The time course of generalization in phonotactic learning

Tal Linzen and Gillian Gallagher
Department of Linguistics, New York University
{linzen,ggillian}@nyu.edu

Phonotactics

- Speakers of a language have implicit knowledge of how sounds can combine to form words in their language (its **phonotactics**)
- Some of these regularities are at the level of **individual sounds**: e.g., English words can't start with *ng*
- Other regularities hold of **classes of sounds**
- **English onsets:** no English words start with either *sr* or *mb*, but *srip* sounds better than *mbip* (English words can't start with a nasal + stop cluster, but can start with *s* + liquid, e.g. *slip*)
- **Hebrew roots** can't start with (any) two identical segments (**ssm*)

Models of generalization

- **Bottom-up learning:** once individual sounds have been learned, the learner notices commonalities and forms generalizations
- **Simultaneous learning:** the acquisition of broad regularities does not depend on the prior acquisition of narrow ones
- What determines the level of granularity of the phonotactic regularities that are acquired?
- We show that bottom-up models, e.g. **STaGe** [1] and **MGL** [2], are incompatible with our results
- We evaluate two simultaneous models: **MaxEnt** [3,4] and a **rational clustering** model [5] applied to this problem for the first time

Experimental methods

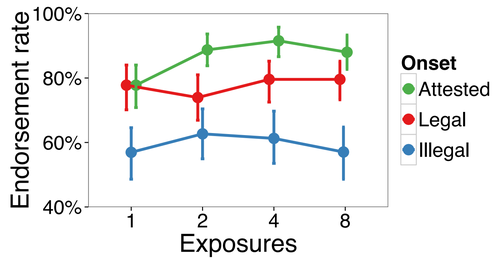
- Adult participants listened to words in an artificial language (**training phase**)
- They then judged whether novel words sounded like they could be part of the same language (**test phase**)
- None of the test words had been heard in training, but some conformed to narrower or broader regularities in the training set
- Participants were recruited on Amazon Mechanical Turk
- Amount of exposure was varied across participants: some participants saw one word of each type, others two words, etc.

Behavioral experiments

Exp 1: Phonological feature (categorical)

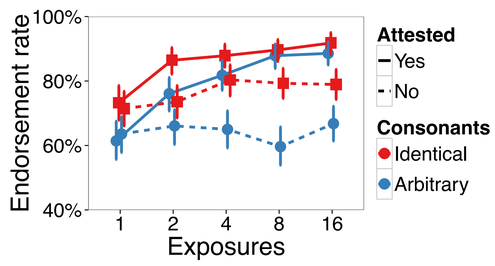
Training: ***f**ula, **θ**omi, **p**inu, **k**elin, **t**anu*

Test:	Attested	Legal	Illegal
	***f**emi*	***s**unu*	***z**oma*



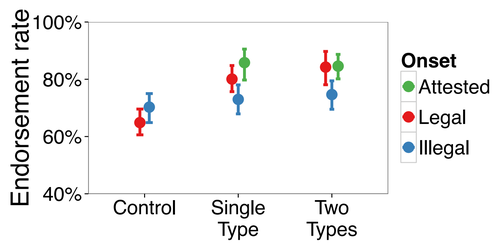
Exp 2: Segment identity (probabilistic)

	Identical		Arbitrary	
Training:	*pipa*		*sunu*	
Test:	Attested	Unattested	Attested	Unattested
	pepu	*gagi*	*senu*	*pagi*



Exp 3: Generalization from a single type

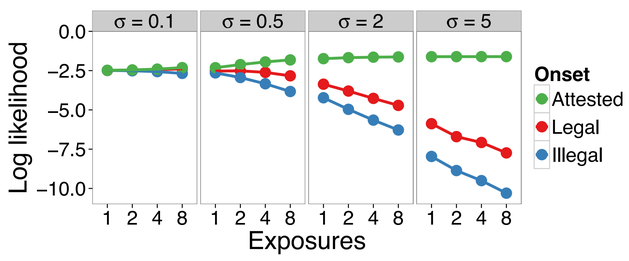
Training:	Control	Single	Two
	(none)	***k**ima, **k**upa*	***k**ima, **p**upa*
	(+ 6 filler items starting with ***l**, **y** or **w***)		
Test:	Attested	Legal	Illegal
	***k**una*	***t**ima*	***z**una*



Computational simulations

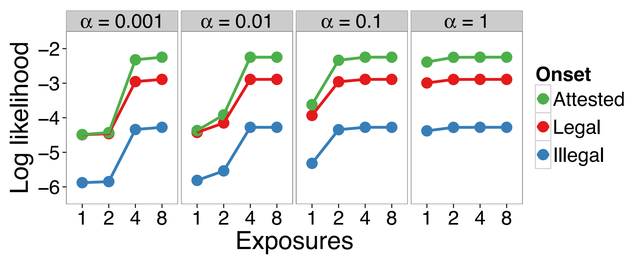
Maximum Entropy (log-linear model)

- Each onset type has a value of 0 or 1 for features such as "is *p*" or "is voiced"
- $y_i \sim Poisson(\lambda_i)$
- $\log(\lambda_i) = \sum_{j=1}^{n} x_{i,j}\beta_j$
- $\beta_j \sim N(0,\sigma^2)$



Nonparametric Bayesian clustering

- Words generated from a mixture of clusters, each associated with rules such as "all onsets are voiced", "*p* is the only allowed onset"
- Specificity bias (Size Principle): $\mathrm{P}(r_c) = \frac{1}{|r_c|}$
- Simplicity bias (Chinese Restaurant Process)
- Likelihood: $\mathrm{P}(\mathbf{T} \mid \mathbf{Z}) = \prod_{C \in \mathbf{Z}} \sum_{r_C} \prod_{t_i \in C} \mathrm{P}(t_i \mid r_c)\mathrm{P}(r_c)$



Conclusions

- Broad regularities were learned before narrow ones, casting doubt on bottom-up models
- Simultaneous models do not necessarily produce this pattern either (e.g. MaxEnt doesn't)
- The granularity of generalization can be captured by a Bayesian simplicity bias

References

[1] Adriaans, F., & Kager, R. (2010). Journal of Memory and Language.

[2] Albright, A. (2009). Phonology.

[3] Goldwater, S., & Johnson, M. (2003). Workshop on variation within Optimality Theory.

[4] Hayes, B., & Wilson, C. (2008). Linguistic Inquiry.

[5] Frank, M. C., & Tenenbaum, J. B. (2011). Cognition.