Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 43-58

HOW WE LEARN VARIATION, OPTIONALITY, AND PROBABILITY*

Paul Boersma

Abstract

Variation is controlled by the grammar, though indirectly: it follows automatically from the robustness requirement of learning. If every constraint in an Optimality-Theoretic grammar has a ranking value along a continuous scale, and the disharmony of a constraint at evaluation time is randomly distributed about this value, the phenomenon of optionality in determining the winning candidate follows automatically from the finiteness of the difference between the ranking values of the relevant constraints; the degree of optionality is a descending function of this difference.

In the production grammar, a symmetrized maximal gradual learning algorithm will cause the learner to copy the degrees of optionality from the language environment. In the perception grammar, even the slightest degree of noise in constraint evaluation will cause the learner to become a probability-matching listener, whose categorization distributions match the production distributions of the language environment. Evidence suggests that natural learners follow a symmetric demotion-and-promotion strategy, rather than a demotion-only strategy.

A typical example of optionality in speech production is place assimilation of nasals at the sentence level, i.e. a word underlyingly ending in |an| and a word starting with |pa| may, when concatenated, be pronounced either as [anpa] or as [ampa]. This poses a problem for a theory with fixed relative constraint rankings, like the original version of Optimality Theory (Prince & Smolensky 1993).

Let's say that the relevant constraints for our example are *GESTURE (tongue tip: close & open) and *REPLACE (place: coronal, labial / nasal / _C), i.e., the choice between [anpa] and [ampa] is the outcome of a struggle between the importance of not performing a tongue-tip opening-and-closing gesture and the importance of honouring an underlying specification for the value [coronal] on the perceptual place tier as conditioned by a nasal environment before a consonant (Boersma 1997). The candidate [anpa] would win if the ranking were *REPLACE (cor) >> *GESTURE (tip):

|an+pa|	*REPLACE (cor)	*GESTURE (tip)
☞ [anpa] /anpa/		*
[ampa] /ampa/	*!	

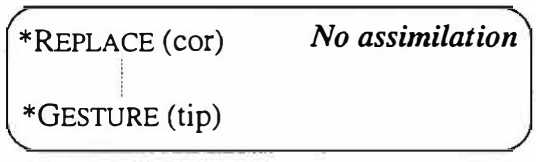
(1)

A short explanation of the notation may be appropriate. According to the ideas of Functional Phonology, the gestural constraint evaluates the articulatory candidate [anpa], and the faithfulness constraint evaluates the difference between the

* An earlier version of this paper appeared as Rutgers Optimality Archive #221.

underlying perceptual specification |an+pa| and the output /anpa/, which is the acoustic result of [anpa] as perceived by the listener; the similarities between these forms are deceptive: the brackets contain shorthand notations for articulatory events, the slashes contain shorthands for perceptual features.

If (1) were the only possible outcome, we could describe it with the following grammar (the dotted line depicts a language-specific crucial ranking):



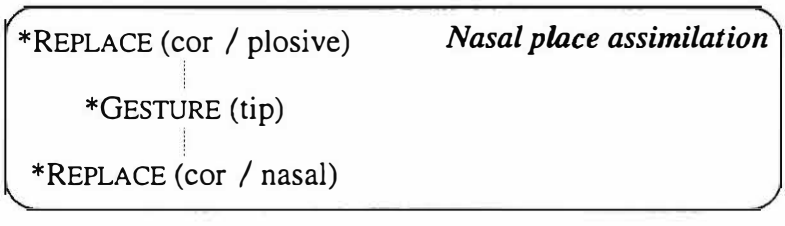
(2)

With the reverse ranking, [ampa] would win:

|an+pa|	*GESTURE (tip)	*REPLACE (cor)
[anpa] /anpa/	*!	
☞ [ampa] /ampa/		*

(3)

With the following grammar, nasals would assimilate, but plosives would not:



(4)

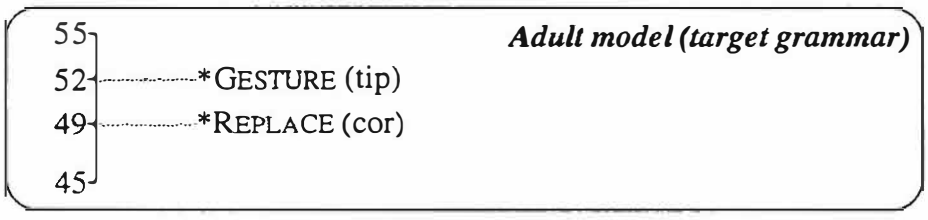
If place assimilation is optional, and if we cannot have both grammars (2) and (4) at the same time, then we have a problem.

One possibility would be to rank *REPLACE (cor) and *GESTURE (tip) equally high. We should then not follow the suggestion by Tesar & Smolensky (1993), who interpret equal ranking in such a way that the violation marks incurred by the two constraints are capable of cancelling each other. Rather, we should interpret equal ranking in a probabilistic manner: if the two constraints are in conflict, either of them could win at evaluation time, both with a probability of 50% (Anttila 1995). However, real life learns us that optionality is often gradient, e.g., one form may occur in 80%, the other in 20% of the cases, and these numbers differ between neighbouring dialects. The proposal of the current paper shows that a continuously ranking OT grammar can maintain any degree of optionality, that speakers will learn to reproduce the degree of optionality of their language environment, and that listeners will learn to match the degree of optionality of their language environment in their categorization systems.

1. Continuous ranking scale and stochastic disharmony

Our optionality problem is solved by a random stochastic element in constraint evaluation (in §2.4, we will see that this random element is independently needed to implement the robustness requirement of a natural language user's learning strategy).

If place assimilation occurs more often than not, we say that *GESTURE (tip) is ranked higher than *REPLACE (cor) along a continuous scale (whose physiological correlate could be synaptic strength), with a real number attached to each constraint:



(5)

In this example, the ranking value of *REPLACE (cor) is 49, and the ranking value of *GESTURE (tip) is 52. In the absence of stochastic evaluation these values would determine the order of the constraints in an evaluation tableau, in which case this ranking would be equivalent to grammar (4). However, with stochastic evaluation (whose physiological correlate could be the noise in the amount of locally available neurotransmitter), this order is determined by the *disharmonies* ("effective" rankings) of the constraints, which are determined at evaluation time from the ranking value and a random variable:

$$disharmony = ranking + rankingSpreading \cdot \mathbf{z} \tag{6}$$

where $\mathbf{z}$ is a Gaussian random variable with mean 0 and standard deviation 1. For instance, a simulation of ten implementations of |an+pa| with a *rankingSpreading* of 2 yielded the following disharmonies:

trial	1	2	3	4	5	6	7	8	9	10
*GESTURE disharmony	50.5	51.2	50.2	49.1	52.9	52.9	52.7	53.8	55.4	54.3
*REPLACE disharmony	50.8	48.3	50.7	51.2	48.9	48.8	48.2	50.3	48.1	48.7
outcome	np	mp	np	np	mp	mp	mp	mp	mp	mp

(7)

We see that in most replications, *GESTURE (tip) was evaluated as higher than *REPLACE (cor), but that *REPLACE (cor) was higher in three of the ten cases. Thus, our simulated speaker would have said [ampa] seven times, and [anpa] three times. The distribution of the disharmony difference between two constraints C_1 and C_2 with rankings r_1 and r_2 is given by

$$disharmony_1 - disharmony_2 = r_1 - r_2 + rankingSpreading \cdot (\mathbf{z}_1 - \mathbf{z}_2) \tag{8}$$

Now if both $\mathbf{z}_1$ and $\mathbf{z}_2$ are Gaussian distributions with standard deviations of 1, their difference $\mathbf{z}_1 - \mathbf{z}_2$ is also Gaussian, with a standard deviation of $\sqrt{2}$, so that the probability that C_1 is evaluated higher than C_2 is

$$P(disharmony_1 > disharmony_2) = \tfrac{1}{2} \cdot \left(1 - erf\left(\tfrac{1}{2}\sqrt{2} \cdot \frac{r_1 - r_2}{rankingSpreading \cdot \sqrt{2}}\right)\right) \tag{9}$$

which for a ranking spreading of 2 can be tabulated as

$r_1 - r_2$	0	1	2	3	4	5	6	7	8	9	10	11	12
P	1/2	36%	24%	14%	7.9%	3.9%	1.7%	0.7%	0.2%	$7 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$1 \cdot 10^{-5}$

(10)

So our speaker will say [anpa] 14% of the times. If the ranking difference is less than 10 (or so), we may talk of *optionality*; if it is greater, of *obligation*. The optionality may still divide into *variation* (for distances below, say, 7) and *error*, though these subjective labels will generally be assigned with more criteria than rate of occurrence alone.

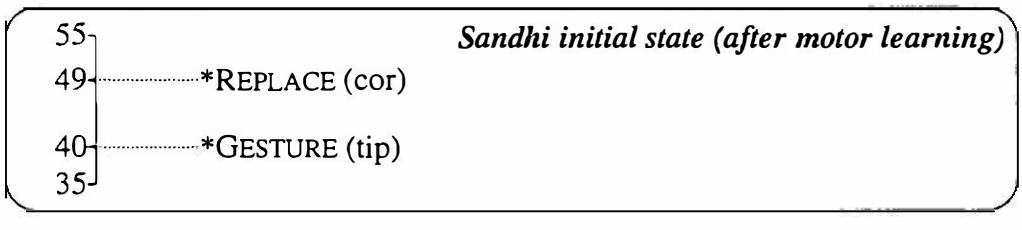
In Boersma (to appear), I show that the continuous ranking scale allows some very simple and robust gradual learning algorithms, and that the current idea of optionality leads to a realistic learning curve. In this paper, I will show that optionality in the production grammar can be learned and that the listener's categorization system automatically adapts to asymmetries in the distributions of variations in production. Finally, I present a shallow proof of the correctness of the "maximal gradual" algorithm for learning stochastic grammars.

2. Learning optionality in production

In this section, I will show that if adults exhibit place assimilation of nasals in 86% of all cases, like with grammar (5), then their children will copy this degree of optionality in their production grammars.

2.1. Learning that faithfulness can be violated in an adult grammar

At four years of age, Dutch children tend to pronounce |an+pa| faithfully as [anpa], though their parents probably say [ampa] most of the time. This is a natural stage in phonological development: the underlying form ends in |-an|, which the learner can easily deduct from the form as spoken in isolation. Because the child perceives her own form [anpa] as /anpa/, no faithfulness constraint is violated. In fact, earlier stages in learning have centred around acquiring all the gestures necessary for implementing the perceptual contrasts of the language, and the adult form, as perceived by the learner, has always been taken to be the underlying specification, with respect to which she evaluates the faithfulness constraints. Thus, the child's grammar is something like



(11)

The next step in phonological development is to learn that faithfulness constraints can be violated: the separation between perceived and underlying forms can begin. The learner will notice that she says /anpa/, but that adults sometimes say /ampa/. The discrepancy within this *learning pair* is shown in the following tableau:

[ampa] /ampa/ |an#pa|	*REPLACE (place: cor)	*REPLACE (place / nas)	*REPLACE (place / _C)	*GESTURE (lip)	*GESTURE (tip)
☞ [anpa] /anpa/				*	*
√ [ampa] /ampa/	*!	*	*	*	

(12)

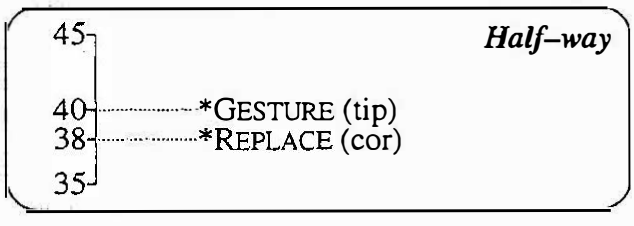
In this tableau, the top left shows the adult production [ampa] and the child's perception of it: /ampa/. Her own production is [anpa], which she perceives as /anpa/. This is the winner of the evaluation, as shown by the pointing finger (☞). However, the learner knows that /ampa/ should have been the winner, and she has

already learned in an earlier stage that she can implement that by saying [ampa]. Therefore, the row with the check mark (√) shows the correct, but losing candidate. Something will have to be done. The learner will take a *learning step.*

2.2. The minimal gradual learning step: demoting the offender

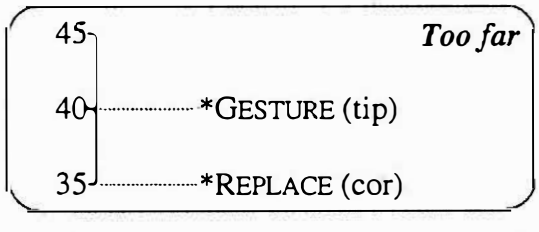
The offending incorrectly ranked constraint is the one with the crucial violation (the exclamation mark) in the evaluation of the correct candidate [ampa] (in the row with the check mark). This offending constraint is *REPLACE (cor). A simple strategy that will eventually prevent the mistake from reoccurring after a number of errors, is to *demote the offender*, i.e., to lower the ranking of *REPLACE (cor) by a small amount (e.g. a step of 0.01) along the continuous ranking scale. In Boersma (to appear), I show that with this and related strategies (*gradual learning algorithms*) any target constraint ranking can be learned within a reasonable time, independently of the initial rankings of the constraints.

Demotion will proceed until the ranking of *REPLACE (cor) is below *GESTURE (tip). But suppose that at a certain moment in time, the ranking is already as follows:



(13)

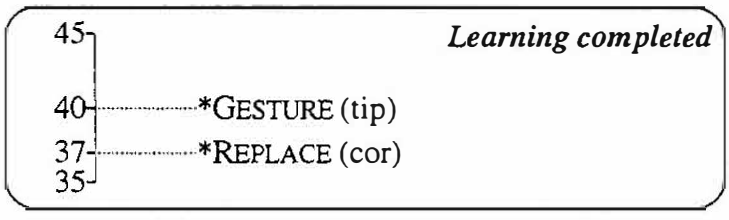
According to table (10), the probability that a subsequent learning pair will contain an adult model /ampa/ and a learner's utterance /anpa/, is still 86%·24% = 21%, and such a case will lead to a further demotion of *REPLACE (cor); the probability that the adult model is /anpa/ and the learner's utterance is /ampa/, is 14%·76% = 11%, and such a case would lead to demotion of *GESTURE (tip). Thus, even now that faithfulness has fallen below the gestural constraint, there will still be more demotions of *REPLACE than of *GESTURE, and this will raise the difference between the ranking values even further. However, if the ranking difference becomes large, there will be more demotions of *GESTURE than of *REPLACE:



(14)

In this case, a demotion of *GESTURE will occur in only 86% · 3.9% = 3.3% of the cases, and a demotion of *REPLACE in 14% · 96.1% = 14% of the cases. The net result is that the two constraints will get closer.

A state of stable equilibrium will be reached when the ranking difference has become such that the demotion probabilities of *GESTURE and *REPLACE are equal, i.e., when they are 86%·14% and 14%·86%, respectively. This, of course, occurs when the ranking difference is 3, as in the adult grammar:



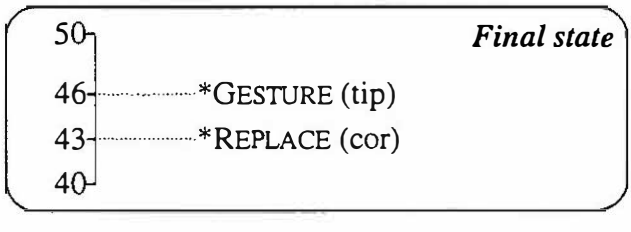
(15)

Thus, stochastically evaluating learners acquire not only the adult ranking order, but also the adult ranking differences and, therefore, the adult degree of optionality in production. In §3, we will see that for a demotion-only learner, this result is valid only if there are only two interacting constraints.

2.3. A remedy for downdrift of constraint pairs: symmetric demotion & promotion

Optionality causes a problem for demotion-only learning. Considered as a whole, the grammar is not very stable, because the finite error probabilities that come with optionality cause the relevant constraint pair to keep on falling down the constraint hierarchy: in (14), learning may be completed but demotion of both constraints will continue. In general, such a movement will push along any constraint that is crucially ranked lower than this pair in the target (adult) hierarchy, and it will drag down any constraint that is ranked higher and has an optionality relationship with one of the members of the pair. For instance, if place assimilation for plosives has a probability of 2%, the constraint *REPLACE (cor / plosive) will be dragged along at a distance of 6 above *GESTURE (tip) (in first approximation, but see §3.7).

Several stabilizing scenarios can be thought of, and one local scenario involves a symmetric combination of demotion of one of the members of the pair, and promotion of the other: when *REPLACE falls by 0.01, *GESTURE will rise by 0.01. More precisely, we should look at the evaluation of the incorrect winner (the row with the pointing finger) and find the highest violated constraint that is not violated by the correct (but losing) candidate. If our constraint set is correct, we know that such an uncancelled mark must exist in the winner, because the winner is obviously not the optimal candidate in the target (adult) grammar. In (12), this constraint is *GESTURE (tip). We now promote this constraint by a small step along the continuous ranking scale. With an original ranking as in (11), the two constraints will end up in the following grammar:



(16)

We see that the centre of the two constraint rankings is still at 44.5, as in the initial state (11). We are justified in calling (16) the "final state" because the rankings will stay in the vicinity of where they are in (16), without joining in a wholesale demotion race.

This combined demotion-promotion scheme seems to be as convergent and robust as that of §2.2, though it is not as "minimal" and local: to implement it, we will have to consider one of the violation marks in the "grey cells" of the tableau (12).

In §3, we will see that the matching of the degree of optionality found in §2.2 for a single constraint pair, extends to larger sets of constraints only if the learner follows the a combined demotion/promotion strategy described here.

2.4. Stochastic evaluation independently needed

We did not introduce random constraint evaluation with the intent of accounting for variation. Rather, this random evaluation is independently needed to guarantee a fundamental property of the natural language user's learning behaviour: robustness. If a minority of errors in the input is to have no dramatic consequences in our grammar, the learner must be allowed to adjust constraint rankings only by an amount that is much smaller than the difference between the rankings of relevant constraints. To prevent a modest number of errors from turning the grammar upside down, a *safety margin* (safe ranking difference) must be maintained. In an error-driven learning scheme, this can only be achieved by stochastic evaluation: only by making a few mistakes herself (or a single one, in the "maximal" algorithm of §4) can the learner refresh a safety margin that has been shrunk by an error. Thus, optionality follows directly from the robustness requirement of learnability.

3. Learning a perception grammar

Consider perceptual categorization along a continuous auditory dimension with values from [0] to [100]. Suppose that a language has the three contrastive categories /30/, /50/, and /70/ along this dimension.

3.1. An OT grammar for perceptual categorization

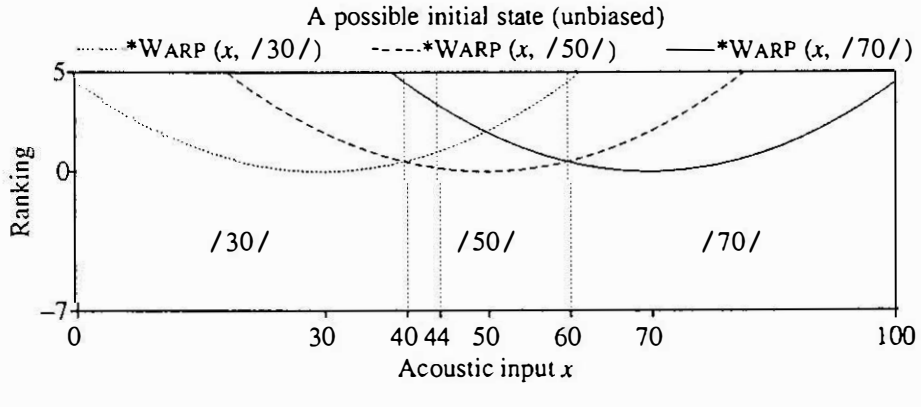
In the listener's perception grammar, the relative fitness of the various categories, given an acoustic input value x, is described by a family of *WARP constraints for each category y:

Def. *WARP (f: x, y)
"An acoustic value x on a perceptual tier f is not categorized into the category whose centre is at y." (17)

This formulation is slightly different from the one in (Boersma 1997: §6.3) because of its dependence on y, so that *WARP is now analogous to the *REPLACE family of the production grammar. Now, a less distorted recognition is preferred over a more distorted recognition, so the *WARP constraints are locally ranked according to

$$\text{*WARP}(\textit{feature}: x_1, y) \gg \text{*WARP}(\textit{feature}: x_2, y) \Leftrightarrow |y - x_1| > |y - x_2| \quad (18)$$

provided that x_1 and x_2 are on the same side of the category centre y. The continuous *WARP families for our three categories could thus be depicted as



(19)

To see how these constraints interact in the listener's categorization system, consider what happens to the datum [44]. The listener has three candidate categories, and the perception grammar determines the winner:

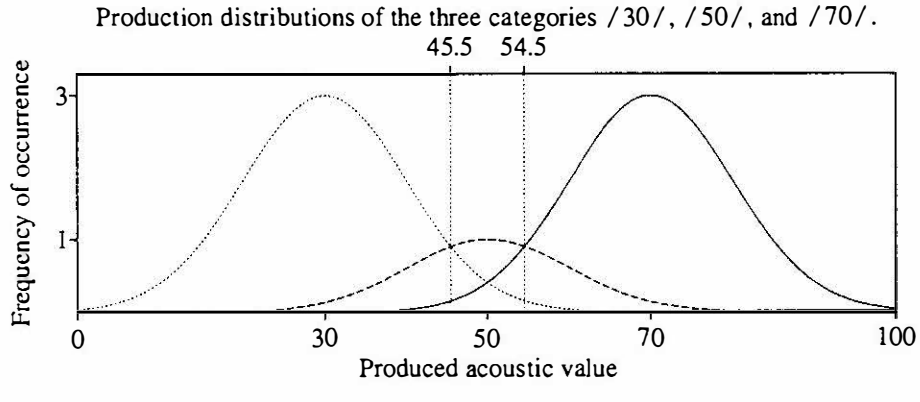
[44]	*WARP ([44], /70/)	*WARP ([44], /30/)	*WARP ([44], /50/)
/30/		*!	
☞ /50/			*
/70/	*!		

(20)

The ranking of the three relevant *WARP constraints can be read from the dotted line at [44] in figure (19): in going from the bottom up, it first cuts the *WARP (x, /50/) curve, then the *WARP (x, /30/) curve, and finally the *WARP (x, /70/) curve. Because the *WARP (x, /50/) curve is the lowest of these curves for x = 44, the listener categorizes the acoustic input into the /50/ class. Given the three equally shaped and equally high curves in (19), the discrimination criteria are obviously at [40] and [60], and if evaluation is not stochastic, these criteria are hard: every input above [60] is classified as /70/, every input below [40] as /30/, and every other input as /50/.

3.2. Production distributions and the optimal listener

Variations within and between speakers will lead to random distributions of the acoustic input to the listener's ear. Suppose that a language has three categories with midpoints at [30], [50], and [70] along a perceptual dimension, and a problematic three-way contrast: the middle category is weaker than the others (e.g. has fewer lexical occurrences). The speaker's productions, which are the inputs to the listener's perception grammar, are then distributed as follows:



(21)

The listener will make the fewest mistakes in initial categorization if she uses the criterion of *maximum likelihood*, i.e., if given the acoustic input x she chooses the perceptual category y that maximizes the a posteriori probability

$$P(prod = y \mid ac = x) = \frac{P(ac = x \mid prod = y) \cdot P(prod = y)}{P(ac = x)} \qquad (22)$$

For instance, if the acoustic input is [44], an optimal listener will choose the /30/ category because the curve of the distribution of the production of /30/ in figure (21) is above the curve associated with the production of the category /50/, although the

value [44] is nearer to the midpoint of the /50/ category than to the midpoint of /30/. Therefore, she will initially categorize all inputs below the criterion [45.5] into the class /30/, all the values between [45.5] and the second criterion [54.5] into the class /50/, and all values above [54.5] into the class /70/. I will now show how an OT listener establishes these criteria.

3.3. The initial state and its inadequacy

Figure (19) shows a possible initial state with unbiased categorization. Given the language environment, the listener will more often have to recognize a [44] input into the /30/ class than into the /50/ class, though she will prefer the /50/ class herself. Therefore, (19) is not an optimal grammar.

3.4. Learning from categorization errors

The categorization according to (19) is independent from what the speaker's intended category was, but if the listener gets to know (in the recognition phase, after lexical access etc.) which category the speaker had meant to produce, she may take a *learning step*. Suppose that the speaker had intended the /30/ category. Tableau (20) can then be enriched in a way analogous to (but somewhat simpler than) the learning tableau for production grammars (12):

/30/ [44]	*WARP ([44], /70/)	*WARP ([44], /30/)	*WARP ([44], /50/)
√ /30/		*!	
☞ /50/			*
/70/	*!		

(23)

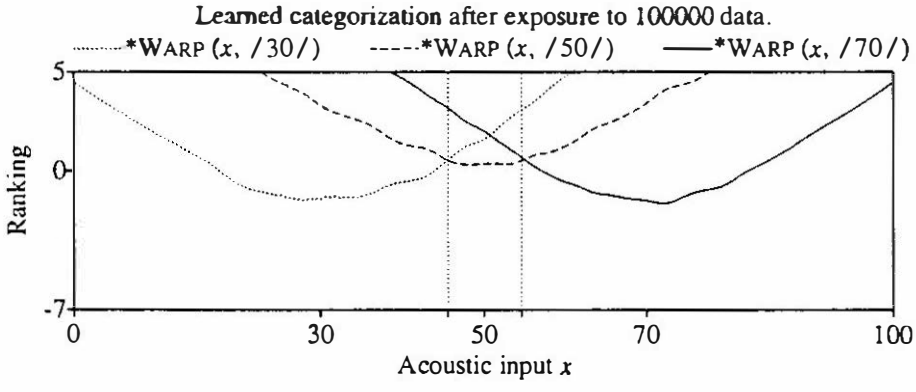
The listener now "knows" that she has made a categorization error. The offending constraint is the one with the crucial violation (the exclamation mark) in the evaluation of the intended category /30/ (in the row with the check mark). This offending constraint is *WARP ([44], /30/). A simple learning strategy (§2.2) is to demote the offender, i.e., to lower the ranking of *WARP ([44], /30/) by a small amount (say 0.01) along the continuous ranking scale, which runs from –7 to 5 in figure (19).[1]

3.5. Stochastic categorization and the optimal criterion

A crucial ingredient for the model is the stochastic constraint evaluation of §1: the ranking of each categorization constraint at evaluation time is drawn from a Gaussian distribution about its ranking in figure (19), again with a spreading of 2. This means that an acoustic input of [44] has a chance of being initially categorized as /30/, /50/, or even /70/, with probabilities that depend on the differences between the heights of the three *WARP ([44], y) curves. Even after *WARP ([44], /30/) has fallen below *WARP ([44], /50/), there is still a chance that a [44] datum will be initially perceived as /50/. This optionality will lead to safety margins between the curves: *WARP

[1] Because the constraint family is continuous, I used a Gaussian *demotion window* in the simulations, i.e., the nearest neigbours (say, [39] through [49]) were also demoted, according to a Gaussian window with a spreading of 1.58 acoustic units.

([44], /30/) will be demoted below *WARP ([44], /50/) until the error probabilities, given the production distributions and the categorization noise, are the same for both classes. After exposure to 100,000 data, the perception grammar of a demotion/promotion learner will look like



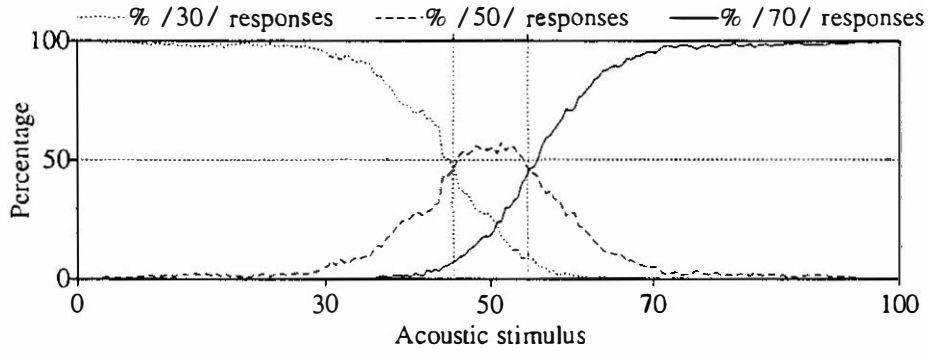
(24)

In the simulation that led from the initial unbiased grammar (19) to the adult grammar (24), the perceptual range was divided into 200 steps of 0.5, the error-driven demotion step (*plasticity*) was 0.01 (also stochastic, with a relative spreading of 0.1), and the categorization spreading was 2, and the local-ranking principle was not enforced[2]. We see that the minimal gradual learning algorithm causes the two criteria between the middle category and its neighbours (the cutting points in the figure) to shift in the direction of the middle category, until they fall together with the optimal criteria identified in §3.2. Thus, *the minimal gradual OT learner will automatically learn to set the criteria in a way that a maximum-likelihood listener would*. Note how the local learning strategy of demoting a single incorrectly invalidating constraint implements the global functional principle of maximizing *the ease of the effort of comprehension*, i.e. minimizing the number of initial perception errors, thus minimizing the number of cases that the initial categorization will have to be repaired by the "higher" parts of the recognition system.

3.6. Probability matching

But our learner does not become a perfect maximum-likelihood listener. This is because the learned criteria are 'soft': because of the stochastic categorization, there will be regions in the acoustic space where more than one category can be initially perceived: even though the acoustic input [44] is most likely to come from an intended /30/ production, there is still a large probability that it is initially perceived as /50/. From grammar (24), we can determine the perception-probability curves for the three categories, by the following simulation. We present 1000 acoustic replications of each of the 200 acoustic stimuli 0.25, 0.75, 1.25, ..., 99.75 to the (simulated, patient) listener who is defined by the grammar (24). We will ask her what she hears and force her to choose from the categories /30/, /50/, and /70/; we will assume that her grammar is fixated, i.e. that she will not adapt her criteria to the uniform distribution of the stimuli (only computerized listeners can be frozen in this way). The 200,000 stimuli gave the following three curves for the percentages of the responses of each of the three categories, as functions of the controlled acoustic stimulus:

[2] The Praat script that performs these simulations and produces the figures (19), (21), (24), (25), (29), (34), (35), and (36), is available from http://fonsg3.let.uva.nl/paul/.



(25)

These curves are very similar to the categorization curves for controlled acoustic stimuli, as have been measured for several ternary categorizations: voice-onset time (the [b]-[p]-[p^h] continuum) in Thai (Lisker & Abramson 1967); vowel height (the [ɪ]-[ɛ]-[æ] continuum) in English (Fry, Abramson, Eimas & Liberman 1962); and place "of articulation" (the perceptual [b]-[d]-[ɡ] continuum) in English (Liberman, Harris, Hoffman & Griffith 1957).

So, the listener does not maintain an accurate maximum-likelihood strategy. We can compute the categorization probabilities from the production probabilities, if we realize that in an equilibrium situation, the demotion frequencies of the two competing categories will be equal. For instance, the acoustic input [40] represents an intended /30/ category in 74% of all cases, and the /50/ category in 25% of all cases. Equilibrium has been achieved (for a demotion/promotion learner, who shows no "downdrift") when the probability of the error of classifying an intended /30/, realized as [40], into the /50/ category, is equal to the probability of the error of classifying an intended /50/, also realized as [40], into the /30/ category:

$$P(prod = 30 \wedge perc = 50/ac = 40) = P(prod = 50 \wedge perc = 30/ac = 40) \quad (26)$$

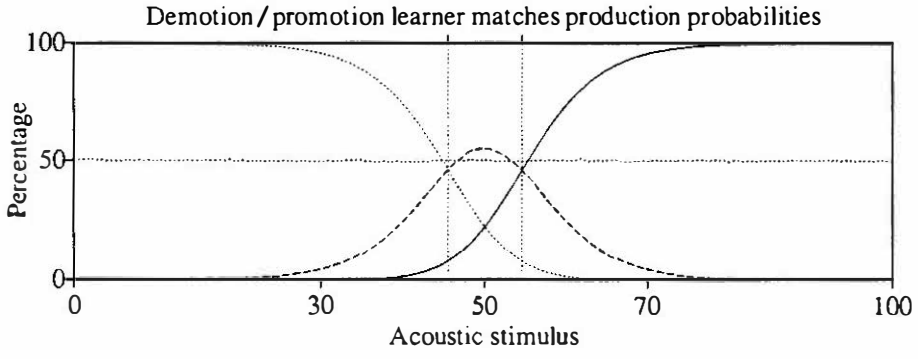
Under the assumption that the initially perceived category does not depend on the speaker's intended category, but only on the acoustic input, we can rewrite the combined probabilities as

$$P(prod = 30 \wedge perc = 50/ac = 40) = P(prod = 30/ac = 40) \cdot P(perc = 50/ac = 40)$$
$$P(prod = 50 \wedge perc = 30/ac = 40) = P(prod = 50/ac = 40) \cdot P(perc = 30/ac = 40) \quad (27)$$

Combining (26) and (27), we get

$$\frac{P(perc = 30/ac = 40)}{P(perc = 50/ac = 40)} = \frac{P(prod = 30/ac = 40)}{P(prod = 50/ac = 40)} \quad (28)$$

Thus, our learner becomes a *probability-matching listener*: her perception bias is going to equal the production bias: she will categorize the input [40] into the /30/ class in 74% of all cases, and into the /50/ class in 24% of the cases. We may note the similarity between (25) and a graph of the posterior production probabilities given any acoustic input, which can be derived easily by dividing the three values for each acoustic value in (21) by the sum of these three values:



(29)

The probability-matching strategy automatically results from OT learning with stochastic evaluation, *no matter how weak the random part of it is, as long as it is greater than the plasticity.*

Note that this strategy does not minimize the global number of perception errors, though it may aid in the recovery from initial errors if the acoustic signal is still in short-term memory.

3.7. Poor categorization performance of a demotion-only learner

The results of §3.6 are valid for demotion-only learners in learning a single constraint pair, and for combined demotion/promotion learners in general. We will now see how a demotion-only learner would mess up the *three* constraint families that are relevant for our categorization problem.

For a given acoustic input, say [40], an equilibrium is reached when all three *WARP constraints are demoted equally often, i.e., when the listener makes an equal amount of "mistakes" in classifying an intended /30/, /50/, or /70/ production. Thus, suppressing the condition clause, (26) expands to

$$\begin{aligned} &P(prod = 30 \wedge perc = 50) + P(prod = 30 \wedge perc = 70) = \\ &P(prod = 50 \wedge perc = 30) + P(prod = 50 \wedge perc = 70) = \\ &P(prod = 70 \wedge perc = 30) + P(prod = 70 \wedge perc = 50) \end{aligned} \tag{30}$$

Again under the assumption of independent categorization, this becomes

$$\begin{aligned} &P(prod = 30) \cdot \big(P(perc = 50) + P(perc = 70)\big) = \\ &P(prod = 50) \cdot \big(P(perc = 30) + P(perc = 70)\big) = \\ &P(prod = 70) \cdot \big(P(perc = 30) + P(perc = 50)\big) \end{aligned} \tag{31}$$

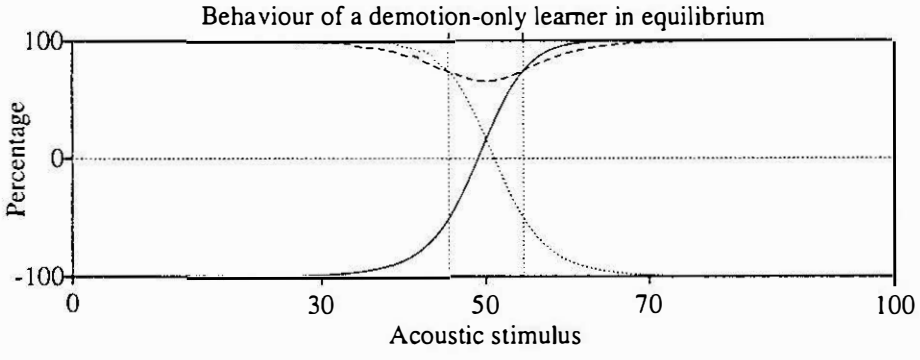
Remembering that

$$P(perc = 30) + P(perc = 50) + P(perc = 70) = 1 \tag{32}$$

we can compute the three unknown perception probabilities by solving the three linear equations (31) and (32). Instead of the probability-matching formula (28), we get (with a notation adapted to the width of the page):

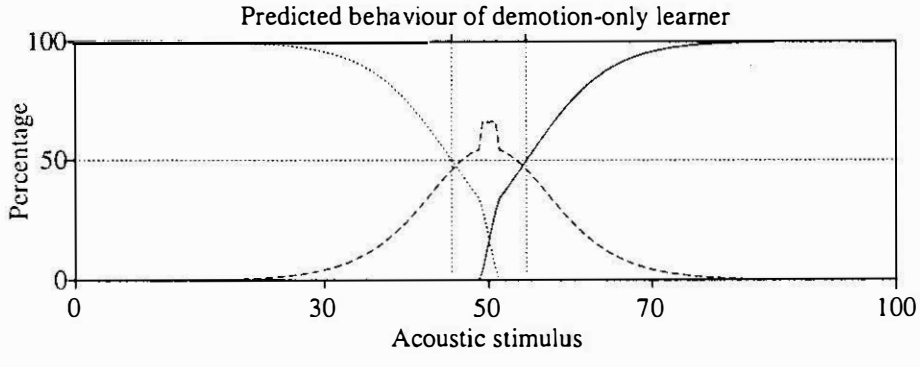
$$P(perc = 30) = 1 - 2\frac{P_{prod}(50) \cdot P_{prod}(70)}{P_{prod}(30) \cdot P_{prod}(50) + P_{prod}(30) \cdot P_{prod}(70) + P_{prod}(50) \cdot P_{prod}(70)} \tag{33}$$

This predicts the following categorization probabilities for each acoustic input:



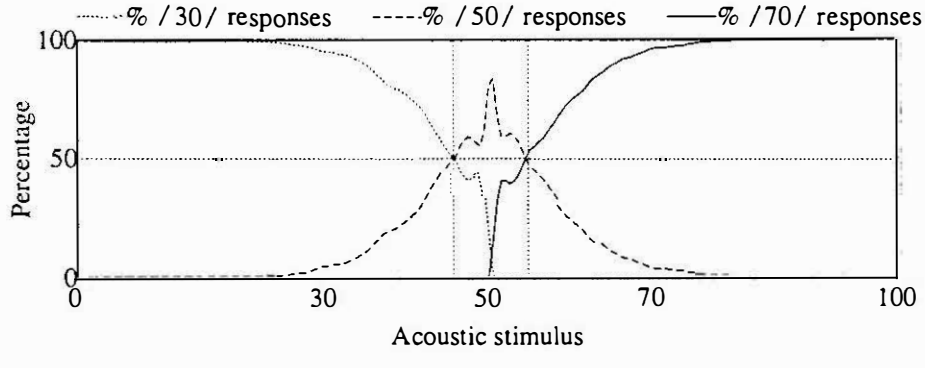
(34)

The situation is clearly pathological: we see negative probabilities except in a small range of acoustic values around [50]. This just means that outside this domain there is no concerted downdrift of the three constraints: at [60], for instance, *WARP (x, /50/) and *WARP (x, /70/) will be drifting down the ranking scale, but *WARP (x, /30/) will be left behind, driving the probability that the listener classifies an acoustic input [60] as /30/ to zero. In the limit, therefore, the listener's perception will seem to follow a two-constraint probability-matching strategy outside the small acoustic domain in the centre:



(35)

A simulated demotion-only learner confirmed this when asked to classify the whole acoustic range after a million learning data[3]:



(36)

To my knowledge, the discontinuities and exact zeroes exhibited by (35) and (36) have not been found in categorization experiments. To the extent that the response distributions (25) and (29) are more realistic, we must conclude that a symmetric demotion/promotion learning model better represents reality than a demotion-only model. This, added to the solution of the grammatical downdrift problem, leads us

[3] The small differences between (35) and (36) arise from using the Gaussian demotion window (fn. 2).

into questioning the validity of demotion-only learning schemes, be they gradual (Boersma, to appear) or not (Tesar & Smolensky 1993, 1996).

4. The correct maximal algorithm for learning a stochastic grammar[4]

Contrary to what §3.6 suggested, the symmetric version of the *minimal* gradual learning algorithm often does not lead to probability matching. Instead, the correct algorithm must demote *all* violated constraints in the adult's utterance, and promote *all* violated constraints in the learner's utterance. In the example of (23), there would be no difference between this *maximal* algorithm and the minimal algorithm, but in a grammar with a larger number of constraints, there would.

Suppose that there are K candidates, each of which has a probability P_k^{L} ($k = 1...K$) of being chosen by the learner, and a probability of P_k^{A} of being chosen by the adult. Suppose that the grammar contains N constraints with rankings r_n ($n = 1...N$). As a result of the demotion of all the adult's violated constraints, the ranking of constraint n will increase upon the next learning pair by a negative amount Δr_n, whose expectation value is

$$\mathrm{E}^{\mathrm{A}}[\Delta r_n] = -p \cdot \sum_{k=1}^{K} P_k^{\mathrm{A}} m_{kn} \tag{37}$$

where p is the plasticity constant, and m_{kn} is 1 if candidate k violates constraint n and 0 otherwise (for now, we consider only constraints that can be violated only once). Likewise, the promotion of all the learner's violated constraints will lead to an expected positive ranking increase of

$$\mathrm{E}^{\mathrm{L}}[\Delta r_n] = p \cdot \sum_{k=1}^{K} P_k^{\mathrm{L}} m_{kn} \tag{38}$$

The total expected change in the ranking is

$$\mathrm{E}[\Delta r_n] = p \cdot \sum_{k=1}^{K} \left(P_k^{\mathrm{L}} - P_k^{\mathrm{A}} \right) m_{kn} \tag{39}$$

We can see that if a candidate occurs with greater probability in the speaker than in the adult, its violated constraints will rise on average, so that the probability of this candidate in the speaker will decrease. Thus, the expected ranking change seems to decrease the gap between the two grammars. Now, we will have to find a more formal proof.

We can see immediately that if the learner's grammar equals the adult's grammar, i.e. if P_k^{L} equals P_k^{A} for all k, the expected ranking change of every constraint n is zero, i.e. the expected change in the learner's grammar is zero. To prove learnability, however, we have to show the reverse, namely the convergence of the learner's grammar upon the adult's grammar. An important part of the proof involves showing that the learner cannot end up in a different grammar from the adult. Suppose the learner does end up in such a *local maximum*, i.e. $\mathrm{E}[\Delta r_n]$ is zero for every constraint n. We can write this situation in vector-matrix notation:

---

[4] This section did not occur in the ROA version of this chapter. The maximal algorithm evolved after a computer simulation of the learning of the extensive optionality data from Hayes & MacEachern (to appear).

$$\mathbf{m}^T\left(\mathbf{P}^L - \mathbf{P}^A\right) = 0 \tag{40}$$

Given a violation matrix **m**, the learner can end up in any grammar $\mathbf{P}^L$ that satisfies (40). As we now from linear algebra, however, the vector $\mathbf{P}^L - \mathbf{P}^A$ must be zero if the matrix **m** behaves well. We can distinguish the following cases of ill-behaved violation matrices:

1. There are two candidates k and l who violate the same set of constraints. Equation (40) is then valid for any **P** for which there is an a so that $P_k^L = P_k^A - a$ and $P_l^L = P_l^A + a$. However, under our evaluation regime, these candidates are equally harmonic in every respect, so they must have equal probabilities in the learner's grammar $\left(P_k^L = P_l^L\right)$ as well as in the adult's grammar $\left(P_k^A = P_l^A\right)$. Combining the four equations, we see that a must be zero.
2. There is a candidate k that violates all constraints violated by candidate l as well as those violated by candidate m. Equation (40) is then valid for any **P** for which there is an a so that $P_k^L = P_k^A - a$, $P_l^L = P_l^A + a$, and $P_m^L = P_m^A + a$. However, if candidate k violates a proper superset of the constraints violated by another candidate, it should always be judged less harmonic than that other candidate in the pairwise evaluation, regardless of the constraint ranking. Therefore, $P_k^L = P_k^A = 0$, so that a must be zero.
3. Candidate k violates constraints A and C, l violates B and D, m violates A and D, and n violates B and C. Equation (40) is then valid for any **P** for which there is an a so that $P_k^L = P_k^A - a$, $P_l^L = P_l^A - a$, $P_m^L = P_m^A + a$, and $P_n^L = P_n^A + a$. This is a genuine case of degeneracy: the constant a will be adjusted so that P_k^L / P_m^L ends up near P_n^L / P_l^L, *irrespectively of the initial constraint rankings*; for instance, if the adult has $P_k^A = 0.1$, $P_l^A = 0.2$, $P_m^A = 0.27$, and $P_n^A = 0.43$, the learner will arrive near $P_k^L = 0.2$, $P_l^L = 0.3$, $P_m^L = 0.17$, and $P_n^L = 0.33$, and she will never reach the adult distribution. But! This adult distribution could never have been derived from a stochastically evaluating OT grammar: there is no constraint ranking that produces it. In fact, with the given candidates and violations, any grammar must satisfy the condition that if $P_k < P_m$ (i.e., C dominates D), then also $P_n < P_l$. This is one of the empirical predictions of our hypothesis of stochastic evaluation: some distributions are impossible.
4. Any more complicated dependencies between the violations of the candidates. Generally, if there are many more candidates than constraints (which is true under most interpretations of the candidate generator in OT), and if these candidates cover the range of possible sets of violations, (40) must lead to the conclusion that $\mathbf{P}^L = \mathbf{P}^A$, i.e. that the algorithm converges upon the adult grammar.

We have made plausible, though not yet rigorously proved, that the maximal symmetrized gradual learning algorithm is capable of learning any stochastically evaluating OT grammar.

5. Conclusion

Optionality follows directly from the robustness requirement of learnability: a demotion/promotion learner will show the same error rate herself as she hears in her environment. To be resistant against 5% errors, you must make 5% errors yourself; 30% variation in your environment will make you produce 30% variation yourself; and if a certain acoustic input has a 30% probability of stemming from an intended

category x, your perception grammar will make you classify this acoustic input into the category x 30% of the times.

These results are exact only for a symmetrized and maximal version of the Gradual Learning Algorithm, i.e., a version in which the learning step involves demotion of all constraints with uncancelled marks in the correct (but losing) candidate, and simultaneous promotion of all constraints with uncancelled marks in the incorrectly winning candidate. There is some evidence that this combined demotion/promotion learning scheme is a better model of learning than the demotion-only scheme: apart from the grammar-internal downdrift problem, the observable quantities of categorization show unrealistic behaviour with the demotion-only scheme.

The account of optionality presented here naturally encapsulates pragmatics-based reranking. For instance, if you want to speak more clearly, you may raise all your faithfulness constraints by, say, 5 along the continuous ranking scale. In this way, an 80%-20% preference *for* place assimilation will turn into a 18%-82% preference *against*. Depending on whether the faithfulness constraint is ranked above or below its rival, slight variation may turn into obligation or the reverse. If the ranking difference is large to begin with, however, nothing happens; so we see that discrete properties of surface rerankability are compatible with, and may well follow from, a general continuous rerankability of all constraints.

Our account of optionality may well extend to other parts of the grammar, including the problem of constituent ordering in syntactical theory, which is a field where optionality is very common. Our account may well explain how the "interacting and possibly competing principles and preferences" of Functional Grammar (Dik 1989: 337) determine the choice between, say, surface SVO and OVS orders in a V2 language: one part of the answer will be pragmatical reranking of the relevant functional principles (like "subject first", "human first"), and another part will be the random variation that occurs at evaluation time; in an obligatory SVO language, one of the constraints is just ranked so far above the other that the degree of variation is essentially zero.

6. References

Anttila, Arto (1995). "Deriving variation from grammar: A study of Finnish genitives", ms. Stanford University. [Rutgers Optimality Archive **63**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1997). "The elements of Functional Phonology", ms. University of Amsterdam. [Rutgers Optimality Archive **173**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (to appear). "Learning a grammar in Functional Phonology", in J. Dekkers, F. van der Leeuw, and J. van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition.*

Dik, Simon C. (1989). *The Theory of Functional Grammar. Part I: The Structure of the Clause.* Foris, Dordrecht.

Fry, D.B., A.S. Abramson, P.D. Eimas, and A.M. Liberman (1962). "The identification and discrimination of synthetic vowels", *Language and Speech* **5**, 171-179.

Hayes, Bruce and Margaret MacEachern (to appear): "Folk verse form in English", *Language.*

Liberman, A.M., K.S. Harris, H.S. Hoffman, and B.C. Griffith (1957). "The discrimination of speech sounds within and across phoneme boundaries", *J. of Experimental Psychology* **54**, 358-368.

Lisker, Leigh and Arthur S. Abramson (1967). "The voicing dimension: some experiments in comparative phonetics", *Proc. Sixth International Congress of Phonetic Sciences*, 563-567.

Prince, Alan and Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar.* Rutgers University Center for Cognitive Science Technical Report **2**.

Tesar, Bruce and Paul Smolensky (1993). "The learnability of Optimality Theory: an algorithm and some basic complexity results", ms. Department of Computer Science & Institute of Cognitive Science, University of Colorado, Boulder. [Rutgers Optimality Archive **2**]

Tesar, Bruce and Paul Smolensky (1996). *Learnability in Optimality Theory (long version).* Technical Report **96-3**, Department of Cognitive Science, Johns Hopkins University, Baltimore. [Rutgers Optimality Archive **156**, http://ruccs.rutgers.edu/roa.html]