“What we can and cannot model about child speech”

Towards more psycholinguistic plausibility in modeling phonological development

Lise Menn, University of Colorado

in conversation with

Paul Boersma, University of Amsterdam

Progress in modeling **cognitive development, including language** development, is coming from progress in several areas:

- **larger corpora** of data to work with
- positing **interactive information processing**
- simulating more **neurally plausible** mechanisms

Progress in phonology includes breaking down the *mystiscism* of 'markedness-as-a-primitive,' thanks to:

- **phonetically-grounded**/ lab phonology (Hayes, Flemming, Pierrehumbert, Beckman…)
- larger **language data bases** (cf. recent revisions of markedness by Keren Rice)

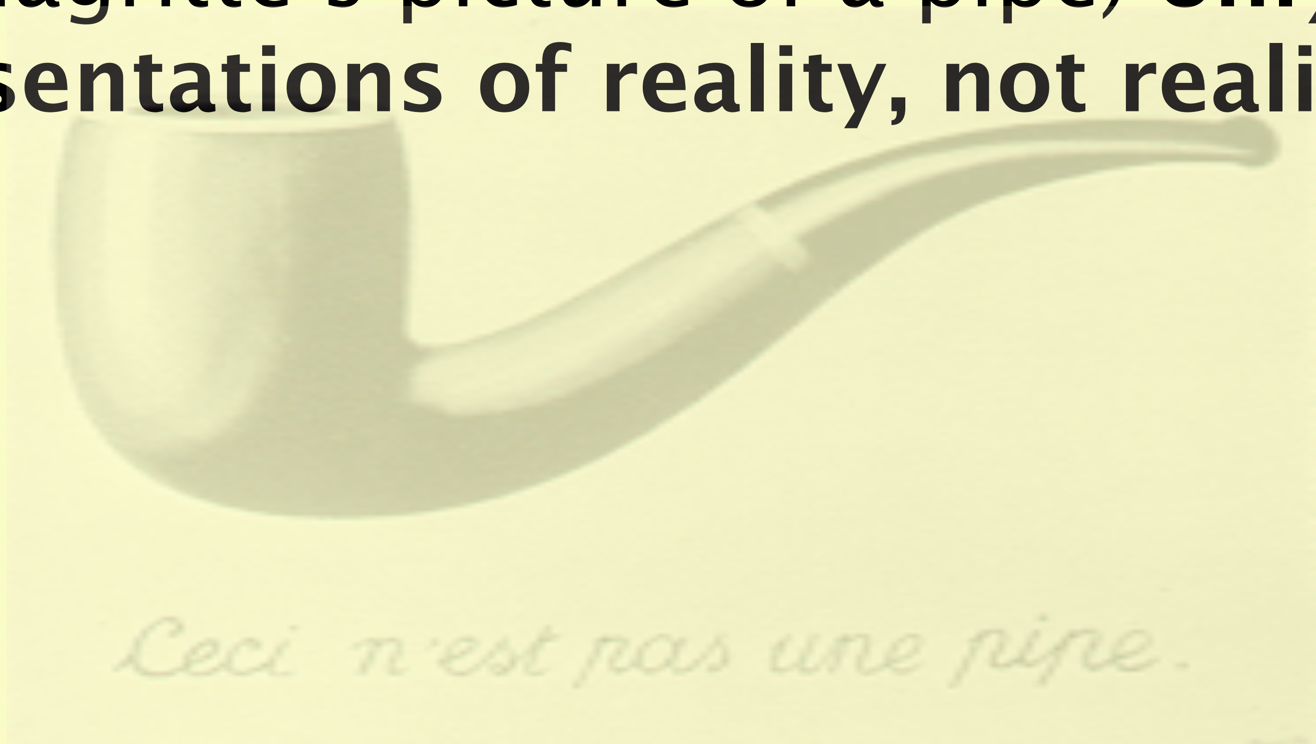
Our group's goal for improving theories of phonological development ("From Mysticism to Mechanism") is to **bring together** this progress in developmental modeling and in phonology, but…

…it might be the case that **not all insights can be incorporated into a single model** (think of Newtonian vs. quantum physics, gas laws vs. statistical mechanics)

– although one hopes that ‘macro’ levels will emerge from ‘micro’ level models.

If not all insights can be incorporated into a single model,

…it's not terrible, as long as we **keep from believing** too passionately **in our theoretical entities**, which are (like Magritte's picture of a pipe) **only representations of reality, not reality itself.**



Longitudinal acquisition data give **three problems** for conventional (maturational/ feature-based, e.g. Jakobson, standard OT) approaches to phonological development:

1. **Differences between children** may be large, but they **vary within a probability envelope**: some sounds/phonotactic patterns are much commoner than others.
2. **Lexical exceptions** (phonological idioms) are always present**; frequency** promotes the **entrenchment** of a form as an idiom**.**
3. **U-shaped curves** are a **developmental norm**

Because they are pervasive, a proper theory should not just **accommodate** these three facts, but **predict** them.

1. **Differences between children** may be large, but they **vary within a probability envelope**: some sounds/phonotactic patterns are much commoner than others.
2. **Lexical exceptions** (phonological idioms) are always present; **frequency** promotes the **entrenchment** of a form as an idiom.
3. **U-shaped curves** are a **developmental norm**

The Boulder model directly addresses these three pervasive phenomena.

Example of differences between children, within language (this one is neatly handled even by early Optimality Theory):

Two ways to make s+stop clusters into one consonant:

-preserve the **stop** (manner & place) (**Daniel**)

-or the **frication,** approximating place (**Stephen**)

	Daniel	Stephen
spill	[pɪl] ‘pill’	[fɪl] ‘fill’
store	[tɔr] ‘tore’	[sɔr] ‘sore’
school	[kul] ‘cool’	[sul] ‘sool’

Example of a difference *within* a child, neatly handled by *Stochastic* Optimality Theory:

A child with an output template (Priestley 1977):
'monster' → [majos], [mejan]

- underlying form: |mVnsV|
- top-ranked: the template /CVjVC/
- top-ranked: preserve the first consonant (C1)
- top-ranked: preserve consonant order
- lower-ranked, equally high: preserve C2, C3

Resulting output: 50% /mVjVn/, 50% /mVjVs/

The Boulder child phonology group proposes the **Linked-Attractor Model,** intended as the precursor of a computationally implementable model for early phonological development.

Central features of the Linked-Attractor Model:

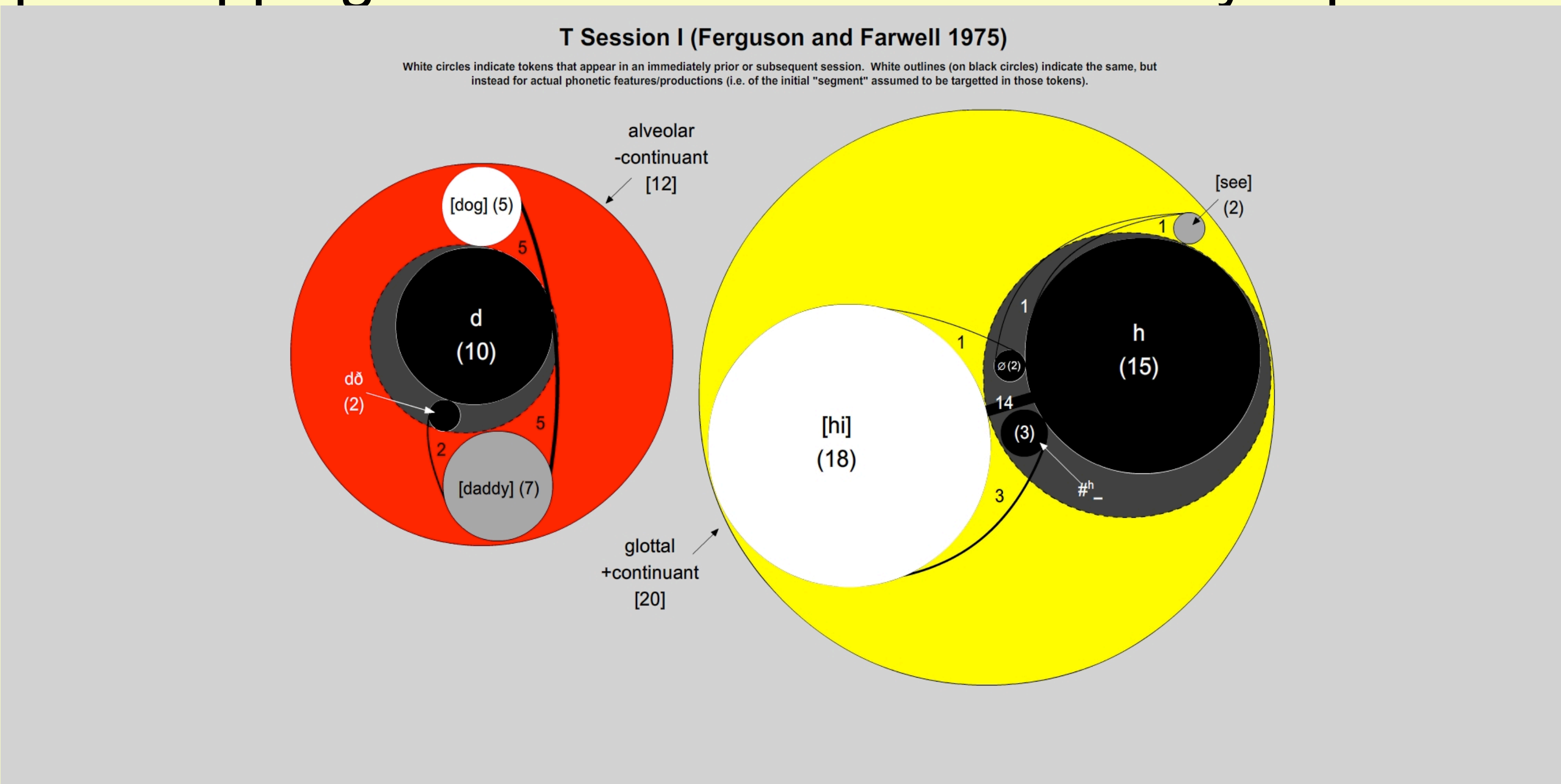
1. **discovery** of articulatory-auditory connections, beginning during babble
2. **instance-based** (lexical, word-by-word) **learning** preceding abstraction of patterns
3. At least four partially **redundant** stored entities: **auditory input, articulatory output,** plus **mappings** between them

Let's look at some data and how a diagram of the Linked-Attractor Model represents them.

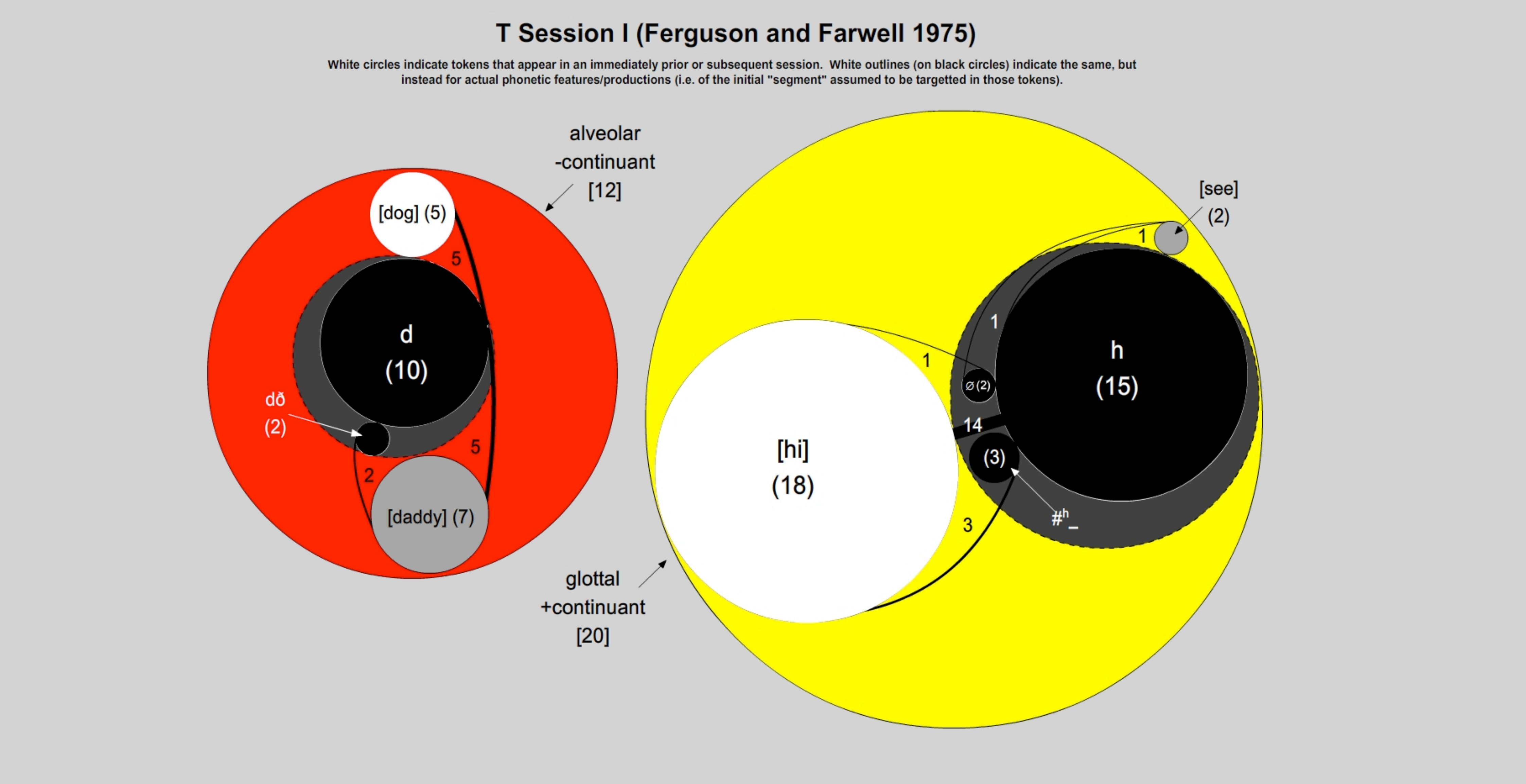
Here are complete data from one child, "T.", at one recording session (Ferguson & Farwell 1973). Look at the numbers and at the variation in the first consonant of each word.

Target/Type	Attempts	Token1	Token2	Token3	Token4	Token5
daddy	7	daeji	daeɪ (2)	daedɪ (2)	dðae	dðaei
dog	5	da (5)				
hi	20	hai (2)	hai (12)	^{h}ai (3)	ai (3)	
see	2	hɪ^{i}	ɪ^{i}			

Here’s a 2-D picture (Nicholas diagram) for how “T.” produced her four words *daddy, dog, see, hi.* Black, white, and gray circles represent the input and output representations that “T.” has formed for each word; the lines represent mappings from input to output. Colored circles represent discrete output areas of articulatory space. Larger circles represent more frequent words/sounds; thicker lines represent more frequent input-output mappings. Constraints are not directly represented.



A Nicholas diagram like this is a very simple representation of how the Linked-Attractor Model conceptualizes “T.”’s knowledge of her four words; we’ll come back to it after we’ve described the model better.



Articulatory-auditory discovery, a component of Boersma's Programme paper (2008), is essential for understanding variation across children.

- In the Linked-Attractor Model, an initial 'landscape' models the **physical biases** of the articulatory system that **shape the probability** that a particular sound or sequence of sounds will be acquired.
- Discovery of **how to make a sound or sound sequence** is a **trial-and-error procedure**.
- The **likelihood of success** is given by the shape of the landscape, without ruling out the possibility that some **rare sounds might be discovered early** by a particular child who happens to figure out how to make them.

So ‘**warping the landscape’** for discovering how to make particular sounds/sound sequences) is the metaphor that our model uses to deal with the **first** of our **three problems** for conventional approaches, again:

The **differences between children** are sometimes large, but they **vary within a probability envelope**: certain sounds and phonotactic patterns are much commoner than others.

But what do we **mean** by ‘warping the landscape’?

Attractors are the basic metaphor of the Linked-Attractor Model, which was visualized by Brent Nicholas of our group, based on many existing precursors.

Metaphor/model: Things you are **likely to do** (to 'fall into') are 'down' in a gravity well (attractor); so, things you tend to do can be considered **attractors**, in the sense of chaos theory.

Developing a skill is **increasing the attraction** of an attractor, i.e., of a place or a series of places in a space of appropriate dimensions.

To model articulatory-auditory space, we need maybe 20+ dimensions, plus a time axis for modeling sequential action.

- Increasing the attraction of an attractor (e.g. of a perceived or produced form) is like **digging a hole** or a ditch in a 2-D surface, or like **increasing** (incrementing) **the mass of an attractor** in a higher-dimensional space.
- Digging holes/increasing attractor mass thus **warps the landscape** – it changes where sensory stimuli or motor impulses are likely to ‘roll’ to.
- For example, **categorical perception** is having a stimulus ‘roll into’ one or the other of two neighboring holes; cf. Jusczyk’s WRAPSA ‘warping’ perceptual acquisition model and Kuhl’s ‘perceptual magnet effect’.

As in Jusczyk's and Kuhl's models, **experience digs the holes.** **Each instance** of items and of sequences entrenches them further, so that more frequent events become stronger attractors, other things being equal.

In a neural model, the **connections** linking sensation to memory of percept, and linking percepts to one another, are **strengthened each time they are activated** (cf. Hebb 1949: "what fires together, wires together").

Again: **experience digs the holes.**
Each instance of items and of sequences entrenches them further, so that more frequent events become stronger attractors, other things being equal.

The **Linked-Attractor Model builds up the strengths of representation and mapping incrementally** (as in Vihman & Croft's Templatic production model). This makes it immediately **tractable for neurally-responsible computational modeling.**

Articulatory-auditory discovery is important in understanding **variation across languages** as well as across children.

Pre-speech **auditory learning** of the sounds and phonotactic patterns of the ambient language sculpts the purely perceptual part of the initial 'landscape'. This auditory learning helps to determine which auditory patterns the child will **attempt** to match/be happy to have matched, and therefore be likely to discover and master.

On to the **2nd problem for conventional theories: Lexical exceptions** (phonological idioms) are always present; **frequency** promotes the **entrenchment** of a form as a phonological idiom.

Warping **by incremental experience** deals with this automatically, because it implies **instance-based learning**: each lexical item is learned separately, although it is affected by the existing attractors. We'll look at an example in a moment.

Now, of course, instead of lexical exceptions being a problem, the **formation of generalizations** becomes a problem.

Fine-grained data (Jacob, 3 days/week from 12:8 to 22+) indicate that the formation of generalizations is in fact a convoluted process whose complexity has been overlooked.

Paucity of sufficiently fine-grained data is a real problem for good theory-building. Have a look at this 2008 *Psych Review* paper when you get a chance:

“What Is the Shape of Developmental Change?”
Karen E. Adolph, Scott R. Robinson, Jesse W. Young, Felix Gill-Alvarez

Developmental trajectories provide the empirical foundation for theories about change processes during development. However, the ability to distinguish among alternative trajectories depends on how frequently observations are sampled.

Continuing the Adolph et al. abstract:

… Data were derived from a set of 32 infant motor skills indexed daily during the first 18 months. Larger sampling intervals (2–31 days) were simulated by systematicallyremoving observations from the daily data and interpolating over the gaps. Infrequent sampling caused decreasing sensitivity to fluctuations in the daily data: Variable trajectories erroneously appeared as step functions... Sensitivity to variation decreased as an inverse power function of sampling interval, resulting in severe degradation of the trajectory with intervals longer than 7 days… Inadequate sampling regimes therefore may seriously compromise theories of development.

Psychological Review 2008, Vol. 115, No. 3, 527–543

For "Ellie", a child with SLI, formation of generalizations is a more serious problem

Diary data from 1;6 (first word) to present (4;3.9)
Phonological development is atypical, as well as some morphology and syntax
Normal cognition and normal hearing
Receives weekly speech therapy

move	[muʃ]	have	[hæs]	Steve	[div]
this	[diʃ]	yes	[jɛ_]	juice	[dus]
please	[piʃ]	ladybugs	[ibʌts]	squeeze	[giz]
out	[aʊʃ]	paint	[peɪn_]	hat	[æt]
black	[bæʃ]	milk	[mot]	yuck	[gʌk]

General rules don't work, constraints can't be ordered

In early production-based two-level Optimality Theory, constraints indeed cannot be ordered

For instance, there is no way of turning underlying |pliz| into surface [piʃ], and at the same time underlying |skwiz| into surface [giz].

The first question to ask about such data, however, is: is the variation perhaps tokenwise, i.e., could the child also say [piz] and [giʃ]? If so, *stochastic* OT would handle the case.

In early production-based two-level Optimality Theory, constraints indeed cannot be ordered

Now, assume that the child *always* says [piʃ] for *please*, and [giz] for *squeeze*, i.e., the variation that was observed is not tokenwise but lexical.

Then such cases are problematic for any single-lexicon theory that assumes adultlike underlying forms, i.e., that the production mappings are |pliz| → [piʃ] and |skwiz| → [giz].

But in e.g. bidirectional three-level OT, there are at least four sources of observed variation

1. Articulatory problems in phonetic implementation:
   [pliz] → /.pliz./ → |pliz| → /.pliz./ → [piʃ]
   (articulatory constraints >> cue constraints)
2. Problems representing the structure in production:
   [pliz] → /.pliz./ → |pliz| → /.piʃ./ → [piʃ]
   (structural constraints >> faithfulness constraints)
3. Problems storing the structure in the lexicon:
   [pliz] → /.pliz./ → |piʃ| → /.piʃ./ → [piʃ]
   (lexical structural constraints >> faithfulness)
4. Perceptual problems:
   [pliz] → /.piʃ./ → |piʃ| → /.piʃ./ → [piʃ]
   (struct >> cue, or no category, or variation)

If the child initially perceived [pliz] randomly as /.piʃ./, and [skwiz] as /.giz./, then she has likely stored them as |piʃ| and |giz|, and will therefore produce them as [piʃ] and [giz].

Just as in loanword adaptation,

e.g. Korean |t^{h}aip| < *type* vs. |p^{h}aip^{h}ɨ| < *pipe*.

Fortunately, there are ways of testing whether the child's underlying forms are adultlike or not.

Most evidence is anecdotical:

[fɪs] ~ [fɪʃ]   (Berko & Brown 1960)

[səːt] ~ [ʃəːt]   (Neil Smith 1973)

What does the **formation** of a generalization look like in a developing phonology?

1. Danny (diary), permanent persistence of old forms.

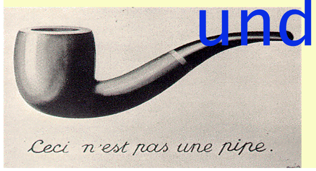
Two of his 4 earliest words, *Hi* and *Hello,* were h-initial and produced with adult-like #/h/: [hai], [h^{w}ow]

All subsequent adult h-initial words (for some months) were produced without the /h/, but [hai] and [h^{w}ow] never lost their [h]s, **remaining as exceptions**.

2. Danny again, temporary persistence of old forms.

Early (and favorite) words were *down* [dæwn] and *stone* [don].

Then a nasal assimilation pattern started, e.g. *beans* = [minz]. [dæwn] and [don] **remained for a while**, then **varied** with [næwn] and [non], then **disappeared** under the pressure of the general pattern.



What does the **formation** of a generalization look like in a developing phonology?

Looking at Jacob without blinking:
Eventually, he developed a pattern of pronouncing certain 1-syllable words containing adult /ei/ using the vowel /i/ instead.
But this rule (or constraint against /ei/ in certain contexts) developed slowly and messily.

And as for problem 3, U-shaped curves are inevitable if learning is instance-based and generalizations spread slowly, but do eventually go beyond the input evidence. Let's watch that starting to happen.

[ei] to [i]l table, p. 125 of IULC diss.; notation revised.							
	16:16	16:30	17:02	17:11	17:16	17:18	17:23
tea /ti/	ti						
tea /ti/						di, dEi	
tape /teip/			dæ		tEi	tEi, dE	
tape /teip/		?tA			ei		tE
key /ki/						khi	khi,xE
key /ki/							
okay /okei/				ki			
gate /geit/							

	17:25	17:27	18:02	18:04	18:09	18:18	
tea /ti/							
tea /ti/		ti	(vowel always i after this)				
tape /teip/							
tape /teip/	?txi	ti, di	gei	tei			
key /ki/							
key /ki/	xE		ki		ki	ki, xi	
okay /okei/							
gate /geit/						gi	

Shaded cells are imitations, yellow are spontaneous; all words known to be in receptive vocabulary

Finally, looking at the model again:

There's redundancy in the Linked Attractor model, because we posit both stored input (perceptual) and output (articulatory) representations **plus** output mappings (grooves in the 2-D Nicholas diagram) that link perceptual to articulatory representations

We argue for this because there's evidence for both on-line swift mapping of new forms to articulatory outputs AND for stored articulatory outputs in the case of entrenched forms like Danny's *hi* and *hello*. (One could describe them with item-specific constraints, but instance-based entrenchment explains WHY those early words are special; item-specific constraints emerge as epiphenomena.)

Linguists historically dislike redundancy, but the only good reason for parsimony is to simplify a theory when doing so is helpful. There's no reason to hold to parsimony when the evidence supports multiple or overlapping representations of information.

(William of Ockham also took this view.)

Redundancy is a basic property of biological systems. It's **essential for maintaining information in noise**, from DNA repair on up to higher cognition. There is a great deal of redundancy in the brain and in what it learns – for example, it **learns co-occurrences** like voice pitch range and speaker identity, or like a string of printed words and the color of paper they are printed on, **without knowing whether the information will be useful** or not.

A rich example to further support the existence of stored output representations: Jacob's latent output forms for #/b/-initial words

Jacob began to shadow *byebye* as /dada/ at **13:25**; all other /b/-initial words were avoided (showing only occasional transient attempts) until about **18:2**. Even on these attempts, no labial was produced (it was assimilated or omitted).

At **19:9** the first correct #/b/ was produced on *box* - but the very first occurrence was /**da**/. Other words long-established in his reception vocabulary also showed a first attempt with #/d/, swiftly corrected to #/b/ — except in the word *ball*, which was produced as /da/ for several weeks, even after *byebye* had switched to a correct form.

Hypothesis: Jacob had formed a **latent and inhibited** #/d/ form for *ball*, and after learning that he could at last say /b/, he forgot to listen to and correct this deeply entrenched form.

Our goals.

We like neural plausability. A ‘connectionist’ type of OT with very many very simple constraints has that to some extent. The linked-attractor model may represent the underlying instance-based mechanism from which the higher-level OT emerges.

We need:

- computational implementations on every level;
- much more densely detailed child data

Thank you for listening!