Learning-Theoretic Linguistics: Some Examples from Phonology

Bruce Hayes

Department of Linguistics

UCLA

Relation of this talk to this session

- I'm not an expert on PDP at all! Inviting me here was quite a stretch.
- Common theme: *learning about language through computational modeling*.

Some PDP models

- Letter perception McClelland and Rumelhart (1981), Rumelhart and McClelland (1982)
- Word recognition and naming Seidenberg, and McClelland (1989)
- Inflectional morphology Rumelhart and McClelland (1986)
- Speech perception McClelland and Elman (1986)
- Phonetic categories Vallabha et al. (2007)

Linguistics in cognitive science

- We're externally famous for taking hard-line nativist views.
- As we view ourselves: obsession with language data and its patterning.

A bit of generative linguistics

- The logical problem of language acquisition is at the core of linguistic theory (see, e.g. Chomsky 1965, Ch. 1)
- Premise: language acquisition is an extraordinary feat
- Proposed explanation: *Universal Grammar* + learning mechanisms

Kinds of UG

- Hard UG: Grammatical principles in the genome
- Soft UG: Aspects of language shaped by human nature, e.g. phonetically-grounded phonology (e.g. Hayes, Kirchner and Steriade 2004)

The modeling program for linguistic theory

- **Learning model** – perhaps with UG in it
- **Data corpus** – approximate what children hear
- **Test the grammars** the model learns – let it act as an experimental subject.

Modeling with constraint-based grammars

- Permits direct access to key ideas of linguistic theory
- Easy diagnosis of the learned system

Embedding constraint-based grammars in a quantitative framework

- I currently use **maxent grammars** (Goldwater and Johnson 2003)
- Every constraint gets a weight reflecting its strength; simple math assigns probabilities to candidates, based on weights and violations.
- Weights are set during learning, using some of the same math as PDP (Smolensky and Legendre 2006)

Why deploy constraints in a quantitative model?

- The data demand it.
- Idealizing away gradience was a useful research strategy in the earlier days of generative linguistics; nowadays we can aim higher. (See e.g. Bod et al. 2003, Fanselow et al. 2006.)

What emerges when modeling is combined with experimentation?

I. The model might learn more about the language than the linguists have.

II. The model might provide a different interpretation of an experiment than the experimentalists proposed.

III. The model might work better when you put some UG in it.

I. The model might learn more about the language than the linguists have

- Example: Islands of reliability

Islands of reliability in English past tenses

- English has ca. 350 verbs ending in a *voiceless fricative* ([f, θ, s, ʃ]).
- *All* are regular.

A model that can learn islands of reliability

- Albright and Hayes (2002): a quantitative model for morphological and phonological learning, applied (2003) to English past tenses.
- Fed 4000 present-past pairs, it learned many constraints, including "Verbs ending in a voiceless fricative must take *-ed*."
- The model's quantitative metric says this is a *great* constraint: 100% accurate, 350 examples.

Experiment

- A classical "wug test" with rating data.
- The participants gave very high ratings to forms in islands, e.g. *blafed*, *wissed*, *teshed*.
- Statistics: a bigger effect than poor performance of competitors (*blofe*) would produce.
- Similar islands have been demonstrated for other languages (Albright 2002)

Consequences

- People **know more** than the linguistically-standard *just-one-rule-for-regulars* analysis (e.g., Pinker 1999).
- In general: I think traditional linguistics often underestimates what people know about their language.

What emerges when modeling is combined with experimentation?

I. The model might learn more about the language than the linguists do.

II. The model might provide a different interpretation of an experiment than the experimentalists proposed.

III. The model might work better when you put some UG in it.

The Sonority Hierarchy Projection Effect

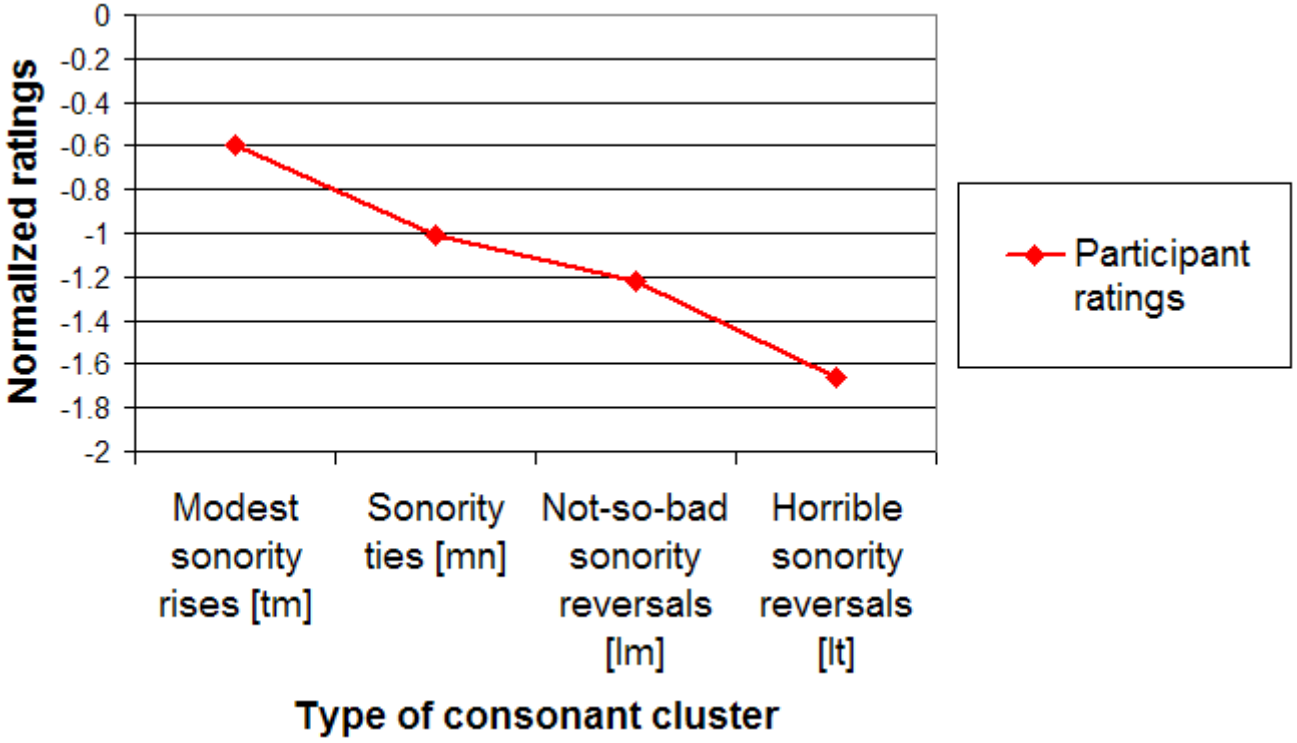
- This is a possible "UG effect" – people know what they couldn't learn. (Berent et al. 2007, Berent et al. 2008)

Sonority sequencing in consonant clusters

- [tra] a good sonority rise, quite common among languages
- [tma] not all that much sonority rise, rare
- [mna] sonority tie, rarer still
- [lma] modest “reversed sonority”; rarer still
- [lba] a horrible sonority reversal, very rare

Ratings of zero-frequency clusters match language typology

- from Daland et al. (in progress):



…and not just in our data

- Other ratings studies Albright (2007)
- Perception/production data (Berent et al. 2007, Berent et al. 2008)

Arguing for UG based on the Sonority Projection Effect

- If the sonority sequencing principle were not in UG, all clusters of zero frequency would be treated equally.

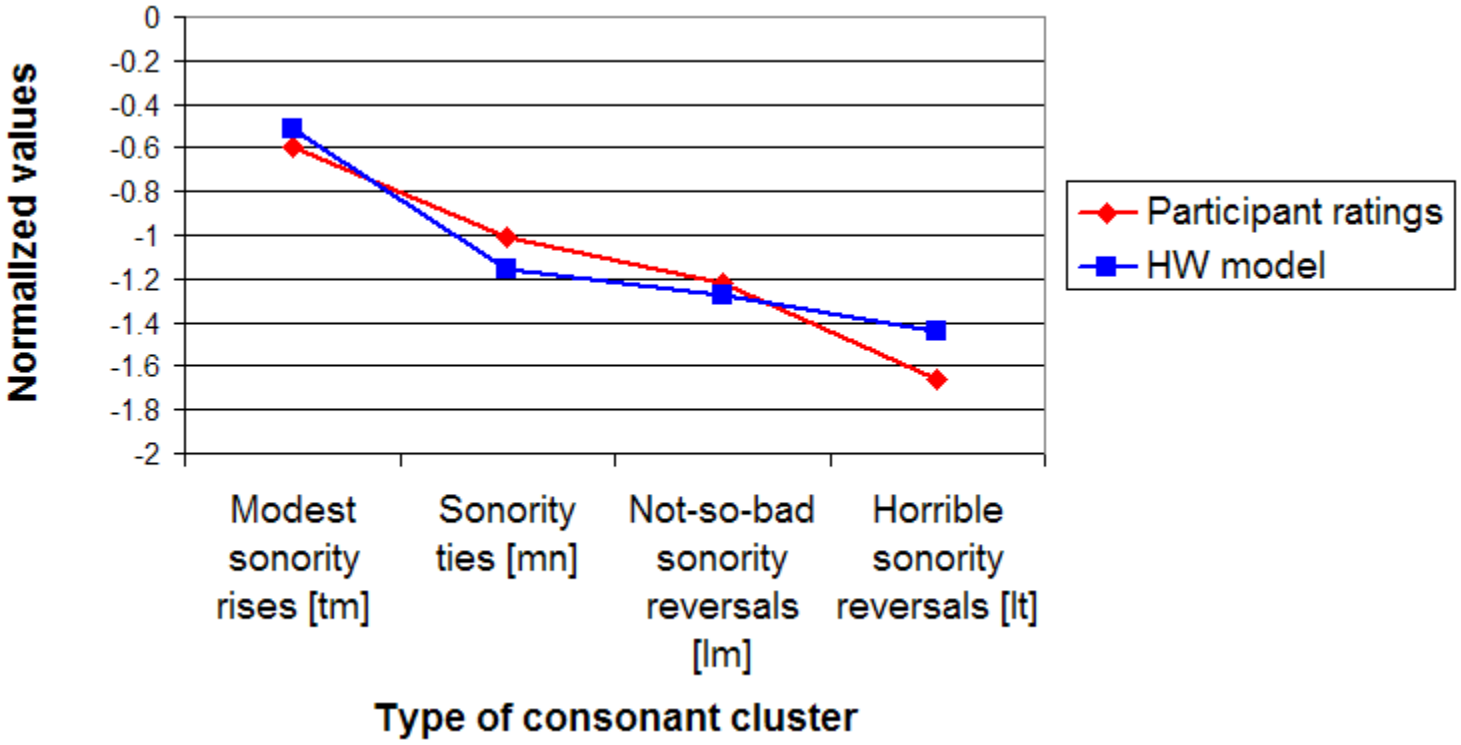
A computational learner for phonotactics

- Hayes and Wilson (2008)
- Finds constraints inductively, learning from a phonetically-transcribed lexicon.
- Weights the constraints by maxent principles.
- Resulting grammar assigns a well-formedness value to any novel phonetic string.

Using the model to study the sonority hierarchy projection effect

- Daland et al. trained this model on 18,000 transcribed English words.
- They tested it on the same forms given to the experimental participants.

Hayes-Wilson model predicts sonority projection



How does it do it?

- It knows the *features* that represent sonority, and generalizes from the real-word clusters ([tr], etc.) it hears.

	[sonorant]	[approximant]	[vocalic]
[t]	–	–	–
[m, n]	+	–	–
[l]	+	+	–
[r]	+	+	+

Upshot

- The UG needed to learn the Sonority Projection Effect is perhaps not as substantial as an innate sonority sequencing principle.
- Knowing *features* let you project the pattern from the sonority-respecting clusters you already have.

What emerges when modeling is combined with experimentation?

I. The model might learn more about the language than the linguists do.

II. The model might provide a different interpretation of an experiment than the experimentalists proposed.

III. The model might work better when you put some UG in it.

Effects of natural vs. unnatural constraints

- Source: Hayes and White (in progress)

Naturalness in phonological constraints

- Phonologists observe the same constraints in phonologies the world over.
- A widespread view: such constraints are part of the language faculty, "in UG." (Prince and Smolensky 1993)
- Soft-UG variant: the relevant constraints are those that help make speech easier to articulate and perceive.

Goal

- Test some very surprising constraints learned by the Hayes/Wilson model
- They look unnatural (no phonetic or typological support).
- They fit the data of English very well.

Sample stimulus pairs

- Test form violates a natural constraint: *canifl* vs. *canift*.
- Test form violates a constraint induced by HW model and lacking in phonetic/ typological support: *foushert* vs. *fousert*. (Constraint: no diphthongs before palato-alveolar fricatives).

Method

- Magnitude estimation (Bard et al. 1996)
- Simultaneous spoken and orthographic input

Main result

- Massive effects for natural constraints (*canifl* rated horrible, *canift* fine)
- Small nonsignificant effect for unnatural constraints (*foushert* and *fousert* rated similarly)

What does it mean?

- We have sought extensively for an inductive explanation: how could the model be modified to devalue the unnatural constraints?
- This doesn't seem to be working…
- Letting the model select from constraints based on Steriade's (1999) phonetics-based UG seems to be working rather well.

Upshot

- We suggest that the ratings reflect not just the patterns of the existing lexicon, but also UG principles based on phonetic naturalness.

In conclusion

- Three practical suggestions

Link experiments to modeling

- Models can locate the stimuli that best test a hypothesis (all three experiments described here)
- Models avoid vagueness in interpreting the results.

Posting data

- Modelers would love to test their models against published experiments – but the data are seldom available.
- The web would make it easy to fix this.

Posting software

- When first programmed, learning models tend to be messy, accident-prone, and user-unfriendly.
- It is toil to turn them into software others can use.
- But doing this would make the models more widely useful and testable.

Thank you

- For a downloadable version of these slides included the cited references, please visit http://www.linguistics.ucla.edu/people/hayes/.

References

My 15-minute talk did little justice to the work of many other linguists working on computational phonological modeling and "UG experiments". A partial list of those unmentioned would include John Alderete, Paul Boersma, Eugene Buckley, Andries Coetzee, Emanuel Dupoux, Paola Escudero, Sara Finley, Stefan Frisch , Karen Jesney, Andrew Martin, Elliott Moreton, Joe Pater, Sharon Peperkamp, Janet Pierrehumbert, Anne Pycha, Amanda Seidl, Anne-Michelle Tessier, and Jie Zhang. All maintain active research web sites and are easy to find.

Albright, Adam (2002) Islands of reliability for regular morphology: evidence from Italian. *Language* 78: 684-709.

Albright, Adam (2007) Natural classes are not enough: Biased generalization in novel onset clusters. *15th Manchester Phonology Meeting*, Manchester UK, May 24–26.

Albright, Adam and Bruce Hayes (2002) Modeling English past tense intuitions with minimal generalization. In Mike Maxwell, ed., *Proceedings of the 2002 Workshop on Morphological Learning, Association of Computational Linguistics*. Philadelphia: Association for Computational Linguistics.

Albright, Adam and Bruce Hayes (2003) Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90: 119-161.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Berent, I., Lennertz, T., Jun, J., Moreno, M., A., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences* 105: 5321-5325.

Berent, I., Steriade, D., Lennertz, T & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104: 591-630.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy (2003) *Probabilistic Linguistics*. Cambridge: MIT Press.

Chomsky, Noam (1965) *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Daland, Robert, Mark Garellek, Ingrid Normann, James White, and Bruce Hayes (in progress) Predicting the Sonority Projection Effect: A Comparison of Models. Ms., Department of Linguistics, UCLA.

Fanselow, Gisbert (2006) *Gradience in Grammar: Generative Perspectives*. Oxford: Oxford University Press.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

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

Hayes, Bruce, and James White (in preparation) Accidentally-true constraints in phonotactic learning. Ms., Department of Linguistics, UCLA.

Hayes, Bruce, Robert Kirchner, and Donca Steriade, eds. 2004. *Phonetically-based phonology*. Cambridge: Cambridge University Press.

McClelland, J. L. and Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, 18, 1-86.

McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings. *Psychological Review*, 88, 375-407.

Pinker, Steven (1999) *Words and Rules: The Ingredients of Language.* Basic Books.

Prince, Alan and Paul Smolensky. 1993. Optimality Theory. Ms., Rutgers University. Published 2004: Oxford, Blackwell.

Rumelhart, D. E., & McClelland, J. L. On learning the past tenses of English verbs. In Rumelhart, D. E., McClelland, J. L., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I*. Cambridge, MA: MIT Press.

Rumelhart, D. E., and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The context enhancement effect and some tests and extensions of the model. *Psychological Review* 89, 60-94.

Seidenberg, M. S. and McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, 96, 523-568.

Smolensky, Paul and Geraldine Legendre (2006) *The Harmonic Mind* (in two volumes). Cambridge: MIT Press.

Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. In *Proceedings of the 1998 Linguistics and Phonetics Conference*, ed. Osamu Fujimura, Brian Joseph, and B. Palek, 205–245. Prague: The Karolinum Press.

Vallabha, G, McClelland, J, Pons, F, Werker, J, and Amano, S. Unsupervised learning of vowel categories form infant-directed speech. *PNAS*, 104, 13272-13278. 2007.