Class 1: Introduction to variation

To do for tomorrow (Tuesday)

- OPTIONAL: read Coetzee & Pater 2011
- Write a one-page description of variation in a project you're working on:
  - Description and examples
  - What kind of variation is it?
  - What are some quantitative questions that you have about it?

0 First-day-of-class items

- introducing ourselves
- syllabus (first page only; we'll do the rest at the end)
- student information sheets

Overview: Types of variation found in phonology: free vs. lexical, multi-site, lexical selection/filtering. For today we will use only simple, non-quantitative models.

1 <u>Idealized</u> free variation—it's possible that no real language has this

Suppose a language has an optional vowel harmony process:

/álkat+i/ → [álk**a**t-i] *or* [álk**e**t-i] (V becomes [-lo] before high vowel)

It's "free" variation because...

- The same speaker can produce both variants, for any word.
- There is no meaning difference between the variants, though they may represent different degrees of formality, different speech rates, etc.
- One variant may be more frequent, but the rate is the same for all target morphemes, and for all triggering morphemes (if target and trigger are in different morphemes):

[álkat-i]	70%	[álket-i]	30%
[móbak-im]	70%	[móbek-im]	30%
[sélab-ik]	70%	[séleb-ik]	30%

- Exception to the above: there might be other phonological factors that affect the rate of variation, but words with the same phonological properties will behave alike:

e.g., suppose that stress matters—stressed V is less likely to undergo harmony

[semát-i]	90%	[semét-i]	10%
[lukár-im]	90%	[lukér-im]	10%
[sikáb-ik]	90%	[sikéb-ik]	10%

Why did I use an imaginary language? Because it's hard (impossible?) to find a real example.

2 Idealized lexical variation

Suppose a language has two different ways to ensure that adjacent obstruents match in voicing ($*\begin{bmatrix}-\text{sonorant}\\ \alpha\text{ voice}\end{bmatrix}\begin{bmatrix}-\text{sonorant}\\ -\alpha\text{ voice}\end{bmatrix}$):

/sif+z/ → [sif-**s**]	change second C
/wof+z/ → [wo**v**-z]	change first C

In the simplest form of lexical variation...

- Each word has just one behavior—the variation is across items, not within items.

3 Modeling idealized free variation

Variable rules

V → [–low] / __ C_0 [+high], *optional*

That is, we just label the rule as optional.

Variable constraint ranking

Jagged line (not standard notation): ranking of these two constraints *varies*

On some occasions, *[+low]C_0[+hi] >> IDENT(low)
On other occasions, IDENT(low) >> *[+low]C_0[+hi]

/álkat-i/	*[+low]C_0[+hi]	IDENT(low)
☞ *a* [álkat-i]	*	
☞ *b* [álket-i]		*

- How is the jagged line different from the dashed/dotted line you often see in tableaux?

Our goal in this seminar: enrich/revise these models that we can say *how frequent* each variant is, and incorporate things like the stress effect in our imaginary example.

4 Lexical variation

Recall: in lexical variation, each word has its own behavior.

Tagalog: Austronesian language from the Philippines with ~17 million native speakers (Ethnologue 2005, data from Zuraw 2009's corpus ; see also Schachter & Otanes 1972)

d → ɾ / V__V :

dunoŋ	'knowledge'	ma-**ɾ**unoŋ	'intelligent'
dinig	'heard'	ma-**ɾ**inig	'to hear'
dupok		ma-**ɾ**upok	'fragile'

But, there are also words like this

daʔig	'beaten'	ma-**d**aʔig	'beaten'
dulas	'slipperiness'?	ma-**d**ulas	'slippery'
daʔan	'road'	ma-**d**aʔan-an	'passable'

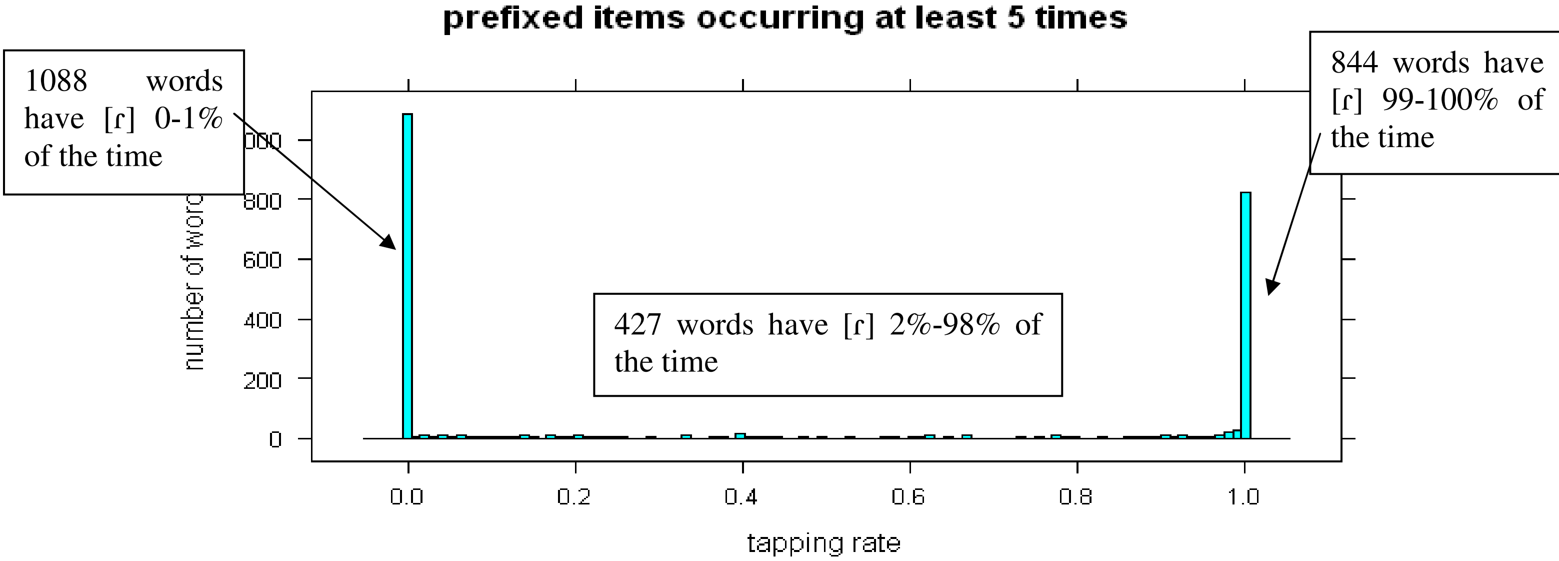
and like this

duŋis	'dirt on face'	ma-**ɾ**uŋis ~ ma-**d**uŋis	'dirty (face)'
dumi	'dirt'	ma-**ɾ**umi ~ ma-**d**umi	'dirty'

How often does each word have each variant?

word	*# with d*	*# with ɾ*	*% ɾ*
ma-_unoŋ	33	9130	99.6%
ma-_inig	97	3517	97.3%
ma-_upok	0	235	100.0%
ma-_ulas	348	23	6.2%
ma-_aʔan-an	132	6	4.3%
ma-_aʔig	102	0	0.0%
ma-_umi	319	708	64.4%
ma-_uŋis	59	52	46.8%

Then, we can count up how many words are 0-<5%, how many 5-<10%, 10%-<15%, etc., and make a **histogram**.



1088 words have [ɾ] 0-1% of the time

844 words have [ɾ] 99-100% of the time

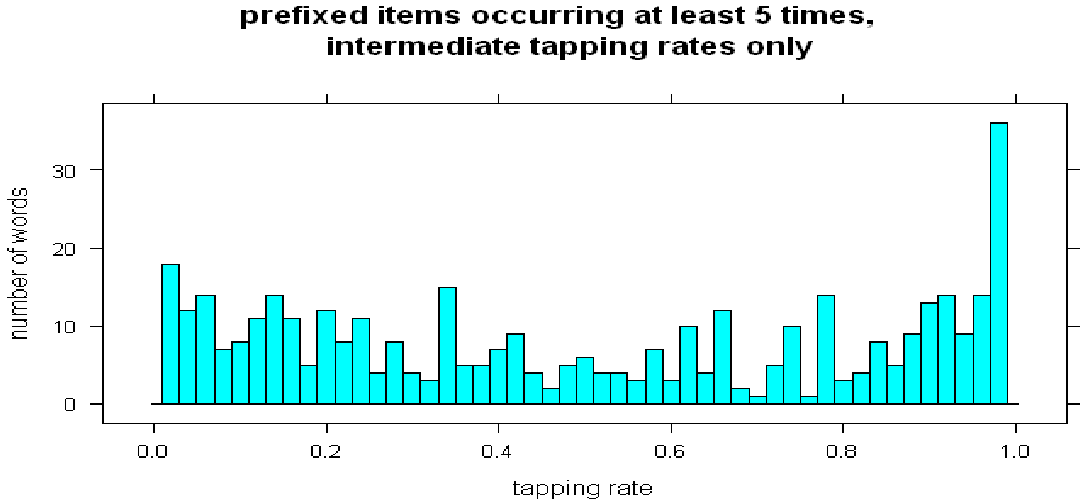
427 words have [ɾ] 2%-98% of the time

==> Most words have a fixed behavior, though some do vary

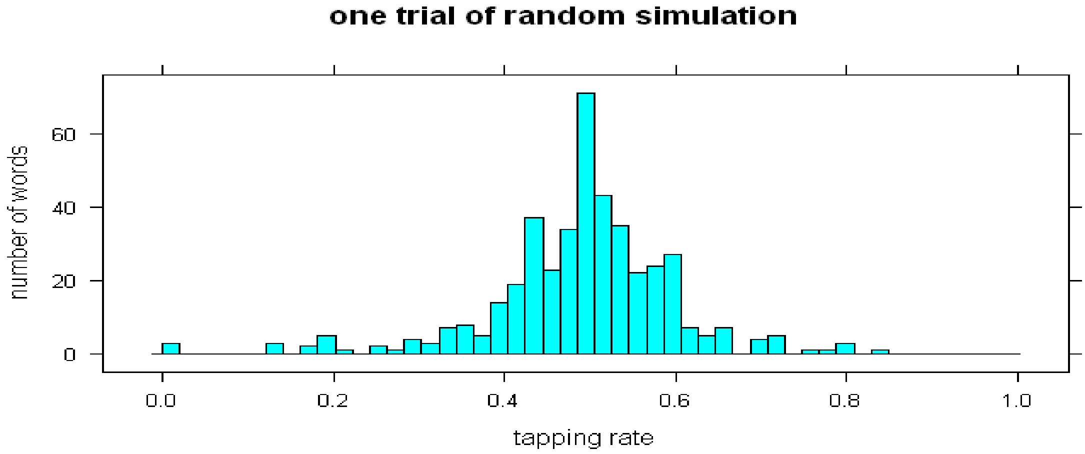
- Let’s sketch out a grammar with variable constraint ranking. What problems do we run into in modeling these data?
- Let’s discuss the pros and cons of simply listing all the prefixed words in the lexicon, with /d/ or /ɾ/ in their lexical entries.

5 Mixed variation

What about the 427 in-between words for Tagalog? Here's a histogram just for them:



If all the words underlying had 50% tapping rate, and we sampled the same number of tokens of each word as found in corpus, we expect a distribution more like this:



Instead of free variation, it looks like different words have their own tapping rates.

Interim summary: Variation can be free (across all words) or lexical (word-to-word), or mixed. Free variation is more straightforward to capture.

6 Multi-site variation

Cases taken from Kaplan 2011, Riggle & Wilson 2005, Vaux 2008. All of these types are hard to find cases of.

6.1 Warao: global optionality

Language isolate of Venezuela, Guyana, and Suriname; 28,100 speakers. From Osborn 1966.

Little raw data, but Osborn is very definite about the generalization:

> "/p/ has allophones [p b]. The voiced allophone [b] is heard more frequently than the voiceless [p] in most words. In every word, except for a few words noted below, alternation between [b] and [p] is presumably possible, since many alternations of this order have been heard. Thus in /paro+parera/ *weak*,

> both the initial and medial phoneme /p/ is heard as [b] generally, and as [p] infrequently. In words like the one cited, with two or more occurrences of /p/, the allophones are consistently [b] or [p] for each utterance of the word. If the first occurrence of /p/ in the word is [b], the following occurrence(s) will be [b]. If the first occurrence is [p], the following occurrence(s) will be [p]. The following are examples of words with two occurrences of /p/: poto+poto *soft*, apaupute *he will put them*, kapa+kapa *kind of banana*." (p. 109)

I.e., [**p**aro-**p**arera] ~ [**b**aro-**b**arera], but not *[**p**aro-**b**arera] or *[**b**aro-**p**arera].
Also, for a non-reduplicative case, [ha**p**isa**p**a] ~ [ha**b**isa**b**a] 'other side'

o Let's make a tableau with variable constraint ranking and see what happens.

6.2 Local optionality (see Riggle & Wilson 2005 for some more cases)

Vaux says that he can produce, for English *marketability*:
[mɑɹkət^{h}əbɪlət^{h}i] ~ [mɑɹkəɾəbɪləɾi] ~ [mɑɹkət^{h}əbɪləɾi] ~ [mɑɹkəɾəbɪlət^{h}i]

o Again, let's make a tableau with variable constraint ranking and see what happens.

6.3 Vata: iterative optionality

Ethnologue classifies as dialect of Lakota Dida, a Niger-Congo language of Côte d'Ivoire with 98,8000 speakers. Data taken from Kaplan 2009; originally from Kaye 1982, which I didn't consult.

[+ATR]: [i,u,e,o,ʌ] [–ATR]: [ɩ, ɷ, ɛ, ɔ, a]

[+ATR] optionally to the final syllable of a preceding word:
/ɔ̀ nɩ sàká **pì**/ → ɔ̀ nɩ sàká **pì** ~ ɔ̀ nɩ sà**kʌ́** **pì** 'he didn't cook rice'

If all the words are monosyllabic, there are various options, all possible...
/ɔ̀ ká zā **pī**/ → ɔ̀ ká zā **pī** ~ ɔ̀ ká **zʌ̄** **pī** ~ ɔ̀ **kʌ́** **zʌ̄** **pī** ~ **ò** **kʌ́** **zʌ̄** **pī** 'he will cook food'

o Let's try a tableau for this one.

6.4 Hypercorrection in Dominican Spanish: unique-target optionality

(Vaux calls this "Basic Optionality")
Dialect of the Indo-European language from Spain with 328 million speakers worldwide. Data from Bradley 2006. See there for original data sources, esp. Núñez-Cedeño 1994, which I didn't get a chance to consult. See also Bullock & Toribio 2010.

/s/ typically deletes in a syllable coda:

Dominican Spanish	*Conservative Spanish*	
se.co	**s**e.co	'dry'
ca.**s**o	ca.**s**o	'case'
e.tú.pi.do	e**s**.tú.pi.do	'stupid'
do	do**s**	'two' (p. 3)

Hypercorrection can insert a coda [s] (in the “hablar fisno” speech style):[1]

Dominican fisno	*Conservative*	
in.vi**s**.tado	in.vi.ta.do	‘guest’
co.mo**s**	co.mo	‘like’
e.tú**s**.pi.do	e**s**.tú.pi.do	‘stupid’
de.de**s**	de**s**.de	‘since’ (p. 4)

And there can be variation of where the [s] is inserted:

Dominican fisno	*Conservative*	
a**s**.bo.ga.do ~ a.bo**s**.ga.do ~ a.bo.ga**s**do ~ a.bo.ga.do**s**	a.bo.ga.do	‘lawyer’ (p. 4)

But, apparently there can only be one inserted *s*:[2] *a**s**.bo.ga.do**s**, etc.

- Let’s try a tableau for this one.

Interim summary: Even free variation becomes problematic when there are multiple targets for variation within a word. Simple variable constraint ranking predicts that all sites behave the same.

7 Lexical selection

There’s one more type of variation we need to consider: not in *how* a form will be pronounced, but in *whether* it will be used at all.

English monosyllables beginning sC and ending with a C, sC{l,ɹ,w,j}*V{l,ɹ,[+nas]}CC*#, as listed in CMU pronouncing dictionary:[3]

C_2 / C_1	p	b	f	v	m	θ	t	d	s	z	n	l	ɹ	tʃ	dʒ	ʃ	k	g	ŋ
p			3		3	3	39	20	14	12	35	27	21	1	5	2	36	6	9
m			2			5	12	3	1			12	5		2	2	13	2	
t	55	25	26	18	30	2	66	31	11	20	39	44	34	13	9	2	80	7	15
n	11	4	6			1	4	4		4		6	8	3			12	5	
l	20	4	3	8	9	4	20	10	5	3	7			1	2	5	8	6	4
k	32	9	16	2	14		33	19	5	16	19	28	20	14	4	2	13	8	
w	24	2	3	3	9	1	15	8	4	5	14	7	5	5		4	4	2	3

small jagged box: C_1 and C_2 are both nasal

box with double lines: C_1 and C_2 are both labial, liquid, or velar

- Certain areas of the chart are underpopulated—discuss.

[1] though not before an otherwise intervocalic tap or trill, which would be phonotactically illegal

[2] See p. 24 for discussion of an apparent counterexample given by Harris.

[3] grep ' S [^AEIOU][^AEIOU]*[AEIOU][AEIOU]*[^AEIOU]*[^AEIUO]$' cmudict_0_6d.txt

Excluded row and columns with totals <10. See spreadsheet for what I did about when to consider {l,ɹ,[+nas]} as pre-C_2 and when as C_2 itself, and some other tricky cases.

How underpopulated? First, determine what we expect if each combination depends just on row and column totals:

C_2 C_1	p	b	f	v	m	θ	t	d	s	z	n	l	ɹ	tʃ	dʒ	ʃ	k	g	ŋ
p	24.1	7.5	10.0	5.3	11.1	2.7	32.1	16.5	7.0	10.2	19.4	21.1	16.3	6.3	3.7	2.9	28.4	6.1	5
m	6.0	1.9	2.5	1.3	2.8	0.7	8.0	4.1	1.7	2.6	4.8	5.3	4.1	1.6	0.9	0.7	7.1	1.5	1
t	53.9	16.7	22.4	11.8	24.7	6.1	71.8	36.8	15.6	22.8	43.3	47.1	36.4	14.0	8.4	6.5	63.4	13.7	11
n	7.0	2.2	2.9	1.5	3.2	0.8	9.3	4.8	2.0	2.9	5.6	6.1	4.7	1.8	1.1	0.8	8.2	1.8	1
l	12.2	3.8	5.1	2.7	5.6	1.4	16.2	8.3	3.5	5.1	9.8	10.6	8.2	3.2	1.9	1.5	14.3	3.1	2
k	26.0	8.1	10.8	5.7	11.9	2.9	34.6	17.8	7.5	11.0	20.9	22.7	17.6	6.8	4.0	3.1	30.6	6.6	5
w	12.1	3.7	5.0	2.6	5.5	1.4	16.1	8.2	3.5	5.1	9.7	10.5	8.2	3.1	1.9	1.4	14.2	3.1	2

Now take the ratio Observed/Expected (O/E)—see Frisch, Pierrehumbert, & Broe 2004. I removed cells where Expected < 5, and shaded cells where O/E ≤ 0.5:

C_2 C_1	p	b	f	v	m	θ	t	d	s	z	n	l	ɹ	tʃ	dʒ	ʃ	k	g	ŋ
p	0.0	0.0	0.3	0.0	0.3		1.2	1.2	2.0	1.2	1.8	1.3	1.3	0.2			1.3	1.0	1.7
m	0.0						1.5					2.3					1.8		
t	1.0	1.5	1.2	1.5	1.2	0.3	0.9	0.8	0.7	0.9	0.9	0.9	0.9	0.9	1.1	0.3	1.3	0.5	1.3
n	1.6						0.4				0.0	1.0					1.5		
l	1.6		0.6		1.6		1.2	1.2		0.6	0.7	0.0	0.0				0.6		
k	1.2	1.1	1.5	0.4	1.2		1.0	1.1	0.7	1.5	0.9	1.2	1.1	2.1			0.4	1.2	0.0
w	2.0		0.6		1.6		0.9	1.0		1.0	1.4	0.7	0.6				0.3		

[We could get deeper into this: if C_1 is [+nasal], is there less likely to be a nasal *preceding* C_2? Similarly for liquids. See Berkley for an in-depth exploration.]

Suppose English speakers have learned this pattern (see Coetzee 2010 for evidence that they have, at least for *s*CVC words; see Frisch & Zawaydeh 2001 for evidence that Arabic speakers know a similar but stronger pattern in Arabic).

- o How could the grammar express the pattern? What happens to an input like *spaff* (not a word)? What about exceptional *spam* (it is a word)?

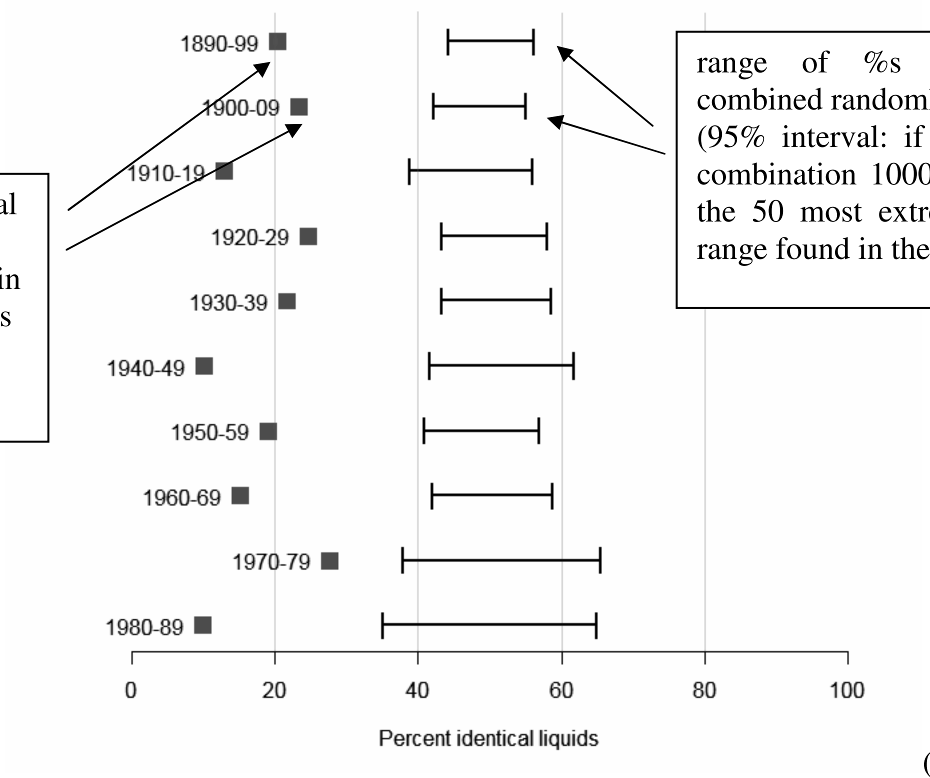
8 Lexical selection as an active shaper of the lexicon

In the English lexicon overall, if a word has two liquids, they're more likely to be *l...r* or *r...l* than *l...l* or *r...r*.
Martin 2007 shows... (pp. 76-77)

- In Old English, about 35% of words with two liquids have identical liquids, compared to ~55% expected by chance.
- In Middle English, it's about 25% (expect ~50%)
- Today, it's about 25% (expect ~50%)

Even though current English retains only ~10-15% of the Old English vocabulary!
English gained and lost many, many words, but always tended to respect the constraint against *l...l* or *r...r*.

Martin 2007 also uses the Oxford English Dictionary, which gives dates of earliest attested use for each word, to look at words newly entering the language. In every decade, new words avoid identical liquids:



% identical liquids observed in new words of each decade

range of %s expected if liquids combined randomly
(95% interval: if you simulate random combination 1000 times, and throw out the 50 most extreme trials, this is the range found in the remaining 950 trials)

(p. 78)

See Martin 2007 for an implemented model of lexical selection.

9 Filters in general

A lot of phonological and paraphonological activity has this character: it's not so much about mapping an input to an output as about deciding how good the resulting output is.

- Which new words enter the language, as we just saw
- Which names people choose for babies, fantasy role-playing characters, and pharmaceuticals (Martin 2007)
- Which first-name/last-name combinations people choose (Shih 2012)
- Which words make a good pun (Fleischhacker 2006)
- Which pairs of words make a good compound (Martin 2004; Martin 2007; Martin 2011)
- Which lines of poetry are legal (Hayes 2009)
- Which words can take which affixes (a big literature, but see Orgun & Sprouse 1999 in particular for the idea of a filter)
- Which words are chosen to combine into a blend (Ahn 2011)

See Martin 2007 for an implemented model of how different options compete. Options like...

- *couch* vs. *sofa* (totally synonymous, as far as I know)
- *carp pond* vs. *koi pond* (*carp* and *koi* are basically synonyms)
- *Mainer* vs. *Mainean* vs. *Maineite* (for 'person from the U.S. state of Maine')

pushing the idea further...

- writing a love song about the moon in June vs. writing one about the sun in August
- drawing a cartoon with the pun *Napoleon Blown-apart* in its caption (Napoleon Bonaparte holding a bomb) vs. drawing a cartoon about some other topic.

10 "Goodness" scores

Crucially, one factor in all the situations above is how phonologically "good" a competitor is.

=> Even if each tableau has just one winner (/kawtʃ/ → [kawtʃ], /sowfʌ/ → [sowfʌ]), the grammar must attach a goodness score to it (see Coetzee & Pater 2007 for a way to do it)

Going back to our English example: a hypothetical input like /spowm/ can surface faithfully, but with a poor score attached to it, so that it's unlikely to become popular as a new word:

/spowm/	MAX-C	IDENT(place)/__{V,#}	*s[labial]...[labial]	SCORE
☞ *a* spowm			*	lower score because of *s[lab]...[lab]
b spow	*!			
c stowm		*!		
compare				
/spown/	MAX-C	IDENT(place)/__{V,#}	*s[labial]...[labial]	
☞ *d* spown				higher score
e spow	*!			
f stown		*!		

Another goal in this course: learn ways to attach goodness scores to winning outputs.

11 Summary

We have seen an overview of variation and gotten an idea of what we want quantitative models of variation to be able to do

- Idealized free variation can be modeled as variable constraint ranking or optional rules
  - But, we need to develop our models of grammar so that they can quantify free variation, including the influence of various factors on a single phenomenon.
- Lexical variation (and mixed variation) is more challenging: how do we allow each word to surface faithfully but still let the grammar capture variation across words?
- Multi-site variation is also a challenge. How can we allow different sites to behave differently, for example?
- Lexical selection and other types of filters require the grammar to be able to assign goodness scores even to winning candidates.

Let's go over the "course outline" page in the syllabus to see how we will model these phenomena quantitatively in the rest of the course.

Next time: Variation in the sociolinguistic tradition—variable rules and logistic regression, a relatively simple and theory-neutral type of quantitative model.

References

Ahn, Suzy. 2011. Seoul National University, Master's thesis.

Becker, Michael. 2009. Phonological trends in the lexicon: the role of constraints.. University of Massachusetts Amherst ph.d. dissertation.

Berkley, Deborah. 2000. Gradient OCP effects. Northwestern University, PhD dissertation.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14. 150-177.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21. 43–58.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32. 45–86.

Ethnologue. 2005. Ethnologue: Languages of the World, Fifteenth edition. (Ed.) Raymond G. Gordon.

Hayes, Bruce, Bruce Tesar & Kie Zuraw. 2003. OTSoft 2.1. http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Kaisse, Ellen M. 1990. Toward a typology of postlexical rules.. In Sharon Inkelas & Draga Zec (eds.), *The Phonology-Syntax Connection*, 127-144. Center for the Study of Language and Information.

Labov, William. 1972. The reflection of social processes in linguistic structures. *Sociolinguistic Patterns*, 110-121. Philadelphia: University of Pennsylvania Press.

Mahanta, Shakuntala. 2009. Morpheme-specific exceptional processes and emergent unmarkedness in vowel harmony.. In Rajendra Singh (ed.), *Annual review of South Asian languages and linguistics: 2009*. Walter de Gruyter.

Pater, Joe. 2009. Morpheme-specific phonology: constraint indexation and inconsistency resolution.. In Steve Parker (ed.), *Phonological argumentation: essays on evidence and motivation*. (Advances in Optimality Theory). Equinox.

Schachter, Paul & Fe T Otanes. 1972. *Tagalog Reference Grammar*.. Berkeley, CA: University of California Press.

Zuraw, Kie. 2009. Frequency influences on rule application within and across words. *Proceedings of CLS (Chicago Linguistic Society) 43*.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28(2). 417-472.