Exemplar-Based Linguistics

Royal Skousen

21 April 2014
University of Illinois
Urbana-Campaign
A Classical Rule

Indefinite article in English

$a$ __ + C

following word begins with a consonant

$a$ boy, $a$ university, $a$ history

$an$ __ + V

following word begins with a vowel

$an$ apple, $an$ honest, $an$ historical
An Exemplar-Based Approach

Use of examples to predict *a* versus *an*

For a given case, hunting for an appropriate exemplar to make a prediction:

___ + coy

the actual occurrence of *coy*

a nearest neighbor: *boy, toy, decoy*

exemplars further away: *cry, … , zebra*
The “Rule” Equivalent

(1) Apply every possible “true” rule:

- $a$ if followed by a consonant (ah, the real rule!)
- $a$ if followed by an obstruent: soy, pie, fried, ...
- $a$ if followed by a velar: goy, guy, grim, ...
- $a$ if followed by a voiceless velar stop: kite, criminal, cute, ...
- $a$ if followed by a velar and a diphthong: coy, cow, kind, ...

etc.

(2) The probability of applying a “true” rule is proportional to its frequency squared.
No “false” rules!

Every subrule of a “true” rule behaves identically:
    a “homogeneous” or “correct” rule

Eliminate the “false” (“heterogeneous” or “incorrect”) rules, such as:

    90% of the time choose \textit{a} \\
    when the first vowel in the following word is \textit{oi}: \\
    \textit{a} \quad \textit{toy, boy, boycott, loiterer}, ... \\
    \textit{an} \quad \textit{oil, oink, ointment}
Missing the significant generalization

You flunk beginning linguistics!
The real question is: What are speakers doing?
Everything is analogical, no rules at all.

How could the brain keep track of all possible “true” rules?
And why would the probability of applying a rule equal the square of its frequency?
Doing it with exemplars, not rules

A dataset based on specific occurrences specifying variables, for example:
phonemes (two before and two after)
consonant-vowel distinction (C versus V)
phrasal break before the article (+ or | )
and the outcome, for each one, *a* or *an*
Sample database

164 instances, 136 with *a* and 28 with *an* with no exceptions to the “rule”
From an actual text, the first 10:

* a  CrVu+CgC1 through a glass darkly
* a  VACs+CmV% across a most interesting book
* a  VeCt+CkVA get a copy
* a  CnCt|CfCy print. a few years
* an VoCr+VICn for an institute class
* an VeCt+V*Cr get an early version
* a  VaCd+CbV@ had a bound copy
* a  VECv+CbVU of a book
* an V$Cd+VICm played an important part
* a  VeCn|CgCr then, a great honor
Predictions for given contexts

Here are 8 (one of which is in the dataset):

? CkCt+VaCp picked __ apple
? CkCt+VoCr picked __ orange
? CkCt+CpVe picked __ pear
? VACn+CtV! upon __ time
a VaCt+CnVa at __ gnat
? CsV$|VACn say. __ honest
? CpV$|CyVu pay. __ university
? ViCz+VECw he’s __ Hawaiian

Virtually always 100% if consonant-initial, from the beginning
Four vowel-initial cases

- `o he’s _ Hawaiian`
- `x picked _ orange`
- `+ picked _ apple`
- `* say. _ honest`

NUMBER OF OCCURRENCES
Children’s errors

I want it go in a upper.
What’s a alligator?
This a end.
No, this a engine.
I wanta make a egg.
Where’s a other one?

Reversions in learning, making “progress”
Transition from *a* to *an*

Initial preference for *a*

Gyrations towards *an*

No place at which the traditional rule is learned

Adult behavior is reached after about 80 instances

A residual window of leakage:

one-way (*an > a*, not *a > an*)

Sapir: “All grammars leak”
“Competence is performance”

No need for a rule independent of usage

No need to find reasons for the one-way leakage

*markedness*

open syllables versus closed syllables

*pronunciation*

harder to pronounce *an boy* than *one boy*?
Robustness

The rule is fragile:
What to do when the initial segment is overlaid with noise?

Influence of a particular word:
we can recognize the intended word

Certain sequences of sounds are expected:
if the second segment is an $s$,
the first is undoubtedly a vowel in English

Need for redundant variables;
not just the crucial ones
A different database

251 instances, 211 with *a* and 40 with *an*

no exceptions to the “rule”
from an actual text, the second 10:

```
a , V I  C th + C b  V e  , with a better
a , V a  C z  + C k  V E  , as a conjectural
a , V a  C z  + C r  V I  , as a result
an , V E  C z  + V o  C f  , was an awful
a , V e  C m  | C r  V i  , them | a reading
a , V I  C n  + C d  V I  , in a difficult
an , V I  C ng + V e  C s  , missing an s
a , V *  C r  + C sh  V o  , after a short
a , V I  C th + C s  V i  , with a single
a , C p  C t  | C n  V E  , manuscript | a number
```
The same basic results!
Historical or Dialectal Drift

First iteration: from no exceptions to a few
Second iteration: put the exceptions in the data set
   A few more leaks!
Continuing with further iterations,
   we get a few relics of an
Eventually, all instances of an go to a
A Neogrammarian-like shift without exception,
   but with Sapir’s drift
The traditional $s$-curve!
Turbulent drift!
“The Social Motivation of a Sound Change”
Martha’s Vineyard Island, 1961
ai > əi
au > əu
effect of linguistic and social variables
turbulent drift for individual speakers
at varying rates for individual words
probabilistic throughout
North Tisbury Fisherman GB

right
night
white
like
sight
quite
striped
swiped
wife
life
knife
spider
side
tide
applied
characterized

Ivory
live
five
I’ve
by
fly in
high
fryin’
why
my
try
I’ll
piles
while
mile
violence
<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>shiners</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kind</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td>about</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iodine</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td>trout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quinine</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td>house</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>south</td>
<td></td>
<td></td>
</tr>
<tr>
<td>line</td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td>mouth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>couch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fired</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>now</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tire</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>how</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sound</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>down</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>round</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hound</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ground</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic variables</td>
<td>Not favoring centralization</td>
<td>→</td>
<td>Favoring centralization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----------------------------</td>
<td>---</td>
<td>-------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sonorants</td>
<td></td>
<td>→</td>
<td>obstruents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nasals</td>
<td></td>
<td></td>
<td>orals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced</td>
<td></td>
<td></td>
<td>voiceless</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>velars</td>
<td></td>
<td></td>
<td>apicals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fricatives</td>
<td></td>
<td></td>
<td>stops</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Centralization by Age

<table>
<thead>
<tr>
<th>Age Range</th>
<th>ai &gt; əi</th>
<th>au &gt; əʊ</th>
</tr>
</thead>
<tbody>
<tr>
<td>over 75</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>61 to 75</td>
<td>0.35</td>
<td>0.37</td>
</tr>
<tr>
<td>46 to 60</td>
<td>0.62</td>
<td>0.44</td>
</tr>
<tr>
<td>31 to 45</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>14 to 30</td>
<td>0.37</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Centralization by Geography

<table>
<thead>
<tr>
<th>Geography</th>
<th>ai &gt; əi</th>
<th>au &gt; əu</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Down-island</strong></td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Edgartown</td>
<td>0.48</td>
<td>0.55</td>
</tr>
<tr>
<td>Oak Bluffs</td>
<td>0.33</td>
<td>0.10</td>
</tr>
<tr>
<td>Vineyard Haven</td>
<td>0.24</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Up-island</strong></td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Oak Bluffs</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>North Tisbury</td>
<td>0.35</td>
<td>0.13</td>
</tr>
<tr>
<td>West Tisbury</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Chilmark</td>
<td>1.00</td>
<td>0.81</td>
</tr>
<tr>
<td>Gay Head</td>
<td>0.51</td>
<td>0.81</td>
</tr>
</tbody>
</table>
# Centralization by Occupation

<table>
<thead>
<tr>
<th>Occupation</th>
<th>ai &gt; øi</th>
<th>au &gt; øe</th>
</tr>
</thead>
<tbody>
<tr>
<td>fishermen</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>farmers</td>
<td>0.32</td>
<td>0.22</td>
</tr>
<tr>
<td>others</td>
<td>0.41</td>
<td>0.57</td>
</tr>
<tr>
<td>Ethnic Group</td>
<td>ai &gt; ei</td>
<td>au &gt; eu</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>English</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>Indian</td>
<td>0.56</td>
<td>0.90</td>
</tr>
</tbody>
</table>
## Centralization by Attitude

<table>
<thead>
<tr>
<th>Attitude</th>
<th>$ai &gt; ei$</th>
<th>$au &gt; eu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td>neutral</td>
<td>0.32</td>
<td>0.42</td>
</tr>
<tr>
<td>negative</td>
<td>0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Combinations of probabilities?

Probabilistic rules?

David Sankoff approach:
- an underlying probability for each variable, linguistic and social
- a logistic formula combines the probabilities for any given collection of variables
- one free factor determined by the data

A static approach:
- does not predict drift nor movement towards exceptionality
Neogrammarian tendencies

Dealing with exceptions?
  ignore, set aside, due to dialect contamination

Strong evidence for drift towards completion

Possibility of different states of final completion,
  found in related languages and dialects
Gemination in Finnish dialects

Southwest dialects: special gemination

obstruent > long / __ VV

\{p, t, k, s\}

no restriction on the preceding syllable

now historical, no longer productive
Another form of gemination

Eastern dialects: general gemination

consonant > long / [stressed V] __ VV
obstruents and sonorants
preceding syllable has some stress
currently productive
Extension of eastern gemination

Eastern dialectal special gemination
consonant > long / __ VV
no restriction on the preceding syllable
spreading independently outward
on the edges of general gemination
Kettunen, Rapola earlier reported as idiosyncratic;
now complete in many places

Analogy works towards regularization
Dialect Map of Gemination
Complex Morphological Systems

Children’s errors:
  regular verbs > irregular
  “Yesterday it snew.” (my son Lawrence, 6 years, 10 months old)
  “He succame to it.” (a teenager, reported by Bruce Derwing)

Analogical modeling: A single-route system
  Regular verbs must be discovered among the mass of all the verbs, irregular and regular
### Past-tense database for English

Based on actual past-tense forms (all verbs are familiar); from speech and writing of children (3rd through 6th grades)

Six stages, exponentially increasing (doubling), from the most frequent to the least:

<table>
<thead>
<tr>
<th>Stage</th>
<th>Count</th>
<th>Range</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>stage 1</td>
<td>30</td>
<td>1 – 30</td>
<td>be, have, … , know, like</td>
</tr>
<tr>
<td>stage 2</td>
<td>60</td>
<td>31 – 60</td>
<td>write, sit, … , grow, pull</td>
</tr>
<tr>
<td>stage 3</td>
<td>122</td>
<td>61 – 122</td>
<td>fly, win, … , miss, invite</td>
</tr>
<tr>
<td>stage 4</td>
<td>244</td>
<td>123 – 244</td>
<td>wonder, last, … , count, shed</td>
</tr>
<tr>
<td>stage 5</td>
<td>488</td>
<td>245 – 488</td>
<td>arise, sign, … , arrange, brag</td>
</tr>
<tr>
<td>stage 6</td>
<td>976</td>
<td>489 – 976</td>
<td>split, trade, … , unite, value</td>
</tr>
</tbody>
</table>
Figure 1. Predicted past tense forms for the verb ‘snow’
overflew

Figure 2. Predicted past tense forms for the verb ‘overflow’
Figure 3. Predicted past tense forms for the verb 'succumb'
Scatalogical past-tense forms

Figure 4. Predicted past tense forms for the verb ‘shit’
Behavior near exceptions

ox ~ oxen

fox ~ foxes, box ~ boxes, tax ~ taxes

effect of nearest neighbors:

nonce test: ux ~ uxen (about 20%)

exceptional plural: ax ~ axen (NPR radio)
Names of the days

*Monday*: /məndei/ ~ /məndi/

and all other names for days of the week

From older speakers:

*Wendy’s* as /wɛndeiz/

*Sandy* as /sændei/, corrected to /sændi/
Exceptional misspellings

Frequent misspelling by school children
GREAD for grade (cf. great)
Historical change

WHO, WHOSE, WHOM: /h/ exceptionally spelled as wh
/hwo:/ > /ho:/ (late Middle English)
as in /two:/ > /to:/ for two

WHOLE, WHORE, WHOOP
all spelled earlier with h, historically not /hw/

Other words misspelled in Early Modern English (1400s – 1600s)
WHORD hoard
WHOLY holy
WHOME home
WHOOD hood
WHOTE hot
Pronunciation effects

consonantal /ɛnt/, not /ænt/
<< continental

nuclear /kjələr/, rather than /kliər/
<< particular, spectacular, circular,
  molecular, secular, perpendicular,
  muscular, vernacular, jocular, ...
standard nuclear << cochlear (transplant)
Will nearest neighbors work?

Effects of regular, strong gangs
further away than the nearest neighbors

Past tense of *sorta-* ‘to oppress’ in Finnish

<table>
<thead>
<tr>
<th>TV-si</th>
<th>sorta- $\sim$ sorsi</th>
<th>cf. murta- $\sim$ mursi ‘break’</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-I</td>
<td>sorta- $\sim$ sorti</td>
<td>cf. souta- $\sim$ sousi ‘row’</td>
</tr>
</tbody>
</table>

Historically, supposed to be *sorsi*
Nearest neighbor is *murta-
Inexplicable until explained by analogical modeling
Evidence for sorti

Nykysuomen Sanakirja
sorta- listed under its own morphological type
citations: 10 sorti, 1 sorsi

Suomen Kielen Perussanakirja
sorti listed as normal, sorsi as rare
citations: 2 sorti, 0 sorsi
Looking for the answer

The regular $V$-$i$ rule just took over
(nearest neighbor doesn’t work)

Dissimilation: avoidance of two $s$’s in a row

sisään ‘in’, sisu ‘courage’, sorsa ‘wild duck’,
susi ‘wolf’, sisältä- ‘to contain’

The answer: it’s the $o$ vowel
sorti
Learning probabilities

Analogical Modeling has no probabilistic rules

Predictions are made “on the fly”,
made in terms of the given context
All we have are exemplars

Hypothetical probabilistic rule for dived ~ dove

Mark Davies’ Corpus of Contemporary American English (COCA), 1990-2012

52.9% dove, 47.1% dived

A random sequence of d’s and o’s:

29 instances of dove, 21 of dived

dodoooodoodooddodoooddddoddododdoddododdod oddododdoddo ...
Storing and accessing exemplars

Problems with probabilistic rules:
  problems learning a probability from data
  problems using a probability to predict an occurrence

The exemplar approach (analogical modeling):
  store exemplars
  randomly select an exemplar

Difficulties replicating random probabilistic behavior:
  artificial neural nets, parallel distributed processing, connectionism
Adults learning probabilities

(Number of trial blocks (twenty trials per block)
(adapted from Messick and Solley 1957:26)
Kids learning probabilities

no-reward experiment (random selection for all ages):

NUMBER OF TRIAL BLOCKS
(adapted from Messick and Solley 1957:30)
Adding a reward

reward experiment (selection by plurality for older ages):

NUMBER OF TRIAL BLOCKS
(0 1 2 3 4 5 6 7 8 9 10)

ages
- 7-8
- 5
- 3-4

(adapted from Messick and Solley 1957:30)
Analogical Modeling Literature

1989, *Analogical Modeling of Language*  
Kluwer: Dordrecht, Netherlands

1992, *Analogy and Structure*  
Kluwer: Dordrecht, Netherlands

2002, *Analogical Modeling: An Exemplar-Based Approach to Language*  
edited with Deryle Lonsdale and Dilworth Parkinson  
John Benjamins: Amsterdam, Netherlands
The AM Website

http://humanities.byu.edu/am/

Articles by various scholars using analogical modeling

The AM computer program, available for download in various formats
The Exponential Explosion

Every new variable added:
  doubled the running time
  doubled the memory requirements

Can currently run up to 70 variables
  basically in linear running time
  after 70 variables, exponentiality sets in
  space requirements are consistently exponential
The “rule” equivalent to AM

All possible rules are considered, but the heterogeneous ones are removed.

The probability of using one of the homogeneous rules is proportional to its frequency squared.

Frequency squared describes language behavior most accurately.

The squaring also follows from a statistical rule of always minimizing uncertainty (a quadratic measure).
Quantum Computing of Analogical Modeling

Quantum mechanics (QM)
  all possible paths
  interference wipes out conflicting paths
  observe one of the remaining paths
    probability is the amplitude squared

Quantum analogical modeling (QAM)
  all possible variable combinations
  remove all cases of heterogeneity
  observe one of the remaining homogenous cases
    probability is the frequency squared
Richard Feynman’s observation

Quantum computing quote from Richard Hughes

Note that input data must typically be carried forward to the output to allow for reversibility. Feynman showed that in general the amount of extra information that must be carried forward is just the input itself.
QAM Literature on arXiv.org

the arguments for why AM is quantum mechanical

the general quantum algorithm for analogical modeling
runs in linear time and linear space

2010, “Quantum Analogical Modeling with Homogeneous Pointers”
arXiv:1006.3308, 16 June 2010
the simplest conceptual version of QAM
runs in quadratic time and quadratic space