ö}-n^\text{ék} \quad \text{‘book-dative’} \]

  \[
  \begin{array}{c}
  \text{target vowel} \\
  \text{three intervening consonants} \\
  \text{trigger vowel}
  \end{array}
  \]

  - Our model cannot ignore the consonants that intervene between vowels, so it could not learn this kind of rule.

(53) Example 2: Prosody

- Prosodic structure often plays a role in defining morphological rules.
  - Syllables: all polysyllabic English verb stems are regular (Pinker and Prince 1988)
  - Syllable weight (e.g. Latin abstract nouns in [-ia][-ie:s]; Mester 1994)
  - Metrical feet (e.g. foot-based allomorphy in Yidiñ; Dixon 1977)

(54) …or maybe not

- See
  

  for a later attempt to find non-local stuff by string-alignment procedures.

(55) Multiple changes

- Recall from (17) that we sought to locate the “changing portion” by maximizing the context terms:

  “Assuming rule format A \rightarrow B / C ___ D, maximize C, D.”

- But in many cases, this fails to locate a generalizable change, because there are two changing portions. Ilokano:

  \[ n\text{waŋ} \quad \text{‘water buffalo’} \quad # \quad n \, w \, a \, ŋ \quad # \]

  \[ p\text{ag-waŋ-an} \quad \text{‘place for water buffalo’} \quad # \quad p \, a \, g \, n \, w \, a \, ŋ \, a \, n \quad # \]

  - The rule obtained by our method is nwaŋ \rightarrow pagnwaŋan / # ___ #.
We need something like the two-rule solution $\varnothing \rightarrow \text{pag} / \# \_\_ \_ \_ X \#$, $\varnothing \rightarrow \text{an} / \# X \_\_ \_ \_ \#$.

- What might help:
  - Use some form of string-edit distance (Kruskal 1983), weighted by phonetic similarity, to determine that -nway- is the string shared by the two forms;
  - Adopt some method of morpheme discovery (e.g. Baroni 2000; Goldsmith 2001; Neuvel, to appear; Schone and Jurafsky 2001; Baroni et al. 2002; Snover, Jarosz and Brent 2002) and use its results to favor rules that prefix pag- and suffix -an.

WEIGHING CONFLICTING EVIDENCE: WHAT IS THE RIGHT WAY?

(56) OT’s method is (probably) not right

- OT (e.g., in the form of Boersma’s GLA) ranks constraints solely on the basis of when they conflict.
- This wrongly lets perfect low-scope constraints totally outrank extremely general constraints.
- See Vsevolod Kapatsinski (forthcoming, linked from course web site) for elegant experimental work suggesting the same conclusion.

(57) Five phenomena we must consider

- Straightforward ranking of large-scale generalizations, with full override.
- Small perfect generalizations making some headway against big imperfect ones.
- Distributional encroachment (above)
- Islands of reliability
- Pseudo-islands of reliability

(58) Straightforward ranking of large-scale generalizations, with full override

<table>
<thead>
<tr>
<th>Input: Xm</th>
<th>Xmd</th>
<th>ADD D</th>
<th>ADD T AFTER VOICELESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xm</td>
<td>Xmd</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Xmt</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Input: Xp</td>
<td>Xpd</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Xpt</td>
<td>1000</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

ADD D 13.98
ADD T AFTER VOICELESS 28.36

<table>
<thead>
<tr>
<th>Input:</th>
<th>Candidate:</th>
<th>Observed:</th>
<th>Predicted:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xm</td>
<td>Xmd</td>
<td>1000</td>
<td>0.999999</td>
</tr>
<tr>
<td>Xm</td>
<td>Xmt</td>
<td>0</td>
<td>8E-7</td>
</tr>
<tr>
<td>Xp</td>
<td>Xpd</td>
<td>0</td>
<td>5.6E-7</td>
</tr>
</tbody>
</table>
(59) Small perfect generalization making (only) some headway against big imperfect ones

<table>
<thead>
<tr>
<th></th>
<th>Xpt</th>
<th>1000</th>
<th>0.999999</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TAKE D</th>
<th>TAKE T AFTER VOICELESS</th>
<th>ING UNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xing</td>
<td>Xung</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Xingd</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xingt</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Xvoiceless</td>
<td>XvoicelessT</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>XvoicelessD</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>X</td>
<td>Xd</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xt</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

- This one yields near-perfect matching if you don’t hobble ING-UNG
- So you need to hobble—the second reason (after unnaturalness, perhaps) for hobbling.
- I tried, purely ad hoc: sigma = 100000 for the general constraints, 10 for ING-UNG
- This yields 62% Xung, 38% regular.
- As noted earlier, we need a principled basis for hobble-size, not an ad hoc adjustment.

(60) Distributional encroachment

- Here, the allomorph for one environment occurs just a few times in the environment of the other.
- Many dialects of English have this in verbs like burnt, spelt, spoilt.
- Presumably, these are irregular and should get little credence.
- Albright/Hayes added a whole extra provision “Impugnment” to their system to handle this—since they create the unwanted constraint “Add t to any stem”.
- Maxent treats the marginal cases straightforwardly as irregulars.

<table>
<thead>
<tr>
<th></th>
<th>TAKE D</th>
<th>TAKE T AFTER VOICELESS</th>
<th>TAKE T ANYWHERE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xvoiced Xd</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Xt</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Xvoiceless Xd</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Xt</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>X-ODDn X-ODDnd</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>X-ODDnt</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TAKE D</td>
<td>5.52</td>
<td></td>
</tr>
<tr>
<td>TAKE T AFTER VOICELESS</td>
<td>20.24</td>
<td></td>
</tr>
<tr>
<td>TAKE T ANYWHERE</td>
<td>1.73</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Xvoiced</td>
<td>0.996 d — i.e. frequency matching</td>
</tr>
<tr>
<td>Xvoiceless</td>
<td>1.000 t</td>
</tr>
</tbody>
</table>
(61) Islands of reliability

- These work fine in maxent, with the island constraint ganging with its regular partner to produce the required boost.
- This is hard to simulate, but here is a rough approximation:

<table>
<thead>
<tr>
<th></th>
<th>Xd</th>
<th>X</th>
<th>1425</th>
<th>1</th>
<th>TAKE D AFTER VOICED</th>
<th>TAKE T AFTER VOICELESS</th>
<th>TAKE T AFTER VOICELESS FRICATIVE</th>
<th>TAKE IRREG (VARIOUS CONSTRAINTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xvoiced</td>
<td></td>
<td>Xd</td>
<td>1425</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XI</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRREG</td>
<td></td>
<td></td>
<td>75</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xvoiceless</td>
<td></td>
<td>Xd</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XI</td>
<td></td>
<td></td>
<td>950</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRREG</td>
<td></td>
<td></td>
<td>50</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xf</td>
<td></td>
<td>Xfd</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xft</td>
<td></td>
<td></td>
<td>300</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRREG</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weights:

| TAKE D AFTER VOICED | 14.9 |
| TAKE T AFTER VOICELESS | 14.9 |
| TAKE T AFTER VOICELESS FRICATIVE | 11.2 |
| TAKE IRREG (VARIOUS CONSTRAINTS) | 11.9 |

This produces what we would hope for:
Frequency matching (95/5) for the regulars.
100% regular for the island.

(62) Pseudo-islands of reliability

- Imagine a dialect of English in which the only verbs that start with [dʒ] are:

  judge, gerrymander, jabber, gel, jumble

- Imagine a learner that infers:

  **TAKE D AFTER dʒX**

  - One imagines that a Wug test with e.g. *joke* will yield voiceless -t, since hundreds of words support this choice.
  - This is a dangerous situation in the Albright/Hayes system of constraint evaluation—a small but perfect generalization ought to have some say.
  - Maxent utterly rejects this spurious environment:
<table>
<thead>
<tr>
<th></th>
<th>TAKE D</th>
<th></th>
<th>TAKE T AFTER VOICELESS</th>
<th>TAKE D AFTER JX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xvoiced</td>
<td>Xd</td>
<td>400</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xt</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xvoiceless</td>
<td>Xd</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xt</td>
<td>200</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>JXVoiced</td>
<td>JXVoicedD</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>JXVoicedT</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>JXVoiceless</td>
<td>JXVoicelessD</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>JXVoicelessT</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Take d 13.1
Take T after voiceless 26.0
Take d after JX 0.3

with virtually 100% regulars derived.

- Why? My guess is that the algorithm “sees” that promoting Take d helps accuracy with 400 words, and promoted Take D after JX only helps a subset of 4 of them.
- The weak prior I used (sigma = 100000) perhaps exaggerated this effect.
- But notice that for independent reasons, Take D after JX would be hobbled.
- Such hobbling is perhaps needed when there are whole hordes of pseudo-IOR’s; Albright and Hayes (2006).

(63) Upshot

- Everything seems to be going swimmingly but for one case; i.e. the need to hobble low-scope perfect constraints.