Linguistics 201A Winter 2019
Phonological Theory I B. Hayes

Class 8, 2/2/2023: MaxEnt II; Type Variation I

1. Current assignments

- Hebrew homeworks due today
- BH Hand back half-page summaries of Labov chapter.
- For Tues. 2/7: read:
    - ➢ Kie Zuraw and Bruce Hayes (2017) Intersecting constraint families: an argument for Harmonic Grammar. *Language* 93: 497-548.
    - ➢ Download from course website.
    - ➢ No summary required
- Think about a term paper topic and come talk with me. The official deadline for this is Friday Feb. 10.

2. MaxEnt: what do we have so far?

- Subspecies of Harmonic Grammar
    - ➢ Constraints are weighted.
    - ➢ Ganging is predicted to be ubiquitous.
- A stochastic theory, meant to model probability distributions in data
- It comes with a (very abstract) learning theory; a way for language learners to match distributions in data optimally by weighting.[1]
- Access to highly competent computer science, implementing the learning theory swiftly and accurately with ubiquitous software.

3. The maxent formula (repeated)

$$\Pr(x) = \frac{\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))}{Z}\text{, where } Z = \Sigma_j\ \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$$

4. Is MaxEnt intuitive? The step-by-step calculations

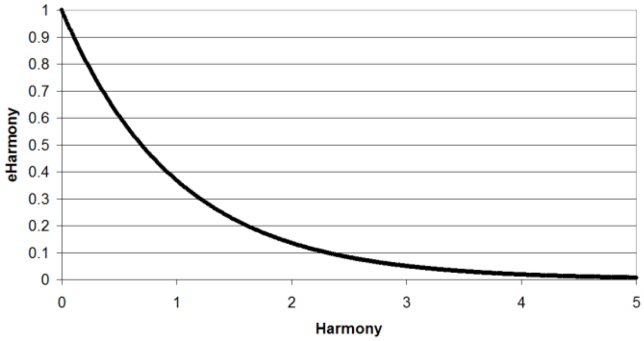
- Imagine that constraint violations are *evidence* for making a *decision.*

[1] The objection "of course, people don't really do this in their head" can be leveled at almost any linguistic theory! For neurally-oriented thoughts about OT and MaxEnt, see Smolensky and Legendre (2006) *The Harmonic Mind.*

Compute this	*Name of what is computed*	*How and why it is computed*
1. $\Sigma_i\, w_i \mathrm{f}_i\,(x)$	Harmony (Smolensky 1986)	Multiply *x*'s violation counts for each constraint (designated $\mathbf{f}_{\boldsymbol{i}}\,\mathbf{(}\boldsymbol{x_j}\mathbf{)}$) by the weight of the constraint ($\boldsymbol{w_i}$), then add up the results across all constraints ($\boldsymbol{\Sigma_i}$). *All available evidence (i.e. constraint violations) bearing on a candidate is considered, in proportion to the constraint weights.*[2]
2. $\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))$	eHarmony (Wilson 2014)	Negate the harmony of *x*, then compute the function **exp( )** on the result, where exp(*x*) is a typographic convenience for e^x, $e \approx 2.72$. *As we consider a series of candidates with ever greater harmony penalties, their probabilities should descend not in linear fashion, but instead asymptote to zero (negative exponential curve) — certainty is evidentially expensive.*
3. $\Sigma_j \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$	Z, the "normalizing constant"	Compute the eHarmony of every candidate derived from the same input as *x* (*x* included), and sum these values.
4. $\dfrac{\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))}{Z}$	Probability of *x*	Divide the eHarmony of *x* by Z (and similarly for all other candidates). *The probability of a candidate depends inversely on the probability of the candidates with which it competes. (Probability of all candidates must sum to one.)*

[2] As noted last time, is *not* so for Optimality Theory, where decisions between candidates are made by the highest-ranking constraint that that distinguishes them, and all the evidence from other constraints is ignored.

5. Reference: graph plotting eHarmony against Harmony (from readings)



MORE ON MODEL ASSESSMENT

6. How do we assess if our analyses are adequate?

- For non-stochastic phenomena, we just pre-sort the data into bins and make sure each bin is derived correctly.
    - ➢ There will be controversy, but the controversy is likely to involve whether the pre-analytic characterization of the data is correct.
- For stochastic phenomena, we are plunged into **model-fitting**, as in many other sciences.

7. Basics of model fitting

- Some quantitative way to assess accuracy.
- Some sort of penalty on model complexity — you can't set up a constraint for every datum.
- Often, an insistence on grounded principles (Archangeli and Pulleyblank, etc.) as the basis of the analysis. Hard to quantify.

8. Simple assessment of the Japanese MaxEnt model from last time

- From last time:
    - ➢ Form adjacent columns of Predicted and Observed, as probabilities.
    - ➢ Plot Predicted and Observed with a scattergram.
- Further:
    - ➢ Compute absolute value of difference (ABS( ) function) to find outliers.
    - ➢ You can do conditional formatting, or sort, to make them stand out.
    - ➢ statistical testing, below

9. One way *not* to assess a MaxEnt model, I think

- **Correlation coefficient** (*r*) of predicted vs. observed
- It is quite possible for a model's predictions to be perfectivel correlated with the observations but not to predict them. [☞ explain]

10. Adapting model evaluation to the needs of linguistics — some qualitative common-sense principles

- If some candidate has frequency 0, and the model gives noticeably above 0, that is quite bad.
  - ➢ Similar to generating an ungrammatical outcome in classical nonstochastic grammar
- This implies that if a candidate has probability 1 (given its input), it would be bad for a stochastic model to fall short of 1 — that would wrongly help the incorrect altenatives.
- Otherwise, rough matching is probably fine for most data sets. (We aren't operating particle accelerators.)

STATISTICAL TESTING

11. This is relatively new to linguistics

- The work in the 1970s of researchers like William Labov or Peter Ladefoged, top quality in its day, did not use statistical testing to evaluate quantitative claims.
- Only in this century, I feel, has linguistics acquired statistical expertise, which is now widespread
  - ➢ Some linguist experts: Harald Baayen, Shravan Vasishth
- Statistics itself is getting better (ANOVA replaced by mixed effects regression, now being replaced by Bayesian statistics).
  - ➢ The old stats had to be done on paper, new stats do ever more crunching.
  - ➢ Statistical models increasingly resemble theories, and give you more interpretable results.
- Here, we cover just a quick test; up to you to gain the expertise you feel you need.

12. The Likelihood Ratio Test

- Remember that log likelihood is our metric of model goodness.
- We can compare log likelihood of *nested* models; e.g. one is the same as the other with an extra constraint.

13. Procedure

- Record the log likelihood with constraint C included.
- Take C out, re-fit the weights, and record the (probably lower) likelihood.
- Compute the difference and double it.
- Do a chi-square test
  - ➢ Excel: =CHIDIST(double-difference, 1)
  - ➢ Gives probability that improvement is not accidental (random variation in data).
- What to throw out is a matter of scientific outlook and journal reviewers. $p < .05$ is loosey-goosey at a social-science level, $p < .00001$ is stricter.
- Let us try this for our Japanese example.

14. Likelihood ratio for Multiple constraints

- You can test two constraints at once if you use =CHIDIST(double-difference, 2)
- Or *n*. This value is known as "degrees of freedom"

15. A caution

- If you have vast amounts of data (e.g., 100,000's of data points) this test is utterly untrustable; don't use it. It can classifying fully-random constraints as significant.
- UCLA has great stats consulting and can help you on this and other matters.

16. Strategies for constructing a grammar using the Likelihood Ratio Test

- I. Build up from the bottom: add the constraint that best improves log likelihood. Keep going until no further constraints test as significant.
- II.Start at the top: keep deleting the least effective constraint until all remaining constraints test as significant.
- Example that uses both:
  - ➢ (2012) Bruce Hayes, Colin Wilson, and Anne Shisko, Maxent grammars for the metrics of Shakespeare and Milton. *Language*, Dec. 2012
- Widely cited reference:
  - ➢ Anderson, D., and K. Burnham. *Model selection and multi-model inference*. NY: Springer-Verlag.
- This procedure can be automated in R: try Kie Zuraw and Connor Mayer's new MaxEnt R implementation (it plugs into all the other goodies available in R).

17. A note on constraints with zero weights

- They are of course useless for explaining your data.
- But before you throw them away, try letting them take negative values (uncheck the box in Solver).
- It may be that the best weight is significantly negative — what you naively thought was a constraint is a credit.

18. MaxEnt itself exists in statistics, under another name

- Multinomial logistic regression
- Same math, but meant as an effort to find causes of patterns in data, not as a model of linguistic competence
- Little kids, likewise, are making an effort to find patterns in the parental language data, so appealing to an effective statistical method is perhaps not far off base (see also work of Laurel Perkins, with Bayesian inference)

HARMONIC BOUNDING IN MAXENT

19. There is none

- Irrespective of harmonic bounding, every candidate gets at least some probability.

- Try this.
- The weaker implication: harmonically bounded candidates never get the most probability.

20. Philosophical position on ill-formedness

- We can accept numbers like 10^{-50} as the equivalent of zero.
- speech error rates offer a floor for how low we can expect a frequency to go

21. Consequences for candidate selection

- You can't just look at "all the plausible repairs for the Markedness violations of the maximally-faithful candidate."
- Try putting in /pa/ → [ba] in Japanese.
- What happens in a happy MaxEnt model is that the constraints punishing the harmonically bounded losers have high enough weights to (essentially) exclude them.
- The method for candidate selection must be based on the combinatorics instead.

22. Views on MaxEnt and Harmonic Bounding

- It's a defect: it permits all sorts of "monsters" to be generated
  - ➢ Kaplan, Aaron. 2021. Categorical and gradient ungrammaticality in optional processes. *Language* 97:703–731.
  - ➢ Anttila, Arto, and Giorgio Magri. 2018. Does MaxEnt overgenerate? Implicational universals in Maximum Entropy grammar. In *Proceedings of the Annual Meetings on Phonology*, vol. 5.
  - ➢ Mai, Anna, and Eric Baković. 2020. Cumulative constraint interaction and the equalizer of OT and HG. *Proceedings of the 2019 Annual Meeting on Phonology.*
  - ➢ Magri, Giorgio. 2015. How to keep the HG weights non-negative: The truncated Perceptron reweighing rule. *Journal of Language Modelling* 3.2:345–375.
- Embrace non-harmonic bounding and use it as a crucial ingredient in analysis
  - ➢ Hayes, Bruce, and Russell Schuh. 2019. Metrical structure and sung rhythm of the Hausa rajaz. *Language* 95:e253–e299.
  - ➢ Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.
  - ➢ Kaplan, Aaron. 2011. Variation through Markedness Suppression. *Phonology* 28.3:331–370.
  - ➢ Bruce Hayes and Claire Moore-Cantwell. (2011) Gerard Manley Hopkins' sprung rhythm: corpus study and stochastic grammar. *Phonology* 28:235-282.
  - ➢ Goldrick, Matt and Robert Daland (2009) Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. Phonology 26: 147-185.

23. Another source of harmonically bounded winners: the GEN-based theory of well-formedness

- This is an alternative to the Rich Base theory of phonotactics.
  - ➢ See Bruce Hayes and Colin Wilson (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379-440.
  - ➢ We will try this out in the next homework.

TYPE VARIATION

24. What is type variation?

- Different stems or words behave differently in the phonology (but each stem or word behaves more or less consistently).

25. Sources of type variation

- **Morphological**: a stem inflects exceptionally for a particular slot in its paradigm (irregular verbs)
- **Phonological exceptionality I**: a stem exceptionally undergoes phonology undergone by only a few other stems (*keep* ~ *kept*)
- **Phonological exceptionality II**: a stem exceptionally fails to undergo phonology it ought to undergo (e.g. *carpenter* [ˈkɑɹpəntɚ], not [ˈkɑɹpəɾ̃ɚ]; cf. *seventy* [ˈsɛvənti, ˈsɛvəɾ̃i]

26. Do phonologists have a warped view of the world? Traditional study vs. corpora

- The laboratory phonologist John Coleman is the only person I know who has denounced phonology problem sets in public,[3] but perhaps others also have qualms?
- Classical phonology problems (paradigms with alternations) often don't acknowledge the exceptions found in the language as a whole; they select exceptionless sets of example forms from the full body of data.
- For example, in Kenstowicz and Kisseberth's (1979) classic textbook:

Language	**Rule**	**Source used to document exceptions**
Tagalog	*Syncope* V → ∅ / VC___CV	Zuraw (2000)
Catalan	*Final /r/ Deletion* r → ∅ / ___ $]_{\text{word}}$	Max Wheeler (2004), *The Phonology of Catalan*
Lardil	*Apocope* V → ∅ / ___ $]_{\text{word}}$ *Final Lowering* V → [+low] / ___ $]_{\text{word}}$	Klokeid, Terry J. (1976) *Topics in Lardil grammar*. MIT dissertation, available https://dspace.mit.edu
Serbo-Croatian	*/l/ Vocalization* l → o / ___ $]_{\text{syl}}$	Bochner 1981[4]

- So these problems are great pedagogically but may create a false impression about the world.

[3] I.e. orally, at a LabPhon conference, not in print.

[4] Harry Bochner (1981) 'The l → o rule in Serbo-Croatian," in N. Clements, ed., *Harvard Studies in Phonology II*, Indiana University Linguistics Club.

27. Traditional data sources in phonology

- Division of labor: the fieldworker gives us generalizations, the analyst assesses their theoretical significance.
- Is there a possible disconnect?
  - ➢ The fieldworker extrapolates from six examples to the whole language; analyst unskeptically takes her at her word.
- Sometimes the fieldworker *has* checked. "I have carefully checked all the generalizations asserted herein through six years of fieldwork and daily conversation with native speakers."[5]
- Another clue to possible issues is the use of synthetic data.
  - ➢ Stanley Newman's classical grammar of Yawelmani is so poor in data that later generative reanalysts have had to synthesize novel forms, just for expository clarity.

28. Finding the data patterns for the whole language

- In well-studied languages you can make use of **traditional scholarship**; e.g. pedagogical material probably can give you all of the English irregular verbs.
- But the future here probably lies in **electronic corpora**.
- One method: give a native speaker a part-time job: "please type into the computer all the paradigms you know".
  - ➢ The beauty of this is, "elicit once, analyze forever"
  - ➢ Turkish Electronic Living Lexicon (http://linguistics.berkeley.edu/TELL/)
  - ➢ I've done this to some extent with Hungarian and Italian.
- Another method (Hayes/Zuraw readings): we **synthesized** Hungarian paradigms by rule (based on a grammar, and allowing multiple variants), then searched for our synthesized forms in a very large electronic text corpus.

29. Traditional approaches to type variation

- Traditional generative grammar typically designated the most frequent pattern as regular, and attached diacritics (e.g., [–Rule x]) to the exceptions.
- Occasionally (see Brame and Bordelois[6] vs. Harris[7], 1970's), it became quite controversial which patterns really were the regular ones.

30. Moving beyond the lexicon to wug testing

- Zuraw, Kie (2000) *Patterned exceptions in phonology*, UCLA dissertation.
- and a variety of work following it; see literature review in this week's reading

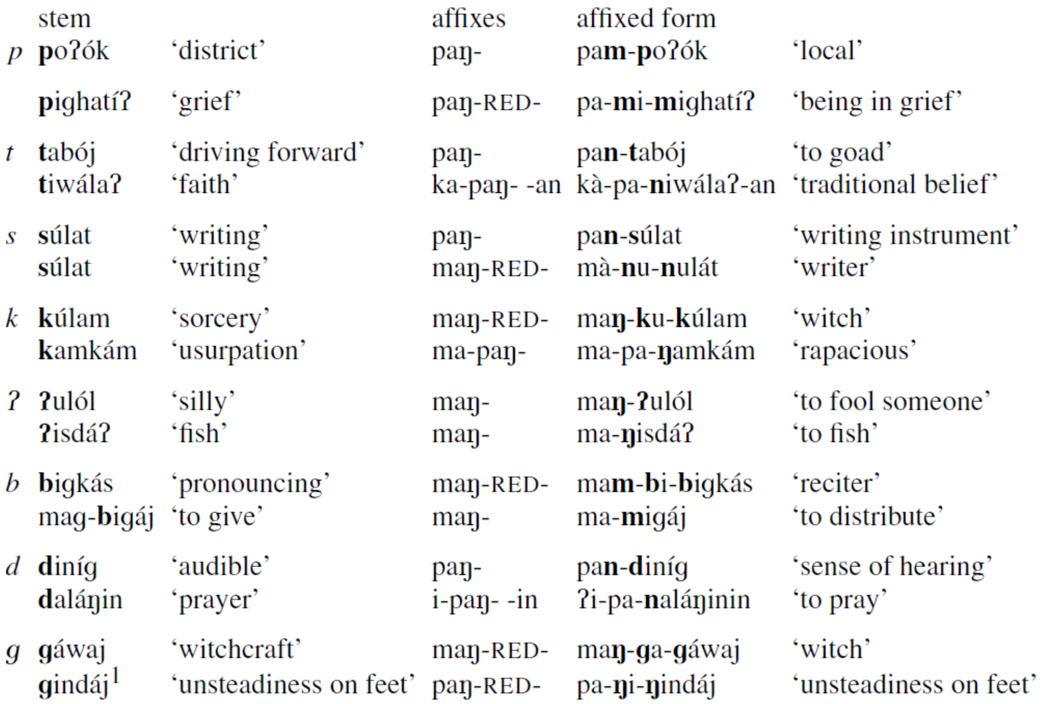
[5] I've read this somewhere and can't recall where — probably Jeffrey Heath.

[6] Brame, Michael K. and Ivonne Bordelois. 1973. Vocalic alternations in Spanish. *Linguistic Inquiry* 4, 111–168.

[7] Harris, James W. (1974) On Certain Claims concerning Spanish Phonology, *Linguistic Inquiry* Vol. 5, No. 2 (Spring, 1974), pp. 271-282

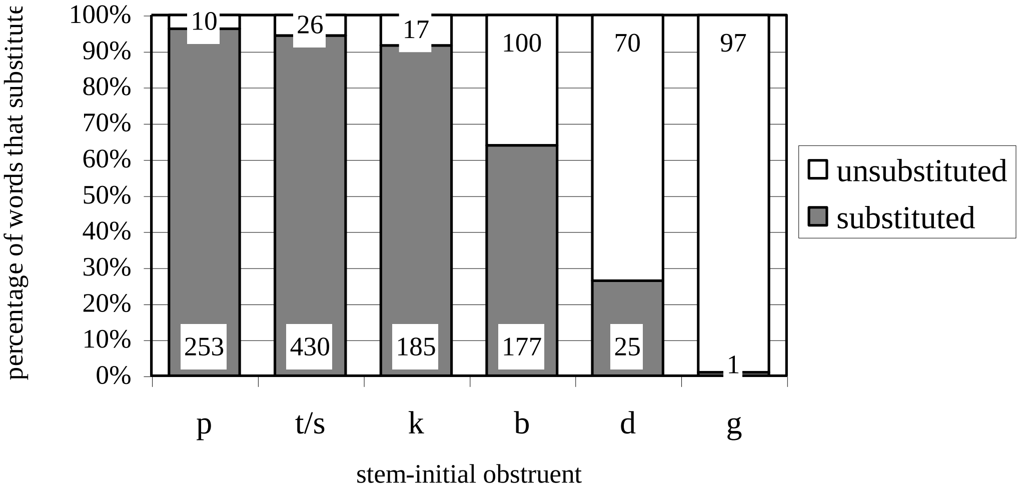
31. Zuraw (2000): lexical study of percent application of Nasal Substitution in Tagalog: N+obstruent → {m,n,ŋ}

mag-bigáj ‘give’, but
/maŋ-bigáj/ → mamigáj ‘distribute’



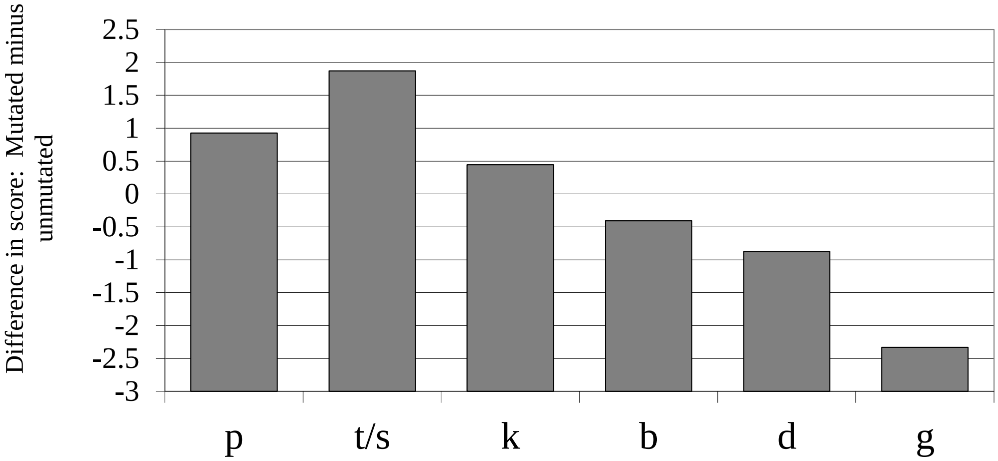
	stem		affixes	affixed form	
p	**po**ʔók	‘district’	paŋ-	pa**m**-**po**ʔók	‘local’
	pighatíʔ	‘grief’	paŋ-RED-	pa-**mi**-**m**ighatíʔ	‘being in grief’
t	**t**abój	‘driving forward’	paŋ-	pa**n**-**t**abój	‘to goad’
	tiwálaʔ	‘faith’	ka-paŋ- -an	kà-pa-**n**iwálaʔ-an	‘traditional belief’
s	**s**úlat	‘writing’	paŋ-	pa**n**-**s**úlat	‘writing instrument’
	súlat	‘writing’	maŋ-RED-	mà-**nu**-**n**ulát	‘writer’
k	**k**úlam	‘sorcery’	maŋ-RED-	ma**ŋ**-**ku**-**k**úlam	‘witch’
	kamkám	‘usurpation’	ma-paŋ-	ma-pa-**ŋ**amkám	‘rapacious’
ʔ	**ʔ**ulól	‘silly’	maŋ-	ma**ŋ**-**ʔ**ulól	‘to fool someone’
	ʔisdáʔ	‘fish’	maŋ-	ma-**ŋ**isdáʔ	‘to fish’
b	**b**igkás	‘pronouncing’	maŋ-RED-	ma**m**-**bi**-**b**igkás	‘reciter’
	mag-**b**igáj	‘to give’	maŋ-	ma-**m**igáj	‘to distribute’
d	**d**iníg	‘audible’	paŋ-	pa**n**-**d**iníg	‘sense of hearing’
	daláŋin	‘prayer’	i-paŋ- -in	ʔi-pa-**n**aláŋinin	‘to pray’
g	**g**áwaj	‘witchcraft’	maŋ-RED-	ma**ŋ**-**ga**-**g**áwaj	‘witch’
	gindáj[1]	‘unsteadiness on feet’	paŋ-RED-	pa-**ŋi**-**ŋ**indáj	‘unsteadiness on feet’

32. Frequency of Nasal Substitution varies in the lexicon according to the stem-initial consonant



33. Native speakers are tacitly aware of this pattern

- Again Zuraw, a "wug" test (following Berko 1958). Preference for the nasally-mutated form (difference between both options, each rated on 1-10 scale)



34. The Law of Frequency Matching

- Hayes, Zuraw et al. (2009) go for broke in their rhetoric:

Speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies.

- Some other phonological experiments whose results support this law are reported in Eddington (1996, 1998, 2004), Berkley (2000), Coleman and Pierrehumbert (1997), Zuraw (2000), Bailey and Hahn (2001), Frisch and Zawaydeh (2001), Albright (2002), Albright and Hayes (2003), Pierrehumbert (2006), and Jun and Lee (2007), Moore-Cantwell (2021).
- Sociolinguistic study demonstrates frequency matching by children during real-life phonological acquisition (Labov 1994, Ch. 20).

35. The Law in (much) broader perspective

- Frequency-matching is known to be a common ability in animals (Gallistel 1990, ch. 11); and in humans for nonlinguistic tasks (Hasher and Zacks 1984).

MODELING TYPE VARIATION

36. Zuraw's theory (2000, 2010): the dual listing/generation model

- Words are memorized—even inflected ones—as they are heard.
- Psycholinguistic work has strongly supported a **huge capacity for word memorization** in humans (contra early generative phonology, which emphasized data compression)
    - ➢ Baayen, Harald, Robert Schreuder, Nivja De Jong, and Andrea Krott "Dutch inflection: The rules that prove the exception," in Sieb Nooteboom, Frank Wijnen and Fred Weerman (eds.), *Storage and computation in the language faculty* (2002, Kluwer)
    - ➢ See Baayen's web site for further work
    - ➢ Basic argument: recognition-ease or speed of fully-inflected forms is dependent on the frequency of the form itself, not its morphological base or paradigm as a whole.
- Back to Zuraw: claims that there is hard analytic work as well as memorization: a stochastic grammar is created from the data — treating them *as if* they were free variation data.
- I.e.: memorize, but be ready to project.
- If you *have* a listed form, you generally use it: USE LISTED

37. Zuraw describes a near-optimal human

- Memorization is just great for producing irregulars accurately.
- Children's memorization capacity is strong but not unlimited.
- So grammar-based back-up is sensible too.

- … and a grammar is essential for production and understanding[8] of novel forms.

38. An alternative: constraint cloning theory

- When you hit a ranking contradiction, make a copy of the relevant Faithfulness constraint, indexing it to the words that are lexically allowed to be more marked.
- Hence the grammar encodes the exceptionality directly.
- … and is non-stochastic
- New forms must be projected — somehow — from the populations of existing forms that violate the various Faithfulness constraints.
- References:
  - Pater, Joe. 2000. Nonuniformity in English stress: the role of ranked and lexically specific constraints. *Phonology* 17:2. 237-274.
  - Pater, Joe. 2009. Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution. In Steve Parker, (ed.) *Phonological Argumentation: Essays on Evidence and Motivation*. London: Equinox.
  - Becker, Michael 2008. *Phonological trends in the lexicon: the role of constraints*. http://becker.phonologist.org/papers/becker_dissertation.pdf

[8] When I first heard [mɪdˈwɪfəɹi] for *midwifery*, I was extremely surprised but knew exactly what was meant, since it is the output of Trisyllabic Shortening (cf. *divine* ~ *divinity*).