The Locus of Variation in Weighted Constraint Grammars

Karen Jesney
University of Massachusetts Amherst
kjesney@linguist.umass.edu

Workshop on Variation, Gradience and Frequency in Phonology, Stanford, CA
Poster session, July 6-8, 2007

1. Background

In most OT work on variation, the output is made to vary from utterance to utterance by varying the constraint ranking (cf. Coetzee 2006). Two ways of doing this:

- Specific constraint values are selected from a normal distribution around a mean with each iteration of EVAL (e.g., Boersma 1998, Boersma & Hayes 2001)
- A full ranking consistent with a partially-stratified hierarchy is selected with each iteration of EVAL (e.g., Anttila 1997, 2002, Kiparsky 1993, Reynolds 1994)

Numeric constraint values have no real meaning in OT. As long as A >> B >> C, the same candidate will be selected as optimal; for a different variant to emerge, the constraint ranking must change.

(1)

/input1/	A = 4	B = 2	C =1
☞ candidate1-a		*	*
candidate1-b	*!		

/input1/	A = 6	B = 4	C =3
☞ candidate1-a		*	*
candidate1-b	*!		

Because the *ordinal* ranking of the constraints is the same in both tableaux, the same candidate is optimal in both cases.

Weighted constraint systems differ from ranked constraint systems by making direct reference to *numeric values* associated with constraints – i.e., to their *weights*.

Here I consider two ways of determining optima and incorporating variation in weighted constraint systems:

- ***"Noisy HG"*** – Linear additive constraint interaction as in Harmonic Grammar (Legendre, Miyata & Smolensky 1990, Smolensky & Legendre 2006) with added noise (per Boersma 1998; in HG, see Boersma & Pater 2007, Pater, Bhatt & Potts 2007)
- ***"MaxEnt/OT"*** – Log-linear constraint interaction as in Maximum Entropy OT (Goldwater & Johnson 2003, Jäger to appear, Jäger & Rosenbach 2006)

Thanks to Gaja Jarosz, John McCarthy, Joe Pater and Matt Wolf for ideas and feedback on various aspects of this project. Thanks also to Paul Boersma for assistance with Praat. This work was supported in part by SSHRC Doctoral Fellowship #752-2005-1708.

While Noisy HG and MaxEnt/OT systems can capture many of the same patterns, they make different predictions about the types of variation patterns that are possible.

2. Noisy HG

Constraint violations are treated as penalties. The harmony score of a candidate *R* is determined by multiplying its violations of each constraint $\{C_1(R),C_2(R),\ldots,C_n(R)\}$ by the weights associated with those constraints $\{w_1,w_2,\ldots,w_n\}$ and summing. The candidate with the highest score is deemed optimal.

(2) $H(R) = (C_1(R)*w_1) + (C_2(R)*w_2) + \ldots + (C_n(R)*w_n)$

Following Prince (2002) and Pater, Bhatt & Potts (2007), constraint weights are limited to positive real numbers.

This system allows cumulativity effects to emerge; multiple lower-weighted constraints can overcome the pressure of a higher-weighted constraint.

(3)

/input1/	A = 4	B = 2	C =1	*H*
☞ candidate1-a		-1	-1	-1(2) + -1(1) = **-3**
candidate1-b	-1			-1(4) = **-4**

/input1/	A = 6	B = 4	C =3	*H*
candidate1-a		-1	-1	-1(4) + -1(3) = **-7**
☞ candidate1-b	-1			-1(6) = **-6**

In (3), Candidate1-a is optimal when *w*A > *w*B + *w*C, and candidate1-b is optimal when *w*B + *w*C > *w*A. The *ordinal* relationship of the constraints has not changed.

Variation is introduced in Noisy HG using GLA-style noise.

- With each iteration of EVAL, a specific value for each constraint is selected from a normal distribution around its mean. This can alter the relevant inequalities, introducing variation.
- Unlike in Stochastic OT (Boersma 1998), the ordinal relationship of the constraints need not necessarily change; changing the inequalities is sufficient.

The locus of variation in Noisy HG is the *constraint values*. Variation emerges over multiple applications of EVAL.

Noisy HG systems can reproduce both categorical and variable patterns.

- Both types of pattern can be learned using the error-driven learner in Praat (Boersma & Weenink 2007; PositiveHG decision strategy; 2.0 noise).

Categorical pattern: The Wolof tongue-root grammar (Boersma 1999) was learned with average **99.997%** accuracy over 10 trials.

Variation Pattern: The Finnish genitive plurals pattern (Anttila 1997, 2002, Boersma & Hayes 2001) was learned with the average % accuracy shown in (4) over 10 trials (columns are ERC patterns – see Goldwater & Johnson 2003):

(4)

	1	2	3	4	5	6	7	8
Target % light ending:	0	100	100	79.0	100	0.2	81.1	49.5
Noisy HG % light ending:	0	100	100	69.4	100	1.0	82.1	49.0

Which output emerges as optimal on a given iteration of EVAL depends upon the weighting conditions that hold once the constraint values have been selected. It is the *weighting conditions* (e.g. *w*A > *w*B+*w*C vs. *w*B+*w*C > *w*A) that are key, not the raw Harmony scores.

3. MaxEnt/OT

Constraint violations are treated as penalties and the summed weighted score (i.e., *H*) is calculated for each candidate. The probability of each candidate is then determined by taking the exponent of its score (e^H) and dividing this by the sum of the exponents of the scores for the full candidate set ($e^{H1}+e^{H2}+\ldots.+e^{Hn}$).

(5) Given a candidate set Y, Probability($cand_j$) = $$\frac{e^{Hj}}{\sum_{cand_X \in Y} e^{Hx}}$$

(6)

/input1/	A = 4	B = 2	C =1	*H*	e^H	***p***
candidate1-a		-1	-1	-3	0.0498	**0.73**
candidate1-b	-1			-4	0.0183	**0.27**

/input1/	A = 6	B = 4	C =3	*H*	e^H	***p***
candidate1-a		-1	-1	-7	0.0009	**0.27**
candidate1-b	-1			-6	0.0025	**0.73**

Probabilities vary depending upon the raw Harmony scores of the candidates.

The locus of variation is in the *candidate probabilities computed by the grammar*. EVAL yields probabilities directly using a single set of constraint values and no noise.

As with Noisy HG, this system allows cumulativity effects to emerge and can reproduce both categorical and variable patterns.

- Both types of pattern can be learned using the error-driven learner in Praat (MaximumEntropy decision strategy; 0.0 noise).

Categorical pattern: The Wolof tongue-root grammar was learned with average **99.996%** accuracy over 10 trials.

Variation Pattern: The Finnish genitive plurals pattern was learned with the average % accuracy shown in (7) over 10 trials:

(7)

	1	2	3	4	5	6	7	8
Target % light ending:	0	100	100	79.0	100	0.2	81.1	49.5
Noisy HG % light ending:	0.4	100	100	70.1	99.8	1.8	80.3	44.5

Candidates' probabilities are directly calculated based on their Harmony scores. Weighting conditions do not determine optimal outputs.

4. Difference 1 - Harmonic Bounding

In MaxEnt/OT, *all* candidates – including harmonically-bounded candidates (marked with ☠) – always receive some portion of the probability mass.

- This probability can be trivial (<0.00000001), or it can be fairly substantial.

(8)

/CV/	*COMPLEX = 2	NOCODA = 1	FAITH = 1	*H*	e^H	***p***
CV				0	1	**0.84**
☠ CVC		-1	-1	-2	0.135	**0.11**
☠ CCV	-1		-1	-3	0.050	**0.04**
☠ CCVC	-1	-1	-2	-5	0.007	**0.005**

The MaxEnt/OT system readily learns a variation pattern based on this distribution including harmonically-bounded candidates; the Noisy HG system does not.

(9)

		MaxEnt /OT	Noisy HG			MaxEnt /OT	Noisy HG
CV	**→ CV**	84	100	**CVC**	**→ CV**	47.6	50.1
☠	**→ CVC**	11.4	0		**→ CVC**	47.8	49.9
☠	**→ CCV**	4	0	☠	**→ CCV**	2.3	0
☠	**→ CCVC**	0.6	0	☠	**→ CCVC**	2.4	0
CCV	**→ CV**	64.7	68.9	**CCVC**	**→ CV**	36.7	31.6
☠	**→ CVC**	17.3	0		**→ CVC**	36.8	34.5
	→ CCV	23.3	31.1		**→ CCV**	13.2	15.5
☠	**→ CCVC**	3.1	0		**→ CCVC**	13.3	15.5

Noisy HG, like Standard HG, never allows simply harmonically-bounded candidates to win, assuming all weights are positive (see Prince 2002).

Why? By definition, simply harmonically-bounded candidates incur a proper superset of the violations of some other candidate.

- E.g., /CVC/→[CCVC] violates NoCoda and Faith, and so it is bounded by /CVC/→[CVC], which violates only NoCoda. No matter what values are selected, it will always be the case that:

  *w*NoCoda + *w*Faith > *w*NoCoda

  Weighting conditions determine optimality in Noisy HG, and so /CVC/→[CCVC] can never win.

MaxEnt/OT calculates probabilities directly, without reference to weighting conditions. Harmonically-bounded candidates are treated like all other candidates.

Positional variation poses a different challenge. Candidates that resolve the same marked structure differently at different loci are normally *collectively harmonically bounded*; however, they are clearly attested (Jäger & Rosenbach 2006, Riggle & Wilson 2005, Vaux 2003).

- MaxEnt/OT systems readily allow collectively-bounded candidates to emerge.

- HG allows collectively-bounded candidates to prove optimal, but only under certain circumstances (Prince 2002). In cases of positional variation, collectively-bounded candidates normally require inconsistent weighting conditions and so are ruled out.

(10)

/ãvid**ə**t**ə**batʁ/		*Schwa	Max-V
a.	ãvid**ə**t**ə**batʁ	-2	
b. ☠	ãvidt**ə**batʁ	-1	-1
c.	ãvidtbatʁ		-2

French variable schwa deletion (based on Riggle & Wilson 2005)

For (10b) to beat (10a) in HG, it must the case that ***w**Schwa > *w*Max-V**
For (10b) to beat (10c) in HG, it must be the case that ***w*Max-V > *w**Schwa**

This typological gap can be resolved in Noisy HG using positional constraints as proposed by Riggle & Wilson (2005), or by making assessment local and noisy (Pater, Bhatt & Potts 2007).

MaxEnt/OT is a *less restrictive* theory than Noisy HG. In MaxEnt/OT, harmonically-bounded candidates readily receive some portion of the probability mass; there is no clear way to disentangle the desirable collectively-bounded candidates from the undesirable simply-bounded ones.

5. Difference 2 - Process Interaction

In MaxEnt/OT, the likelihoods of variable processes applying are always independent of one another – even when the sets of constraints governing them overlap.

In Noisy HG, the likelihoods of variable processes applying are influenced by one another when the processes share some subset of relevant constraints.

(11) Onset cluster simplification: *w**Complex > ***w*Max**
Coda consonant deletion: *w*NoCoda > ***w*Max**

I trained the Noisy HG and MaxEnt/OT systems on the inputs /CCV/ and /CVC/, with the following input-output distribution:

- Onset cluster simplification: /CCV/→[CCV] 50 /CCV/→[CV] 50
  Coda consonant deletion: /CVC/→[CVC] 50 /CVC/→[CV] 50

In both systems, the three constraints must all have the same value for these patterns to obtain.

- In MaxEnt/OT, for /CCV/→[CCV], which violates *Complex, and /CCV/→[CV], which violates Max, to have equal probability, they must have the same *H* value. The weights of *Complex and Max must therefore be equal. The same holds of NoCoda and Max, based on /CVC/→[CVC] and /CVC/→[CV]. By transitivity, then, all three constraints must have the same value.

- In Noisy HG, for /CCV/→[CCV] and /CCV/→[CV] to have equal probability, it must be the case that *w*Max > *w**Complex and *w**Complex > *w*Max are equally likely. This can only occur if Max and *Complex have the same mean. The same holds of NoCoda and Max, based on /CVC/→[CVC] and /CVC/→[CV]. By transitivity, then, all three constraints must have the same value.

To see how the resulting grammars generalize, I gave them input /CCVC/, where both processes could apply. The average result over 10 trials is given in (12).

(12)

	MaxEnt/OT	Noisy HG
/CCVC/→[CCVC] (neither process)	24.4	33.2
/CCVC/→[CVC] (cluster simplification)	25.1	16.8
/CCVC/→[CCV] (coda deletion)	24.8	16.5
/CCVC/→[CV] (both processes)	25.7	33.3

In **MaxEnt/OT**, there is no dependence between the two processes; all output candidates are equally likely.

- **Why?** Each candidate incurs two marks, and, because all constraints are equally weighted here, the resulting *H* value for each candidate is the same. Candidates with the same harmony scores have equal probability in MaxEnt/OT.

(13)

/CCVC/	*COMPLEX =100	NOCODA =100	MAX =100	H	e^{H}	*p*
CV			-2	-200	1.38^{-87}	**0.25**
CVC		-1	-1	-200	1.38^{-87}	**0.25**
CCV	-1			-200	1.38^{-87}	**0.25**
CCVC	-1	-1		-200	1.38^{-87}	**0.25**

In **Noisy HG**, it is most likely that neither process will apply or that both processes will apply.

- **Why?** If all constraints have the same value, each of the 6 (3!) possible weighting conditions is equally likely. Two of the weighting conditions favour /CCVC/➔[CCVC], and two favour /CCVC/➔[CV]. Only one favours each of the other mappings.

(14)

***w*MAX** > *w**COMPLEX > *w*NOCODA	/CCVC/ ➔ [CCVC]
***w*MAX** > *w*NOCODA > *w**COMPLEX	/CCVC/ ➔ [CCVC]
*w**COMPLEX > *w*NOCODA > ***w*MAX**	/CCVC/ ➔ [CV]
*w*NOCODA > *w**COMPLEX > ***w*MAX**	/CCVC/ ➔ [CV]
*w**COMPLEX > ***w*MAX** > *w*NOCODA	/CCVC/ ➔ [CVC]
*w*NOCODA > ***w*MAX** > *w**COMPLEX	/CCVC/ ➔ [CCV]

Both MaxEnt/OT and Noisy HG systems predict that variable processes that are governed by *complementary* sets of constraints will pattern independently.

MaxEnt/OT predicts that processes will always pattern independently. Noisy HG predicts that related processes will pattern together to at least some extent. The two theories thus impose *different restrictions* upon the interaction of variable processes; these should be empirically tested.

6. Summary

In MaxEnt/OT, the grammar yields candidate probabilities directly using a single set of weights. The probability of a variant is computed by the grammar based on its *H* score and the summed *H* score of the full candidate set. This system allows harmonically-bounded candidates to emerge, and predicts that the likelihoods of variable processes will always be independent of one another, even when they share relevant constraints.

In Noisy HG, variation emerges through perturbation of constraint weights over multiple iterations of EVAL. The probability a variant is based on the probability that the associated weighting conditions will hold. This system generally prevents harmonically-bounded candidates from emerging, and predicts that the likelihoods of variable processes will be interdependent if they share relevant constraints.

7. References

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout and W. L. Wetzels (eds.), *Variation, Change and Phonological Theory.* Amsterdam: John Benjamins.

Anttila, Arto. 2002. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory* 20: 1-42.

Boersma, Paul. 1998. *Functional Phonology: Formalizing the Interactions Between Articulatory and Perceptual Drives*. PhD dissertation. Amsterdam: University of Amsterdam.

Boersma, Paul. 1999. Optimality-theoretic learning in the Praat program. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 23, 17-35.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45-86. [ROA-348].

Boersma, Paul & Joe Pater. 2007. Testing gradual learning algorithms. Ms, University of Amsterdam and University of Massachusetts Amherst.

Boersma, Paul & David Weenink. 2007. Praat: Doing Phonetics by Computer v.4.6.08. [www.praat.org].

Coetzee, Andries W. 2006. Variation in accessing "non-optimal" candidates – a rank ordering model of EVAL. Ms., University of Michigan. [ROA-863].

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Workshop on Variation within Optimality Theory*, 111-120. Stockholm University.

Kiparsky, Paul. 1993. An OT Perspective on phonological variation. Paper presented at *Rutgers Optimality Workshop*.

Reynolds, William T. 1994. *Variation and Phonological Theory*. PhD dissertation. University of Pennsylvania.

Jäger, Gerhard. to appear. Maximum entropy models and stochastic Optimality Theory. In Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson & Annie Zaenen (eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford: CSLI.

Jäger, Gerhard & Anette Rosenbach. 2006. The winner takes it all - almost: cumulativity in grammatical variation. *Linguistics* 44(5): 937-971.

Legendre, Geraldine, Yoshiro Miyata & Paul Smolensky. 1990. Harmonic Grammar – a formal multi-level connectionist theory of linguistic wellformedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Cambridge, MA: Lawrence Erlbaum.

Pater, Joe, Rajesh Bhatt & Chris Potts. 2007. Linguistic Optimization. Ms., University of Massachusetts Amherst.

Prince, Alan. 2002. Anything goes. In Takeru Honma, Masao Okazaki, Toshiyuki Tabata & Shinichi Tanaka (eds.), *New Century of Phonology and Phonological Theory*, 66–90. Tokyo: Kaitakusha. [ROA-536].

Riggle, Jason & Colin Wilson. 2005. Local optionality. In Leah Bateman & Cherlon Ussery (eds.), *Proceedings of NELS 35.* Amherst, MA: GLSA.

Smolensky, Paul & Geraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press.

Vaux, Bert. 2003. Why the phonological component must be serial and rule-based. Paper presented at the *77th Meeting of the Linguistic Society of America*.