Linguistics 219 Spring 2018
Phonological Theory III B. Hayes

Class 9, 4/30/2018: Acquisition III: Learning the parental phonology

1. Assignments

- Hand in Homework #3
- Read: Sharon Goldwater, Thomas L. Griffiths, Mark Johnson (2009) A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112:21–54.
    - ➢ This is the last technical paper we will cover; no homework yet so you can spend a little more time with it.
    - ➢ Feel free to skip appendices.
- Come talk with me about your term paper.

2. Where are we?

- Bifurcated system of acquisition, grammars for production, perception
- Turning now to the task of learning the parental system.

3. The plausible course of learning the parental pattern

- **Segmentation**, done distributionally at first, later with world-knowledge and grammar-knowledge.
- Treatment of the discovered words and allomorphs:
    - ➢ phonotactic analysis (which might in turn help segmentation, in virtuous circle)
    - ➢ discovery of alternations and underlying forms

4. Learning alternations

- Bifurcate:
    - ➢ productive phonology needs to be treated with some kind of GEN + EVAL architecture, which would permit generalization of alternations to novel morphemes (*blitting* [ˈblɪɾɪŋ], reluctantly done by Albright/Hayes 2003 subjects).
    - ➢ Else learn to deploy the listed allomorphs properly.

5. Learning good old fashioned phonology

- We have a theory to derive outputs and algorithms to rank/weight the constraints.
    - ➢ Though we sometimes wish we could discover the constrints themselves.
- It probably would help to use alignment to find the particular segmental alternations:
    - ➢ Example: Polish [vut͡ɕ] ‘lead-imp.sg.’ ~ [vod͡z-e] ‘lead-1 sg.’

```
a. v  u  t͡ç          not:     b.   v  ut͡ç  ∅
   |  |  |                         |  | |   |
   v  o  d͡z                        ∅  vo   d͡z
```

Therefore:

[u] ~ [o] is an attested alternation.
[t͡ɕ] ~ [d͡z] is an attested alternation.
[u] ~ [v] is not an attested alternation.

- Some research in this area:
  - Gaja Jarosz (2006 dissertation, later work)
  - Ryan Cotterell, Jason Eisner et al. (big ACL bake-off with computer scientists and large data sets. Connectionism wins!)
  - Tesar (2014 book with Cambridge)

6. Learning distribution of listed allomorphs

- I don't know of any work in this area.

7. Allomorphs in contemporary linguistics

- Recent WCCFL presentations of McPherson and Zhang for phrasal phonology-as-allomorphy.
- When we do bases later, we will study **lexical conservatism**, a theory that presupposes a *lot* of allomorph-memorization.

8. Learning to take the wug test

- Classical phonological analysis does not equip you for this!
  - It only *rationalizes* the data pattern, showing how the data could be derived from a set of underlying forms.
  - To wug test, you must go from surface data to surface data.
- How to fix this?
  - Albrightianism: there are privileged forms in the paradigm that always permit the UR to be inferred (e.g. by grabbing the relevant allomorph and undoing the allophonic rules). E.g. Adam Albright (2010) Base-driven leveling in Yiddish verb paradigms. *NLLT* 28:475-537.
  - Perception grammars, part of a large bidirectional program by Boersma.
  - Bayesianism: evaluate UR's on the basis of the probability with which they would yield observed SR's in general, then predict the SR's you want by applying the grammar in the forward direction from the distribution of UR's you deduced.

BASICS OF DISTRIBUTIONAL SEGMENTATION

9. A key empirical paper

- Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning in 8-month-old infants. *Science*, 274, 1926–1928.
- This showed that 8-month-old babies can extract and later recognize "words" in the form of horrible-sounding synthesized CVCVCV sequences from monotone, unmodulated "speech".
- It's widely cited (4400 citations), often for ideological reasons (i.e. that purely-statistical learning is possible, Chomsky is wrong wrong wrong, etc.)
- More evidence bearing on ideology:
  - Elissa L. Newport and Richard N. Aslin (2000) Innately constrained learning: blending old and new approaches to language acquisition. In S. C. Howell, S. A. Fish, and T. Keith-Lucas (eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development.* (pp. 1-21).
  - Somerville, MA: Cascadilla Press, 2000.
  - It doesn't have to be sounds; it can be tones, lights, colors …
  - It doesn't have to be humans, it can be tamarin monkeys,[1] …
- They say "infants have access to a powerful mechanism for the computation of statistical properties of the language input"
  - But *what is this mechanism?*
  - Goldwater and her colleagues on this line of work: "This research, however, is agnostic as to the mechanisms by which infants use statistical patterns to perform word segmentation."

10. Acquired knowledge about the domain quickly facilitates further segmentation (virtuous circle)

- Lots of results from Anne Cutler and other psycholinguists show word segmentation is aided by phonotactics.
  - E.g. English iambic words (*balloon, believe*) are rare.
  - Children tend to split them up in segmenting: *the gui | tar is.*[2]
  - Not so for other languages.
- Finnish kids can use "vowel harmony breaks"; Suomi, McQueen, & Cutler, 1997[3]
- For a big literature review, see Kim diss., cited below.

[1] Caution: this work done with the now-defrocked academic fraudster Marc Hauser, so you can't trust at least the original citations.

[2] Jusczyk, P. W., Houston, D. W., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. Cognitive Psychology, 39(3-4), 159-207.

[3] Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, 36, 422-444.

11. Real life feats of segmentation: more

- Getting common morphemes, like -*s*, very early
  - ➢ Kim, Yun Jung (2015) 6-month-olds' segmentation and representation of morphologically complex words, UCLA dissertation.
- Segmenting even when the result is obscured by alternations
  - ➢ last readings, with paradigm uniformity effect

AN EXAMPLE OF DISTRIBUTIONAL LEARNING: GOLDWATER ET AL. (2009; READINGS)

12. Our goal in studying it

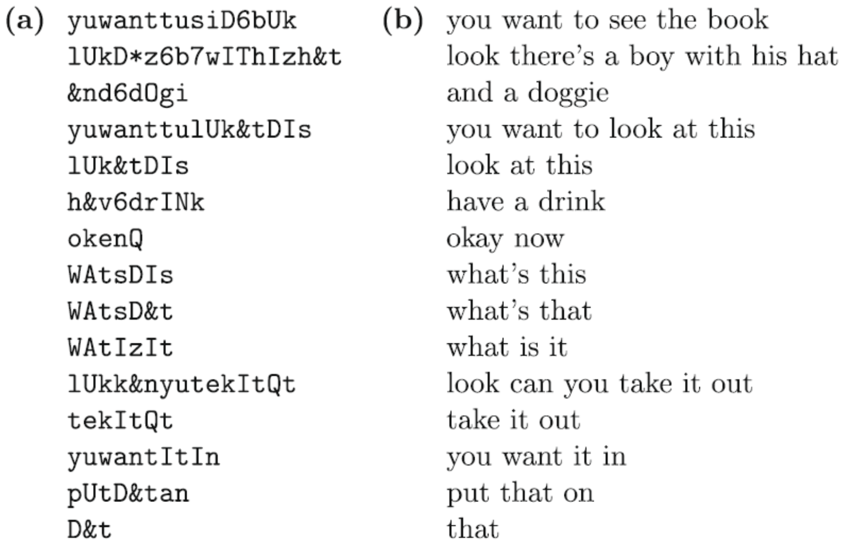
- This *sort* of model — learning something useful from sheer distributions — may become increasingly common in formal linguistics; as with Wilson and Obdeyn, let us wrap our heads around it as much as we can.

13. Concrete goal of the work

- Develop a system that can distributionally segment phonetic corpora into their words, without understanding them in the slightest.
  - ➢ Mounting evidence suggests that this is more or less what is happening in the mind of the 8-month-old.[4]
- Of course, the parsed corpus itself is of no importance; surely it will be forgotten.
- The real payoff will be in:
  - ➢ the candidate lexicon
  - ➢ word frequencies
  - ➢ (later on:) preliminary statistical information about what word precedes what

[4] Presumably the easy words like *Mommy* are learned with meaning pretty early.

14. A chunk of the corpus used (30K words, child-directed speech)



(a)	(b)
yuwanttusiD6bUk	you want to see the book
lUkD*z6b7wIThIzh&t	look there's a boy with his hat
&nd6dOgi	and a doggie
yuwanttulUk&tDIs	you want to look at this
lUk&tDIs	look at this
h&v6drINk	have a drink
okenQ	okay now
WAtsDIs	what's this
WAtsD&t	what's that
WAtIzIt	what is it
lUkk&nyutekItQt	look can you take it out
tekItQt	take it out
yuwantItIn	you want it in
pUtD&tan	put that on
D&t	that

15. Some lessons that will be promulgated on the way

- An example of work that uses Bayes's Theorem as a starting point.
- Even the crudest conception of a higher-order grammar helps a lot ("What words like to come after this word?"
    - "There will … be word boundaries with relatively high transitional probabilities (where two words are highly associated, as in *rubber ducky* or *that's a*)."
    - Goldwater has gone on to write further papers saying: learning two things at once can be easier than learning one at a time, if the tasks inform each other.
- The **algorithmic level** vs. the **computational level** (cf. GLA vs. maxent).
    - Algorithmic level: find a way to move forward such that you think it will move you toward the best answer.
    - Computational level: define an objective function that when optimized would characterize the best answer. Use whatever method words best to optimize.
    - They demonstrate that earlier work on the algorithmic level succeeds better only due to defects in the searching.

16. Conception of probabilistic learning

- There is a large hypothesis space.
- Each hypothesis has a probability, at every stage of learning.
- The probabilities existing after learning are the **posterior distribution**, which you can use in various ways.
    - Sometimes they use the MAP, "maximum a posteriori" hypothesis; i.e. the best one.
    - Sometimes they sample from the distribution of hypotheses.
- Updating is done with Bayes' Rule (often called Bayes' Theorem).

- Often, as the learned output we pick the hypothesis with the highest posterior distribution.

17. Bayes' Theorem

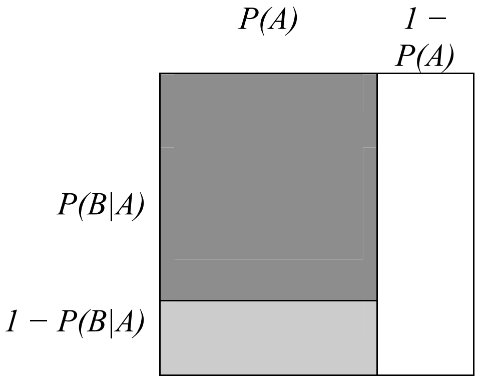
- $P(B|A) = \dfrac{P(A|B)*P(B)}{P(A)}$

- Scenario: "If you currently think the probability of B is P(B), and know evidence A, then you may update your belief to P(B|A) in the manner shown."
    - Hence the terms **prior** distribution and **posterior** distribution, often used.
- Usefulness: updating the value of a hypothesis if you know how to compute the likelihood of the observed data A (both under the hypothesis, and in general).
    - I.e. use your ability to compute the forward direction to get the backward direction.

18. Let's take a minute to prove Bayes' Theorem

- Probability theory has two axioms (like Euclid's for geometry)
- Sum Rule
    - P(A *or* B) = P(A) + P(B) − P(A *and* B)
- Product Rule
    - P(A *and* B) = P(A) * P(B|A)
    - where | means "given"
    - "The probability of both A and B occurring is the probability of A occurring, multiplied by the probability of B occurring given that A occurs."

19. Graphic demonstration of the Product Rule



20. The proof

1. $P(A \textit{ and } B) = P(A) * P(B|A)$ Product Rule
2. $P(B \textit{ and } A) = P(B) * P(A|B)$ Product Rule (applied to A and B in the other way)
3. $P(A \textit{ and } B) = P(B) * P(A|B)$ Commutative property of *and*
4. $P(A) * P(B|A) = P(B) * P(A|B)$ Equality is transitive
5. $P(B|A) = \frac{P(A|B)*P(B)}{P(A)}$ Dividing both sides by P(A)

21. Restating with data and hypothesis

$$P(h|d) = \frac{P(d|h)*P(h)}{P(d)}$$

- This is used as an update rule.
- P(d) is often computed by producing a weighted sum of the probability of the data under all hypotheses.

22. Goldwater et al's set of hypotheses

- Just take the whole corpus and put in word boundaries where you please.[5]
- Socrates: *n* phonetic segments in the corpus; how many hypotheses exist?
- Adjustments are made for utterance boundaries, which must be word boundaries; ignored here.

23. Degenerate Bayes's Theorem

- When we do the above, P(d|h) is always one! Every segmentation of the data, when concatenated, yields the data.
- P(d), the weighted sum of the probability of the data under all hypotheses, is also one.
- So really, we just seek the most probable *prior* hypothesis — P(h).
- That will depend on making sensible prior assumptions about what word sequences are like.

24. What would be a sensible prior hypothesis? Several ingredients:

- We expect that the same words will appear over and over.
- The more data we look at, the fewer words new to us there will be.
- Word frequency in corpora follows a characteristic pattern statistically explored long ago by Zipf.
- About new words: if we encounter a new word, it likely obeys the phonotactics of the language.
- New words cannot be unreasonably long.[6]

[5] Extra stuff goes on because the corpus is divided into small utterances.

25. These ideas rendered with mathematical formulae

- Overall scheme: to evaluate a particular assignment of #'s, pretend you are generating the corpus left to right, one word at a time; and keep multiplying until you get the prior probability of this word-parse.
- Let's do this one aspect at a time.

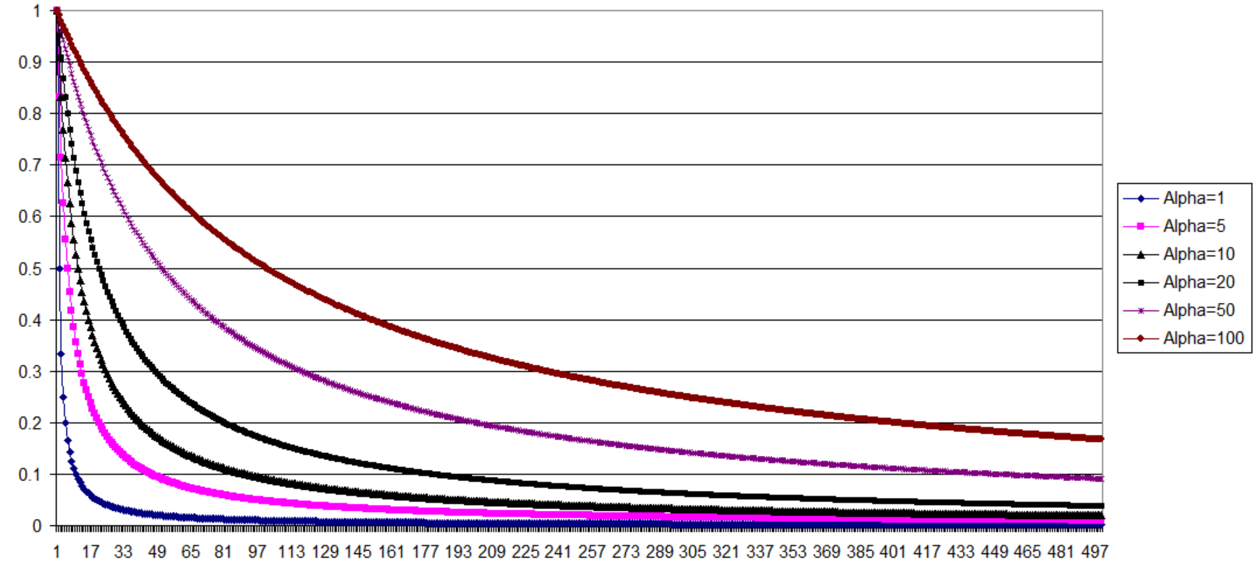
- *The same words appear over and over.*
- *The longer we look at data, the fewer the new words will be.*
- *Word frequency follows a characteristic statistically explored long ago by Zipf.*

$$P(w_i \text{ is novel}) = \frac{\alpha_0}{n+\alpha_0}$$

- ➢ This makes the first word (n = 0) novel with certainty.
- ➢ Keeps on sinking; after an infinite corpus we expect no new words.
- ➢ Otherwise, we use simple counting to estimate the probability of hearing a known word.
- ➢ Socrates: what is the effect of increasing the model parameter α? Turn the page to verify.

[6] I think this can actually be phonology: English has zero monomorphemic words of more than three metrical feet, despite abundant sources in Native American language borrowings: *Okaloacoochee*, but ?*Okaloamoacoochee*.

26. Probability of new word as you proceed, for various values of alpha



27. Existing words

$$P(w_i = \ell \mid w_i \text{ is } \breve{\text{not}} \text{ novel}) = \frac{n_\ell}{n}$$

- Just use the known counts to estimate probability.

28. Continuing with the model: phonotactics

- *If we encounter a new word, it likely has a reasonable length and obeys the phonotactics of the language.*

$$P(w_i = x_1 \ldots x_M \mid \breve{w}_i \text{ is novel}) =$$
$$p_{\#}(1 - p_{\#})^{M-1} \prod_{j=1}^{M} P(x_j)$$

where

- M is the length of the word you are considering.
- $p_{\#}$ is the probability of the choice, “hey, you’re done, end the word here”.
- Socrates: what distribution of word lengths is defined? Is this realistic?
- $P(x_j)$ is the fraction of total phonemes taken up by phoneme x_j
- Socrates: what kind of phonotactics is this? Is this realistic?

29. Being clear on this point: there are two divinely-set parameters.

- As you did in the generality bias homework, Prometheus tries out various versions of the language faculty to give to humanity, and picks the one with good parameters for learning words from continuous speech.
- Word-novelty parameter
- Word-length parameter

30. An unacknowledged debt

- The authors only fit their parameters to English.
- Socrates: what properties of other languages might be more comfortable with different parameters?

31. More Socrates: sanity check on rejecting "obviously stupid" parses

- One bad segmentation is to split the corpus into its phonemes. How will this rack up a low probability score?
- Another is to take the whole corpus as one gigantic word. How will this rack up a low probability score?
- Another is to carefully go through your parse and make sure it never employs the same word twice. How will this rack up a low probability score?

32. Finding the most probable segmentation(s)

- This is done with the technique known as Gibbs sampling.
- It generates an infinite journey that after an initial settling-in period visits each hypothesis in proportion to its probability.
- From this the best hypothesis (MAP) may (in principle) be selected.
- OR you can sample from the distribution, trying each hypothesis according to their probability.
- The authors do both.

33. Local maxima

- There is no guarantee that the method will find the best answer, and the authors tried to increase their chances by using different starting points:
    - ➢ Every phoneme a word
    - ➢ Corpus is a word
    - ➢ Random

34. Evaluation of model performance

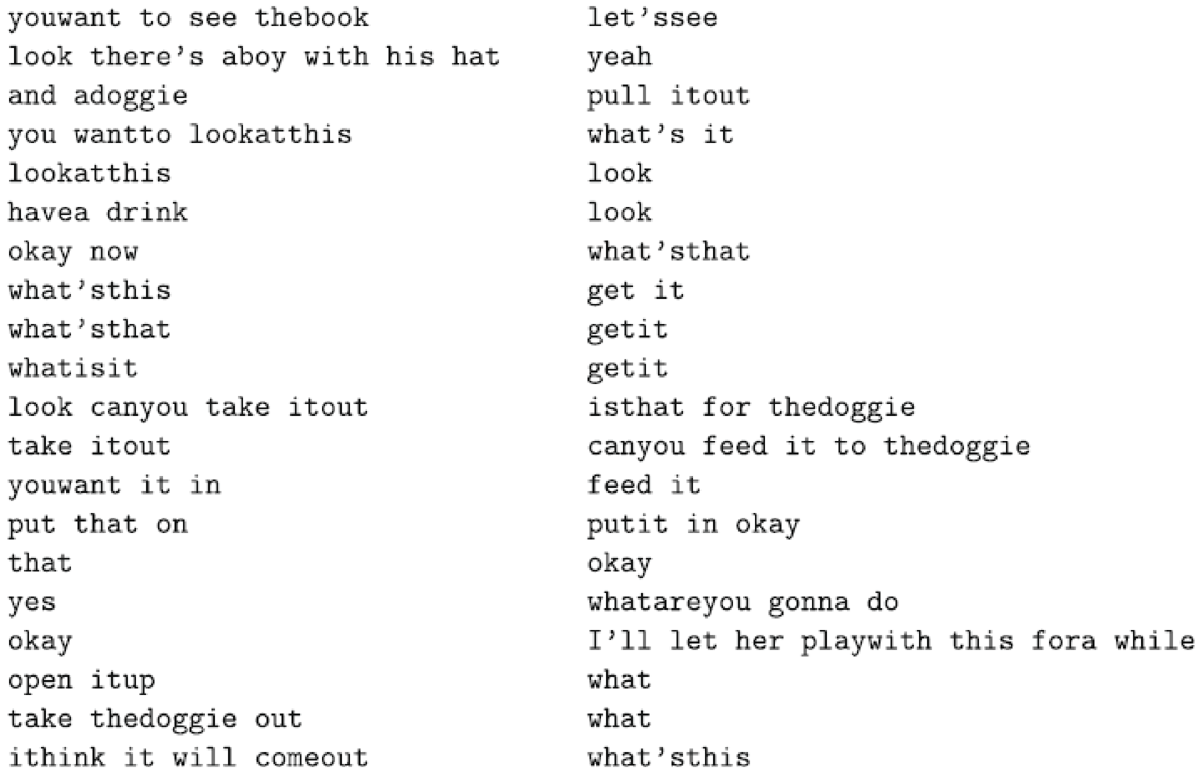
- These terms appear frequently in such work.

Computer scientists	*Cognitive scientists*	*formula*
precision	accuracy	number of correct items/total items found
recall	completeness	number of correct items/total of correct items

- Composite measure: F_0 or F-score = $\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$
- This can be computed on the set of words, the lexicon, and the set of boundaries.
- They can use this for diagnosis in various ways.

35. Their first-pass model was not great

- It learned tons of two-word sequences as words, no matter how they set the new-word and word-length parameters.



```
youwant to see thebook              let'ssee
look there's aboy with his hat      yeah
and adoggie                         pull itout
you wantto lookatthis               what's it
lookatthis                          look
havea drink                         look
okay now                            what'sthat
what'sthis                          get it
what'sthat                          getit
whatisit                            getit
look canyou take itout              isthat for thedoggie
take itout                          canyou feed it to thedoggie
youwant it in                       feed it
put that on                         putit in okay
that                                okay
yes                                 whatareyou gonna do
okay                                I'll let her playwith this fora while
open itup                           what
take thedoggie out                  what
ithink it will comeout              what'sthis
```

36. Initially, their rivals' earlier models looked better

- … until they tried checking their own best solutions as candidates *within* the rival models, and they won.
- It turned out that the putative good performance of the rival models was due to inferior, 1970's searching — the true predictions of the earlier models were just as crummy.
- So the old models were a clear, if inadvertent, case of the computational vs. the algorithmic levels (see above): algorithmic success, computational failure.

GOLDWATER ET AL. SCALE UP: THE BIGRAM MODEL

37. Words are not emitted at random

- … but are generated by a grammar!
- … plus also, probably, some lexical listing of phrases …

38. A barbaric CS kind of grammar

- Bigram model: each word is emitted with a probability dependent solely on the preceding word.

39. Scaling up the Goldwater et al. model to take the preceding word into account

- This discussion paraphrases the article.
- $\mathbf{w_{i-1}}$ is the last word you've gotten to, ready to move on.

Decide, by flipping a special coin, whether the pair < w_{i-1}, w_i> will be a novel bigram type.

```
if novel bigram type  (probability parameter associated with this)
    i. Decide whether $w_i$ will be a novel unigram type.
            if $w_i$ is novel unigram type (probability parameter associated with this)
                    pick its phonemic form as before (probability parameter associated with this)
            if $w_i$ is not novel unigram type
                    pick $w_i$ following the probabilities of words that you have seen
if not novel bigram type
                    pick $w_i$ following the probabilities of bigrams that you have seen
```

- So there are three parameters in the model now, which they set against data.

40. Are things better?

- Definitely, and their model is now beating the competition.
- Things are not perfect:
    - Still some word sequences learned as words. *look.at, can.you, is.it*, etc.
    - Trickily, *affixes* are now getting segmented: -s, -z, -d, -t, -ɪŋ.

   - Even worse, they evidently are segmented in **weird places**: *Have a d rink*. Morphology needed, I suspect.

41. The "linguistics" in these models is simply appalling

- Phonotactics = product of probabilities of the segments
- Syntax and diction: a Markov chain, known since the 50's to be a flop for syntax.
- The authors aren't dumb; they know this, and presumably are trying to walk before they run — the conceptual apparatus for distributional learning needs to be put in place.

42. The concept of *ideal learning analysis* and the competence of human children

- Scheme:
  - Set up a formal system with strong foundations and high performance level.
  - This is in principle informative about people because people are close to optimized, or so it is thought.
- Assessment:
  - Certainly in linguistics, there is some belief that language learners are, to the contrary, not especially competent.
    - -- Obsession by many with on-line rather than batch processing, assuming poor memory.
    - -- Proposing totally blind hill-climbing as an acquisition model. ("Hey, I couldn't parse this sentence, better pick a parameter at random to switch!")
  - Goldwater et al. are more optimistic about humanity; p. 22: "To date, there is little direct evidence that very young language learners approximate ideal learners. Nevertheless, this suggestion is not completely unfounded, given the accumulating evidence in favor of humans as ideal learners in other domains or at other ages [citations]"

A LITTLE BIT ON HIDDEN STRUCTURE

43. What is hidden structure?

- = aspects of representations not inferable from surface form
- Examples:
  - underlying representation (German [rat] = /rad/, /rat/)
  - metrical feet (two ways to bracket a trisyllables with penultimate stress)
  - syllabification ([ab.ra] vs. [a.bra], with consequences for stress, metrics

44. Why is hidden structure hard to learn?

- If you make an assumption about feet, then all the rest of the grammar must be tailored to that assumption.
- But most ranking/weight algorithms blindly try to optimize all the constraints at once.

45. A toy example: mini-German

- Example drawn from:
  - Pater, Joe, Robert Staubs, Karen Jesney and Brian Smith (2012) Learning probabilities over underlying representations. In the Proceedings of the Twelfth Meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology. 62-71.
- There only four data:

advice-plain	/rat/	→	[rat]
advice-suffixed	/rat-a/	→	[rata]
wheel-plain	/rad/	→	[rat]
wheel-suffixed	/rad-a/	→	[rada]

  - N.B. -*a* is not a suffix in German but it is easy to type.

- For learning, let's explore the larger set of candidates that arises if we are trying to learn UR's.
  - No particular reason to think 'advice' is anything other than /rat/.[7]
  - But 'wheel' has two candidates, /rat/, /rad/.

advice-plain	/rat/	→	☞ [rat]
advice-suffixed	/rat-a/	→	☞ [rata]
wheel-plain	/rad/	→	☞ [rat]
	/rad/	→	[rad]
	/rat/	→	☞ [rat]
	/rat/	→	[rad]
wheel-suffixed	/rad-a/	→	☞ [rada]
	/rad-a/	→	[rata]
	/rat-a/	→	☞[rada]
	/rat-a/	→	[rata]

46. What defines success?

- We must derive *at least one* of the observed ☞ candidates for each input.
- We must impose **consistency** on the UR's, since we need a good UR to pass a wug test on future forms.
  - It will not do, as Pater et al. suggest, to let the UR vary freely in its paradigm.

47. Constraints

- Let's not bother with constraints that would derive Intervocalic Voicing, since /rat-a/ → [rata] will straightforwardly remove this possibility.

[7] Actually, people occasionally override the "what you see is what you get" principle for non-alternating morphemes when they do "set up as": set up all [h] as /x/, so it can trigger velar place assimilation (Toba Batak); then revert all /x/ to [h] on the surface. This is not so commonly done as it used to be …

- We do need the standard constraints for Final Devoicing:
  - *[−sonorant, +voice] $]_{\text{word}}$
  - Ident(voice)
- We need, following Boersma, Appousidou, Pater et al., constraints that force a particular allomorph as the UR.
  - Wheel is /rad/ — correct!
  - Wheel is /rat/ — wrong!
  - Apoussidou, D. (2007). The learnability of metrical phonology Utrecht: LOT

48. A fancier kind of tableau: collating over hidden structures

- Observed candidates sum over all their sources.
- You win if the frequency of the correct observed candidate is 1.
- The weights were established by me, using thought.
  - Socrates: justify them, remembering that this is maxent.

	Hidden	Overt freq	wheel /rad/	wheel /rat/	*Coda Voiced Obs	Ident (voice)			
			50.0	0.0	50.0	25.0	H	p	p(overt)
advice	rat --> rat	1					0	1.000	1.000
advice-a	rat-a --> rat-a	1					0	1.000	1.000
wheel	rad --> rat	1		1		1	25	1.000	1.000
	rat --> rat		1				50	0.000	
	rad --> rad	0		1	1		50	0.000	0.000
	rat --> rad		1		1	1	125	0.000	
wheel-a	rad-a --> rad-a	1		1			0	1.000	1.000
	rat-a --> rad-a		1			1	75	0.000	
	rad-a --> rat-a	0		1		1	25	0.000	0.000
	rat-a --> rat-a		1				50	0.000	

49. The fiasco: hand-ranking is easy, but algorithmic-search ranking crashes and burns!

	Hidden	Overt freq	wheel /rad/	wheel /rat/	*Coda Voiced Obs	Ident (voice)			
			0	0	20	0	H	p	p(overt)
advice	rat --> rat	1					0	1.000	1
advice-a	rat-a --> rat-a	1					0	1.000	1
wheel	rad --> rat	1		1		1	0	**0.5**	**1**
	rat --> rat		1				0	**0.5**	
	rad --> rad	0		1	1		20	0	
	rat --> rad		1		1	1	20	0	

wheel-a	rad-a --> rad-a	1		1			0	**0.25**	**0.5**
	rat-a --> rad-a		1			1	0	**0.25**	
	rad-a --> **rat-a**	0		1		1	0	**0.25**	**0.5**
	rat-a --> **rat-a**		1				0	**0.25**	

- Wrong UR, wrong outputs.
- This is if you take 0 as starting point weights for the Solver.
- If you take a very big starting weight for WHEEL = /rad/, then everything works.
  - ➢ This is making it innate that the word for wheel is /rad/, not a hopeful strategy.

50. Why fiasco?

- The summing over hidden structures evidently removes the beautiful **convexity** that makes maxent learning so appealing.
- *If* you are in the region when WHEEL = /RAD/ is high, then the best ranking of Markedness and Faithfulness is the one that yields final devoicing.
  - ➢ IDENT(voice) rightly wants to be high, protecting /rad-a/ and /rat-a/ from undesired random changes.
- *If* you are in the region when WHEEL = /RAD/ is low, then you are in danger of deriving (from wrong UR) /rat-a/ → *[rata] 'wheel'
  - ➢ Now IDENT(voice) only wants to get out of the way! Being Faithful can only do harm, as it *encourages* the bad outcome.
  - ➢ But if IDENT(voice) is near zero, then promoting WHEEL = /rad/ does no good; the UR won't get enforced.
  - ➢ "Hey, I thought it was your job, so I decided to just nap."
  - ➢ They nap on the couch of a wrong local maximum.
- More generally, we are letting the *violations* of IDENT(voice) be dependent on the values of the UR constraints, a context-dependency that seem responsible for defeating convexity.

51. Efforts to learn hidden structure

- Tesar and Smolensky (2001) *Learnability in Optimality Theory*. An approach called Robust iterative parsing; non-stochastic OT.
- Tesar book (2014) *Output-Driven Phonology*, Cambridge University Press.
- Appousidou, cited above
- Gaja Jarosz paper in progress, with a whole new version of OT, evidently best of the lot but not perfect. I would love to try out her system.
  - ➢ Jarosz, Gaja. 2015 / in revision. Expectation driven learning of phonology. University of Massachusetts manuscript.

52. The exterminationist approach to hidden structure

- Pending further progress in learning theory, perhaps hidden structure is more trouble than it's worth?
- It's been tried for **metrical stress theory** a number of times (no feet):

  - ➢ Alan Prince (1983 *LI*, “Relating to the grid”)
  - ➢ Gordon, Matthew (2002) A factorial typology of quantity insensitive stress, *Natural Language and Linguistic Theory* 20, 491-552.
- Donca Steriade is an exterminationist w.r.t. **syllables**, a strikingly non-traditionalist point of view, but she has replacement theories in hand for both
  - ➢ phonotactics (phonetic cue-based theory)
  - ➢ metrical structure (interval theory)
- For **underlying representations**, there is a modest contingent who want to do phonology just with allomorphs, no UR’s inferred from allomorphs. Harry Bochner, Luigi Burzio are examples.
- Exterminationists are thinner on the ground in **syntax** (e.g., trees with fewer nodes) but perhaps categorial grammar is an example. Here is an automated-learning-of-syntax paper:
  - ➢ Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater and Mark Steedman (2017) Bootstrapping language acquisition. *Cognition* 164, pp. 116–143.
- Remember always that *complete* extermination of hidden structure is certainly not feasible; there’s various stuff I can’t imagine we could do without.