Modeling competition and uncertainty in word class decisions

Adam Albright

Department of Linguistics and Philosophy
Massachusetts Institute of Technology

28 Jan, 2005 / IPAM Probabilistic Models of Cognition

The big question

How do speakers of a language learn to inflect words of their language, in the midst of variable, contradictory, and sometimes sparse evidence?

Conflicting evidence about word classes

Example: verb classes in English

- Most English verbs form their past tense by adding a version of the *-ed* suffix
  - [t]: *rip*[t], *kick*[t], *miss*[t], *wish*[t], *laugh*[t], etc.
  - [əd]: *wait*[əd], *land*[əd]
  - [d]: *rub*[d], *drag*[d], *save*[d], *seize*[d]
- There are numerous other classes, which do other things
  - Change *i* to *u*: *dug*, *spun*, *clung*
  - Change *i* to *a*: *swam*, *sang*
  - Add nothing: *hit*, *shut*, *shed*
  - Change something to *aught*: *taught*, *caught*, *thought*, *brought*
  - Weirder things: *went*, *was*, *had*

Conflicting evidence about word classes

- The past tense of a verb is not always predictable based on the sound of the present tense
  - Minimal pairs: *ring* ∼ *rang* vs. *wring* ∼ *wrung*
  - Optionality for a single verb: *shrunk* ∼ { *shrank*, *shrunk* }
- At some level, speakers must simply learn/memorize the behavior of individual words
- Yet there is evidence that they do more than this
  - Generalize about characteristic properties of word classes, to try to predict the behavior of words

Evidence for generalized knowledge about class membership

- Behavior on novel (nonce) words:
  - John likes to *mip*. Yesterday he ____.
  - John likes to *spling*. Yesterday he ____.

Evidence for generalized knowledge about classes

- "Wug tests"
  - Berko, J. (1958) The child's acquisition of English morphology. *Word* 14, pp. 150-177.
  - Derwing, B. (1979) English Pluralization: A Testing Ground for Rule Evaluation. In Prideaux, Derwing & Baker, eds., *Experimental Linguistics*.
  - Bybee, J. and D. Slobin (1982) Rules and Schemas in the Development and Use of the English Past Tense. *Language* 58, pp. 265-289.
  - Bybee, J. and C. Moder (1983) Morphological Classes as Natural Categories. *Language* 58, pp. 265-289.
  - Prasada, S. and S. Pinker (1993) Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8, pp. 1-56.
  - Albright, A. and B. Hayes (2003) Rules vs. analogy in English past tenses. *Cognition* 90, pp. 119-161.

(and many others)

Evidence for generalized knowledge about class membership

- Errors (esp. among children):
  - *eated*, *goed*, *holded*
  - *brung*
- "U-shaped learning"
  - Marcus, et al (1992) *Overregularization in language acquisition.*
  - Maratsos, M. (2000) *More overregularizations after all: new data and discussion on Marcus & al. J Child Lang.* 27, pp. 183-212.
  - Xu and Pinker (1995) *Weird past tense forms J Child Lang* 22, pp. 531-56.

Evidence for generalized knowledge about class membership

- Historical changes
    - *molt*, *holp*, *clamb*/*clumb* → *melted*, *helped*, *climbed*
    - *stringed*, *dived*, *sneaked* → *strung*, *dove*, *snuck*

Goal

What is the best model of how people actually decide what class a word should belong to?

- Do they try to think of other, similar-sounding words, and decide based on
  - Exemplar-based approach
- Do they extract generalizations about features of words tend to reveal class membership?
  - Feature, or rule-based approach
- Or some combination of the two?

Outline

- Data: verb classes in Spanish
- Models: an exemplar-based model, and a rule-based model
- Comparison of the two approaches
  - Modeling behavior on novel (nonce) words
  - Modeling behavior on a curious class of real words
    - For some existing words, intuitions about class membership fail, and speakers are surprisingly unable to inflect the word
- Preview of the results
  - Classification of novel words is a "well-behaved" problem; classes compete, and words are assigned to the most probable class
  - For uninflectable words, however, we need to explain why no class seems probable enough to choose
  - Claim: there is sometimes too little evidence to go on, because of constraints on how learners search for generalizations

Outline

Background on Spanish verb classes

Spanish verbs fall into classes along a number of dimensions

- Differences that affect practically all inflected forms
- Differences that affect just certain present tense forms[1]
- Differences in how past tenses are formed
- Other minor irregularities

Background on Spanish verb classes

Spanish verbs fall into classes along a number of dimensions

- Differences that affect practically all inflected forms
- Differences that affect just certain present tense forms[1]

Background on Spanish verb classes

The conjugation classes

Spanish verbs fall into (roughly) three idiosyncratic conjugation classes, defined by their "theme vowel" ([a], [e], or [i])

Class 1: [a]		Class 2: [e]		Class 3: [i] ∼ [e]	
'to speak'	*hablar*	'to eat'	*comer*	'to live'	*vivir*
hábl-o	habl-ámos	cóm-o	com-émos	vív-o	viv-ímos
hábl-as	habl-áis	cóm-es	com-éis	vív-es	viv-ís
hábl-a	hábl-an	cóm-e	cóm-en	vív-e	vív-en

- 1sg suffix is always *-o*
- 3sg suffix is *-a* or *-e* (depending on conjugation class)
- Class 1 ([a]) is the largest class (≈85% of all verbs), and is used for new words

Background on Spanish verb classes

Vowel alternations within the present tense

Vowel sometimes changes when the verb root is stressed (1,2,3sg and 3pl)

e ∼ ie

‘to seat’	*sentar*
sient-o	sent-ámos
sient-as	sent-áis
sient-a	sient-an

o ∼ ue

‘to count’	*contar*
cuent-o	cont-ámos
cuent-as	cont-áis
cuent-a	cuent-an

- “Diphthongization”
- There are diphthongizing verbs in all conjugation classes; higher concentration in classes 2 and 3 ([e] and [i])

Background on Spanish verb classes

How predictable is diphthongization?

- Not completely predictable based on surrounding consonants

Diphthongizing		Not diphthongizing	
sentar	'sit'	*rentar*	'rent'
tender	'stretch'	*vender*	'sell'
poder	'be able'	*podar*	'prune'
morir	'die'	*morar*	'dwell'
contar	'count'	*montar*	'mount'

Background on Spanish verb classes

How predictable is diphthongization?

- However, numerous possible generalizations
  - Some conjugation classes more likely to diphthongize (2,3)
  - Different likelihood in different contexts:
    - Never in roots ending in [rx] (Brame & Bordelois 1973, p. 157)
      emerger 'emerge', *asperjar* 'sprinkle', *forjar* 'forge', *sumergir* 'submerge', *divergir* 'diverge'
    - Frequently in roots ending in [r], [rC], or [nC]
      morir 'die', *sugerir* 'hint', *hervir* 'boil', *mentir* 'beg', *sentir* 'feel', etc.
    - Countless other possible generalizations

Background on Spanish verb classes

Another type of vowel alternations

A competing pattern: *e* becomes *i* when stressed

‘to request’	*pedir*
píd-o	*ped-ímos*
píd-es	*ped-ís*
píd-e	*píd-en*

- Verbs with mid-vowel raising are confined to class 3 ([i])

Background on Spanish verb classes

Asymmetries between conjugation classes

Among verbs that have *e* or *o*:

- Class 1 ([a]), diphthongization is a minority pattern (roughly 10%; no raising
- Class 2 ([e]), diphthongization is more prevalent; no raising
- Class 3 ([i]), almost every verb either diphthongizes or raises under stress; almost no non-alternating verbs

Background on Spanish verb classes

Insertion of *k* or *g* in the 1sg

Some verbs add *k* or *g* in just the 1sg form:

Insertion of [k]	
‘to grow’	*crecer*
cré[sk]-o	cre[s]-émos
cré[s]-es	cre[s]-éis
cré[s]-e	cré[s]-en

Insertion of [g]	
‘to put’	*poner*
pó[ng]-o	po[n]-émos
pó[n]-es	po[n]-éis
pó[n]-e	pó[n]-en

- Limited to conjugation classes 2 and 3 ([e] and [i])
- [k] insertion is always after [s]
- [g] insertion in a wider variety of contexts (*sal-g-o* ‘I leave’, *trai-g-o* ‘I bring’)

Background on Spanish verb classes

A goal of the next two sections

- Attempt to discover which generalizations speakers actually make about the factors that encourage these alternations

Outline

1. Background on Spanish verbs
2. Two classification models for word class decisions
   - Classification by similarity to exemplars
   - Classification by features
3. Modeling diphthongization in novel verbs
4. Modeling uncertainty in existing verbs

Outline

Similarity-based classification
The GCM

Generalized Context Model (Nosofsky 1986)

- Novel items are classified by to similarity to existing items
- For novel item i, compare all members j of a class c
- Similarity of i to class $c = \sum e^{(-d_{i,j}/s)}$, where
  - $d_{i,j}$ = psychological distance between i and j
  - s = sensitivity (free parameter)
- Probability of assigning i to class c

$$= \frac{\text{similarity of } i \text{ to } c}{\text{total similarity of } i \text{ to all classes}}$$

Similarity-based classification
The GCM

Generalized Context Model (Nosofsky 1986)

- Applications to language:
  - Modeling regular vs irregular morphology (Nakisa and Hahn 1996; Albright and Hayes 2003)
  - Modeling artificial language learning (Pothos and Bailey 2000)
  - Modeling acceptability of nonsense words (Bailey and Hahn 2001)

Similarity-based classification
Calculating similarity of words

Intuition: words are similar if their component sounds are similar

- That is, if the sounds of one word are well-matched to those of the other
- In order to calculate this, we need:
  - Similarity values for arbitrary pairs of sounds
  - A method of determining optimal alignment of sounds in the words being compared

Similarity-based classification

Calculating similarity of sounds

Broe (1993); Frisch (1996); Frisch, Broe, and Pierrehumbert (2004):

- Similarity of sounds can be calculated from their features
- Relies on fact that sounds are standardly classified according to features (voicing, place of articulation, manner)

Similarity-based classification

Calculating similarity of sounds

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

Similarity-based classification
Calculating similarity of sounds

Properties of phonological features

- Based on a mix of articulatory and acoustic properties
- Primary role of features is to define *natural classes*

Definition

A natural class is a group of sounds that share a particular set of feature specifications

Similarity-based classification

Natural classes

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

Similarity-based classification
Natural classes

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

- [+syllabic] (= vowels)

Similarity-based classification
Natural classes

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

- [−syllabic] (= consonants)

Similarity-based classification
Natural classes

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

- [−syllabic, −sonorant] (= "obstruents")

Similarity-based classification
Natural classes

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

- [+syllabic, −sonorant, +labial] (= “labial obstruents”)

Similarity-based classification

Natural classes

- Classes defined by any possible combination of values for any subset of the feature space
- Subset/superset classes:
  - $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ -\text{voice} \end{bmatrix}$: { p, t, k}
  - $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \end{bmatrix}$: { p, t, k, b, d, g }

Similarity-based classification

Natural classes

- Classes defined by any possible combination of values for any subset of the feature space
- Orthogonally overlapping classes:
  - $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ -\text{voice} \end{bmatrix}$: { p, t, k}
  - $\begin{bmatrix} -\text{sonorant} \\ +\text{labial} \end{bmatrix}$: { p, b, f }

Similarity-based classification

Natural classes

- Features crafted to capture exactly those classes of sounds that have been observed to behave as a group in the world's languages
- Monovalent features: avoid spurious classes based on negative values

Similarity-based classification
Calculating similarity of sounds

	syllab.	sonor.	contin.	nasal	voice	labial	coronal	strid.	lateral	dorsal	high	low	back
p	−	−	−		−	+			−				
t	−	−	−		−		+		−				
k	−	−	−		−				−	+			
f	−	−	+		−	+			−				
s	−	−	+		−		+	+	−				
b	−	−	−		+	+			−				
d	−	−	−		+		+		−				
g	−	−	−		+				−	+			
m	−	+	−	+	+	+			−				
n	−	+	−	+	+		+		−				
ŋ	−	+	−	+	+				−	+			
l	−	+	+		+		+		+				
r	−	+	+		+		+		−				
w	−	+	+		+	+			−	+			
j	−	+	+		+				−	+			
a	+	+	+		+				−	+	−	+	+
e	+	+	+		+				−	+	−	−	−
i	+	+	+		+				−	+	+	−	−
o	+	+	+		+				−	+	−	−	+
u	+	+	+		+				−	+	+	−	+

- E.g., no [−labial] specification

Similarity-based classification
Natural classes

Feature set is universal; individual languages “underutilize” the feature space

- Feature values must distinguish and classify all possible sounds in human languages
- Individual languages use only a fraction of the possible speech sounds
- Consequence: feature values are often redundant

Similarity-based classification

Natural classes

Example

Vowel features

	high	low	back	round
i	+	−	−	−
e	−	−	−	−
a	−	+	+	−
o	−	−	+	+
u	+	−	+	+

Similarity-based classification

Natural classes

Example

Vowel features

	high	low	back	round
i	+	−	−	−
e	−	−	−	−
a	−	+	+	−
o	−	−	+	+
u	+	−	+	+

- [+round]

Similarity-based classification

Natural classes

Example

Vowel features

	high	low	back	round
i	+	–	–	–
e	–	–	–	–
a	–	+	+	–
o	–	–	+	+
u	+	–	+	+

- [+back, −low]

Similarity-based classification

Natural classes

Phonological features define a complexly structured space

- Contingent relations caused by the definition of features
- Contingent relations caused by the fact that each language uses only a subset of available contrasts

Similarity-based classification
Natural classes

- The feature set is not always aptly tailored to the inventory of sounds that a particular languages employs
- This is an obstacle to calculating similarity directly from the feature specifications
- The solution: calculate similarity based on natural classes

Similarity-based classification

Calculating similarity of sounds

Similarity of pairs of sounds: metric based on natural classes (Frisch, Broe, and Pierrehumbert 2004)

- Try all possible combinations of feature values
  - All subsets of features, all combinations of values
- See which combinations result in distinct natural classes
  - Collect set of distinct classes
- Sounds are similar if they are grouped together in many natural classes
  - *m* and *n* are both voiced, both nasal, both sonorant, both stops, etc.
- Similarity of sounds $s_1, s_2 = \frac{\text{shared natural classes}}{\text{shared + unshared natural classes}}$

Calculating alignments

Optimal alignment of sounds in two words calculated with minimum string edit distance algorithm (Kruskal 1983)

- Alignment that permits transformation from one word to the other in as few steps as possible
  - Bailey and Hahn (2001); Hahn, Chater and Richardson (2003); Albright and Hayes (2003)
- Cost of substituting sounds = dissimilarity of the sounds
- Cost of insertions and deletions: parameter, found by fitting (here, .7)

Similarity-based classification

Example: likelihood to diphthongize novel form *lerrar*

- Aggregate similarity of *lerrar* to diphthongizing verbs = 4.936
  - Top contributors: *errar* (0.493), *cerrar* (0.446), *aserrar* (0.110), *helar* (0.105), *aferrar* (0.093)
- Aggregate similarity of *lerrar* to non-changing verbs = 15.551
  - Top contributors: *reglar* (0.268), *orlar* (0.240), *ahorrar* (0.213), *forrar* (0.211), *lograr* (0.203)
- Probability of diphthongizing $= \frac{4.936}{4.936+15.551} = 24.09\%$

Similarity-based classification

A note on token frequency

One might expect token frequency to play a role in weighting attracting to various classes

- More frequent words come to mind more easily, and exert more pull?
- Bailey and Hahn (2001): non-monotonic effect, with mid-frequency words getting most weight
  - But a really tiny effect; leaving frequency out altogether does almost as well
- Bybee (1995), Albright (2002), Pierrehumbert (2003): token frequency does not seem to play a role in this type of decision
  - Certainly, words with high token frequency do not seem attract novel items (*was*, *gave*, *saw*, etc.)
- Parameter fitting in modeling the Spanish data: best fit obtained by ignoring token frequency

Similarity-based classification

Summary:

- Attraction to a class based on similarity to words in that class
- The more words, and the greater their similarity, the stronger the attraction
- Novel items attracted to the class with the greatest aggregate similarity

Outline

The Minimal Generalization model

Albright and Hayes (2002)

The Minimal Generalization Model (Albright and Hayes 2002)

- Explores contexts not by exemplar similarity, but by shared combinations of features
- Inductively learns generative grammar of stochastic rules

The Minimal Generalization model

The phonological rule formalism

$$\boxed{\text{A} \rightarrow \text{B} \;/\; \text{C} \;\_\_\; \text{D}}$$

- Change: A $\rightarrow$ B
- Context: C ___ D
- CAD $\rightarrow$ B

Example

One variant of the English past tense suffix:

- $\varnothing \rightarrow$ əd / $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{coronal} \end{bmatrix}$ ___
- "Add [əd] after coronal stops (= t,d))"

The Minimal Generalization model

Goal: compare words that share the same change, to see what contextual features they have in common

The Minimal Generalization model

Input: pairs of related forms, in phonetic transcription

Stressless	Stressed	Gloss	Orthography
jeg	jeg	‘arrive’	(*llegar*)
dex	dex	‘leave’	(*dejar*)
jeb	jeb	‘bring’	(*llevar*)
ked	ked	stay	(*quedar*)
enkontr	enkwentr	‘find’	(*encontrar*)
pens	pjens	‘think’	(*pensar*)
kont	kwent	‘tell, count’	(*contar*)
entr	entr	‘enter’	(*entrar*)
tom	tom	‘take’	(*tomar*)
kre	kre	‘create’	(*crear*)
empes	empjes	‘start’	(*empezar*)
esper	esper	‘wait, hope’	(*esperar*)
rekord	rekwerd	‘remember’	(*recordar*)
tembl	tjembl	‘tremble’	(*temblar*)

The Minimal Generalization model

Step 1: analyze individual pairs

Factor related forms into changing and unchanging portion

Example

Pair (tembl, tjembl) has vowel change:

- e $\rightarrow$ je / t ___ mbl
- "stressless [e] corresponds to stressed [je] when preceded by [t], and followed by [mbl]"

The Minimal Generalization model

Result: word-specific rules

Stressless	Stressed	Gloss	Rule
jeg	jeg	'arrive'	e → e / j ___ g
dex	dex	'leave'	e → e / d ___ x
jeb	jeb	'bring'	e → e / j ___ b
ked	ked	stay	e → e / k ___ d
enkontr	enkwentr	'find'	o → we / enk ___ ntr
pens	pjens	'think'	e → je / p ___ ns
kont	kwent	'tell, count'	o → we / k ___ nt
entr	entr	'enter'	e → e / ___ ntr
tom	tom	'take'	o → o / t ___ m
kre	kre	'create'	e → e / kr ___
empes	empjes	'start'	e → je / emp ___ s
esper	esper	'wait, hope'	e → e / esp ___ r
rekord	rekwerd	'remember'	o → we / rek ___ rd
tembl	tjembl	'tremble'	e → je / t ___ mbl

The Minimal Generalization model
Step 2: Generalize to discovered shared properties

The challenge: determine what arbitrary pairs of words have in common

- E.g., [entr] and [esper]

e	n	t	r
+syllab	−syllab	−syllab	−syllab
+sonor	+sonor	−sonor	+sonor
+contin	−contin	−contin	+contin
	+nasal		
+voice	+voice	−voice	+voice
	+coronal	+coronal	+coronal
−lateral	−lateral	−lateral	−lateral
+dorsal			
−high			
−low			
−back			

e	s	p	e	r
+syllab	−syllab	−syllab	+syllab	−syllab
+sonor	−sonor	−sonor	+sonor	+sonor
+contin	+contin	−contin	+contin	+contin
+voice	−voice	−voice	+voice	+voice
		+labial		
	+coronal			+coronal
	+strid			
−lateral	−lateral	−lateral	−lateral	−lateral
+dorsal			+dorsal	
−high			−high	
−low			−low	
−back			−back	

The Minimal Generalization model

Strict locality assumption

Locality restriction

Comparing *tembl*/*tiembl-* 'tremble', *desmembr-*/*desmiembr-* 'dismember':

Residue	Shared feats	Shared segs	Change loc.	Shared segs	Shared feats
	t		___	mb	l
des	m		___	mb	r
X	$\begin{bmatrix} -\text{syl} \\ -\text{cont} \end{bmatrix}$		___	mb	$\begin{bmatrix} -\text{syl} \\ +\text{sonor} \\ +\text{contin} \\ +\text{voi} \\ +\text{coronal} \\ +\text{anterior} \end{bmatrix}$

(In other words, after a stop and before [mb] + a liquid)

The Minimal Generalization model

Structural constraints on possible generalizations:

- Locality: words are compared with respect to what sounds are immediately adjacent to the sound in question
  - Here, the sounds immediately before and after the *e*, *o*
- Feature theory
- Rule format: A $\rightarrow$ B / C ___ D
  - C, D further constrained by set of possible terms

The Minimal Generalization model

Increased generality through iterative generalization

- Pairwise comparisons performed iteratively through entire training set
- Comparison across diverse forms yields very broad generalizations
- At the same time, comparison of similar forms yields very narrow generalizations

The Minimal Generalization model

Just a few of many possible generalizations:

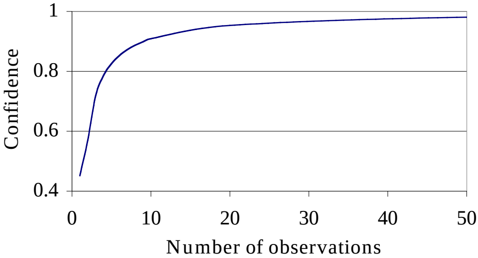
o $\rightarrow$ we / $[+\text{consonantal}]$ ___ rs
o $\rightarrow$ we / $\begin{bmatrix} -\text{continuant} \\ -\text{voice} \end{bmatrix}$ r ___ $\begin{bmatrix} -\text{continuant} \\ -\text{syllabic} \end{bmatrix}$
o $\rightarrow$ we / $\begin{bmatrix} -\text{syllabic} \\ +\text{consonantal} \end{bmatrix}$ ___ $[-\text{syllabic}]$
o $\rightarrow$ o / $\begin{bmatrix} -\text{syllabic} \\ -\text{sonorant} \\ +\text{consonantal} \end{bmatrix}$ ___ $\begin{bmatrix} -\text{syllabic} \\ +\text{consonantal} \\ \text{-continuant} \end{bmatrix}$
o $\rightarrow$ o / $\begin{bmatrix} -\text{syllabic} \\ +\text{voice} \end{bmatrix}$ ___ $[-\text{syllabic}]$
o $\rightarrow$ o / ___ $[-\text{syllabic}]$

The Minimal Generalization model

Evaluating generalizations

Generalizations are evaluated according to accuracy in training set

- Reliability $= \dfrac{\text{cases successfully explained}}{\text{successful + unsuccessful cases}}$
- Reliability values then adjusted downwards, using confidence limit statistics



The Minimal Generalization model

Just a few of many possible generalizations:

o → we / [+cons] ___ rs	4/4	.786
o → we / $\begin{bmatrix} -\text{contin} \\ -\text{voice} \end{bmatrix}$ r ___ $\begin{bmatrix} -\text{contin} \\ -\text{syll} \end{bmatrix}$	6/8	.610
o → we / $\begin{bmatrix} -\text{syll} \\ +\text{cons} \end{bmatrix}$ ___ [−syll]	68/545	.116
o → o / $\begin{bmatrix} -\text{syll} \\ -\text{sonor} \\ +\text{cons} \end{bmatrix}$ ___ $\begin{bmatrix} -\text{syll} \\ +\text{cons} \\ \text{-contin} \end{bmatrix}$	101/106	.934
o → o / $\begin{bmatrix} -\text{syll} \\ +\text{voice} \end{bmatrix}$ ___ [−syll]	19/22	.795
o → o / ___ [−syll]	588/668	.871

The Minimal Generalization model

Using generalizations to make predictions about novel forms

For a novel form *i*, and applicable pattern *p*:

- Find all of the generalizations for *p* which pertain to *i*
- Among these, find the generalization with the highest confidence
- This is the confidence of applying pattern *p* to form *i*

Using best available rule provides good fit to ratings of nonce words (Albright 2002, Albright and Hayes 2003)

The Minimal Generalization model

Using generalizations to make predictions about novel forms

Predicting the *probability* of applying a pattern:

- Production probability:

$$\frac{\text{confidence of the pattern for this form}}{\text{summed confidence of all patterns for this form}}$$

The Minimal Generalization model

Example: likelihood to diphthongize novel form *lerrar*

- Best diphthongization generalization:

  $\boxed{\text{e} \rightarrow \text{je} \,/\, \text{coronal consonant} \,\_\_\_\, \text{voiced consonant}}$

  - Coronal consonants = {t, d, n, l, r, s}
  - Voiced consonants = {b, d, g, v, m, n, ñ, ng, l, r}

  Reliability = 10/29; Confidence = .290

- Best no-change generalization:

  $\boxed{\text{e} \rightarrow \text{e} \,/\, \text{Voiced ``non-vowels''} \,\_\_\_\, \text{sonorants}}$

  - “Non-vowels” = {b, d, g, v, m, n, ñ, ng, l, r w, j}
  - Sonorants = {m, n, ñ, ng, l, r, w, j}

  Reliability = 86/86; Confidence = .989

- Production probability(*lierro*) = $\frac{.290}{.290+.989} = .227$

Contrasting the two models

Minimal Generalization model

- Rules are structured representations
- Classification decisions based on strictly matching a structural description
- Structural descriptions are "definitional" statements
  - Stochastic confidence values allow for probabilistic classification

GCM-based model

- Less structured representation (sounds in a sequence)
- Classification based on sum of individual similarity relations
- Not necessarily any interpretable definition of class
  - E.g., top 5 exemplars encouraging diphthongization in novel verb *solcar*: *solar*, *soltar*, *soldar*, *volcar*, *holgar*

Outline

1. Background on Spanish verbs
2. Two classification models for word class decisions
   - Classification by similarity to exemplars
   - Classification by features
3. Modeling diphthongization in novel verbs
4. Modeling uncertainty in existing verbs

Testing the models

A nonce-probe experiment on diphthongization

Albright, Andrade, and Hayes (2001) Segmental Environments of Spanish Diphthongization.

- asked 96 native Spanish-speakers to inflect nonce words

33 nonce verbs

retolbar	*entar*	*prear*
sendrar	*norrar*	*guechar*
botrar	*chostrar*	*dechar*
fotar	*derrar*	*lecar*
delar	*solmar*	*quertar*
tebrar	*guemblar*	*soltrar*
tojar	*chellar*	*lopar*
rempar	*nomar*	*fostrar*
chortar	*mobrar*	*bectar*
lerrar	*solcar*	*debrar*
boldar	*lorrar*	*colbar*

Testing the models

A nonce-probe experiment on diphthongization

Example: novel verb *lerrar*

Cada verano mi familia y yo *lerramos* durante las vacaciones.
'Every summer my family and I *lerr* during our vacation'

Mi fascina *lerrar*.
'I like to *lerr*.'

Tengo seis meses que yo no ____.
'It's been six months since I have ____.'

Testing the models

A nonce-probe experiment on diphthongization

Example: novel verb *lerrar*

Cada	verano	mi	familia	y	yo	*lerramos*	durante	las	vacaciones.
‘Every	summer	my	family	and	I	*lerr*	during		our vacation’

Mi fascina *lerrar*.
‘I like to *lerr*.’

Tengo seis meses que yo no *lierro*/*lerro*/???.
‘It’s been six months since I have . . . .’

Testing the models

A nonce-probe experiment on diphthongization

- Participants volunteered a form, then provided ratings of diphthongized and non-diphthongized variants
  - Production probability:
    $$\frac{\text{number of times a form was volunteered}}{\text{total number of forms volunteered for that verb}}$$
  - Ratings: 1 (worst) to 7 (best)

Example

For *lerrar*,

- 19/95 participants said *lierro* (prod prob = 20%)
- 76/95 participants said *lerro* (prod prob = 80%)

- For details, see Albright, Andrade and Hayes (2001)

Testing the GCM

Using GCM to predict probability of diphthongization:

- Exemplars: database of Spanish verbs with *e*, *o* in stressless syllables
- Classes: diphthongize, raise, no-change
- Values of insertion/deletion, sensitivity, frequency effect set by fitting
    - indel = .7, s = .4, no effect of token frequency

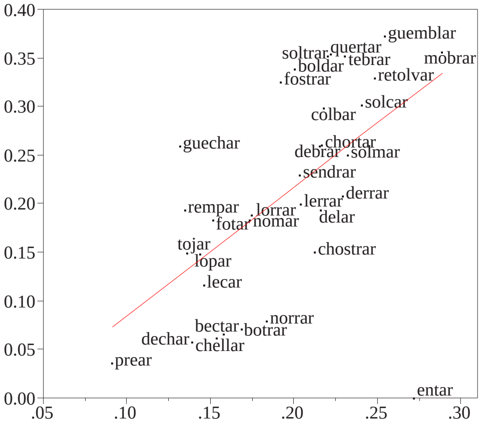
Testing the GCM

An issue that arises:

- Very different rates of diphthongization in the three conjugation classes (*-ar*, *-er*, *-ir*)
- All of the nonce words in this study were *-ar* words (by design)
- When deciding about novel *-ar* words, should the model consider only other *-ar* words, or all roots?
- Tried both models both ways; GNM achieves better fit when all verbs are included, Minimal Generalization model better when broken down by class.
- Treated as a parameter: trained each model using its optimal training set

Testing the GCM

Results ($R^2 = .31$)



Testing the GCM
Results

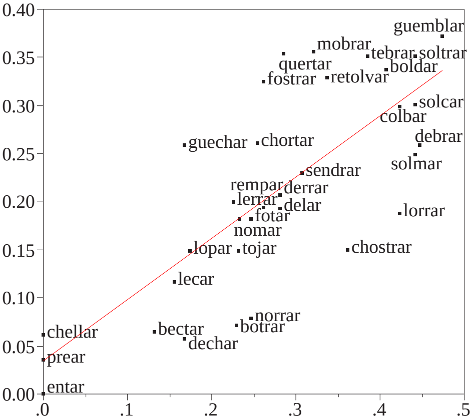
- Overall, a modestly good fit ($R^2 = .31$)
- One salient gaffe: *entar*
  - This verb is quite similar to a number of diphthongizing verbs (*sentar*, *mentar*, *tentar*, *dentar*, *ventar*), so is predicted to be very likely to diphthongize
  - The model has no way of noticing that diphthongization practically never occurs word-initially
  - Human speakers evidently notice and obey this generalization

Testing the Minimal Generalization model

- Trained on same database of Spanish verbs; for this task, limited to just the *-ar* verbs
- Same feature set used to calculate similarity in GCM-based model

Testing the Minimal Generalization model

Results: also a reasonably good fit ($R^2 = .59$)



Testing the models on nonce words

The upshot: both models perform reasonably well on this data

- Rule-based model slightly better
    - Perhaps much better, if wider variety of words like *entar* were included
- One-syllable nonsense words are probably not the best way to differentiate the models, however
- Too few "free parameters" (consonants before the vowel, the vowel, and consonants after the vowel)

Testing the models on nonce words
A caveat

This does not necessarily mean that both models perform well in general on the task of modeling novel words

- Here, we are dealing with very short verb roots (few degrees of freedom)
- Competition between two relatively well-instantiated classes
- This batch of nonce words not designed to tease apart the predictions of the models
- Elsewhere, we have found that unstructured similarity can lead a GCM-based model seriously astray (Albright and Hayes 2003)

Testing the models on nonce words
Pitfalls of the GCM

Global similarity metric leads to "irrelevant" comparisons

- What is the past tense of *whisper*?
- GCM guess: *whispert*
  - On analogy with words like *whip*, *wish*, *whisk*, *wince*, *quip*, *lisp*, *swish*, *rip*, *work*, and *miss*.
  - Form of the past tense suffix depends solely on features of the last sound in the verb
  - Requires a structured representation in which "last sound of the word" is a possible variable

Testing the models on nonce words

Only relevant factors:

- Preceding sound is [+voice]
- Preceding sound is [−voice]
- Preceding sound is [−syllabic] and [−continuant] and [+coronal] and [−strident] (=t,d)

Summary so far

Decisions about inflecting novel words can be modeled as a classification task

- GCM and rule-based classification both do reasonably well modeling this particular data
- Choice rule to turn relative classification weights into behavioral probability

Next section: a harder type of data, in which this type of model fails

Outline

1. Background on Spanish verbs
2. Two classification models for word class decisions
   - Classification by similarity to exemplars
   - Classification by features
3. Modeling diphthongization in novel verbs
4. Modeling uncertainty in existing verbs

A surprising phenomenon

Fill in the blank

"John loves to dive. Every afternoon, he goes diving at the neighborhood pool. He must have ____ off the board there hundreds of times."

Definition

Paradigm gaps (= *ineffability*): forms which one would expect to exist are, for some reason, missing or awkward.

A surprising phenomenon

Another example from English

What is the past tense of *smite*?

- `http://www.geocities.com/mierda_de_vaca/smote.html`
  I hung up with a clatter, leaped into my slippers, pondered over a grapefruit my awful fate. **Smited**! **Smote**! **Struck down**!
- `http://www.trashcity.org/ARTICLES/IBFS0007.HTM`
  ...but the Arab planes are **smote** (**smited**? **smut**? **smeet**?) mysteriously from the sky to the bafflement of everyone.
- `http://heather.tadma.net/archives/00000661.html`
  So I think I've been **smote**. **Smited**? What is the past tense of "to smite"? Anyway, **God got me**!

Paradigm gaps in Spanish

"What is the 3pl of the verb *abolir*?" ('to abolish')

- "um. . . abol. . . abuelen. . . abuelen? oh my god. . . abolir, ellos. . . abolen. . . no. . . oh. . . "
- "abuel. . . abuelen. . . ab. . . ellos abuelen? abolen. . . wow."

"What is the 1sg of *distender*" ('to distend')

- "distendo. . . distiendo. . . no, momento, momento. . . problema. . . distendo? distiendo?"

Here, gaps seem to be caused by uncertainty about whether or not to diphthongize

Possible causes of paradigm gaps

Why does intuition falter? Various possibilities...

- The form is unpronounceable in some way
- The form is never actually used, and speakers are reluctant to synthesize new forms without hearing them first
- Fear of being "wrong" (intuitions point to one form, but speaker has a suspicion that the other form is prescriptively correct)
- Competition between possible outcomes is "too close to call"
- The generative model fails to produce *any* outcome with high enough confidence to sound good

Possible causes of paradigm gaps

Not a pronounceability problem

- Two possibilities for *abolir*: *abolen*, *abuelen*
- Diphthongize: *suelen* ‘used to’, *duelen* ‘hurt’, *vuelan* ‘fly’, *huelen* ‘smell’,
- Don’t diphthongize: *controlan* ‘control’, *violan* ‘violate’, *inmolan* ‘immolate’, *tremolan* ‘flutter’

Pronounceability *does* play a role in some cases:

- 1sg of *roer* ‘gnaw’: should be *roo*, but many speakers find awkward, and avoid (either by fixing: *royo*, *roigo*, or by using a different verb)

Possible causes of paradigm gaps

Reluctance to say new forms?

- A common response when asked about *abolir*:
  - "People just don't use that form. It sounds weird because no one would say it"

Probably not:

- "What is the 1sg of *italianizar*?" ('to italianize')
  - Immediate the universal agreement: *italianizo*
- "What is the 1sg of *descafeinar*?" ('to decaffeinate')
  - Immediate the universal agreement: *descaffeino*
- Even more extreme: nonce probes

Possible causes of paradigm gaps

More evidence that it's not due to scarcity alone:

- Verbs like *abolir* used have inflected forms
- At some point, people stopped using them, even though they had heard their parents saying them
- The historical loss of these forms cannot be explained by reluctance to say unfamiliar word forms

Maybe reluctance specifically when speaker suspects grammar books say otherwise

- A common response when asked about *abolir*:
  - "I'm not sure. You should ask someone who knows more about Spanish grammar."
- Cost-utility analysis: cost of sounding uneducated > cost of not being able to speak

Possible causes of paradigm gaps

Fear of grammar books does not usually create such an extreme reluctance to speak

- Speakers often produce prescriptively banned forms
- Gap effect cannot be modulated by instructions
  - "There are no right or wrong answers—just say what sounds most natural to you."
- Evidently no effect of participant's educational background

Possible causes of paradigm gaps

More suggestive evidence that gaps are linked to uncertainty about word classes: two types of a paradigm gaps

- Type 1: ANTI-STRESS VERBS (Missing forms in which stress falls on the root)

‘to abolish’	*abolir*
—	*abol-imos*
—	*abol-ís*
—	—

- Verbs of this type include: *abolir* ‘abolish’, *agredir* ‘assault’, *aguerrir* ‘harden for battle’, *arrecirse* ‘get stiff with cold’, *aterirse* ‘be numb with cold’, *colorir* ‘color’ (=*colorar*, *colorear*), *denegrir* ‘blacken’, *descolorir* ‘de-color’ (=*descolorar*), *empedernir* ‘harden’, *garantir* ‘guarantee’ (=*garantizar*), *transgredir*/*trasgredir* ‘transgress’, etc.

Possible causes of paradigm gaps

More suggestive evidence that gaps are linked to uncertainty about word classes: two types of a paradigm gaps

- Type 1: ANTI-STRESS VERBS (Missing forms in which stress falls on the root)

'to abolish'	*abolir*
abolo? abuelo?	*abol-imos*
aboles? abueles?	*abol-ís*
abole? abuele?	abolen? abuelen?

- Verbs of this type include: *abolir* 'abolish', *agredir* 'assault', *aguerrir* 'harden for battle', *arrecirse* 'get stiff with cold', *aterirse* 'be numb with cold', *colorir* 'color' (=*colorar*, *colorear*), *denegrir* 'blacken', *descolorir* 'de-color' (=*descolorar*), *empedernir* 'harden', *garantir* 'guarantee' (=*garantizar*), *transgredir*/*trasgredir* 'transgress', etc.

Possible causes of paradigm gaps

More suggestive evidence that gaps are linked to uncertainty about word classes: two types of a paradigm gaps

- Type 2: ANTI-EGOTISTIC VERBS (Missing the 1sg)

'to grasp'	*asir*
—	*as-imos*
as-es	*as-ís*
as-e	*as-en*

  - Also: *balbucir* 'stammer', *pacer* 'graze'

Possible causes of paradigm gaps

More suggestive evidence that gaps are linked to uncertainty about word classes: two types of a paradigm gaps

- Type 2: ANTI-EGOTISTIC VERBS (Missing the 1sg)

'to grasp'	*asir*
aso? azco?	*as-imos*
as-es	*as-ís*
as-e	*as-en*

  - Also: *balbucir* 'stammer', *pacer* 'graze'

Possible causes of paradigm gaps

Claim: gaps are due to low-level breakdown of intuition about which class the word should belong to

- It does not seem to be the case that the speaker decides on classification, but the output is subsequently blocked by unpronounceability, or some metalinguistic realization
- The linguistic system fails to produce a form confidently enough to proceed

Why does classification fail?

Two plausible explanations:

- Two equally good possibilities—system can't adjudicate
  - Similarity/structural overlap with some diphthongizing verbs, some no-change verbs
  - Numbers are too close to provide clear winner
- No good possibilities—not enough evidence to do anything confidently
  - Similarity/structural overlap with very few verbs
  - Numbers are too low to trust

How could speakers possibly lack data?

On the face of it, Spanish provides abundant evidence that could be generalized to these verbs

- Overall, strong preference *not* to diphthongize
- Within *-ir* class, most verbs either diphthongize or raise
- Plenty of similar-sounding verbs that pose no trouble (*soler*, *doler*, *volar*, etc.)

How could the evidence possibly be so evenly tied?

Not only are there many available generalizations, but these are quantifiable

- Preference for non-changing vowels covers hundreds (. . . thousands) of verbs
- Preference for vowel changes in the *-ir* class affects a small number of high frequency verbs, and is not absolute
- Probabilities of vowel-change and no-change are never equal, or even that close

People do not usually avoid “close calls”

Usually when two forms sound equally good, they are both used in free variation

- English examples: *shrank*/*shrunk*, *speeded*/*sped*, *pleaded*/*pled*
- Competing variants sometimes get associated with new meanings (differentiation)
- . . . but they seem not to kill each other off, leaving no usable output

People do not usually avoid “close calls”

- Cases like *-eeded* vs. *-ed* seem like close competition between two well supported patterns
- The observed behavior is different choices on different occasions (but no overwhelming feelings of doubt or uncertainty)
- This type of behavior is also the most straightforward prediction of probabilistic models of competition

So what causes ineffability?

Suggestion:

- Although it seems like many generalizations are potentially available to learners that could cover these cases, in fact the generalization system is constrained so that it cannot find them
- Different conjugation classes involve different suffixes, so generalization proceeds separately for each class
- Locality bias restricts search to features that are adjacent to the change—in this case, focusing attention on the end of the root
- There are rather few *ir* verbs with the vowel *o* in the root; and *abolir* is the only one with an *l* in between

What causes ineffability?

Consequence:

- When generalization is restricted in this way, the number of possible analogs/data points available for mid-vowel *-ir* verbs drops to just a few
  - In most cases, a half dozen or less
- Generalizations based on such small numbers are inherently untrustable, and lead to low confidence
  - Interestingly, ≈6 words is also approximately the threshhold for making past tense patterns moderately plausible in English (cf. the $i \rightarrow a$ change)

What causes ineffability?

An unintended consequence:

- Operating over highly structured representations restricts the hypothesis space so we can search it
- Yet in a few cases, it seems to restrict the hypothesis space so much that no generalization exists

How can we model this?

- Minimal Generalization model already uses lower confidence statistics
- Parallel to the "frequency illusion"; Griffiths & Tenenbaum-style model of causal support

A quantitative test of paradigm gaps

Albright (2003) A quantitative study of paradigm gaps. *WCCFL 22 Proceedings.*

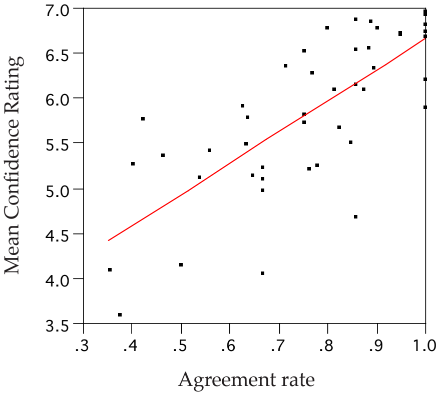
- Preliminary attempt to gather data on reluctance to produce forms
- Constructed a list of 37 existing Spanish verbs, of varying lexical frequencies
    - 27 with potential diphthongization
    - 10 with potential velar insertion

A quantitative test of paradigm gaps

- Pretest: participants asked to rate subjective familiarity (0 = completely unknown, 7 = word and meaning very well known)
- Inflection task: verbs presented in the infinitive form in a frame sentence; participants required to provide an inflected form (1sg or 3sg)
- Rating: at end of each trial, participants rate how confident they are about their reponse

A quantitative test of paradigm gaps

Results, part 1: confidence is correlated with agreement ($r(35) = .75$)



A quantitative test of paradigm gaps

Testing possible sources of doubt

- Trained Minimal Generalization model on Spanish verbs, two different training regimes:
  - Trained on all conjugation classes simultaneously (generalize contexts across all classes)
  - Trained on each class separately (restricted to generalizations within each class)
- Tested model on experimental items
  - Confidence of best output
  - "Winning margin": difference between best output and next best competitor
  - Total summed confidence of all outputs

A quantitative test of paradigm gaps

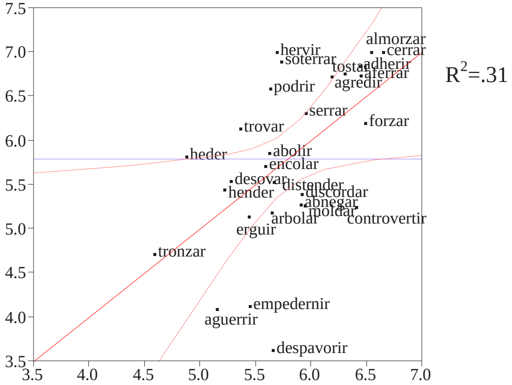
Stepwise multiple regression to determine best predictors of uncertainty ratings:

- Subjective familiarity of the verb
- Model's confidence in the best output
- (Small effect of total summed confidence in all outputs)

Crucially, "winning margin" (closeness of competition) plays no role; uncertainty is better explained by low confidence in winners

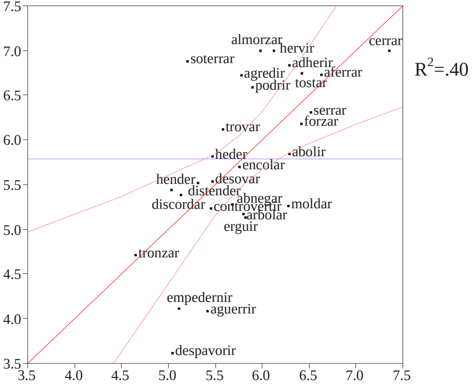
A quantitative test of paradigm gaps

Overall model: $R^2 = .31$



A quantitative test of paradigm gaps

Incidentally: similar results from GCM, if we look at total support for classification (rather than probability of the winning class)



Summary

- Paradigm gaps seem to be related to uncertainty of inference from small numbers of examples—*even though the language appears to provide plenty of examples*
- Highly structured representations too rigid to find broader generalizations?
- Potential to explain why competition sometimes yields variation, and sometimes yields doubt

Conclusion

- A generative model of how words are inflected
  - Scheme for induction of stochastic grammars of rules
  - Competition between best applicable rule from each class
- Support from behavior on nonce-words
  - Probabilistic assignment of words into classes, based on contexts present
- An interesting challenge to this model
  - Cases where classification fails
  - A threshhold effect: too little data to be confident in any generalization
  - Lack of data created by constraints on how words are compared when forming generalizations

Acknowledgments

- Bruce Hayes
- Argelia Andrade