Gang effects: the perspective from variation

Edward Flemming & Giorgio Magri
MIT & CNRS

1 Introduction

Labov (1969) hypothesized that the rates of application of variable phonological processes obey an intriguing generalization that we dub the “Strict Domination Generalization” (SDG) to highlight its intuitive connection with strict domination in categorical OT. Informally, the SDG says that, where multiple grammatical factors affect the rates of application of a process, the factors can always be ordered in a hierarchy such that the effect on the rates of “each [factor] in the hierarchy outweighs the effects of all [factors] below it” (Labov 1969:page 741). Although this generalization has faded into obscurity, we think it is worth bringing it back to the fore because our initial investigation reveals that the generalization might be empirically well supported despite imposing a significant restriction on the range of possible probabilistic processes. Furthermore, the generalization has significant theoretical implications because it does not follow from any current model of probabilistic phonology.

Section 2 offers an explicit formulation of the SDG. Section 3 provides some empirical support for the SDG from recent studies of variable schwa realization in French (Smith & Pater 2020; Storme 2021). Turning to the theoretical implications of the SDG, section 4 shows that its force can be understood as drawing a line between two types of “gang effects”. On the one hand, the SDG permits “superset” gang effects, in which adding factors that favor an optional phonological process can result in a higher rate of application of that process. On the other hand, the SDG bans “disjoint” gang effects, in which two or more individually weaker factors “gang up” to produce higher rates of process application than one stronger factor. Finally, section 5 demonstrates that neither Stochastic OT (SOT; Boersma 1998; Boersma & Hayes 2001) nor MaxEnt grammars (ME; Goldwater & Johnson 2003; Hayes & Wilson 2008; Wilson 2025) makes the desired distinction between these two types of gang effects: SOT cannot derive either, while ME derives both. We end by considering how ME can be restricted to exclude just the disjoint gang effects banned by the SDG via restrictions on possible weighting vectors. Discussion of the SDG in noisy HG (NHG; Boersma & Pater 2016; Hayes 2017) is left for future work.

2 The Strict Domination Generalization (SDG)

In this section we offer an explicit formulation of the SDG and make explicit the connection with Labov’s original formulation. We begin by introducing the concepts and notation required for this discussion.

2.1 *Rates of application*

The SDG is a generalization about the rates of application of phonological processes that display token variation. For convenience, we describe these processes in terms of SPE rules A→B/X_Y. A **target underlying string** is any string that contains an instance (for simplicity: only one instance) of the structural description XAY. We make the assumption (known as the **principle of multiple causes**; Young & Bayley 1996; Bayley 2002) that the rate with which the process applies to various target underlying strings depends on a certain number n of **factors** $X_1, \ldots, X_n$. We focus exclusively on grammatical factors (such as phonological environments) and ignore external factors (such as speaker

Proceedings of *AMP 2025*

gender and social class). We denote by $x_1, x_2, \ldots$ the **values** of a factor X_k.[1] Factor values are mutually incompatible: no target underlying string has two different values x_1 and x_2 for the same factor X_k. We denote by $\mathbb{R}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ the rate with which the variable process considered applies to target underlying strings that have value x_1 for factor X_1, value x_2 for factor X_2, and so on. Whenever possible, we simplify this notation as $\mathbb{R}(x_1, x_2, \ldots, x_n)$.

To illustrate, we consider the classical t-deletion process [−sonorant, −continuant] → ∅/C_##, that deletes a word final alveolar stop preceded by a consonant (Guy 1980). Examples of phrases containing target underlying strings are *cost me* or *cost us*. A factor known to influence the rate of t-deletion is the phonological context (*FC*) that follows the word final stop. This factor takes values such as *con*, *vow*, and *pau* for pre-consonantal, pre-vocalic, and pre-pausal (phrase-final) stops. $\mathbb{R}(FC = con)$ denotes the rate of t-deletion for target underlying strings that have value *con* for factor *FC* because the final stop targeted by deletion is followed by a consonant.

We make the assumption (known as the **principle of quantitative modeling**; Young & Bayley 1996; Bayley 2002) that the rates of application of variable phonological processes are replicable: "different speech samples from an individual will reveal the same patterning" (Wolfram 1973). Berdan (1975) offers empirical evidence for the stability of rates (we are not aware of more recent evidence). One of the goals of probabilistic phonology is to discover the mathematical generalizations that govern the rates of application of variable phonological processes. The next subsection introduces one such generalization, the SDG.

2.2 *Strict domination for rates of application* We consider some ranking $\gg$ of the grammatical factors $X_1, \ldots, X_n$ that are known to influence the application of the variable phonological process of interest. We interpret the inequality $X_h \gg X_k$ as saying that the factor X_h is "ranked higher", namely has a **stronger** effect on the application of the process considered than the factor X_k. Furthermore, for each factor X_k, we consider some ranking $\gg_k$ of its values. We interpret the inequality $x \gg_k \widehat{x}$ as saying that the value x of the factor X_k is "larger", namely promotes the application of the process considered more than the value $\widehat{x}$.

These rankings $\gg$ and $\gg_k$ induce a **lexicographic order** $\succ_{\text{lex}}$ of the tuples of factor values through condition (1). This condition says that a tuple $(x_1, \ldots, x_n)$ of factor values is larger than another tuple $(\widehat{x}_1, \ldots, \widehat{x}_n)$ provided there exists some factor X_k that satisfies two requirements. The first requirement (1a) is that the two tuples compared share the same value for every factor X_h that is ranked above factor X_k according to $\gg$. The second requirement (1b) is that the two tuples have instead different values x_k and $\widehat{x}_k$ for factor X_k and that the value x_k of the larger tuple is larger than the value $\widehat{x}_k$ of the smaller tuple according to $\gg_k$. Crucially, any factor that is ranked underneath X_k according to $\gg$ is irrelevant for ordering the two tuples. In this sense, this highest ranked relevant factor X_k **strictly dominates** any lower ranked factors.

(1) $(x_1, \ldots, x_n) \succ_{\text{lex}} (\widehat{x}_1, \ldots, \widehat{x}_n)$ if and only if there exists some factor X_k such that:

a. $x_h = \widehat{x}_h$ for every factor X_h such that $X_h \gg X_k$ (if any)

b. $x_k \gg_k \widehat{x}_k$

The **Strict Domination Generalization** (SDG) says that, for any phonological process and for any speaker of any dialect in any register where the process applies variably, there exists some ranking $\gg$ of the factors $X_1, \ldots, X_n$ that govern its rate of application and some ranking $\gg_k$ of the values of each factor X_k such that *the inequalities among the rates of application of the process to target underlying forms with different tuples of factor values are consistent with the lexicographic order $\succ_{lex}$ among tuples of factor values*: lexicographically larger tuples cannot have smaller rates of application, as stated in (2). Thus, the effect on the rates of application exercised by "each [factor] in the hierarchy [specified by the ranking $\gg$] outweighs the effects of all [factors] below it" (Labov 1969:page 741), because the lexicographic order in (1) only depends on the highest ranked relevant factor X_k, which strictly dominates all lower ranked factors.

(2) If $(x_1, \ldots, x_n) \succ_{\text{lex}} (\widehat{x}_1, \ldots, \widehat{x}_n)$, then $\mathbb{R}(x_1, \ldots, x_n) \geq \mathbb{R}(\widehat{x}_1, \ldots, \widehat{x}_n)$.

2.3 *Historical remarks* The SDG was introduced as a corollary of a stronger generalization proposed by Labov (1969). To formulate Labov's original stronger generalization, we assume without loss of generality

[1] The variable rule literature uses **factor group** or **constraint family** instead of factor; furthermore, it uses **factor** or **constraint** instead of factor value.

lexicographic order ↓	*PM* ≪ *ST* ≪ *FV*	rates ↓
	(*no*, *no*, *no*)	.297
	(*yes*, *no*, *no*)	.296
	(*no*, *yes*, *no*)	.334
	(*yes*, *yes*, *no*)	.588
	(*no*, *no*, *yes*)	.607
	(*yes*, *no*, *yes*)	.737
	(*no*, *yes*, *yes*)	.814
	(*yes*, *yes*, *yes*)	.829

Figure 1: English variable word-final d-preservation after a vowel (Fasold 1978).

lexicographic order ↓	*CLA* ≪ *CLI* ≪ *CCC*	rates ↓
	(*no*, *no*, *no*)	.09
	(*yes*, *no*, *no*)	.12
	(*no*, *yes*, *no*)	.56
	(*yes*, *yes*, *no*)	.65
	(*no*, *no*, *yes*)	.68
	(*yes*, *no*, *yes*)	.83
	(*no*, *yes*, *yes*)	.91
	(*yes*, *yes*, *yes*)	.94

Figure 2: French optional word-final schwa realization (Smith & Pater 2020).

lexicographic order ↓	*CLA* ≪ *MOR* ≪ *CLU*	rates ↓
	(*no*, *infl*, *lcl*)	.50
	(*yes*, *infl*, *lcl*)	.57
	(*no*, *deriv*, *lcl*)	.64
	(*yes*, *deriv*, *lcl*)	.76
	(*no*, *infl*, *olc*)	.83
	(*yes*, *infl*, *olc*)	.90
	(*no*, *deriv*, *olc*)	.86
	(*yes*, *deriv*, *olc*)	.89

Figure 3: French optional schwa realization between roots and suffixes (Storme 2021).

that the factor ranking $\gg$ is $X_1 \gg X_2 \gg \cdots \gg X_n$. Furthermore, we assume for simplicity that each factor X_k has only two values + and −. Labov's original generalization then says that there exists a coefficient p between 0 and 1/3 (that depends on the process, the language, the speaker, and the register considered) such that the rate $\mathbb{R}(x_1, \ldots, x_n)$ with which the variable process considered applies to target underlying forms with factor values $x_1, \ldots, x_k$ is equal to $\frac{1}{2} + \sum_{k=1}^{n} \pm p^k$, where the sign + or − of the power p^k depends on whether the value x_k of factor X_k is equal to + or −. Labov refers to this generalization as the **axiom of geometric ordering** because the sum $\sum_k p^k$ is called a geometric series. This axiom makes sense because the quantity $\frac{1}{2} + \sum_{k=1}^{n} \pm p^k$ is bounded between zero and one (when p is bounded between 0 and 1/3), no matter the sign of each power.[2] The SDG (2) follows as a non-quantitative corollary of Labov's axiom.[3] The SDG played a prominent role in the early variationist literature (Bickerton 1971; Fasold 1978; Kay & McDaniel 1979). Since this generalization has been disappeared from the more recent literature, we are reviving it under a new name, intended to stress the obvious analogy with strict domination in categorical OT.

3 Some empirical support for the SDG

This section reviews an example of the SDG from Fasold (1978) and provides additional examples from rates of French schwa realization. Magri & Flemming (2025) discuss the SDG for t-deletion. These exhaust the test cases that we are aware of, but the identification of additional tests of the SDG is a priority.

3.1 *SDG and variable d-deletion* Our first illustration of the SDG is taken from Fasold (1978). He looks at the process that variably deletes a word final voiced alveolar stop /d/ after a vowel in some dialects of English, namely /d/ → ∅/V_##. Fasold focuses on three grammatical factors that affect the rate of d-deletion, with binary values *yes* and *no*:

F(ollowing)V(owel): has value *yes* when the word final /d/ is followed by a vowel (as in 'The food is wonderful' [fud##ɪ]) and value *no* otherwise (as in 'The food smells wonderful' [fud##s]).

ST(ress): has value *yes* when the syllable that hosts the word final /d/ is stressed (as in 'We agreed [əˈgri#d] on that') and value *no* otherwise (as in 'We carried [ˈkʰæri#d] a box').

P(ast)M(orphology): has value *yes* when the word final /d/ is a past tense morpheme (as in 'I wonder why he sighed [saɪ#d]') and value *no* otherwise (as in 'That's the wrong side [saɪd]').

[2] In fact, Labov's rate $\frac{1}{2} + \sum_{k=1}^{n} \pm p^k$ is larger than or equal to zero no matter how we choose the sign of each power provided $\frac{1}{2} - \sum_{k=1}^{n} p^k \geq 0$. Analogously, this rate $\frac{1}{2} + \sum_{k=1}^{n} \pm p^k$ is smaller than or equal to one no matter how we choose the signs provided $\frac{1}{2} + \sum_{k=1}^{n} p^k \leq 1$. In either case, we need $\sum_{k=1}^{n} p^k \leq 1/2$. Since $\sum_{k=1}^{n} p^k = \frac{p-p^{n+1}}{1-p}$ (a well known fact about geometric series), we conclude that $\frac{p-p^{n+1}}{1-p} \leq 1/2$ in particular when $0 \leq p \leq 1/3$.

[3] In fact, Labov's rates $\frac{1}{2} + \sum_{k=1}^{n} \pm p^k$ are consistent with the lexicographic order among tuples of factor values corresponding to the factor ranking $X_1 \gg \cdots \gg X_n$ and the ranking of the value + above the value − for each factor X_k. To see that, we suppose that two tuples of factor values satisfy the lexicographic inequality $(x_1, \ldots, x_n) \succ_{\text{lex}} (\widehat{x}_1, \ldots, \widehat{x}_n)$ because they share the same first $k-1$ entries $x_1 = \widehat{x}_1, \ldots, x_{k-1} = \widehat{x}_{k-1}$ while x_k is equal to + but $\widehat{x}_k$ is equal to −. The SDG then requires Labov's rates to satisfy the inequality $\mathbb{R}(x_1, \ldots, x_n) \geq \mathbb{R}(\widehat{x}_1, \ldots, \widehat{x}_n)$ no matter the values $x_{k+1}, \ldots, x_n$ and $\widehat{x}_{k+1}, \ldots, \widehat{x}_n$ of the lower ranked factors. That is the case if and only if $p^k \geq \sum_{h=k+1}^{n} p^k$. And the latter inequality indeed holds whenever the positive constant p is smaller than one.

Fasold ranks these three grammatical factors as in (3a) and their two values as in (3b). The lexicographic order corresponding to these rankings in (3) orders the eight tuples of factor values as in figure 1, with the smallest (largest) tuple at the top (bottom). For each tuple of factor values, the figure also provides the corresponding rate of d-preservation[4] from Wolfram's (1974) study of male adolescents of Puerto Rican ethnicity from New York City.[5] The rates of d-preservation increase from top to bottom, so that tuples that are larger relative to the lexicographic order correspond to larger rates of d-preservation. The only glitch is that the rates are not totally ordered, because of the identity between the rates in the top two rows (highlighted in yellow). Yet, this rate tie is consistent with the SDG because the consequent rate inequality in (2) only requires that the order of rates does not reverse the lexicographic order. Fasold concludes that these rates of d-preservation comply with the SDG.

(3) a. *PM* ≪ *ST* ≪ *FV* b. *no* ≪ *yes*

According to (3b), being followed by a vowel (*FV* = *yes*), sitting in a stressed syllable (*ST* = *yes*), and expressing regular past morphology (*PM* = *yes*) each increase the rate of d-preservation. According to (3a), the factor *FV* outranks the other two factors *ST* and *PM*. Indeed, whenever we compare the rates of d-preservation for two tuples of values that differ for the value of *FV*, the rate is larger for the tuple with value *FV* = *yes* than for the tuple with the opposite value *FV* = *no*, regardless of the values of the other two factors *ST* and *PM*. This is so even for the tuples (*PM* = *yes*, *ST* = *yes*, *FV* = *no*) versus (*PM* = *no*, *ST* = *no*, *FV* = *yes*) (in the fourth and fifth rows): the rate of d-preservation is larger for the latter tuple because the final /d/ is followed by a vowel (*FV* = *yes*), despite it sitting in an unstressed syllable (*ST* = *no*) and not expressing regular past tense (*PM* = *no*). Analogously, the factor *ST* outranks the factor *PM*: whenever we compare the rates of d-preservation for two tuples of factor values that share the same value of the top ranked factor *FV* but differ for the value of *ST*, the rate is larger for the tuple with value *ST* = *yes* than for the tuple with the opposite value *ST* = *no*, no matter the value of the bottom ranked factor *PM*.

3.2 *SDG and variable schwa realization* Our next two illustrations of the SDG involve variable realization of schwa in French. Many schwa vowels are optional in French ([gæʁdəʁi~gæʁdʁi] 'kindergarten'). Smith & Pater (2020) investigate three grammatical factors that affect the rate of realization of schwa at the end of a word, all with values *yes*/*no*: *CCC* has value *yes* when omission of schwa yields a cluster of three consonants; *CLI(tic)* has value *yes* when the potential word-final schwa belongs to a clitic; *CLA(sh)* has value *yes* when omission of schwa yields a clash between stresses of adjacent words. We rank these factors as in (4a) and their values as in (4b). Figure 2 lists the tuples of factor values from top to bottom in increasing lexicographic order together with the corresponding rates of schwa realization. As the rates increase from top to bottom, we conclude that this test case complies with the SDG.

(4) a. *CLA* ≪ *CLI* ≪ *CCC* b. *no* ≪ *yes*

Storme (2021) investigates three grammatical factors that affect the rate of realization of French schwa at the boundary between roots and suffixes: *CLU(ster)* has value *lcl* when omission of schwa yields a liquid+C+liquid cluster ([ˌgæʁd(ə)+ˈʁi]) and value *olc* when it yields an obstruent+liquid+C cluster ([ˌʁegl(ə)-ˈʁa] 'adjust-FUT.3SG'); *MOR(phology)* encodes the morphological status of the suffix as *infl(ection)* ([ˌgæʁd(ə)-ˈʁa] 'keep-FUT.3SG') versus *deriv(ational)* ([ˌgæʁd(ə)+ˈʁi]); *CLA* is as above. We rank these factors as in (5a) and their values as in (5b). Figure 3 lists the tuples of factor values from top to bottom in increasing lexicographic order together with the corresponding rates of schwa realization. Again, we observe that the rates increase from top to bottom. The only glitch is that the rate is larger for (*yes*, *infl*, *olc*) than for (*no*, *deriv*, *olc*). Since this reversal is not statistically significant, this test case also complies with the SDG.

(5) a. *CLA* ≪ *MOR* ≪ *CLU* b. *lcl* ≪ *olc*, *infl* ≪ *deriv*, *no* ≪ *yes*

[4] For consistency with the examples of schwa realization in figures 2-3, figure 1 reports the rates of d-preservation rather than d-deletion. These rates of d-preservation are obtained by subtracting from one the rates of d-deletion reported by Fasold. To illustrate, the rate of d-preservation corresponding to (*no*, *no*, *no*) in figure 1 is equal to 0.297 because the rate of d-deletion corresponding to (*no*, *no*, *no*) according to Fasold is equal to 0.723 and 1 − 0.703 = 0.297.

[5] For the feature value combination (*yes*, *no*, *no*), Wolfram (1974:table 19, page 119) reports that 40 out of 54 tokens underwent deletion, yielding a rate of deletion of 0.741. Fasold (1978:figure 1, page 88) reports instead that 38 out of 54 tokens underwent deletion, yielding a rate of 0.704. We follow Fasold, who adds the specification "errors corrected" after his quote of Wolfram's study. The rate of d-preservation for (*yes*, *no*, *no*) in figure 1 is therefore 1 − 0.704 = 0.296.

4 Implications of the SDG for the theory of gang effects

If the SDG holds generally, it imposes significant restrictions on possible rate functions which need to be derived by our model of probabilistic phonology. This section elucidates the implications of the SDG by showing that it can be restated as a restriction on the types of gang effects among factors that are attested in variable phonological processes.

4.1 *Background Independence Generalization* To set the stage for our reformulation of the SDG in terms of gang effects, we consider an arbitrary variable phonological process sensitive to n grammatical factors $X_1, \ldots, X_n$. We partition them into **target** and **background** factors. Without loss of generality, we choose the first k factors $X_1, \ldots, X_k$ as targets and the remaining factors $X_{k+1}, \ldots, X_n$ as background. Both the antecedent and consequent inequalities in (6) compare the rates of application to target underlying forms that only differ for the values $x_1, \ldots, x_k$ (on the lefthand side) versus $\widehat{x}_1, \ldots, \widehat{x}_k$ (on the righthand side) of the target factors $X_1, \ldots, X_k$. The two tuples compared share the same values for the background factors $X_{k+1}, \ldots, X_n$. The two inequalities only differ in the shared values of the background factors $x_{k+1}, \ldots, x_n$ (in the antecedent inequality) versus $\widehat{x}_{k+1}, \ldots, \widehat{x}_n$ (in the consequent inequality).

$$
\text{(6)}\qquad
\begin{array}{rl}
\text{If} & \mathbb{R}(\overbrace{x_1, \ldots, x_k}^{\text{target}}, \overbrace{x_{k+1}, \ldots, x_n}^{\text{background}}) < \mathbb{R}(\overbrace{\widehat{x}_1, \ldots, \widehat{x}_k}^{\text{target}}, \overbrace{x_{k+1}, \ldots, x_n}^{\text{background}})\\
\text{then} & \mathbb{R}(\underbrace{x_1, \ldots, x_k}_{\text{target}}, \underbrace{\widehat{x}_{k+1}, \ldots, \widehat{x}_n}_{\text{background}}) \leq \mathbb{R}(\underbrace{\widehat{x}_1, \ldots, \widehat{x}_k}_{\text{target}}, \underbrace{\widehat{x}_{k+1}, \ldots, \widehat{x}_n}_{\text{background}})
\end{array}
$$

We say that the rate function $\mathbb{R}$ satisfies the **Background Independence Generalization** (BIG) provided this implication (6) holds no matter how we choose the target factor values $x_1, \ldots, x_k$ and $\widehat{x}_1, \ldots, \widehat{x}_k$ and the background factor values $x_{k+1}, \ldots, x_n$ and $\widehat{x}_{k+1}, \ldots, \widehat{x}_n$ (and no matter how we split the factors into target and background). In other words, the BIG says that, changing the shared background values cannot invert (turn < in the antecedent to > in the consequent) the inequality between the rates corresponding to two patterns of target values. Changing the shared background values can only neutralize the inequality into a tie (turn < in the antecedent to = in the consequent), because the consequent of (6) features a loose inequality.

To illustrate, let us revisit the example of d-preservation from figure 1. We choose *PM* as the target factor and *ST* and *FV* as the background factors. The antecedent inequality in (7) compares the rates corresponding to the values *no* versus *yes* of the target factor *PM* when the background factors *ST* and *FV* have values *no* and *yes*, respectively. This antecedent inequality holds strictly because its lefthand side is equal to 0.607 while the righthand side is equal to 0.737 according to figure 1. The BIG then requires the inequality to hold (at least loosely) for the other three pairs of background factor values, yielding the three consequent inequalities in (7), which indeed all hold (at least loosely: we assume that the first consequent inequality $0.297 \leq 0.296$ holds as an identity). Since we know of no counterexamples to this BIG and since all constraint-based models of probabilistic phonology (SOT, NHG, ME) predict it, we assume the BIG as background to our discussion.

$$
\text{(7)}\qquad
\begin{array}{rlll}
\text{If} & \mathbb{R}(\overbrace{\mathit{PM} = \mathit{no}}^{\text{target}}, \overbrace{\mathit{ST} = \mathit{no}, \mathit{FV} = \mathit{yes}}^{\text{background}}) & < \mathbb{R}(\overbrace{\mathit{PM} = \mathit{yes}}^{\text{target}}, \overbrace{\mathit{ST} = \mathit{no}, \mathit{FV} = \mathit{yes}}^{\text{background}}) & 0.607 < 0.737\\
\text{then} & \mathbb{R}(\mathit{PM} = \mathit{no}, \mathit{ST} = \mathit{no}, \mathit{FV} = \mathit{no}) & \leq \mathbb{R}(\mathit{PM} = \mathit{yes}, \mathit{ST} = \mathit{no}, \mathit{FV} = \mathit{no}) & 0.297 \leq 0.296\\
 & \mathbb{R}(\mathit{PM} = \mathit{no}, \mathit{ST} = \mathit{yes}, \mathit{FV} = \mathit{no}) & \leq \mathbb{R}(\mathit{PM} = \mathit{yes}, \mathit{ST} = \mathit{yes}, \mathit{FV} = \mathit{no}) & 0.334 \leq 0.588\\
 & \mathbb{R}(\mathit{PM} = \mathit{no}, \mathit{ST} = \mathit{yes}, \mathit{FV} = \mathit{yes}) & \leq \mathbb{R}(\mathit{PM} = \mathit{yes}, \mathit{ST} = \mathit{yes}, \mathit{FV} = \mathit{yes}) & 0.814 \leq 0.829
\end{array}
$$

4.2 *Monotonicity Generalization* From now on, we assume that each factor X_k takes binary values *yes* and *no*. Let us consider a factor X_k as the target and all other factors as background. Under the BIG, if the inequality $\mathbb{R}(\ldots X_k = \mathit{yes} \ldots) > \mathbb{R}(\ldots X_k = \mathit{no} \ldots)$ holds for some background factor values (represented by the dots), it holds (possibly loosely) for all background factor values. In this case, we say that the value *yes* of the target factor X_k promotes the application of the process considered while the value *no* inhibits it, independently of the background factor values. Without loss of generality, we assume that is the case for all n factors. To illustrate, this assumption is verified in the examples from Fasold (1978) and from Smith & Pater (2020) in figures 1-2. This assumption is also verified for the example from Storme (2021) in figure 3 when the values *infl* and *lcl* are identified with *no* and the values *deriv* and *olc* with *yes*.

Under this assumption that all factors are binary and that the value *yes* is the promoting value for each factor, the BIG entails the loose inequalities (8) for any two factors X_h and X_k. These inequalities say that the process considered cannot apply with a smaller rate to forms where both factors have value *yes* (on the lefthand side) and with a larger rate to forms where only one of the two factors has value *yes* (on the righthand side). This corollary of the BIG is called the **Monotonicity Generalization** (MG) because it ensures that the rate of application cannot decrease when we replace *no*'s with *yes*'s.

(8)
$$\begin{aligned} \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{yes} \ldots) &\geq \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{no} \ldots) \\ \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{yes} \ldots) &\geq \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{yes} \ldots) \end{aligned}$$

To illustrate, we refer to the rates where a single factor has value *yes* as **monovalent**; the rates where two factors have value *yes* as **bivalent**; and so on. To illustrate, when there are $n = 3$ factors, $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{yes})$ is a bivalent rate while $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no})$ is a monovalent rate. In these terms, the MG ensures that a bivalent rate can never be smaller than the subset monovalent rates.

4.3 *Two types of gang effects* In the rest of this section, we use the BIG (and the MG that follows from it) in order to reinterpret the SDG in terms of gang effects among factors. To this end, we start here by distinguishing between two formally different types of gang effects. We say that a rate function $\mathbb{R}$ displays a **super-set gang effect** between two factors X_h and X_k provided it satisfies condition (9). This condition says that the process considered applies with a strictly larger rate to forms where both factors have value *yes* (on the lefthand side) and with a lower rate to forms where only one of the two factors has value *yes* (on the righthand side). That is, the two factors X_h and X_k "gang up" to derive a higher rate of application than either does in isolation. Because of the MG, the only way these strict inequalities (9) could fail is that an identity holds instead: there is no difference between the rates corresponding to one versus two values *yes*.

(9)
$$\begin{aligned} \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{yes} \ldots) &> \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{no} \ldots) \\ \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{yes} \ldots) &> \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{yes} \ldots) \end{aligned}$$

Next, we say that a rate function $\mathbb{R}$ displays a **disjoint gang effect** between two factors X_h, X_k and a third factor X_ℓ provided it satisfies condition (10). The first two inequalities say that the two monovalent rates where factors X_h or X_k are equal to *yes* (on the righthand side) are smaller than the monovalent rate where factor X_ℓ is equal to *yes* (on the lefthand side). Yet, the third inequality says that the latter monovalent rate where X_ℓ is equal to *yes* (on the lefthand side) is in turn smaller than the bivalent rate where both factors X_h and X_k are equal to *yes* (on the righthand side). That is, X_ℓ is stronger than factors X_h and X_k individually, but these two weaker factors "gang up" to derive a higher rate of application than X_ℓ alone.

(10)
$$\begin{aligned} \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{no} \ldots X_\ell = \mathit{yes} \ldots) &\geq \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{no} \ldots X_\ell = \mathit{no} \ldots) \\ \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{no} \ldots X_\ell = \mathit{yes} \ldots) &\geq \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{yes} \ldots X_\ell = \mathit{no} \ldots) \\ \mathbb{R}(\ldots X_h = \mathit{no} \ldots X_k = \mathit{no} \ldots X_\ell = \mathit{yes} \ldots) &< \mathbb{R}(\ldots X_h = \mathit{yes} \ldots X_k = \mathit{yes} \ldots X_\ell = \mathit{no} \ldots) \end{aligned}$$

4.4 *Gang effects and the SDG* There is no doubt that superset gang effects (9) are well attested in probabilistic phonology. We observe them in every single test case we have looked at, for any two factors considered. To illustrate, let us consider again the example of d-preservation from Fasold (1978) described in figure 1. We focus on the factors *PM* and *ST*. They display a superset gang effect independently of the value of the remaining factor *FV* because the inequalities (11) all hold strictly. It is also clear that superset gang effects are compatible with the SDG: if one factor tuple has a superset of the *yes* values contained in another, then the first tuple is higher in the lexicographic order.

(11) a.
$$\begin{aligned} \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) &> \mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}) && 0.588 > 0.297 \\ \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) &> \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{no}) && 0.588 > 0.334 \end{aligned}$$

b.
$$\begin{aligned} \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{yes}) &> \mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{yes}) && 0.829 > 0.737 \\ \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{yes}) &> \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes}) && 0.829 > 0.814 \end{aligned}$$

On the other hand, we now show that the SDG can be equivalently restated as the generalization that disjoint gang effects (10) are unattested. So the SDG draws a line within the theory of gang effects among the factors that control the rates of application of variable phonological processes: it predicts that superset gang effects are attested but disjoint gang effects are unattested. If the SDG is empirically correct, we need a model of quantitative probabilistic phonology that predicts the attested superset gang effects but avoids the unattested disjoint gang effects. The rest of the paper discusses what such a model might look like.

To establish this connection between the SDG and lack of disjoint gang effects, we suppose for simplicity that there are only $n = 3$ factors X_1, X_2, and X_3.[6] Furthermore, we assume without loss of generality that the three monovalent rates satisfy the inequalities in (12), whereby the monovalent rate where only factor X_1 has value *yes* is smallest while the monovalent rate where only factor X_3 has value *yes* is largest.

(12) $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes})$

To start, we assume that the rate function $\mathbb{R}$ yields no disjoint gang effects (and furthermore satisfies the BIG and therefore also the MG that follows from it). We now argue that this rate function $\mathbb{R}$ then satisfies all the inequalities in (13) and therefore complies with the SDG because the rates track the lexicographic order corresponding to the ranking $X_1 \ll X_2 \ll X_3$ of the three factors and the ranking *no* $\ll$ *yes* of their values. In fact, all of the inequalities in (13) other than (13d) follow independently of the SDG. The inequalities (13a), (13c), (13e), and (13g) follow from the MG, which says that the rates cannot decrease when we add *yes*'s. The inequality (13f) follows from the inequality $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{no})$ between monovalent rates in (12) together with the BIG, when we choose X_1 and X_2 as target factors and X_3 as background factor. The inequality (13b) holds because of (12). Finally, if the crucial inequality (13d) failed, we would have the disjoint gang effect $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}), \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes}) < \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no})$.[7]

(13) $$\mathbb{R}(\mathit{no}, \mathit{no}, \mathit{no}) \overset{(a)}{\leq} \mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}) \overset{(b)}{\leq} \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{no}) \overset{(c)}{\leq} \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) \leq$$
$$\overset{(d)}{\leq} \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes}) \overset{(e)}{\leq} \mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{yes}) \overset{(f)}{\leq} \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes}) \overset{(g)}{\leq} \mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{yes})$$

Conversely, let us suppose that the rate function $\mathbb{R}$ complies with the SDG because the rate inequalities track the lexicographic order among the triplets of factor values corresponding to some ranking of the factors and their values. Because of the assumption that the value *yes* of each factor promotes the application of the process more than the value *no*, the values of each factor must be ranked as *yes* $\gg$ *no*. Furthermore, because of the assumption that the monovalent rates satisfy the inequalities (12), the three factors must be ranked as $X_1 \ll X_2 \ll X_3$. The triplets of factor values are then lexicographically ordered as in figure 1. Since the rate function satisfies the SDG, the rates track this lexicographic order and therefore satisfy all the inequalities in (13). It follows that this rate function yields no disjoint gang effects. In conclusion, the SDG is satisfied if and only if there are no disjoint gang effects.

4.5 *A digression about overlap gang effects* According to subsection 4.3, superset gang effects compare rates such as $\mathbb{R}(\mathit{yes}, \mathit{yes})$ versus $\mathbb{R}(\mathit{yes}, \mathit{no})$ corresponding to tuples of *yes*'s that stand in a superset relation. Furthermore, disjoint gang effects compare rates such as $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no})$ versus $\mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes})$ corresponding to disjoint tuples of *yes*'s. What about comparisons between rates such as $\mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes})$ versus $\mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no})$ corresponding to tuples of *yes*'s that overlap without one being a superset of the other? It turns out that the ordering between these rates follows from the SDG and so does not need to be considered separately. Assume again without loss of generality that the monovalent rates satisfy the inequalities (12). The assumption that the rate function $\mathbb{R}$ yields no disjoint gang effects (which has just been shown to be equivalent to the SDG) then ensures the inequality (13d), repeated below as (14a). Furthermore, the trivial MG ensures the inequality (14a). In conclusion, we have obtained the inequality $\mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes})$ between rates corresponding to overlapping patterns of *yes*'s.[8]

(14) $$\mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) \overset{(a)}{\leq} \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes}) \overset{(b)}{\leq} \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes})$$

[6] The SDG is equivalent to the trivial BIG of subsection 4.1 when there are only two binary factors (see Magri & Flemming 2025 for discussion). It thus only makes sense to study the SDG when there are more than two factors or when some factor is non-binary. In this paper, we only discuss the former case.

[7] In order to derive the SDG from the ban against disjoint gang effects, we need to cast our net wide when defining disjoint gang effects, by allowing the first two inequalities in (10) among monovalent rates to hold loose. If we were instead to require the first two inequalities in (10) to hold strict in order to have a disjoint gang effect, a ban against disjoint gang effects would not suffice to ensure the crucial inequality (13d) and thus to derive the SDG.

[8] The assumption that the monovalent rates satisfy the loose inequalities (12) can be made without loss of generality. Suppose that we could actually assume (this time with loss of generality) that the monovalent rates also satisfy the strict inequality $\mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{no}) < \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes})$. Then, the assumption that the rate function $\mathbb{R}$ yields no disjoint gang effects

	Example in figure 1	Example in figure 2	Example in figure 3
C_{dont}:	*Vd#	*SCHWA	DEP
C_{apply}:	MAX	*CC	*CCC
C_1:	MAX/Morph	*CLASH	*CLASH
C_2:	MAX/Stress	MAX	*CCC/deriv
C_3:	MAX/V	*CCC	*OLC

Figure 4: Baseline and factor-specific constraints for the three examples described in figures 1-3.

5 How SOT and ME fare with respect to the SDG

This section shows that SOT complies with the SDG, namely it predicts no disjoint gang effects, but it complies for the wrong reason: it predicts virtually no gang effects at all. ME presents the converse problem: while it can generate superset gang effects, it also can generate disjoint gang effects, making it incompatible with the SDG. We first demonstrate these points, then explore how the space of possible ME weight vectors can be pruned to eliminate disjoint gang effects while retaining superset gang effects.

5.1 *Constraint-based set up* We explore the properties of constraint-based probabilistic grammar frameworks through consideration of a variable phonological process that is sensitive to three grammatical factors X_1, X_2, and X_3. Furthermore, we assume that each factor takes only binary values *yes* and *no*. Finally, we assume without loss of generality that the value *yes* promotes application more than the value *no* for each of the three factors. To model this scenario in constraint-based phonology, we assume that the underlying target forms corresponding to the eight feature value combinations have only two candidates each, corresponding to application and non-application of the process considered. Furthermore, we assume that the constraint set consists of five constraints. Two constraints C_{apply} and C_{dont} are violated whenever the process applies and whenever it does not apply, respectively. These are baseline constraints because they are insensitive to the factor values of a target underlying form. The remaining three constraints C_1, C_2, and C_3 are sensitive to the factor values. C_1 is violated when the process does not apply to underlying forms that have value *yes* for the factor X_1. Constraints C_2 and C_3 are defined analogously, only with factors X_2 and X_3 in place of factor X_1. We denote the ranking values/weights of the baseline constraints C_{apply} and C_{dont} as v and w and those of the factor-specific constrains C_1, C_2, and C_3 as w_1, w_2, and w_3. To illustrate, the three examples in figures 1-3 fit into this constraint-based scheme through the constraints in figure 4.[9]

5.2 *SOT predicts no gang effects* SOT more or less complies with the SDG in the sense that it predicts minimal disjoint gang effects. Yet, that is only because it predicts very limited gang effects of any kind, and thus fails to derive well-attested superset gang effects (Jäger & Rosenbach 2006). All of the variable processes discussed in section 3 exhibit superset gang effects. For example, they all show superset gang effects between factors X_1 and X_3. That is, the bivalent rate $\mathbb{R}_{1,3} = \mathbb{R}(\mathit{yes}, \mathit{no}, \mathit{yes})$ is always greater than the monovalent rate $\mathbb{R}_3 = \mathbb{R}(\mathit{no}, \mathit{no}, \mathit{yes})$. In other words, there is a higher rate of application when both factors X_1 and X_3 favor application of the process than when X_3 alone does. To illustrate with English word-final d-preservation, the bivalent rate is $\mathbb{R}_{1,3} = .737$ whereas the monovalent rate $\mathbb{R}_3 = .607$ is strictly smaller.

SOT can only derive minimal superset gang effects, as illustrated in figure 5. In constructing this figure, we assume, without loss of generality, that the ranking values w_1, w_2, and w_3 of the factor-specific constraints C_1, C_2, and C_3 satisfy the inequalities (15). The plot was then generated by considering all ranking values of all five constraints between 0 and 10 at intervals of 1.5, subject to these inequalities (15). For each ranking vector thus obtained, we estimate the monovalent and bivalent SOT rates $\mathbb{R}_3$ and $\mathbb{R}_{1,3}$ (by sampling noise

would not be needed to derive the loose inequality $\mathbb{R}(\mathit{yes}, \mathit{yes}, \mathit{no}) \leq \mathbb{R}(\mathit{no}, \mathit{yes}, \mathit{yes})$ between rates corresponding to overlapping patterns of *yes*'s. In fact, the latter loose inequality would follow from the former strict inequality simply through the BIG, with X_1 and X_3 as target factors and X_2 as the background factor (with shared background values *no* in the antecedent strict inequality and shared values *yes* in the consequent loose inequality).

[9] The most natural constraint for Fasold's example are obviously MAX-type constraints. Interpreting the relevant variable process as d-preservation rather than d-deletion as discussed in footnote 4 allows for these natural MAX-type constraints to be fitted straightforwardly into the general constraint-based scheme adopted here.

vectors 7,500 times). For each monovalent rate $\mathbb{R}_3$ we find the largest corresponding bivalent rate $\mathbb{R}^*_{13}$. These pairs of rates are plotted as red dots with abscissa $\mathbb{R}_3$ and ordinate $\mathbb{R}^*_{13}$. The heights of these red points above the diagonal represent the maximum superset gang effects predicted by the sampled ranking vectors. The fact that these red dots lie very close to the diagonal shows that SOT predicts virtually no superset gang effects between the two factors X_1 and X_3.

(15) $w_1 \leq w_2 \leq w_3$.

5.3 *ME predicts disjoint gang effects* We now turn to ME. The fact that it predicts disjoint gang effects (as well as superset gang effects) follows from basic properties of the ME rates given the constraint-based set up of subsection 5.1. In fact, because of the assumption that each underlying form has only two candidates (corresponding to application and non-application of the variable process considered), the rates of application predicted by a ME grammar for the eight tuples of factor values can be easily expressed in terms of the corresponding constraint weights as in (16). Here, $\mathbf{S}$ is the **sigmoid** function that takes any number x and returns the value $\mathbf{S}(x) = \frac{1}{1+e^{-x}}$ between zero and one (both excluded).

(16)
$$\begin{array}{rclrcl}
\mathbb{R}(no, no, no) & = & \mathbf{S}(w - v) & \mathbb{R}(yes, yes, no) & = & \mathbf{S}(w_1 + w_2 + w - v) \\
\mathbb{R}(yes, no, no) & = & \mathbf{S}(w_1 + w - v) & \mathbb{R}(yes, no, yes) & = & \mathbf{S}(w_1 + w_3 + w - v) \\
\mathbb{R}(no, yes, no) & = & \mathbf{S}(w_2 + w - v) & \mathbb{R}(no, yes, yes) & = & \mathbf{S}(w_2 + w_3 + w - v) \\
\mathbb{R}(no, no, yes) & = & \mathbf{S}(w_3 + w - v) & \mathbb{R}(yes, yes, yes) & = & \mathbf{S}(w_1 + w_2 + w_3 + w - v)
\end{array}$$

These expressions (16) for the ME rates reveal that the weights v and w of the baseline constraints C_{dont} and C_{apply} are irrelevant: only their difference $w - v$ matters. Furthermore, these expressions can be easily used to verify that ME satisfies the BIG from subsection 4.1. To illustrate, let us choose the factors X_1 and X_2 as target and the remaining factor X_3 as background. Furthermore, let us compare the rates corresponding to the target values $X_1 = yes, X_2 = no$ versus $X_1 = no, X_2 = yes$. The BIG demands an inequality between these rates to be independent of the value of the background factor X_3 in the sense of the implication (17), that is a special case of the scheme (6). This implication (17) does indeed hold because the expressions (16) of the ME rates (together with the monotonicity of the sigmoid function $\mathbf{S}$) ensure that the antecedent inequality boils down to $w_1 + w_3 + w - v < w_2 + w_3 + w - v$ (namely $w_1 < w_2$) and the consequent inequality boils down to $w_1 + w - v \leq w_2 + w - v$ (namely $w_1 \leq w_2$).

(17)
$$\begin{array}{rl}
\text{If} & \mathbb{R}(\overbrace{yes, no,}^{\text{target}} \ \overbrace{yes}^{\text{background}}) < \mathbb{R}(\overbrace{no, yes,}^{\text{target}} \ \overbrace{yes}^{\text{background}}) \\
\text{then} & \mathbb{R}(\underbrace{yes, no,}_{\text{target}} \ \underbrace{no}_{\text{background}}) \leq \mathbb{R}(\underbrace{no, yes,}_{\text{target}} \ \underbrace{no}_{\text{background}})
\end{array}$$

Since ME satisfies the BIG, the reasoning in subsection 4.4 applies, showing that ME complies with the SDG if and only if it predicts no disjoint gang effects. Yet, ME does predict massive disjoint gang effects. In fact, the expressions of the ME rates in (16) show that we can order the eight ME rates in any way which is consistent with the BIG by choosing appropriate non-negative constraint weights. In other words, ME satisfies the BIG but no stronger generalization and thus in particular does not satisfy the SDG. In subsection 5.4 we then ask which weights need to be pruned in order for ME not to predict unattested disjoint gang effects and thus to comply with the SDG. Then in subsection 5.5, we address the question whether the ME grammars that survive this weight pruning still manage to predict the attested superset gang effects.

5.4 *Which weights need to be pruned to prevent disjoint gang effects?* We assume without loss of generality that the weights w_1, w_2, and w_3 of the factor-specific constraints C_1, C_2, and C_3 satisfy the inequalities (15). Because of the expressions (16) for the ME rates (together with the monotonicity of the sigmoid function $\mathbf{S}$), these weight inequalities (15) entail that the monovalent rates are ordered accordingly as in (12). Following the reasoning in subsection 4.4, we conclude that an ME grammar predicts no disjoint gang effects as required by the SDG if and only if it satisfies the crucial inequality $\mathbb{R}(yes, yes, no) \leq \mathbb{R}(no, no, yes)$ in (13d) between the bivalent rate corresponding to the weaker factors X_1 and X_2 and the monovalent rate corresponding to the stronger factor X_3. That is, there is a higher rate when the stronger factor alone favors application than if both of the weaker factors favor application. We now unpack this rate inequality as in (18).

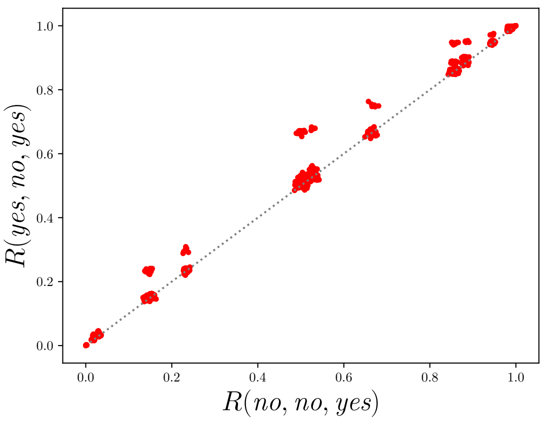

Figure 5: Plot of the OT monovalent rate $\mathbb{R}_3$ (on the horizontal axis) and the largest OT bivalent rate $\mathbb{R}_{1,3}$ on the vertical axis.

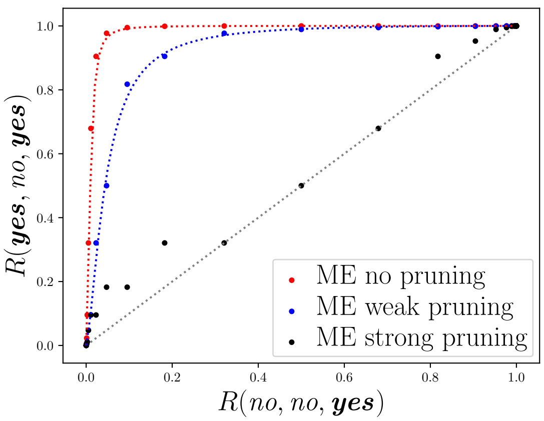

Figure 6: Plot of the ME monovalent rate $\mathbb{R}_3$ (on the horizontal axis) and the largest ME bivalent rate $\mathbb{R}_{1,3}$ on the vertical axis.

Step (18a) holds because of the expressions (16) for the ME rates. Step (18b) holds because the sigmoid function $\mathbf{S}$ is monotonically increasing and therefore respects all inequalities.

$$\text{(18)} \quad \begin{aligned} \mathbb{R}(yes, yes, no) \leq \mathbb{R}(no, no, yes) &\overset{(a)}{\iff} \mathbf{S}(w_1 + w_2 + w - v) \leq \mathbf{S}(w_3 + w - v) \\ &\overset{(b)}{\iff} w_1 + w_2 + w - v \leq w_3 + w - v \\ &\iff w_1 + w_2 \leq w_3 \end{aligned}$$

In conclusion, an ME grammar corresponding to the weights in (15) predicts no disjoint gang effects if and only if these weights actually satisfy the stronger condition (19). In other words, the weight w_3 which is larger than both of the other two weights w_1 and w_2 by (15) is actually larger than their sum $w_1 + w_2$. To secure the SDG in ME, we need to prune all weights that flout condition (19). We refer to this weight pruning condition (19) as **weak** because it only applies to the weights w_1, w_2, and w_3 of the factor-specific constraints C_1, C_2, and C_3, but it does not apply to the weights v and w of the baseline constraints C_{dont} and C_{apply}. In subsection 5.6, we will explore a stronger pruning condition that applies to the latter weights as well.

$$\text{(19)} \quad w_1 + w_2 \leq w_3$$

5.5 *Do the weights that survive pruning predict subset gang effects?* The preceding subsection has shown that, in order to prevent disjoint gang effects in ME, we need to prune the weights that flout the weak pruning condition (19). Do the weights that survive pruning suffice? Do they manage to predict the superset gang effects that are instead attested? To address this question, we start by observing that (weak) weight pruning is consistent with the three data sets described in figures 1-3. To see that, we compute the ME Maximum Likelihood Estimation (MLE) weights for the three sets of rates given the constraints in figure 4 as reported in figure 7. Since the ME rates do not depend on the weights v and w but only on their difference $w-v$, MLE only specifies the difference between the weights of the two baseline-constraints C_{dont} and C_{apply}. These MLE weights satisfy the weak pruning assumption (19), as verified at the bottom of figure 7.

To take a more abstract look, we now probe ME as we have done for SOT in subsection 5.2. Figure 6 is obtained as follows. We consider all five weights between 0 and $\Delta = 10$ at intervals of 0.75 that satisfy the inequalities (15) that we have imposed without loss of generality. For each weight vector thus obtained, we compute the monovalent and bivalent ME rates $\mathbb{R}_3$ and $\mathbb{R}_{1,3}$. For each monovalent rate $\mathbb{R}_3$, we find the largest corresponding bivalent rate $\mathbb{R}^*_{13}$. These pairs of rates are then plotted as red dots with abscissa $\mathbb{R}_3$ and ordinate $\mathbb{R}^*_{1,3}$. A simple calculation (omitted here for reasons of space) shows that these red points fall roughly on the graph of the function $f(x) = \mathbf{S}(2\mathbf{L}x + \Delta)$ plotted as a dotted red curve. The blue dots are obtained analogously, apart from the fact that we exclude the weight vectors that do not comply with the weak pruning condition (19). Again, a simple computation shows that these blue points fall roughly on the graph of the function $g(x) = \mathbf{S}(\frac{3}{2}\mathbf{L}x + \frac{\Delta}{2})$, plotted as a blue dotted curve. The red dots indicate that unpruned ME

	Example in figure 1		Example in figure 2		Example in figure 3	
C_{dont}:	*Vd#	1.0	*Schwa	2.1	Dep	-0.01
C_{apply}:	Max		*CC		*CCC	
C_1:	Max/Morph	0.4	*Clash	0.5	*Clash	0.4
C_2:	Max/Stress	0.4	Max	2.1	*CCC/deriv	0.5
C_3:	Max/_V	1.9	*CCC	2.8	*OLC	1.4
	$0.4 + 0.4 \leq 1.9$		$0.5 + 2.1 \leq 2.8$		$0.4 + 0.5 \leq 1.4$	

Figure 7: The ME MLE weights for the test cases in figures 1-3 satisfy the weak pruning condition (19).

can derive superset gang effects of almost any magnitude. The blue dots bound a somewhat smaller space, showing that (weak) pruning does restrict the range of possible superset gang effects between the monovalent and bivalent rates $\mathbb{R}_3$ and $\mathbb{R}_{13}$. Yet, the blue dots are close to the red dots, showing that weak pruning does not substantially impair ME's ability to predict superset gang effects, as desired.

5.6 *Strong pruning* Weak pruning of ME weights results in a typology that conforms to the SDG while retaining a wide range of superset gang effects, including all of the examples from our case studies. However weak pruning is not straightforwardly applicable as a condition on constraint weights because it relies on a distinction between factor-specific and baseline constraints: weak pruning only restricts the weights of factor-specific constraints. This is problematic because this distinction has no theoretical basis, instead it emerges from the analysis of individual processes. One approach to converting weak pruning into an applicable condition on weights is to generalize it to all constraints: in any subset S of constraints, the highest constraint weight is at least as large as the sum of the weights of the other constraints in S, as stated in (20).

(20) For any constraint subset $S \subseteq \boldsymbol{C}$: if $w_{C_0} = \max_{C \in S} w_C$, then $w_{C_0} \geq \sum_{C \in S \setminus \{C_0\}} w_C$

This strong pruning condition entails the weak pruning condition (19) and thus ensures conformity with the SDG. Furthermore, it does not rely on the problematic distinction between baseline and factor-specific constraints. However it is much stronger than weak pruning precisely because it applies to all constraints. We tentatively suggest that it is indeed too strong. First, it is not satisfied by the analysis of Smith and Pater's schwa realization data: according to figure 7, the highest MLE constraint weight is 2.8 (*CCC), but the remaining constraint weights sum to at least 4.7 (if *CC is assigned a weight of 0). Second, strong pruning of ME weights imposes severe limits on superset gang effects. This is illustrated in figure 6, where the black dots indicate the maximum gang effects under strong pruning for the sampled values of $\mathbb{R}_3$. These black dots are close to the diagonal, leaving only a very narrow range for superset gang effects.

6 Conclusions

In this paper, we have explored the implications of the SDG, a generalization due to Labov (1969) about possible patterns of interaction between the phonological factors that condition the rate of application of variable phonological processes. Informally, the generalization says that these factors and their values can always be ranked in a strict domination hierarchy such that higher ranked factors outweigh the effects of all lower ranked factors in determining the rate of application of a process.

We have shown that the SDG effectively makes a distinction between two kinds of gang effects among factors: superset gang effects, in which addition of factors favoring application of a process increases its rate of application; and disjoint gang effects, in which two individually weaker factors combine to yield a higher rate of application than a single stronger factor. The SDG bans disjoint gang effects while permitting superset gang effects.

The SDG has significant implications for theories of probabilistic phonological grammar because no current model (SOT, ME, NHG) derives it. SOT predicts minimal gang effects of either type, while ME predicts unrestricted gang effects. ME can be restricted to exclude disjoint gang effects while permitting a wide range of superset gang effects by restricting the possible weight vectors. Specifically, the weight assigned to a higher-weighted factor-specific constraint must be equal to or greater than the sum of the weights of all lower-weighted factor-specific weights (weak pruning). However, we do not currently have a way to impose weak pruning via a general condition on constraint weights that does not rely on a process-specific distinction between factor-specific and baseline constraints.

References

Bayley, Robert (2002). The quantitative paradigm. Chambers, Jack K., Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, Blackwell, Oxford, 117–141.

Berdan, Robert (1975). The necessity of variable rules. Fasold, Ralph W. & Roger W. Shuy (eds.), *Analyzing variation in laguage*, Georgetown University school of language and linguistics, 11–26.

Bickerton, D. (1971). Inherent variability and variable rules. *Foundations of Language* 7:4, 457–492.

Boersma, Paul (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam, The Netherlands. The Hague: Holland Academic Graphics.

Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:1, 45–86.

Boersma, Paul & Joe Pater (2016). Convergence properties of a gradual learning algorithm for Harmonic Grammar. McCarthy, John & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, Advances in Optimality Theory, Equinox Publishing, Sheffield, UK and Bristol, CT.

Fasold, R. W. (1978). Language variation and linguistic competence. Sankoff, D. (ed.), *Linguistic variation: Models and methods*, Academic Press, New York, 85–95.

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

Guy, Gregory R. (1980). Variation in the group and the individual: The case of final stop deletion. Labov, William (ed.), *Locating language in time and space*, Academic Press, New York, 1–36.

Hayes, Bruce (2017). Varieties of Noisy Harmonic Grammar. Jesney, Karen, Charlie O'Hara, Caitlin Smith & Rachel Walker (eds.), *Proceedings of the 2016 Annual Meeting in Phonology*, Linguistic Society of America, Washington, DC.

Hayes, Bruce & Colin Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379–440.

Jäger, Gerhard & Anette Rosenbach (2006). The winner takes it all—almost. Cumulativity in grammatical variation. *Linguistics* 44, 937–971.

Kay, Paul & Chad McDaniel (1979). On the logic of variable rules. *Language in Society* 8, 151–187.

Labov, William (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762.

Magri, Giorgio & Edward Flemming (2025). Variation. Jardine, Adam & Paul de Lacy (eds.), *The Cambridge Handbook of Phonology (2nd edition)*, Cambridge University Press.

Smith, Brian W. & Joe Pater (2020). French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5, 1–33.

Storme, Benjamin (2021). Not only size matters: limits to the Law of Three Consonants in French phonology. *Glossa: a journal of general linguistics* 6:1, 1–37.

Wilson, Colin (2025). Maximum entropy grammars. Jardine, Adam & Paul de Lacy (eds.), *The Cambridge Handbook of Phonology (2nd edition)*, Cambridge University Press.

Wolfram, Walt (1973). On what basis variable rules? Bailey, Charles-James & Roger Shuy (eds.), *New ways of analyzing variation in English*, Georgetown University Press, Washington, DC, 1–12.

Wolfram, Walt (1974). *Sociolinguistic aspects of assimilation. Puerto Rican English in New York City*. Center for Applied Linguistics, Arlington, VA.

Wolfram, Walt (1975). Variable constraints and rule relations. Fasold, Ralph W. & Roger W. Shuy (eds.), *Analyzing variation in language. Papers from the second colloquium of new ways of analyzing variation*, Georgetown University Press, Washington, DC, 70–88.

Young, Richard & Robert Bayley (1996). VARBRUL analysis for second language acquisition research. Bayley, Robert & Dennis R. Preston (eds.), *Second language acquisition and linguistic variation*, John Benjamins, Amsterdam, 253–306.