Linguistics 219 Spring 2020
Phonological Theory III B. Hayes

Class 8, 4/23/2020: Acquisition III

1. Assignments

- Homework #3 due on Tuesday
- Read:
  - Claire Moore-Cantwell (2019) The new status of exceptions when phonology is probabilistic; slides for invited lecture at Manchester Phonology Meeting.
  - On course web site.
- Due date for appointment with me discussing your term paper topic is end of next week (which is Week 5).

2. Where are we?

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

WHAT MIGHT BE A PLAU SIBLE ACQUISITION ROUTE?

3. Theme

- Learning stages that can already be done, albeit only mediocrely, with existing procedures.

- Distributional acquisition of (mostly meaningless) words from the speech stream.
- Distributional acquisition of morphemes from the word set.
- Identification (probably with meaning) of allomorphs from the morpheme set.
  - Lexically-listed allomorphs: distributionally discover their environment
  - Phonologically derived allomorphs: see next stage

4. Once you have allomorphs, learn the phonology

- Use alignment (last item last time) to find the alternations.
- Use the alternations to keep the grammar small (GEN?)
- Find the underlying forms (hmm…?)
- Weight the universal constraint set (hmm …) to derive the surface forms with their characteristic frequencies.

5. Cases of "hmm…"

- See below on the hardness of finding underlying forms, and a possible route.
- Many phonologists would hope to discover the constraints themselves, perhaps distributionally.

- and perhaps aided with bias (Becker et al. readings, similar work)
- and perhaps aided with phonotactic learning: Chong, Jo[1]

6. This might best be done all at once rather than stagewise

- Sharon Goldwater, SCIL, 2018 — can't find

7. Even when we're done, we might not yet have testable results: learning to take the wug test

- Classical phonological analysis does not equip the child 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**.
  - And of course this is the real-life wug test, too.[2]

8. How to fix theory so it makes wug-test predictions

- **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.

RESUMING THE THREAD

9. Allomorphs in contemporary linguistics

- Laura McPherson's articles in *Language* and *NLLT* on phrasal allomorphy in Tommo So and other Dogon languages.
- Jie Zhang's work on the limited productivity of phrasal phonology in varieties of Chinese
  - https://linguistics.ku.edu/jie-zhang#link3
- When we do bases later, we will study **lexical conservatism**, a theory that presupposes a *lot* of allomorph-memorization.

---

[1] **Chong, A.J.** (2019). Exceptionality and derived-environment effects: A comparison of Korean and Turkish. *Phonology,* 36(4), 543-572; Jinyoung Jo, UCLA M.A. in progress.

[2] Caution: there is evidence children are reluctant to take it. Vsevolod Kapatsinski thinks novel inflected forms are introduced into the speech community only by phonologically talented outlier people.

BASICS OF DISTRIBUTIONAL SEGMENTATION

10. 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,[3] …
- 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."

11. 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.*[4]
  - Not so for other languages.
- Finnish kids can use "vowel harmony breaks"; Suomi, McQueen, & Cutler, 1997[5]
- For a big literature review, see Kim diss., cited below.

---

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

[4] 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.

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

12. 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

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

13. Ref.

- Sharon Goldwater, Thomas L. Griffiths, Mark Johnson (2009) A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112:21–54.

14. 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.[6]
- 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

15. 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

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

16. Goldwater et al's set of hypotheses

- Just take the whole corpus and put in word boundaries where you please.[7]
  - ☞ *n* phonetic segments in the corpus; how many hypotheses exist?
- A hypothesis is good if its words have characteristic length distribution (given as parameter of model) and recur to a specific degree.
- What makes this possible is massive improvements in ability to search the hypothesis space — Gibbs sampling, now used throughout the sciences.

17. There are two divinely-set parameters.

- We must be Prometheus and give language to humanity.
- Part of this gift is the value of two parameters:
  - Word-novelty parameter
  - Word-length parameter
- The authors only fit their parameters to English.

18. Sanity check: rejecting "obviously stupid" parses

- One bad segmentation is to split the corpus into its phonemes. What does this look like in terms of word length and word novelty?
- Another is to take the whole corpus as one gigantic word. What does this look like in terms of word length and word novelty?
- Another is to carefully go through your parse and make sure it never employs the same word twice. What does this look like in terms of word length and word novelty?

19. 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.

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

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

20. Words are not emitted at random

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

21. A barbaric CS kind of grammar

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

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

- So, a tiny-pseudo syntax; words have probabilities based on preceding words; again with a distribution that favors repeated bigrams.
- So, somewhat better:

```
you want to see the book
look there's a boy with his hat
and a doggie
you want to lookat this
lookat this
have a d rink
okay now
what's this
what's that
what isit
look canyou take itout
take itout
you want itin
put that on
that
yes
okay
open itup
take thedoggie out
i think it will comeout
let'ssee
yeah
pull itout
```

23. 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.
- So, room for improvement, with linguist participation!

HIDDEN STRUCTURE

24. 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
- Linguist theorists love hidden structure, creating a strange prestige system …

25. 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.

DEMO: FAILURE OF STANDARD MAXENT ON A VERY SIMPLE CASE

26. The Pseudo-German Final Devoicing 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]

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

27. Candidate sets when you are learning UR's

- They multiply!

- 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/.[8]
  - But 'wheel' has two candidates, /rat/, /rad/.
- Since [t] and [d] are observed to alternate, let us include all "opposite voicing" candidates in our ad hoc GEN (per above).
- Since we will use MaxEnt, we must take care to use combinatorics in setting up GEN, not relying on harmonic bounding.

28. The candidate set

wheel	/rad/	rad
		☞ rat
	/rat/	rad
		☞ rat
wheel-infl	/rad-a/	☞ rada
		rata
	/rat-a/	☞ rada
		rata
advice	/rad/	rad
		☞ rat
	/rat/	rad
		☞ rat
advice-infl	/rad-a/	rada
		☞ rata

[8] 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 …

	/rat-a/	rada
		☞ rata

29. 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.

30. Constraints

- We do need the standard constraints for Final Devoicing:
  - *[−sonorant, +voice]]$_{word}$
  - IDENT(voice)
- The data are not wildly incompatible with an INTERVOCALIC VOICING constraint, so let's put that in.
- 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

31. Do spreadsheet demo

- 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.

APPROACHES TO HIDDEN STRUCTURE I: EXTERMINATIONISM

32. 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.
  - Adam Albright, to be covered later, is not an exterminationist re. UR's, but his theory is very compatible with UR exterminationism.
- 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 a lot of stuff we probably could do without.
  - ☞ Speculate on this.
  - Hint: nothing is really unhidden except waveforms …

APPROACHES TO HIDDEN STRUCTURE II: COMPUTATION

33. Some literature

- 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.

34. There is a standard recipe in computer science, but I'm not sure how much it has been tried.

- Expectation-Maximimization

- The standard reference is:
  - Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B. 39 (1): 1–38.
- Jarosz, who has mainstream training, uses this in her dissertation.
- To my knowledge, no one has ever tested it out systematically against a broad set of phonological data. what's wrong with this picture?
- This is new to me (I learned in Tim Hunter's Winter 2020 course), so let's just teach it by example.

35. Rad/Rat with EM: new premise

- For now, let's simply assign *probabilities* to /rad/ and /rat/ (as UR's for "wheel", then as UR's for advice.
- The probability of an outcome is calculated thus:
  - For each UR, multiple the probability of that UR by the probability assigned to that outcome by the phonological grammar.
  - Sum the result across all UR's.
- Now look at the spreadsheet.

36. Starting point issues

- Search space is not convex.
  - The theorems proven by the computer science gods say only that this procedure produces local improvements.
- And indeed, I found that with different constraint weights at the starting point, you can fail.
- However, you can follow standard practice and try different starting points, keep the best outcome.
- I find the procedure at least vaguely intuitive; for it is sensibly trying to navigate the search space more flexible, zigzagging its way around the local maxima.