The role of computational modeling in the study of sound structure

Bruce Hayes
Department of Linguistics
UCLA

Conference on Laboratory Phonology
Stuttgart
27 July, 2012

Part I: general comments

Why modeling?

- From the **experimental** point of view: a way of finding fully-explicit hypotheses to test
- From the **theoretical** point of view, particularly in phonology: establishing contact between abstract theories and experimental data

A fourfold program for modeling in phonology

I. Corpus compilation
II. Classical analysis
III. Design of learning algorithms
IV. Paired testing of algorithms and humans

I. Compilation of a realistic data corpus

- E.g. phonetic dictionary, speech database
- Try to match the dialect/vocabulary of future experimental participants.

II. Classical phonological analysis

- Goal is to avoid naïve inquiry by first finding and understanding the essential generalizations.
- Use the corpus to check them.

III. Develop a learning procedure

- This is implemented in software.
- It incorporates the theoretical assumptions in explicit form.
- It digests the corpus data.
- It creates (or, fine-tunes) a phonological **grammar**.

IV. Paired testing

- The grammar and experimental participants take **the same test**.
- In many experiments, the stimuli are **nonce forms**, so we can test how learners have **generalized** from the data they encountered.

How a computational model can be used to engage with phonological theory

- Fairly “deep” aspects of the theory can be embodied in the software code.
- We can also provide **different theories** in the same software.
- The software computes the **concrete consequences** of these changes for experimental data.

Modeling and experimentation can stimulate one another

- With new models we **reinterpret old data**; they can shed new light on (newly explicit) theory.
- Explicit models lead us to ponder what data would be needed to test them
- They can help us construct the most informative nonce forms for experiments — **model-guided stimulus design**. Examples:
  - Albright and Hayes (2001)
  - Hayes and White (in press)

Practical consequence: sharing is essential

- **Sharing of models by modelers to experimentalists** — in easy-to-use, downloadable or online versions.
- Sharing of **full subject data** by experimentalists to modelers —no data ever lose their usefulness.

An example of model-sharing worth emulating

- Vitevitch and Luce's **Phonotactic Probability Calculator** (2004)
- **On line**: (www.people.ku.edu/~mvitevit/PhonoProbHome.html)
- Extensively **used in experimentation** (133 cites on Google Scholar)
- **Easy to use**
- Embodies an **explicit phonotactic theory** — discussed here.

Part II: Case study

Topic: Gradient phonotactic well-formedness

- Examples with English nonce words:
  - [blɪk] (sounds great)
  - [pɔɪk] (sounds odd)
  - [lbøːpr] (sounds horrible)
- Similar terms for (roughly) the same thing: **wordlikeness**, **phonotactic probability**, **phonotactic grammaticality**

Two phonological theories of gradient phonotactic well-formedness

- The theory underlying Vitevitch and Luce's Phonotactic Probability Calculator (abbreviation: **VL**)
- A second theory, combining **traditional phonology + maxent grammars**.

Vitevitch and Luce's phonotactic theory

- Words are made of sequences of speech sounds that occupy **slots**, arranged left-to-right.

***brick*:**

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	etc.
b	r	ɪ	k			

- Phonotactic well-formedness is based on the **relative frequency with which slots are filled by different sounds in the vocabulary of the language as a whole**.

Sample calculation

- /b/ fills the first slot about 5.2% of the time.
- A (modest) adjustment is made for **token frequency** of the slot-sharing words
- Then you **add the values in all slots**; here .052 + .090 + .023 + .041 = **.206**.

Vitevitch and Luce, Version II

- Same approach, but instead divide the word into overlapping **bigrams**.

***brick*:**

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	etc.
b + r	r + ɪ	ɪ + k				

Vitevitch and Luce's claim: slot theory is "neutral"

- "The method of calculating phonotactic probability that we employed was relatively neutral with regard to linguistic theory." (2004, 485)

What would phonologists think of this claim?

- My impression is that the model would be considered **extremely controversial** — why?

Two principles phonologists typically subscribe to

- **Phonology doesn't count large numbers of things**

  "4th X", "5th X" doesn't seem to happen in phonology — small numbers are the limit.

- **Phonology doesn't count segments**.

  Of the things that get counted, segments don't seem to be included — instead, phonology counts hierarchical units like syllables, moras, feet.

  - For clear enunciations of both principles, see McCarthy and Prince (1986, 1).

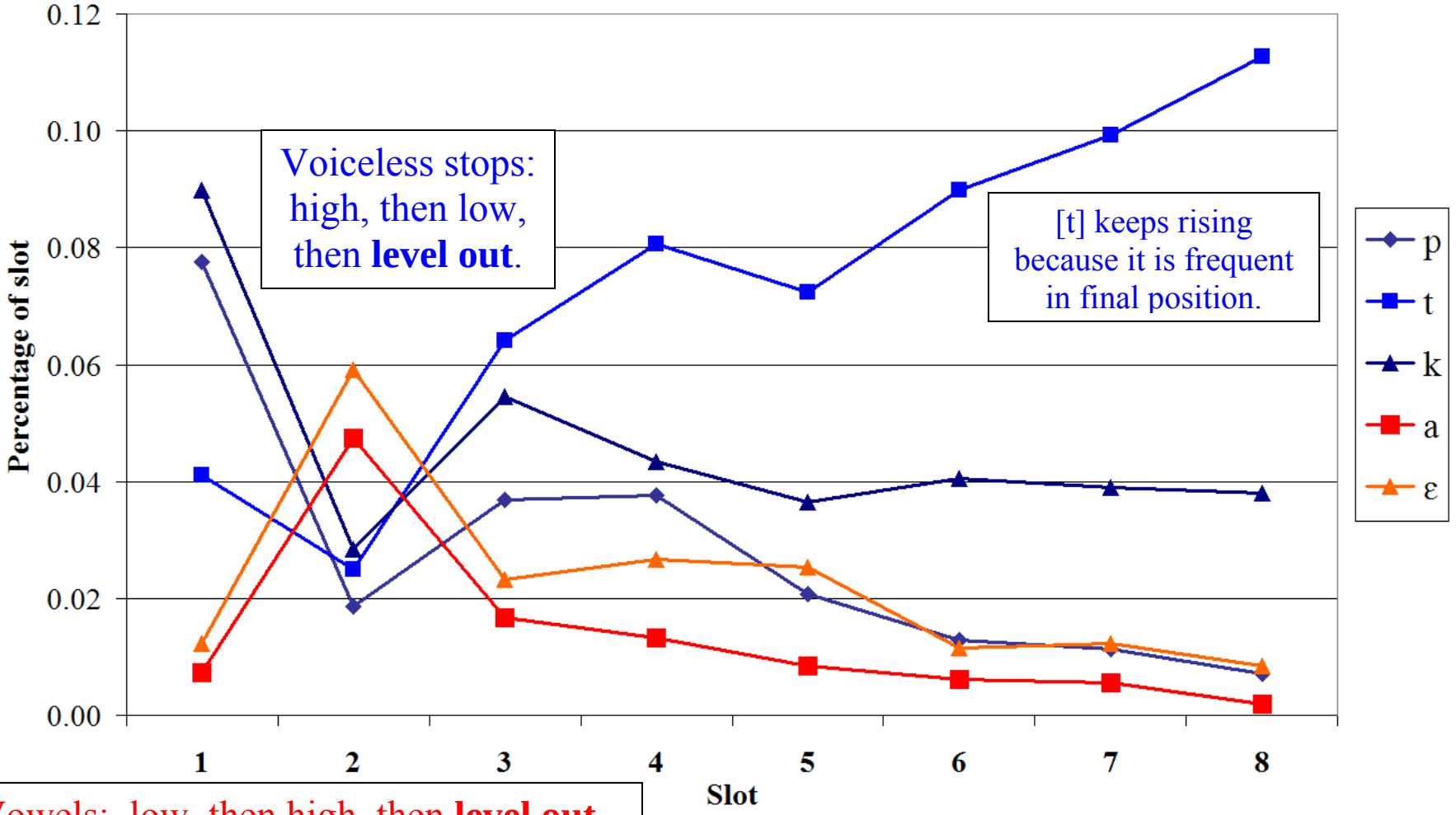
Diagnosing the behavior of the VL model

How do slot frequencies vary across a word? Tracking /p, t, k, ɑ, ɛ/ through their slots

- **Note**: these are based not on the VL lexical corpus but on an edited version of CMU dictionary (www.linguistics.ucla.edu/people/hayes/BLICK/)

[*chart next slide*]

Left-to-right slots in the VL model: Tracking five phonemes



Voiceless stops: high, then low, then **level out**.

[t] keeps rising because it is frequent in final position.

Vowels: low, then high, then **level out**.

“Leveling out” is likely to cause trouble

- As you “move rightward,” predictions of the VL theory reduce to **“prefer frequent phonemes”** (or bigrams) — **sequencing information** will be missed.

A random-word test

- I made 60,000 random 4-phoneme words, constrained to contain one stressed vowel.
  - Examples: [ɜˑɜˑˈo͡ʊʃ], [ˈdɪrd͡ʒ], [ˈɔəvθ], [ˈt͡ʃt͡ʃkɛ], [ˈθo͡ʊfw]
- I obtained VL scores for all of these words
  - CAVEAT: I used my own replication of VL, due to 500 word max on original version. Replication not yet validated.

Words that the VL model scores in the top 100

- Unigram version: [kɛsr], [səˈnɛ], [pætt], [ˈsʊəə], [ˈpnɝə]
- Bigram version: [præh], [ɔʃst], [prmɛ], [dɪˈsʊ], [ŋʊst]
- There is little question that native speakers would rate these words as **bad**.
- About **30%** of each "top 100" list has something comparably wrong with it.
- Full results:
  - www.linguistics.ucla.edu/people/hayes/BLICK/CompareModelsWithRandomWordTest.xls

The VL model evidently fails a plausible adequacy criterion

- “Don’t assign good scores to garbage.”

A phonotactic model based on phonological principles

My model

- I have named it **BLICK**.
- Available in the form of a downloadable phonotactic probability calculator:
  - www.linguistics.ucla.edu/people/hayes/BLICK

Basis of BLICK

- **Traditional phonology**, with ideas like
  - **Syllabification** and notions derived from it: **onsets, codas**
  - **Constraints** based on **natural classes** defined by a **feature system** (e.g., *[+sonorant][−sonorant] in onset position)
- The **maxent** theory of phonotactic grammars — Hayes and Wilson (2008)

A bit on the Hayes/Wilson (2008) theory

- Constraints are **weighted**.
- For any input word, it uses the weights and constraint violations to calculate a **penalty score**, related to probability.
- **Source of weights**: an algorithm finds the weights that **best fit the frequency distribution in a lexical corpus**, forming the most restrictive grammar.
  - I used my edited CMU corpus as the training data.

Where should we get our constraints from?

- Previous work has used the model's own **search heuristics** to select the constraints
  - Hayes and Wilson (2008), Daland et al. (2011), Hayes and White (in press)
- This leads to problems, no time to discuss here (see Hayes and White, in press)
- For now let us try a model in which constraints are **human-selected**, as a kind of baseline — can we do better than machine-selection?

Where I got my constraints from

- **Research literature** (Clements and Keyser 1983, Halle and Mohanan 1985, Hammond 1999, Harris 1994, McClelland and Vander Wyck 2006, Hayes 2011).
- **Trial and error:** Test draft grammars on synthetic lists (onsets, codas, whole words)
  - Keep adding plausible constraints until the system stops failing to penalize garbage.
- For **fine tuning** (small differences among basically well-formed words): unigram constraints.
  - One for each vowel
  - Two for each consonant (onset, coda position)

For the analyst, maxent is a very helpful critic

- Many constraints I tried were, to my surprise, **weighted at zero** — it turns out that their work is already done by overlapping constraints.
- Such constraints have no effect and so I discarded them.

The grammar used by BLICK

- … has 190 constraints
- … is posted in annotated form at
  - www.linguistics.ucla.edu/people/hayes/BLICK/BLICKGrammarMasterFile.xls

Giving BLICK the same test as before (60K random words)

- Words **chosen at** random from the top 100:
  - [ˈɑrɪt], [ˈælək], [ˈɑkʃə], [ˈvæpə], [ˈrɛpə], [ˈstɪn], [ˈkɛʃə], [ˈstɪm], [ˈtɪtə], [ˈɪvəm], [ˈspɛd]
- The **three worst-sounding words in the top 100** (my judgment):
  - [ˈplæl], [ˈfɪrt], [ˈhɪrk]
  - Full data at BLICK web page
- I believe that in general, BLICK meets the "don't fail to penalize garbage" criterion fairly well

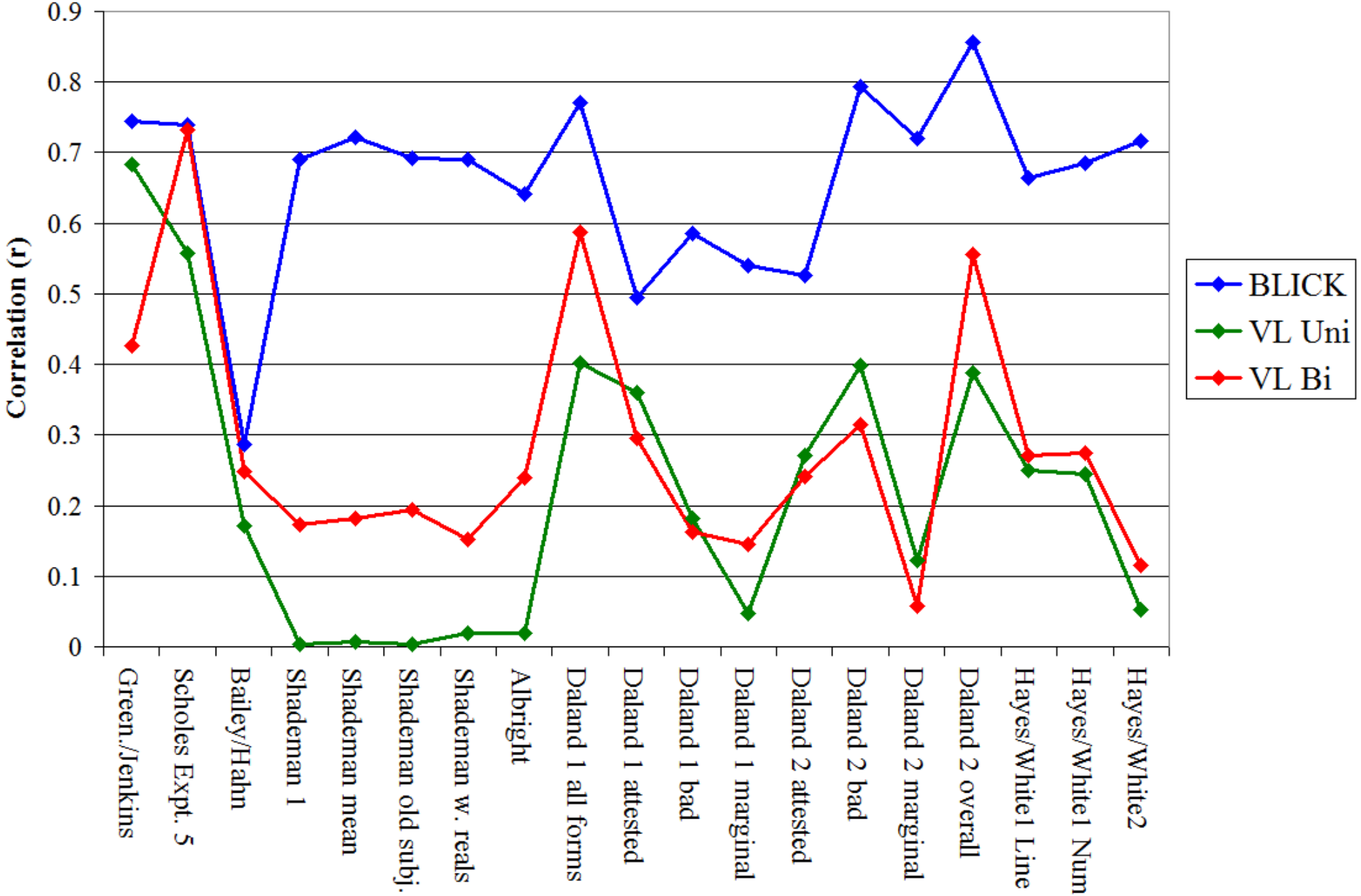
Benchmarking the two models with experimental data

Seven experimental studies

- Greenberg and Jenkins (1964), Scholes (1966), Albright (2009), Bailey and Hahn (2001), Shademan (2007), Daland et al. (2011), Hayes and White (in press)
- All were **nonce-probe rating studies**, with different types of stimuli selected for a variety of purposes.
  - Participant data were either published or shared with me by the authors.

BLICK vs. VL Model: correlation with participant ratings in seven studies

chart next slide


Correlation (r)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BLICK
VL Uni
VL Bi
Green./Jenkins
Scholes Expt. 5
Bailey/Hahn
Shademan 1
Shademan mean
Shademan old subj.
Shademan w. reals
Albright
Daland 1 all forms
Daland 1 attested
Daland 1 bad
Daland 1 marginal
Daland 2 attested
Daland 2 bad
Daland 2 marginal
Daland 2 overall
Hayes/White1 Line
Hayes/White1 Num
Hayes/White2

Result

- BLICK always has a higher correlation than VL; often much higher.

Where does the VL model do best?

- Greenberg and Jenkins (1964) and Scholes (1966): they used a **rigid CCVC template** for all stimuli.
- Here, the left-to-right slots align with syllable structure, so performance goes up.

Model comparison: summing up

- Both VL and BLICK are **statistical** models and assign their parameters based on a **dictionary corpus**.
- They are based on fundamentally different hypotheses about the **phonological mechanisms** involved:
  - VL: **left-to-right slot theory**
  - BLICK: **constraints** based on natural classes; syllabification
- Since left-to-right slots are patterned only at the **left edge** (slide 23), the VL model ultimately reduces in other regions to "use common unigrams/bigrams" — leading to predictably poor performance in most ratings studies.

Model comparison: summing up (cont.)

- Maxent also has problems (details on request) — but these are far less obvious in the data examined so far.
- The difference in performance of the models is directly traceable to the difference in **the phonological theories they assume**.

Moral of the story

- It is by **implementing the core ideas of the theories in explicit computational models** that the consequences of these ideas become clear and testable.

Thank you

Thanks to **Adam Albright**, **Todd Bailey** and **Shabnam Shademan** for sharing their experimental data with me, and to the **members of the UCLA Phonetics Laboratory** for help given at a rehearsal version of this talk.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26: 9–41.

Bailey, Todd M. and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44, 568–591.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno & Paul Smolensky. 2008. Language universals in human brains. Proceedings of the National Academy of Sciences 105: 5321–5325.

Berent, Iris, Donca Steriade, Tracy Lennertz & Vered Vaknin (2007). What we know about what we have never heard: evidence from perceptual illusions. Cognition 104: 591–630.

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

Chomsky, Noam and Morris Halle (1965) Some controversial issues in phonological theory. *Journal of Linguistics.*

Clements, George N., and Samuel Jay Keyser. 1983. *CV phonology: A generative theory of the syllable*. Cambridge, MA: MIT Press.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andreas Davis, and Ingrid Normann. Explaining sonority projection effects. *Phonology* 28: 197–234.

Greenberg, Joseph H. and James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20:157–177.

Halle, Morris and K. P. Mohanan. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16: 57-116.

Hammond, Michael. 1999. *The phonology of English: A prosodic optimality-theoretic approach*. Oxford: Oxford University Press.

Harris, John. 1994. *English sound structure*. Oxford: Blackwell.

Hayes, Bruce. 2011. Interpreting sonority-projection experiments: the role of phonotactic modeling. *Proceedings of the 2011 International Congress of Phonetic Sciences, Hong Kong*, pp. 835-838.

Hayes, Bruce and James White. In press. Phonological naturalness and phonotactic learning. To appear in *Linguistic Inquiry*, vol. 44, no. 1.

Hayes, Bruce and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379-440.

McCarthy, John and Alan Prince. 1996. Prosodic Morphology 1986. Ms. http://works.bepress.com/john_j_mccarthy/54/.
McClelland, James and Brent C. Vander Wyk. 2006. Graded constraints on English word forms. Ms., Stanford University. http://psych.stanford.edu/~jlm/papers/GCEWFs_2_18_06.pdf.
Scholes, Robert J. 1966. *Phonotactic Grammaticality*. The Hague: Mouton.
Shademan, Shabnam. 2007. Grammar and analogy in phonotactic well-formedness judgments. Ph.D. dissertation, Department of Linguistics, UCLA. www.linguistics.ucla.edu/general/dissertations/ShademanDissertationUCLA2007.pdf
Vitevitch, Michael and Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words

and nonwords in English. *Behavior research methods, instruments, & computers* 36: 481-487.

Extra slides

(in case these things come up; or just saving them as notes for the written version)

Why experimentalists need models

Modeling for experimentalists

- I wish to tread softly here because I am not one.
- But, as a consumer of experimental work, I am very often frustrated by the phrase

  "we would predict that …"

- Where can the reader appeal, who disagrees with such claims? I often am that reader.
- Models make precise predictions and put the experimentalist — beneficially —out on a limb.

Phonotactically bad things are usually bad for a reason

Properties English shares with many other languages

- Preference for well-sequenced sonority in onsets and codas
- Dispreference for long consonant clusters
- Special license for coronals
- Preference for nasals to be homorganic with what follows
- Dispreference for homorganic onset clusters (*[pw], *[tl])
- Dispreference for stressless heavy/superheavy syllables
- This is the leading idea behind the (often despised) universal constraint set of OT — I think there are idiosyncratic constraints too, but the strong resemblance

of phonologies cross-linguistically deserves to be explained.

Modeling vindicates particular phonological theories

Some examples from my experience—the Hayes/Wilson (2008) Phonotactic learner

- **Underspecification** → needed; else model can't cope with huge search space
- **Autosegmental tiers** → needed; else "nonlocal" processes like vowel harmony can't be learned
- **Metrical grids** → needed; same for nonlocal stress processes

Evaluating BLICK

BLICK is not always the closest-fit model for any experimental dataset

- The author's own favorite model outperforms it, somewhat, in Greenberg and Jenkins (1964), Albright (2009), and Bailey and Hahn (2001).

BLICK is versatile

- Daland et al. did a rating study of nonce words, which explores a broad continuum:

  I. a range of **awful** words

  ['rdɑsɪp], ['kɑsɪp] (Testing "sonority projection"; Berent et al. 2007)

  II. **marginal** words

  ['vlɑsɪp], ['ʃwɑsɪp]

  III. **pretty good** words

  ['plɑsɪp], ['frɑsɪp]

- BLICK achieves a better correlation than any of the models Daland et al. examined, both on these three populations, and overall.

Versatility is hard to achieve

- Hayes and Wilson's model does well on I, poorly on II, III
- Coleman and Pierrehumbert's model does well on I, II, poorly on III.

What are the main issues with maxent phonotactic grammars in general?

Naturalness and the validity of frequency-matching as the sole weighting criterion

- Hayes, Zuraw et al. (2009) and Hayes and White (in press) suggest that the constraints that have typological/phonetic support play a stronger role in phonotactic well-formedness than constraints that lack such support — even when the grammar gives them the same weight.
- If so, what is the "bias" mechanism that enforces this?

Recapitulating constraint-ranking effects

- Phonotactics has the occasional "except when" pattern that motivated the ranking principle of Optimality Theory.
  - [s] is the fricative that forms clusters, except when the following segment is [r]; then use [ʃ].
  - Obstruent + obstruent and obstruent + nasal onsets are impossible, except when the first obstruent is [s].
  - Sonorant + glide onsets are impossible, except when the glide begins the sequence [juː].
  - Final stress is disfavored, except in monosyllables.

- Maxent as employed for phonotactics has had to do complicated “work-arounds” for these problems.
- Might a system that chose between an output and the “null parse” let us get the benefits of ranking back again?

Granularity

- Treiman and Kessler's (20xx) work suggested that even particular VC rhymes may have to have constraints assigned to them.
- But at this level, the language may be too small to sample the data properly (see Pierrehumbert 20xx on "granularity".
- We need to be able to distinguish meaningful zeros in the data frequencies from nonmeaningful ones.

More criticism of the VL model

A relevant observation?

- The bad words classified as good by the VL model violate principles well-known to phonologists, e.g.

[ŋʊst]	[ŋ] never occurs in English syllable onsets.
[səˈnɛ], [prmɛ], [dɪˈsʊ]	English words never end in stressed lax vowels.
[ˈsʊəə]	Lax vowels never occur prevocalically in English.

An example of a phenomenon phonologists believe will never be found in a real language

- A special suffix allomorph that occurs after **four-segment stems**.

Suffix allomorphs in Fictionalese	
Stems that take -ti	*Stems that take -bu*
[tata-ti]	[trata-bu]
[arat-ti]	[parat-bu]
[trap-ti]	[tap-bu]
[atia-ti]	[patisa-bu]
[tarp-ti]	[tar-bu]

Another example of implausible segment-counting phonology

- a rule that lengthens the second segment of the word

/tata/ → [taːta]

/trap/ → [trːap]

/atia/ → [atːia]

/aita/ → [aiːta]

McCarthy and Prince (1986) on counting

- "Consider first the role of counting in grammar. **How long may a count run**? General considerations of locality, now the common currency in all areas of linguistic thought, suggest that the answer is probably '**up to two**': a rule may fix on one specified element and examine a structurally adjacent element and no other."

  - I personally think, "three", but I suspect no phonologist would want to go much higher.

"What elements may be counted? It is a commonplace of phonology that rules count moras (μ), syllables (σ)), or feet (F) but **never segments**."