Glossa a journal of general linguistics

Smith, Brian W. and Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5(1): 24. 1–33. DOI: https://doi.org/10.5334/gjgl.583

RESEARCH

French schwa and gradient cumulativity

Brian W. Smith[1] and Joe Pater[2]

[1] University of Southern California, Los Angeles, CA, US

[2] University of Massachusetts Amherst, Amherst, MA, US

Corresponding author: Brian W. Smith (bwsmith.linguist@gmail.com)

We explore the interaction of two phonological factors that condition schwa–zero alternations in French: schwa is more likely after two consonants than a singleton; and schwa is more likely between stressed syllables than elsewhere. Using new data from a judgment study, we show that both factors play a role in schwa epenthesis and deletion, and that the two factors interact cumulatively: they have a stronger effect together than individually. Treating each factor as a constraint, we find that their cumulative interaction is better modeled with weighted than with ranked constraints. We provide a characterization of patterns of cumulativity in probability space in terms of the effect of constraint on its own versus its effect in a cumulative interaction with another constraint. Stochastic OT can model cumulative interactions, but only sublinear ones, where the effect of a constraint is weaker in the cumulative context than on its own. Weighted constraint models, MaxEnt and Noisy HG, can model the full range of cumulativity — sublinear, linear, and superlinear. In examining the ability of these models to fit our experimental data, we find that Stochastic OT is hampered by the fact that the data displays superlinear cumulativity. Noisy HG and MaxEnt fare better on this dataset, with MaxEnt yielding the best fit.

Keywords: harmonic grammar; noisy harmonic grammar; maximum entropy grammars; stochastic optimality theory; French; gradient cumulativity

1 Introduction

In his landmark study originally published in 1973, Dell (1985) provides a remarkably thorough description of the complex set of phonological factors conditioning the schwa-zero alternation in the "standard" variety of Parisian French of which he is a native speaker, and proposes an analysis in terms of the phonological framework presented in Chomsky and Halle (1968). One of the central claims of his analysis is that both deletion of underlying schwa and epenthesis are involved in producing the surface distribution. Examples of deletion of underlying schwa are shown in ((1)a) and ((1)b), and a case of epenthesis is provided in ((1)c). We return to the question of how to distinguish between epenthetic and underlying schwa later in the paper. French schwa is transcribed here as [œ], although there is variation across dialects with respect to its phonetic realization (Durand et al. 1987; Fougeron et al. 2007). Some words contain an [œ] that never alternates with zero, even in phonological environments where deletion is usually likely (Dell 1985). We set aside words with exceptional non-alternating [œ], and only consider words that exhibit [œ]~Ø alternations, such as the examples below.

(1) Schwa deletion and epenthesis
  a. /dœvrɛ/ → [dvrɛ]
    Tu **devrais** partir.
    'You should go.'

b. /mœ/ → [m]
Tu **me** dois de l'argent.
'You owe me money.'

c. /film/ → [filmœ]
un **film** danois
'a Danish film'

In all of these examples, the process applies variably: ((1)a, b) could be produced with a surface schwa, and ((1)c) without one. Although speech rate and speech register affect the probability of deletion in these contexts, according to Dell both variants are possible in what might be described as a neutral rate and register.

In this paper, we focus on the interaction of two phonological factors that affect the probability of schwa deletion and epenthesis. The first factor is whether a singleton consonant or a consonant cluster precedes the schwa. Deletion is less likely, and epenthesis more likely, when schwa is preceded by a cluster. Dell's rule of schwa deletion applies only when a single consonant precedes, as in the examples in ((1)a, b), and his rule of epenthesis applies only after a morpheme that ends in a cluster, as in ((1)c). Dell's analysis abstracts away from the fact that schwa deletion can also apply when a cluster precedes, as in ((2)a, b). Deletion almost certainly applies in these examples with lower probability than in ((1)a, b), but it seems possible for most if not all speakers of this variety.

(2) Examples of deletion in the CC_ context in *devrait* and *me*

a. [ʒak dvʁɛ paʁtiʁ]
Jacques devrait partir.
'Jacques should leave.'

b. [ʒak m dwa dœ laʁʒɑ̃]
Jacques me doit de l'argent.
'Jacques owes me money.'

A second factor that plays a role in conditioning the probability of both deletion and epenthesis is the position of the schwa in the phrase. Deletion is less likely, and epenthesis more likely, when the schwa is followed by a stressed monosyllabic word, and schwa's presence avoids a stress clash (see Section 2 for references). For example, *film* is more likely to be followed by a schwa in *un film russe* [ɛ̃ ˈfilmœ ˈʁys] than *un film danois* [ɛ̃ ˈfilmœ daˈnwa].

Examples like these raise both empirical and theoretical challenges. On the empirical side, data on the relative frequency of outcomes are harder to collect than data on categorical differences. Single speaker intuitions and observations like those of Dell (1985) are invaluable as a starting point, but as we will show, they do not provide the fine-grained data needed to evaluate and compare probabilistic models. On the theoretical side, many phonological frameworks have no way to express the greater probability of the schwa-less realization in ((1)a, b) than in ((2)a, b), let alone explain why particular contexts favor schwa. For example, the standard SPE framework adopted by Dell (1985) allows rules to apply categorically or optionally, but not with some specified probability. One could of course describe the patterns in a Variable Rules model (Labov 1969) by writing the conditioning factors into separate deletion and epenthesis rules, but this model would not make particularly strong predictions. For example, there seems to be no reason that a preceding cluster could not increase the probability of deletion and decrease the probability of epenthesis, the opposite of observed facts.

Constraint-based models address both of these theoretical challenges. Such models allow a single factor, or constraint, to play a role across multiple processes, and as we will discuss below, there are several probabilistic constraint-based models that can generate degrees of optionality as required by the schwa data (see Coetzee & Pater 2011 for an overview of such models, and a comparison with Variable Rules). To model the two phonological factors discussed above, our analysis posits a constraint against stress clash and a constraint against consonant clusters.

(3) Constraints on schwa deletion and epenthesis
a. *Clash
Definition: Assign one violation for every two adjacent stressed syllables.
Effect: Schwa is more likely to be realized in σ́_σ́ than in σ́_σσ́.
b. *CCC
Definition: Assign one violation for every sequence of three consonants.
Effect: Schwa is more likely to be realized in VCC_C than in VC_C.

In French, these two constraints appear to interact *cumulatively*. Schwa is more likely to be realized in contexts where it's favored by both constraints, relative to contexts where it's favored by just one. This is shown in Table 1, which reports the probability of realizing an underlying schwa in the relevant contexts, using probability estimates from our experiment. The rows show schwa between stressed syllables *vs.* elsewhere, and the columns show schwas with a preceding singleton consonant *vs.* a cluster. Contexts like CC_σ́ (e.g., the schwa in *se* in *la terre se vend* 'the land is selling') have a greater probability of realized schwa than C_σ́ (e.g, *le vin se vend* 'the wine is selling') and CC_σσ́ (e.g., *la terre se vend bien* 'the land is selling well').

We use these differences in predictions to compare three constraint-based models of variation in detail: Stochastic OT (Boersma 1997), Noisy Harmonic Grammar (Boersma & Pater 2016), and Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003). All three can capture the fact that CC_σ́ has the highest probability of schwa realization, but the three models permit different patterns of relative probability across the cells. Stochastic OT produces some cumulativity in variable patterns (Jäger & Rosenbach 2006). Cumulative constraint interaction is one of the major predictions of Harmonic Grammar (HG; Smolensky & Legendre 2006) whose weighted constraints produce *gang effects*, and probabilistic variants of HG, such as Noisy HG and MaxEnt, can produce gradient cumulativity. We characterize differences between models in terms of how cumulativity affects probability: sublinearly, linearly, or superlinearly. We show that Stochastic OT produces only sublinear cumulativity, where the effect of a constraint is weaker in the cumulative context than on its own, while the other theories have more subtle restrictions on the patterns they predict.

To compare the three frameworks, we report and model experimental data on French schwa, using judgments from multiple native speakers on the acceptability of realized schwa across contexts. Of the three models, MaxEnt provides the best fit to our data. Our

Table 1: Probability of realizing an underlying schwa extimatedd from experiment.

Following context	Preceding context	
	C_	CC_
_σ́	0.65	0.94
_σσ́	0.56	0.91

results add to a growing body of work showing that weighted constraints provide a better fit to probabilistic natural language data than ranked constraints, particularly when it comes to cumulativity (Guy 1997; Goldwater & Johnson 2003; Benor & Levy 2006; Jäger & Rosenbach 2006; Zuraw & Hayes 2017).[1] The French data also illustrate a prediction of weighted constraints that Zuraw and Hayes (2017) call across-the-board effects, which occur when a constraint has an effect on probabilities in every environment in which the constraint is relevant. In the case of French schwa, the effects of the conditioning factors are mirrored in both epenthesis and deletion contexts, modulo floor and ceiling effects, even in contexts previously reported to show no difference. To our knowledge, this is the first model of variation in French schwa to simultaneously account for both probabilistic epenthesis and probabilistic deletion.

The paper is structured as follows. In Section 2, we provide a brief review of the two phonological factors conditioning French schwa and formalize these factors as phonological constraints. After the presentation of the experiment in Section 3, we present a full model of the data in Section 4, using the probabilities from the experiment to compare different constraint-based models of phonological variation.

2 Schwa epenthesis and deletion

In this section, we provide background on the two phonological factors, repeated in (4), which play a role in both schwa epenthesis and deletion, and define the constraints for the formal analysis.

(4) Phonological conditions on schwa realization
 a. The cluster factor: schwa is more likely to be realized in CC_C than in C_C
 b. The stress factor: schwa is more likely to be realized in σ́_σ́ than in σ́_σσ́

There are three morphological environments where schwa alternates with zero: clitic boundaries, word boundaries, and morpheme-internally. Our analysis assumes that underlying schwas are found morpheme-internally, such as the one in *devrais* in (5a), or at clitic boundaries, such as the one in *me* in (5b). Epenthetic schwas are found at word boundaries, such as the schwa that appears after *film* in (5c) and *veste* in (5d).

(5) Schwa deletion and epenthesis
 a. /dœvʁɛ/ → [dvʁɛ] Tu **devrais** partir.
 b. /mœ/ → [m] Tu **me** dois de l'argent.
 c. /film/ → [filmœ] un **film** danois
 d. /vɛst/ → [vɛstœ] un **veste** rouge

The treatment of underlying and epenthetic schwa is not universal. Dell (1985), for example, treats some word boundary schwas as underlying, such as the one in (5d), consistent with the orthography. Schwas at clitic boundaries are especially controversial, and authors are divided as to whether to treat them as epenthetic (Côté 2000; Côté & Morrison 2007; Kaplan 2016) or underlying (Tranel 1981; Lyche & Durand 1996; Jetchev 1999). We'll focus on the distinction between schwas at word boundaries and clitic boundaries, since those are the two contexts we consider in our experiment and model.

The justification for treating schwas at word boundaries as epenthetic is the alternation's productivity. Schwa can appear at *any* morpheme boundary, given the right phonological context. As shown in the examples below, schwa occurs at word boundaries

[1] Two of these papers — Guy (1997) and Benor and Levy (2006) — compare ranked constraint models to logistic regression, which is nearly equivalent to MaxEnt when there are two candidates per candidate set. All of the papers include Stochastic OT as a ranked constraint model, except Guy (1997), who instead considers Anttila's (1997) model of partially ordered constraints.

(6a) and suffix boundaries (6c), even if there's no orthographic *e* in these contexts (6e). In examples, we follow the notation of Dell (1985) when possible: obligatory (or nearly obligatory) schwas are underlined, orthographic *e*'s that are never (or rarely) pronounced are written ɇ, and relatively optional schwas are in parentheses.[2]

(6) Data from Dell (1985): schwa is realized in the context CC_σ́

a. [yn vɛstœ̲ ʁuʒ] (Dell 1985: 224)
une veste̲ rouge
'a red jacket'

b. [yn vɛst ʁuʒ e blɑ̃ʃ] (Dell 1985: 224)
une vestɇ rouge et blanc
'a red and white jacket'

c. [ɛgzaktœ̲-mɑ̃] (Dell 1985: 228)
exacte̲ment
'exactly'

d. [masiv-mɑ̃] (Dell 1985: 228)
massivɇment
'massively'

e. [ɛ̃ ʃɔʁtœ̲ vɛʁ] (Dell 1985: 237)
un short vert
'a green pair of shorts'

We treat schwas at clitic boundaries as underlying because schwa-zero alternations only occur in a subset of clitics. For example, schwa is optional in the object clitic *te* [tœ] in the context VC_CV, as shown in (7a). The object clitic *leur* [lœʁ], on the other hand, is never followed by a schwa, even when the schwa would be in the same context: VC_CV (7b). A similar restriction can be found by comparing the subject clitics *je* and *elle* in (7c) and (7d). The clitic *je* alternates between [ʒ] and [ʒœ], but *elle* [ɛl] never alternates with [ɛlœ].

(7) Schwa-zero alternations are lexically restricted

a. [sœ kœ ʒo t(œ) di]
ce que Joe t(e) dit
'what Joe told you'

b. [sœ kœ ʒo lœʁ di], *[sœ kœ ʒo lœʁœ di]
ce que Joe leur dit
'what Joe told them'

c. [si ʒ(œ) kuʁ]
si j(e) cours
'if I run'

d. [si ɛl kuʁ], *[si ɛlœ kuʁ]
si ellɇ court
'if she runs'

A model in which schwas at clitic boundaries are epenthetic must prevent epenthesis in contexts such as (7b) and (7d), while motivating optional epenthesis in (7a) and (7c).

[2] Using our experimental data, we can roughly estimate how these notational devices correspond to the probability of schwa realization. Contexts for which schwa realization is described as forbidden, "ɇ", have a probability of schwa realization of up to 0.12 in our experiment, contexts for which realization is described as obligatory, "e̲", have a probability of schwa realization of at least 0.83, and contexts with optional schwa "(e)" range from 0.56–0.68.

To solve this problem, we posit that alternating [œ]'s in clitics are underlying. This analysis has the added benefit of straightforwardly accounting for the generalization that schwa is realized more often in clitics than at word boundaries (see e.g. Côté 2000), a generalization also found in our experimental data. In our analysis, the asymmetry between schwas at clitic boundaries and schwas at word boundaries follows from the faithfulness constraints Max and Dep (McCarthy & Prince 1995).

(8) Max: Assign one violation for every segment in the input without an output correspondent.

(9) Dep: Assign one violation for every segment in the output without an input correspondent.

Max prefers schwa to be realized when it is underlying, while Dep prefers schwa to be absent when it would need to be inserted. Since our model only accounts for schwa in clitics and at word boundaries, these two constraints could be replaced with any set of constraints that favors the realization of schwa at clitic boundaries and disfavors the realization of schwa at word boundaries.[3]

2.1 The cluster factor

For both underlying and epenthetic schwa, schwa is realized more often after two or more consonants than after a singleton consonant. The examples in (10) show this for deletion, while controlling for phrase position. In all examples, schwa is also followed by a consonant, since schwa is rarely realized adjacent to a vowel.

(10) Deletion and the cluster factor (Dell 1985: 228–229)

a. [mɑ̃ʒ lœ̲ gato] CCe̲ σσ
mange le̲ gateau

b. [mɑ̃ʒɛ l(œ) gato] C(e) σσ
mangez l(e) gâteau

c. [ʒak dœ̲vʁɛ paʁtiʁ] CCe̲ σσσ
Jacques de̲vrait partir

d. [ɑ̃ʁi d(œ)vʁɛ paʁtiʁ] C(e) σσσ
Henri d(e)vrait partir

The number of preceding consonants also plays a role in epenthesis, as shown in (11).

(11) Epenthesis and the cluster factor (Côté 2007)

a. [la sɛkt(œ) paʁtɛ] CC(e) σσ
la sect(e) partait

b. [l astɛk paʁtɛ] Cɇ σσ
l' Aztèquɇ partait

Support for the probability judgments above are found in our experimental results, previewed in Table 2. The data show that across rhythmic and morphological contexts, schwa

[3] One such possibility is presented in Kaplan (2016), who analyzes schwa at clitic boundaries as epenthetic, driven by constraints requiring alignment between clitic boundaries and syllable boundaries, e.g. Align(accusative, L; σ, L). This account is compatible with our data, and can address the asymmetry in (8). The difference between word boundary schwas and clitic schwas would follow from the fact that alignment favors schwa in clitics (but not at word boundaries), while Dep disfavors schwa generally.

Table 2: Probability of schwa realization from our experiment.

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.65	0.94
	_σσ́	0.56	0.91
Epenthetic schwa	_σ́	0.12	0.83
	_σσ́	0.09	0.68

is realized more often in the context CC_ than in the context C_. These results also show that schwa is generally realized more often when it's underlying (at a clitic boundary) than when it's epenthetic (at a word boundary), as described in the previous section.

In the constraint-based models that follow, we model the cluster factor with the constraint *CCC, which militates against a sequence of three consonants (Grammont 1914).

(12) *CCC: Assign one violation for every sequence of three consonants.

Under the *CCC analysis, schwa is inserted in phrases such as *la secte partait* [la sɛktœpaʁtɛ] because schwaless [la sɛktpaʁtɛ] contains the cluster [ktp]. Similar constraints have been used in previous analyses of French schwa, such as Côté's (2000; 2007) constraint *C↔V, which requires every consonant to be next to a vowel, and Kaplan's (2011) constraint *NTN, which militates against stops flanked by non-approximants. For the data we model, *NTN and *CCC are interchangeable, since all of the CCC sequences in our data set contain a stop as the second consonant in the cluster.

2.2 The stress factor

An effect of phrase position on the probability of realizing schwa has been observed at least since Léon (1966), who describes schwa as more likely to be realized in the penultimate syllable (see also Morin 1974; Dell 1985; Tranel 1987). The examples below show this for schwa at word boundaries (13) and clitic boundaries (14).

(13) Epenthesis: position plays a role when schwa is after two consonants
(Morin 1974: 77)
a. [lœ gaʁdœ mɑ̃] CC e σ
le garde ment
'the guard lies'
b. [lœ gaʁd(œ) mɑ̃ˈtɛ] CC (e) σσ
le gard(e) mentait
'the guard was lying'

(14) Deletion: position plays a role when schwa is after two consonants (Dell 1985: 231)
a. [la tɛʁ sœ vɑ̃] CC e σ
la terre se vend
'the land is selling'
b. [la tɛʁ s(œ) vɑ̃ bjɛ̃] CC (e) σσ
la terre s(e) vend bien
'the land is selling well'

In both (13) and (14), schwa occurs after two consonants in the context VCC_C. In the context VC_C, it has been claimed that there is no effect of the number of following syllables, regardless of whether the schwa is at a word boundary or clitic boundary. Côté and Morrison (2007: 169) report that the probability of schwa realization is the same in (15a) and (15b), as well as the same in (15c) and (15d).

(15) Realization of schwa in the context VC_C is unaffected by the number of following syllables (Côté & Morrison 2007: 169)

a. [lo s(œ) vɑ̃] C (e) σ
l'eau s(e) vend
'water sells'

b. [lo s(œ) vɑ̃ bjɛ̃] C (e) σσ
l'eau s(e) vend bien
'water sells well'

c. [il dɔn pø] C e̸ σ
il donne̸ peu
'he gives little'

d. [il dɔn boku] C e̸ σσ
il donne̸ beaucoup
'he gives a lot'

Contrary to the generalization in (15), our experimental data show an effect of the number of following syllables even after a single consonant. Across segmental and morphological contexts, schwa is realized more often before one syllable than before two syllables, although the effect is very weak at floor and ceiling, when probabilities are close to 0 (as in C_ at a word boundary) or 1 (as in CC_ at a clitic boundary).

In our model, the fact that schwa is realized more often before one syllable than two follows from stress clash avoidance. The constraint *Clash favors schwa when it occurs between two stressed syllables.

(16) *Clash: Assign one violation for every two adjacent stressed syllables.

This approach is similar in spirit to the analysis of Mazzola (1991; 2014), who proposes a stress-based analysis of both schwa and stress assignment, but does not formalize it using constraints.

Stress in French is not fixed at the word level, but falls on the last non-schwa syllable of the phonological phrase (Grammont 1914; Delattre 1939; Jun & Fougeron 2000), with additional stresses on the final full syllable of every lexical word, unless they result in a stress clash (Post 2000). This is shown by the examples in (17), which assume the phonological phrasing and stress assignment rules of Post (2000). According to Post, phonological phrasing is variable, and for a N + Adj sequence, both (17b) and (17c) are possible. We use the term "stress" here for convenience, although some of the references we cite treat prominence in French as pitch-accent (Post 2000) or tone (Jun & Fougeron 2000).

(17) Examples of stress assignment in French

a. (le garde)$_{PP}$ (mentait)$_{PP}$ 'the guard was lying'
σ σ́ σ σ́

b. (une veste marron)$_{PP}$ 'a brown jacket'
σ σ̀ σ σ́

c. (une veste)$_{PP}$ (marron)$_{PP}$ 'a brown jacket'
σ σ́ σ σ́

Given that stress always occurs on the last full syllable of the phonological phrase, when schwa is followed by a phrase-final monosyllabic word, it's also followed by a stressed syllable. As shown in (18), *CLASH prefers schwa before a monosyllabic word in (18a) and (18c), where schwa's realization avoids a stress clash, but not before a disyllabic word in (18b) and (18d), where stress clash is impossible.

(18) Schwa insertion avoids a stress clash
a. (le garde)$_{PP}$ (ment)$_{PP}$ [lœ ˈgaʁdœ̲ ˈmɑ̃] σ́e̲ σ́
b. (le garde)$_{PP}$ (mentait)$_{PP}$ [lœ ˈgaʁd(œ) mɑ̃ˈte] σ́(e) σσ́
c. (une veste rouge)$_{PP}$ [yn ˌvɛstœ̲ ˈʁuʒ] σ̀e̲ σ́
d. (une veste marron)$_{PP}$ [yn ˌvɛst(œ) maˈʁɔ̃] σ̀(e) σσ́

*CLASH has been used in previous analyses of French to account for the realization of both primary and secondary stress (Jun & Fougeron 2000; Post 2000). Primary stress is less likely to be realized when the following syllable is stressed (Jun & Fougeron 2000), and secondary stress is not realized before a stressed syllable, optionally surfacing on an earlier syllable in the phrase (Verluyten 1982; Tranel 1987; Mazzola 1991; 2014; Post 2000). An example of this alternation is shown in (19). Before an unstressed syllable, *l'ami* is realized with final stress (19a), but before a stressed syllable, stress retracts (19b). When schwa is realized in (19c), stress falls on the final syllable of *l'ami*, since the realization of schwa prevents a stress clash.

(19) Clash resolution, stressed syllables are in small caps (Tranel 1987: 200)
a. [laˌ midalˈ fʁɛd] σσ̀ σσ́
l' aMI d' AlFRED
'Alfred's friend'

b. [ˌlamidˈ pjɛʁ] σ̀σ σ́
L'Ami d' PIERRE
'Pierre's friend'

c. [laˌ midœˈ pjɛʁ] σσ̀e σ́
l' aMI de PIERRE
'Pierre's friend'

Crucially, the example in (19c) shows that schwa can serve as a buffer between stresses, avoiding a stress clash and making retraction unnecessary.

Côté (2007) presents a number of arguments against a clash-based account like the one outlined above. She points out that the number of syllables following schwa has an effect on its realization, even when there is only one stressed syllable.

(20) A position effect without stress clash (Côté 2007)
a. [dœ̲ ˈlo] e̲σ́
de̲ l'eau
'some water'

b. [d(œ) loˈdas] (e)σσ́
d(e) l'audace
'some audacity'

(21) *venez* in Dell (1985: 227), schwa is more likely in (a) than (b)
a. [v(œ)ˈne] (e)σ́
v(e)nez
'come'

b. [v(œ)ne iˈsi] (e)σσ́
v(e)nez ici
‘come here’

Côté (2007) argues that these data cannot be accounted for *Clash, and instead require a different constraint, such as one enforcing prosodic minimality. As further support for an account with prosodic minimality, Côté (2007) reports pairs like *jette de̲ l’ortie* and *achète d(e) l’ortie*, which show that schwa realization is also conditioned by the number of *preceding* syllables. Schwa is more likely to be realized when it’s preceded by one syllable than when it’s preceded by two. Although Côté’s data seem to require prosodic minimality, such an analysis is not inherently incompatible with our clash-based analysis. Given the seemingly contradictory data in (20) and (21), it is likely that both prosodic minimality and stress clash avoidance are independently necessary to account for the distribution of stress and schwa in French, and a complete analysis of all of the French schwa facts would include both contraints. Since our focus is on the interaction of a few select factors, we leave the fuller account for future work.

2.3 Other restrictions on schwa

One last restriction on schwa, which is relevant to our experimental design, is that schwa generally doesn’t occur next to another vowel, even in contexts where the stress factor favors its realization.

(22) No schwa next to a vowel
a. [ɛma tɛd], *[ɛma tœ ɛd]
Emma t’aide
‘Emma helps you’
b. [ɛma tœ gid]
Emma te̲ guide
‘Emma guides you’
c. [uvʁ œf], *[uvʁœ œf]
ouvre̸-oeuf
‘egg opener’
d. [uvʁœ bwat]
une ouvre̲-boîte
‘can opener’

The exception to this generalization is h-aspiré words, which phonetically begin with a vowel (or glottal stop), but pattern in many ways as if they begin with a consonant (see e.g. Boersma 2007 on how h-aspiré words differ from both C-initial words and V-initial words). We set those aside here.

3 Experiment

As shown in the previous section, the realization of schwa is conditioned by segmental context and rhythmic context, which we analyze using the constraints *CCC and *Clash. Additionally, the realization of schwa is conditioned by whether it occurs at a clitic boundary or word boundary, which we analyze as the result or Max and Dep.

This section reports the results of a judgment experiment designed to estimate the probability of schwa across contexts, and determine how the four constraints contribute to the probability of schwa realization.

3.1 Experimental design

We conducted the experiment over the internet, using the web-based psycholinguistics experiment platform Ibex (Drummond 2013). The experiment employed a forced choice paradigm, in which participants were asked to imagine that they were speaking with a friend, and choose between two variants of a phrase: one with a pronounced schwa and one without the schwa. Choices were presented in French orthography. Pronounced schwa was indicated with an orthographic *e*, and unpronounced schwa was indicated with an apostrophe, which is sometimes used to mark deleted schwas in songs to aid rhythmic parsing, or in some colloquially written words (e.g. *p'tit* for *petit*). For forms that didn't contain an *e* in the orthography, a pronounced schwa was indicated with an *e* in parentheses (e.g. *un toast(e) chaud vs. un toast chaud* 'a warm piece of toast'). During a pre-experiment practice phase, participants received extra instructions for these forms, and were instructed to treat an *e* in parentheses as a pronounced schwa, and given the example of *film(e) russe vs. film russe*.[4]

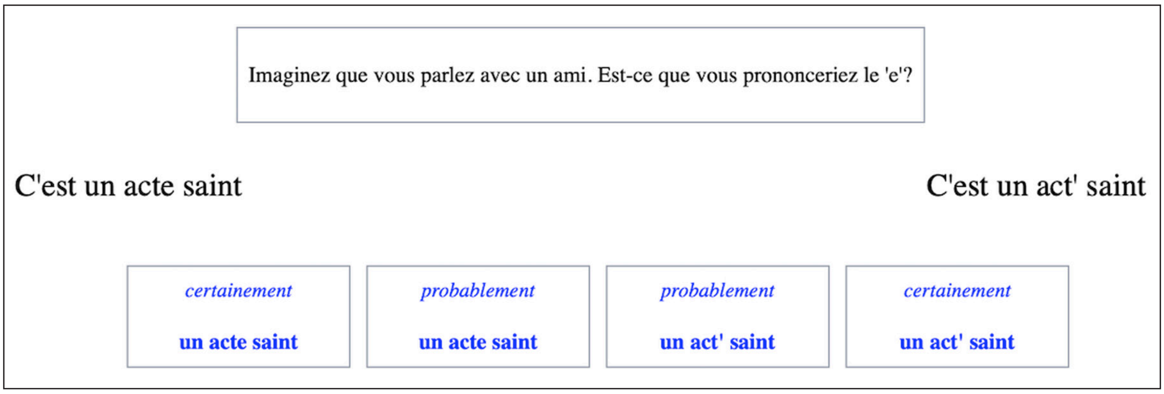
In addition to choosing between schwa and no schwa, participants indicated their confidence in the answer as *certainement* or *probablement*. *Probablement* and *certainement* responses are pooled in statistical analyses of the data, which model the probability of schwa realization.[5] A screen capture of the experiment in progress is shown in Figure 1.

Previous work has shown that French speakers are capable of estimating the frequency of schwa realization in this manner. For example, Racine (2008) asked speakers to complete a written questionnaire, rating the acceptability of a list of words pronounced with and without schwa. Racine reports that the results of the rating task are very strongly correlated with probabilities of schwa obtained from a production experiment (r = 0.79).

The experiment followed a 2 × 2 × 2 factorial design, with 8 conditions.

(23) Factorial design

a.	Cluster before schwa site	C_ *vs.* CC_
b.	Position of schwa site	_σ́ *vs.* _σσ́
c.	Underlying or epenthetic schwa	clitic boundary *vs.* word boundary



Imaginez que vous parlez avec un ami. Est-ce que vous prononceriez le 'e'?

C'est un acte saint

C'est un act' saint

certainement **un acte saint**

probablement **un acte saint**

probablement **un act' saint**

certainement **un act' saint**

Figure 1: Screenshot of experiment in progress.

[4] There were three nouns requiring an *e* in parentheses (*toast, lac*, and *bec*), included out of necessity due to a scarcity of nouns that fit the criteria for our experimental items. While it's possible that *lac(e)* (intended as *lac* + œ, [lakœ]), was interpreted as *lace* (*lace* + œ, [lasœ]), no qualitative differences were observed between these three nouns with parenthetical *e* and other experimental items.

[5] *Certainement* is usually translated as "definitely", and *probablement* as "probably", but an anonymous reviewer points out that for many speakers, *certainement* indicates a probability less than "definitely" but greater than "probably". Since we pool *certainement* and *probablement* responses, these subtle differences should not have an effect on the results.

The construction of items differed for underlying and epenthetic schwas. Items with epenthetic schwas were constructed according to the template in (24), consisting of a noun followed by a post-nominal adjective, with the site of the epenthetic schwa at the boundary between them.

(24) C'est un <Noun> <Adjective>
<Noun>: C-final or CC-final, all final consonants are obstruents, mostly monosyllabic
<Adjective>: σ́ or σσ́, all obstruent-initial

Depending on the condition, nouns ended in either one or two consonants, and adjectives were one or two syllables long. We controlled for segmental and prosodic context as much as possible. All but two nouns were monosyllabic, and the disyllabic nouns were balanced across conditions. All nouns in the experiment ended in stops, and all adjectives began with obstruents. This ensured that all clusters in the experiment consisted of only obstruents, and in three-consonant clusters, the middle consonant was always a stop, controlling for the influence of sonority on the rate of schwa realization. Examples of the four epenthesis conditions are in (25), with parentheses indicating the alternating *e*. Each participant saw each noun and adjective only once. The full list of items is included in the appendix.

(25) Examples of epenthesis items with alternating schwa in parentheses

a. C_σ́
[yn bɔt(œ) ˈʒon]
une bott(e) jaune
'a yellow boot'

b. CC_σ́
[yn vɛst(œ) ˈʒon]
une vest(e) jaune
'a yellow jacket'

c. C_σσ́
[yn bɔt(œ) ʃinˈwaz]
une bott(e) chinoise
'a Chinese boot'

d. CC_σσ́
[yn vɛst(œ) ʃinˈwaz]
une vest(e) chinoise
'a Chinese jacket'

Deletion items contained the clitic *te*, the 2nd person object clitic, which we assume to be underlyingly /tə/. In these items, *te* was preceded by a name and followed by a verb, e.g. *Maurice te cite* ('Maurice cites you'). We used only one type of clitic to control for the fact that clitics may differ in their propensity to undergo deletion (Malécot 1974), and also to ensure that all CCC clusters had similar sonority profiles, with a stop as the medial consonant.

(26) <Name> te <Verb>
<Name>: C-final or V-final, all final consonants are obstruents, disyllabic
<Verb>: σ́ (present) or σσ́ (imperfect), all obstruent-initial

All of the names that occurred before *te* were disyllabic, and ended in either a consonant or a vowel, depending on the condition. The schwa in *te* is preceded by one consonant when the name is V-final, and two consonants when the name is C-final. Position of schwa was manipulated by using different tenses of verbs. In the present tense, these verbs are monosyllabic. In the imperfect tense, the suffix *-ait* /-ɛ/ creates a disyllabic verb. Examples of the four deletion conditions are in (27). Each participant saw every name and verb lexeme only once.

(27) Examples of deletion items with alternating schwa in parentheses
a. C_σ́
[eva t(œ) ˈʃɔk]
Eva t(e) choque
'Eva shocks you'
b. CC_σ́
[mɔʁiz t(œ) ˈsit]
Maurice t(e) cite
'Maurice cites you'
c. C_σσ́
[eva t(œ) ʃɔˈkɛ]
Eva t(e) choquait
'Eva shocked you'
d. CC_σσ́
[mɔʁiz t(œ) siˈtɛ]
Maurice t(e) citait
'Maurice cited you'

Each participant saw 6 items per condition, 24 for deletion and 24 for epenthesis, in addition to 30 fillers. Fillers consisted of tenses (simple past, simple future) and phonological environments that differed from the test items. Most importantly, some fillers contained phrases with schwa adjacent to vowels, which we used as catch trials. We excluded from analysis any participant who judged that schwa should *certainement* be realized when adjacent to a vowel. The design is summarized in (28).

(28) Summary of experimental design
78 judgments per participant
24 deletion: 6 per type in (25), no name or verb repeated
24 epenthesis: 6 per type in (27), no adjective or noun repeated
20 fillers for deletion (e.g. Anna s(e) est levée)
10 fillers for epenthesis (e.g. un iguan(e) solitaire)

3.2 Participants and exclusions

Participants were recruited over the internet through word of mouth. We excluded any participant who did not self-identify as a native speaker of French or chose *certainement* for the realization of schwa adjacent to vowels in catch trials once or more, leaving data for 27 participants after exclusions. Most participants were either from Île-de-France (7/27), Pays de la Loire (7/27), or Auvergne-Rhône-Alpes (6/27), and participant ages ranged from 21 to 48 (mean = 35.11). Location data is included in the appendix.[6]

[6] Location is relevant because French schwa is subject to regional variation. Interspeaker differences (e.g., region, gender, age, social class) are discussed in the context of the statistical model in the next section, where we use random effects in our statistical model to minimize the influence of interspeaker differences on our conclusions.

3.3 Results

The proportion of schwa responses for both underlying and epenthetic contexts are presented in Table 3 and the barplot in Figure 2.

Across all four phonological contexts, schwa is judged as better in deletion contexts than in epenthesis contexts. Schwa is also generally judged as better after two consonants than one consonant (the cluster factor), and better before one syllable than two syllables (the stress factor).

To evaluate the statistical significance and effect size of the factors, we fit a mixed effects logistic regression model in R (R Core Team 2017) using the package lme4 (Bates et al. 2015). The dependent variable in the model is the probability of schwa (expressed as log-odds). The model contains the fixed effects in the table in Table 4, each of which corresponds to an experimental condition, in addition to an interaction term for Stress × Seg. The model also contains a maximal random effects structure, with random intercepts for subject and item, and random slopes by subject for all of the fixed effects (including the interaction term).[7]

Given interspeaker variation in the production of schwa, the use of a random intercepts and slopes ensures that the model generalizes across speakers. The inclusion of random

Table 3: Proportion of schwa realization from experiment. The values in parentheses indicate the range of the 95% confidence interval, specifically the Wilson score interval.

	Following context	**Preceding context**	
		C_	**CC_**
Underlying schwa	_σ́	0.65 (0.57–0.72)	0.94 (0.89–0.97)
	_σσ́	0.56 (0.48–0.64)	0.91 (0.86–0.95)
Epenthetic schwa	_σ́	0.12 (0.08–0.18)	0.83 (0.76–0.89)
	_σσ́	0.09 (0.05–0.14)	0.68 (0.61–0.75)

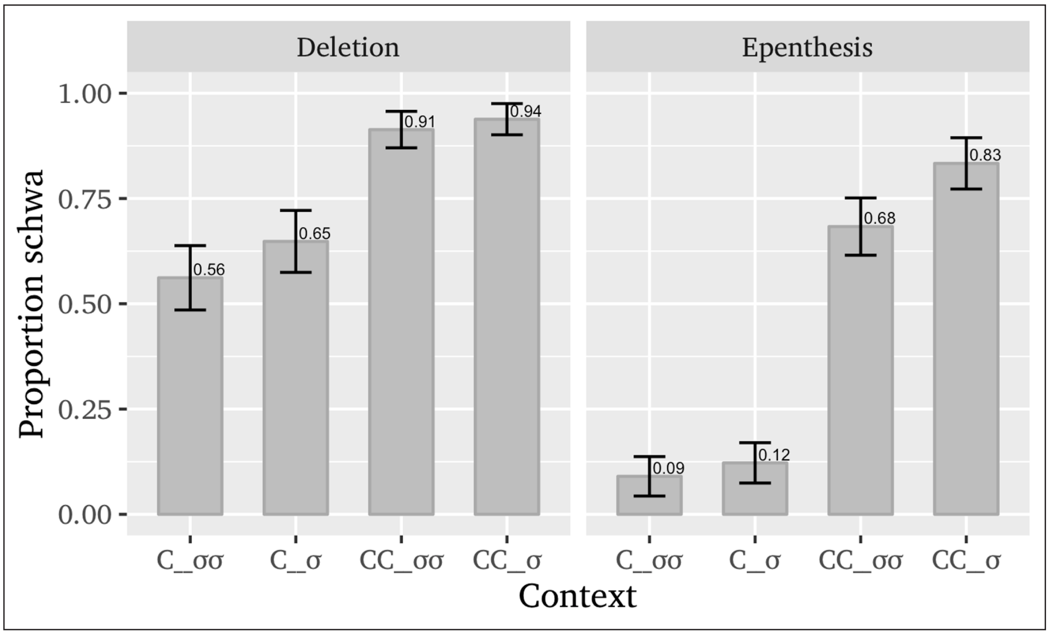


Figure 2: Proportion of schwa realization from experiment. Whiskers show Wilson score intervals.

[7] The glmer equation in R: Schwa ~ Ep/Del + Stress * Seg + (1 | Item) + (1 + Ep/Del + Stress * Seg | Subject).

intercepts means that speakers with exceptionally high or low baseline rates of schwa will have less of an influence on the estimates of the model predictors, while the inclusion of random slopes means that some speakers can be exceptionally sensitive or insensitive to phonological context. In this way, the random effects structure controls for dialectal and sociolinguistic variation, both of which are well-documented for French schwa. Previous work modeling French schwa has taken the same approach. In Bayles et al. (2016) and Bürki et al. (2011), the inclusion of random effects is shown to help control for inter-speaker differences in the realization of French schwa and significantly improve model fit.

The coding of the fixed effects is shown in Table 4. All of the categorical variables in the model were sum coded, as shown in the Coding column. For each variable, the higher level (+1) is the context predicted to favor the realization of schwa.

The fitted values for the model are shown in Table 5. A positive coefficient means the probability of schwa realization increases when the predictor is +1 and decreases when the predictor is –1. A negative coefficient means the probability of schwa realization decreases when the predictor is +1. The rightmost column, Pr > |Z|, shows p-values for the Wald test.

All fixed effects are significant, except the interaction of Stress × Seg. The presence of a preceding cluster has the biggest effect on the realization of schwa; as shown by the coefficient of Seg ($\beta = 1.75$), schwa is more likely after clusters than singletons. Schwa is also more likely in deletion contexts than epenthesis contexts ($\beta = 1.48$), and more likely when followed by one syllable than when followed by two ($\beta = 0.31$). Although the effect of stress is relatively small, it's significant in the model. The lack of significance for Stress × Seg suggests that the effect of stress is not limited to one segmental context (or vice versa). Both Stress and Seg exhibit independent effects on the probability of schwa realization.

4 Presentation of modeling results

In this section, we compare the ability of three models of variation to fit our experimental data: MaxEnt, Stochastic OT and Noisy HG. In the first section, we introduce the models by discussing some of the distributions that each one can generate for a

Table 4: Coding of fixed effects for regression model.

Fixed effect	Level	Coding
Stress (stress factor)	_σ́	+1
	_σσ́	–1
Seg (cluster factor)	CC_	+1
	C_	–1
Ep/Del (epenthesis or deletion)	Deletion	+1
	Epenthesis	–1

Table 5: Mixed effects model: logistic regression (positive = greater likelihood of schwa).

	Coefficient (β)	S.E.	Z	Pr > |Z|
(Intercept)	0.94	0.26		
Stress = _σ́	0.31	0.11	2.70	<0.01
Seg = CC_	1.75	0.15	11.51	<0.001
Ep/Del = deletion	1.48	0.24	6.25	<0.001
Stress × Seg	–0.06	0.11	0.55	0.59

subset of the French contexts, and some of the restrictions that each model places on the distributions it can generate relative to the other models. We then show how the models fare in fitting the actual French data. There has been some previous comparison of these theories (see Hayes & McPherson 2016 and Pater 2016 and references therein); the following discussion draws in particular on Jäger & Rosenbach's (2006) comparison of Stochastic OT and MaxEnt, Pizzo's (2015) discussion of sublinearity in MaxEnt phonotactics, and Zuraw and Hayes' (2017) comparison of Noisy HG and MaxEnt with Stochastic OT.

4.1 The models

4.1.1 Constraint set and violation profiles

To illustrate how the models function, we will consider the set of contexts that we analyze as environments for schwa deletion, as opposed to epenthesis. The constraints are given in (29–31). For simplicity, we omit faithfulness constraints here, but include them below when needed.

(29) *CCC: Assign one violation for every sequence of three consonants.

(30) *CLASH: Assign one violation for every two adjacent stressed syllables.

(31) NOSCHWA: Assign one violation for every [œ] in the output.

The contexts are illustrated in Table 6. the schwa is either between stressed syllables (top row) or not (bottom row); and the schwa follows either a singleton (left column) or a cluster (right column).

We consider two candidates for each context: faithful realization of an underlying schwa, and deletion. The tableau in (32) shows violations for the two candidates in the context where two constraints are violated by deletion. Violations are marked with negative integers.

(32) Constraint violations marked with negative integers

la terre se̲ vend	NOSCHWA	*CCC	*CLASH
Deleted schwa: [laˈtɛʁsˈvɑ̃]		–1	–1
Realized schwa: [laˈtɛʁsœˈvɑ̃]	–1		

The table in (33) uses the more compact representation of difference vectors, which result from subtracting the deletion candidate's violations from the faithful candidate's violations. Positive values indicate constraints that prefer schwa's presence, negative values indicate constraints that prefer schwa's absence, and zeroes indicate constraints that are indifferent.

Table 6: Examples of schwa in the four phonological contexts to be modeled.

Following context	**Preceding context**	
	C_	**CC_**
_σ́	le vín s**e** vend	la térre s**e** vénd
_σσ́	le vín s**e** vend bíén	la térre s**e** vend bíén

(33) Difference vectors for constraint scores: negative values favor schwa's absence, positive values favor schwa's presence

	NoSchwa	*CCC	*Clash
la terre s**e** vend	–1	+1	+1
la terre s**e** vend bien	–1	+1	0
le vin s**e** vend	–1	0	+1
le vin s**e** vend bien	–1	0	0

The table of difference vectors in (33) clearly shows the trade-offs in constraint violations in each context. Faithful realization of the schwa always violates NoSchwa, and deletion always satisfies it, so all contexts have a value of –1 for NoSchwa, indicating a penalty for schwa presence. This penalty trades off against a reward for schwa realization that depends on the environment. In all of the models we consider, the probability of schwa realization will always be greatest in the environment in which both *CCC and *Clash are relevant (the topmost row), and will always be the lowest in the environment in which neither is relevant (the bottom row). This sets these constraint-based models apart from a Variable Rules model. As we mentioned in the introduction, such a model could in principle make a schwa deletion rule apply with higher probability in any of the environments (see Coetzee & Pater 2011 for further related discussion). As we will shortly examine in detail, the three probabilistic constraint-based models differ in exactly how rewards can accumulate in terms of differences in probability as we move up the rows.

In Optimality Theory (OT: Prince & Smolensky 2004), schwa realization is optimal *iff* a schwa-preferring constraint is ranked above NoSchwa. For example, given the ranking *CCC ≫ NoSchwa ≫ *Clash, schwa realization will be optimal in just the top two rows of (33), in which *CCC prefers it, but not the bottom two, in which *CCC is indifferent.

In a deterministic version of Harmonic Grammar (HG; see Smolensky & Legendre 2006; and Pater 2016 and references therein), the optimal candidate is the one whose weighted sum of constraint scores, or *Harmony*, is the highest. In the tableau in (34), we show the Harmonies that result with a set of constraint weights shown beneath the constraint names. The summed weights of *CCC and *Clash are greater than that of NoSchwa, so the candidate that violates them – the deletion candidate in (34a) – receives a greater penalty: –4 *vs.* –3. Candidate (34b) has higher Harmony (the negative number closer to zero) and is optimal.

(34) Cumulativity in deterministic Harmonic Grammar

la terre se vend	NoSchwa w = 3	*CCC w = 2	*Clash w = 2	Harmony
a. [laˈtɛʁsˈvɑ̃]		–1	–1	–4
b. → [laˈtɛʁsœˈvɑ̃]	–1			–3

In terms of our difference vectors, schwa realization is optimal when the sum of the difference scores, each times its constraints' weight, is above zero (see further Pater 2016). For example, with NoSchwa having a weight of 3, and each of the other constraints having a weight of 2, schwa presence would be optimal in only the corresponding top row of (35), where the result of the just-described equation is +1. In the middle two rows and bottom row, the equation would yield –1 and –3, indicating that deletion is optimal.

(35) Weighted Harmony differences

	NoSchwa	*CCC	*Clash	*Sum of weighted difference scores*
	3	2	2	
la terre s**e** vend	–1	+1	+1	+1
la terre s**e** vend bien	–1	+1	0	–1
le vin s**e** vend	–1	0	+1	–1
le vin s**e** vend bien	–1	0	0	–3

This gang effect, or cumulative constraint interaction, cannot be modeled in standard deterministic OT: no ranking of these constraints will produce an output schwa in only the top row; it will always be accompanied by an optimal output schwa in one of the middle rows. In the next section, we will see that a probabilistic version of OT will give the top row higher probability than the middle ones.

4.1.2 Sublinear cumulativity in Stochastic OT

Stochastic OT (Boersma 1997; Boersma & Hayes 2001) is a probabilistic variant of OT. Each constraint is given a real numbered *ranking value*, and when the grammar is used to evaluate a candidate set, the ranking values are converted to an ordinal OT ranking. Variation occurs because the ranking values are perturbed by noise before conversion to ranking: each constraint value has a real number added to it that is sampled from a Gaussian distribution centered on zero (resampled for each constraint). As Jäger & Rosenbach (2006) point out, this model predicts greater probability in a gang effect context like the top row of the table in (33). To see this, consider the case when the constraints are tied in value (e.g. 1, 1, 1). In such a case, the probability of one constraint being ranked above another is 0.5, which is the probability of realized schwa in each of the middle rows. In the top row, the realized schwa is optimal if *either* *CCC or *Clash ranks above NoSchwa, which obtains in 4/6 rankings, thus yielding a probability of 0.67.

Jäger & Rosenbach (2006) identify two differences between the patterns of gradient cumulativity that can be generated by Stochastic OT and MaxEnt. One is that if the two violations in the cumulative case come from a single constraint, in what they call counting cumulativity, Stochastic OT will not show an increase in probability when there are two violations being avoided instead of just one. The other (Jäger & Rosenbach 2006: 939) is an observation they attribute to Paul Boersma: that there are patterns of "strong" cumulativity that cannot be represented by Stochastic OT, but can be represented by MaxEnt. Our formalization of the "weakness" of Stochastic OT is that its cumulativity is always *sublinear*.

To get to a definition of sublinear cumulativity, we must first explain what we mean by the *contribution* of a constraint. A constraint's contribution to the probability of an outcome is the difference between the probability of that outcome in a context in which the constraint applies, compared to a minimally different context in which it is irrevelant. In Table 6, we show the four contexts we have been discussing, with labels A through D. A is the context where neither constraint is relevant (neutral), B and C are the ones where

Table 6: Contexts: neutral (A), non-cumulative (C, B), and cumulative (D).

	_σσ́	_σ́
CC_	C	D
C_	A	B

just one constraint is relevant (non-cumulative) and D is the one where both are (cumulative). We write the contribution of a constraint on its own as ΔConstraint. Δ*CCC is the probability of schwa in C minus the probability in A, and Δ*Clash is the probability in B minus the probability in A. The results of these calculations are shown in Table 7. We can compare the contribution of a constraint alone to the contribution of the constraint in conjunction with another (its *cumulative contribution*). We write the contribution of a constraint (Con1) in conjunction with another (Con2) as ΔCon1|Con2. ΔCon1|Con2 is the difference between the cumulative context and the non-cumulative context for Con2. In Table 7, Δ*CCC|*Clash is (D – B), and Δ*Clash|*CCC is (D – C).

Whether a case of cumulativity is *sublinear, linear*, or *superlinear* depends on how the cumulative contribution of a constraint compares with its independent contribution. If the cumulative contribution is less than the independent contribution, the pattern is *sublinear*. If the cumulative contribution is greater than the independent contribution, the pattern is *superlinear*. If the two contributions are equal, the pattern is *linear*. The Stochastic OT pattern is thus a case of sublinear cumulativity; we will see examples of the others after we introduce the other probabilistic models.

To see why Stochastic OT can only represent sublinear cumulativity, we can consider the differences across environments in terms of the summed probabilities of constraint rankings. In Table 8, an X indicates that a realized schwa occurs in the environment specified in the column heading given the ranking in that row. The environments are those in which neither *Clash nor *CCC is relevant (C_σσ́, corresponding to the bottom row in Table 7), those in which only one constraint is relevant (CC_σσ́ and C_σ́, like the middle rows in Table 7), and those in which both constraints are (CC_σ́, like the top row in Table 7).

In Table 8, the independent contribution of *CCC in the non-cumulative context CC_σσ́ is the sum of the probabilities of the rankings c., d., and f. The contribution of *CCC in conjunction with *Clash is the difference between C_σ́ and CC_σ́, which is just the probability of c., the only ranking that does not also yield schwa in C_σ́. Therefore, the cumulative contribution of *CCC can never be greater than its independent contribution, since the probability of c. cannot be greater than the probability of c., d. and f. If rankings could have zero probabilities, then we could have a limited form of linear cumulativity. Zero

Table 7: Proportion realized schwa in output distributions with constraints set to ranking value 1: sublinear cumulativity in Stochastic OT.

Context	P(schwa)	List of constraint contributions
la terre se vend (CC_σ́)	0.67	Δ*Clash|*CCC = 0.17
la terre se vend bien (CC_σσ́)	0.5	Δ*CCC|*Clash = 0.17
le vin se vend (C_σ́)	0.5	Δ*Clash = 0.5
le vin se vend bien (C_σσ́)	0	Δ*CCC = 0.5

Table 8: Illustration of Stochastic OT cumulativity.

	Ranking	C_σσ́	CC_σσ́	C_σ́	CC_σ́
a.	NoSchwa >> *CCC >> *Clash				
b.	NoSchwa >> *Clash >> *CCC				
c.	*CCC >> NoSchwa >> *Clash		X		X
d.	*CCC >> *Clash >> NoSchwa		X	X	X
e.	*Clash >> NoSchwa >> *CCC			X	X
f.	*Clash >> *CCC >> NoSchwa		X	X	X

probability rankings are not possible in Stochastic OT, but could be in other OT models of variation, such as Partially Ordered Constraints (Anttila 1997) or Pairwise Ranking Grammar (Jarosz 2015). Those alternative models of assigning probabilities to rankings still seem unable to represent superlinear cumulativity, for the reasons we have just discussed for Stochastic OT. In Stochastic OT, ranking values are sampled from a normal distribution over the unbounded space of possible ranking values, so no ranking ever has zero probability. Therefore, cumulativity will always be sublinear in Stochastic OT.

4.1.3 Sublinearity through superlinearity in MaxEnt and Noisy HG

We now turn to patterns of cumulativity in Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003) and Noisy HG (Boersma & Pater 2016), two probabilistic variants of HG. In MaxEnt the probability of a candidate is proportional to the exponential of the weighted sum of violations. In terms of the difference vectors, the probability of the realized schwa is $e^n/(1 + e^n)$, where *e* is the base of the natural logarithm (approximately, $e = 2.71828$) and *n* is the weighted sum of difference scores.[8] This means that the Harmony difference between two candidates, *candidate a* minus *candidate b*, is the log-odds of *candidate a*. A Harmony difference of 0 produces 0.5 probability, 1 → 0.73, 2 → 0.88, 3 → 0.95, 4 → 0.98, 5 → 0.99, and 6 → 1.0, all rounded to 2 decimal points. Negative Harmony differences equal one minus the positive value (–1 → 0.27, –2 → 0.12, –3 → 0.05 and so on).

So, given the weights (3, 2, 2) used for illustrative purposes in (35), the probability of realized schwa would be 0.73 in the top row (since the difference in harmonies is 1), 0.27 in each of the middle rows (since the difference is –1), and 0.05 in the bottom row (since the difference is –3). This is shown in Table 9. This is *superlinear* because the cumulative contributions of *Clash and *CCC (0.46) are greater their independent contributions (0.22).

Noisy HG is like Stochastic OT, except the values of the constraints are used in a weighted constraint evaluation of the candidate set. Like MaxEnt, it can generate superlinear cumulativity, though as we will see, the patterns the two models predict are not identical.

To begin our comparison of the three models, we first consider the probability distributions they produce when constraints values are set at 1, shown in Table 10. For Noisy HG and Stochastic OT, the noise — the Standard Deviation of the Gaussian — is set to 0.2. For the Noisy HG model, any resulting negative weights were converted to zero (this is called Linear OT in Boersma & Weenink's 2017 Praat, which we used to explore these models). All probabilities in the table are rounded to two decimal points. The rightmost column shows contributions of the constraints under MaxEnt. This is a case of linear cumulativity, in which the change in probability due to the introduction of a factor (e.g. a preceding cluster) is the same whether or not the other factor (e.g., adjacent stressed syllables) is already present.

Table 9: Proportion realized schwa in output distributions with NoSchwa = 3, *CCC = 2, and *Clash = 2.

Context	P(schwa)	List of constraint contributions
la terre se vend (CC_σ́)	0.73	Δ*Clash|*CCC = 0.46
la terre se vend bien (CC_σσ́)	0.27	Δ*CCC|*Clash = 0.46
le vin se vend (C_σ́)	0.27	Δ*Clash = 0.22
le vin se vend bien (C_σσ́)	0.05	Δ*CCC = 0.22

[8] The usual MaxEnt calculation for the probability of one of two candidates with Harmony H1 and H2 respectively is $e^{H1}/(e^{H1} + e^{H2})$. Because we have subtracted out the constraint scores for one of the candidates, its probability in the equation can be represented as $e^0 = 1$. See Zuraw and Hayes (2017) for another derivation.

Table 10: Proportion realized schwa in output distributions with constraint weights set to 1.

Context	Stochastic OT	Noisy HG	MaxEnt	List of contributions in MaxEnt
la terre se vend	0.67	1	0.73	Δ*Clash|*CCC = 0.23
la terre se vend bien	0.5	0.5	0.5	Δ*CCC|*Clash = 0.23
le vin se vend	0.5	0.5	0.5	Δ*Clash = 0.23
le vin se vend bien	0	0	0.27	Δ*CCC = 0.23

The MaxEnt probabilities arise because realized schwa is preferred by a Harmony score of 1 in the top row, dispreferred relative to deletion by 1 in the bottom row, and the two outcomes have equal Harmony in the middle. Noisy HG also assigns equal probability in the middle rows (as does Stochastic OT). For the top row in Noisy HG, a noise value of 0.2 has a very low probability of subverting the pre-noise preference for the faithful candidate in the top row by making the sum of the weights of *CCC and *Clash lower than NoSchwa (less than 0.005, hence rounded to zero). In the final row, no constraint prefers the faithful candidate. In Noisy HG, if a value of zero were sampled for NoSchwa, the two candidates would be tied, and the tie would be broken with a random choice, which could yield the faithful candidate. The probability of this happening is less than 0.005. To produce a distribution like this in MaxEnt, one could increase the weight values by some constant factor. As the weights get higher, the probability in the top row would approach 1, and the probability in the bottom row would approach 0. Therefore, MaxEnt is capable of representing the more peaked distribution that Noisy HG produces with the current weights, at least to the degree of resolution we are examining.

Because cumulativity in Stochastic OT is predictably sublinear, we know that there is no set of constraint values that will allow it to model the linear cumulativity produced in Noisy HG and MaxEnt with values of 1. It is also the case that Noisy HG is unable to match the distribution produced by Stochastic OT. For Noisy HG, if the weights are small enough to allow NoSchwa to overcome the cumulative effects of *CCC and *Clash with 0.67 probability when noise is added, a non-negligible number of faithful schwas will be produced in the bottom row (through random selection in a tie when both candidates have Harmony zero). For example, with the ranking values set to 0.2, the top row gets close to the Stochastic OT value at 0.72, and the middle rows are at 0.50, but the bottom is at 0.16. MaxEnt cannot match this Noisy HG distribution, for reasons we will now discuss.

When the probability in the middle rows is at 0.5, MaxEnt is necessarily strictly linear. This can be understood based on Zuraw & Hayes' (2017) observation that the contribution of a given weighted constraint violation difference to probability is highest with a baseline at 0.5, and weakens as it approaches 0 and 1. This is due to the logistic function relating Harmony differences between pairs of candidates and probabilities. In Figure 3, we plot probability on the vertical axis, and Harmony difference on the horizontal axis. Note that when we move from –3 to –2 the change in probability (0.07) is smaller than-from –2 to –1, (0.13), which itself is smaller than –1 to 0 (0.23).

The contribution on either side of probability 0.5 is equal: if adding a violation difference increases probability from a baseline of 0.4 to 0.5, it will also increase probability from 0.5 to 0.6. This is the situation we have looked at in the tables thus far, and this explains why MaxEnt cannot match the Stochastic OT (0.67, 0.5, 0.5, 0) distribution in the table, nor the Noisy HG (0.72, 0.5, 0.5, 0.12) distribution discussed in the text.

To escape the clutches of linearity in MaxEnt, we can change the probability of faithful schwa in the non-cumulative context. For example, if we give *CCC and *Clash a higher value than NoSchwa, such as 2 *vs.* 1 in Table 11, the result of adding one of the constraints is a probability higher than 0.5 as in the middle rows, and the effect of adding the other (0.22, the difference with the top row) will be smaller than its effect on its own (0.46, the difference with the bottom row). This is sublinear cumulativity, displayed here by all theories.

MaxEnt can of course match the Stochastic OT and Noisy HG distributions to the degree of resolution we are examining. With the current constraint set, the MaxEnt distribution is completely out of reach of the other frameworks because the faithful schwa gets non-negligible probability in the bottom row, and it is harmonically bounded by deletion. To give them a chance to match it, we can add McCarthy and Prince's (1995) Max to the constraint set, which assigns a violation to deletion in every context. To find weights, we used the learning procedure from the next section, with the MaxEnt distribution as the learning target. In a typical run, Noisy HG was able to come close to the MaxEnt distribution with this larger constraint set (0.94, 0.73, 0.73, 0.25), but the Stochastic OT distribution remained fairly distant (0.89, 0.78, 0.78, 0.25), presumably because of its weaker cumulativity.

Finally, Noisy HG and MaxEnt can display superlinear cumulativity in probability differences, as shown in Table 12 in which NoSchwa is given a higher value than *CCC and *Clash (again 2 *vs.* 1). In MaxEnt, we get predictable superlinearity when the result of adding a single constraint is probability less than 0.50. Here, the probability increase from the bottom to the middle rows is 0.15, and the increase from middle to top is 0.23.

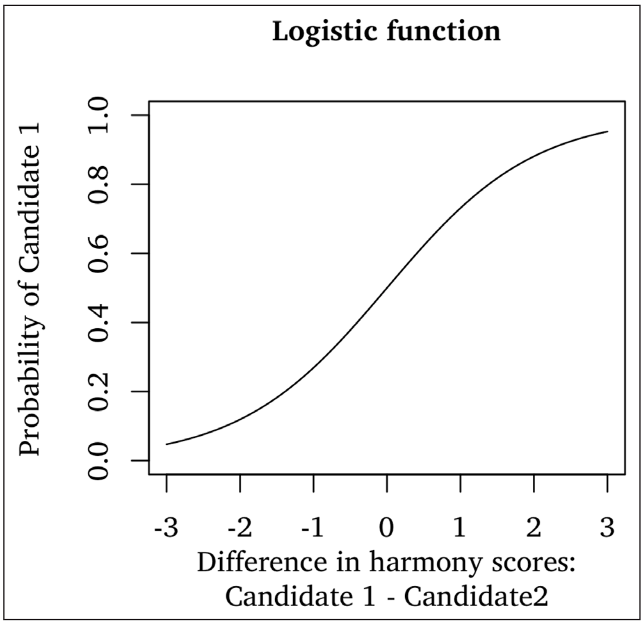


Figure 3: Probability of a candidate relative to Harmony difference.

Table 11: Proportion realized schwa in output distributions with NoSchwa set to 1, and *CCC and *Clash set to 2.

Context	Stochastic OT	Noisy HG	MaxEnt	List of contributions in MaxEnt
la terre se vend	1	1	0.95	Δ*Clash|*CCC = 0.22
la terre se vend bien	1	1	0.73	Δ*CCC|*Clash = 0.22
le vin se vend	1	1	0.73	Δ*Clash = 0.46
le vin se vend bien	0	0	0.27	Δ*CCC = 0.46

Since Stochastic OT is predictably sublinear, superlinear patterns are predictably beyond its scope. MaxEnt and Noisy HG can model the Stochastic OT pattern by assigning NoSchwa sufficient weight relative to the other constraints. With MaxEnt, we can model the Noisy HG pattern by scaling the weights used in the table, which will keep the top row at 0.50, and can bring the other rows as close to 0 as desired, and Noisy HG can in turn model the MaxEnt pattern, at least with the addition of Max.

In sum, we have shown that each model has restrictions on the types of probabilistic patterns it can model. This means that we should be able to test them in their relative ability to match natural language cumulativity. The biggest difference amongst the models appears to be Stochastic OT's weaker cumulativity with respect to the other two: it is always sublinear. MaxEnt's degree of cumulativity, sublinear, linear, or superlinear, was shown to be related to where the effect of a single competing constraint lands in probability space, below 0.50, at 0.50, or above. Noisy HG's degree of cumulativity is less predictable in that it can model sublinear patterns out of reach of MaxEnt, and in that respect, seems like it falls between the two other theories, as might be expected as it combines Stochastic OT's noise with MaxEnt's weighted evaluation.[9]

4.2 Models fitted to French data

Along with cases of underlying schwa discussed in the previous section, our judgment experiment examined four parallel epenthesis contexts, illustrated in Table 13, with the potential schwas underlined.

We assume that the vowels in these cases are not underlying, but are supplied through epenthesis. In the contexts in the rightmost column, the epenthetic schwa avoids a consonant cluster, and in those in the top row, it avoids a stress clash.

The grand means of realized schwa from the experiment are repeated in Table 14, rounded to three decimal points (more precise values were used for finding constraint values). For both underlying and epenthetic schwa, the lowest rate of schwa is in C_σσ́, where the schwa is in the antepenultimate syllable with only a single preceding consonant, and

Table 12: Proportion realized schwa in output distributions with NoSchwa set to 2, and *CCC and *Clash set to 1.

Context	Stochastic OT	Noisy HG	MaxEnt	List of contributions in MaxEnt
la terre se vend	0	0.5	0.5	Δ*Clash|*CCC = 0.23
la terre se vend bien	0	0	0.27	Δ*CCC|*Clash = 0.23
le vin se vend	0	0	0.27	Δ*Clash = 0.15
le vin se vend bien	0	0	0.12	Δ*CCC = 0.15

Table 13: Examples of epenthetic schwa contexts.

Following context	Preceding context	
	C_	CC_
_σ́	la bott**e** jaune 'the yellow boot'	mets ta vest**e** rouge 'put on your red jacket'
_σσ́	la bott**e** chinoise 'the Chinese boot'	mets ta vest**e** marron 'put on your brown jacket'

[9] Edward Flemming (p.c.) points out that one can characterize the difference between MaxEnt and Noisy HG in terms of MaxEnt, but not Noisy HG, being linear in log space.

Table 14: Experimental results (proportion realized schwa).

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.648	0.938
	_σσ́	0.562	0.914
Epenthetic schwa	_σ́	0.122	0.833
	_σσ́	0.090	0.683

the highest rate is in CC_σ́, where schwa is in penultimate syllable with two preceding consonants. Intermediate values obtain when only the constraint against clusters is relevant (CC_σσ́), or the constraint against singletons (C_ σσ́). The presence of an underlying vowel leads to a higher rate of schwa in all contexts.

The constraint set for these models includes the three markedness constraints introduced in the last section for the deletion cases: NoSchwa disprefers schwa across the board, and *CCC and *Clash prefer it in the CC_ and _σ́ respectively. The faithfulness constraint Max prefers the realized schwa when it is underlying, and Dep prefers its absence when it would need to be supplied through epenthesis (see McCarthy & Prince 1995 on Max and Dep). We also include *Cluster, penalizing a CC or CCC sequence, because schwa is preferred by none of the other constraints in the epenthesis context C_σσ́, so Stochastic OT would be unable to grant it any probability, and would be unable to match the empirical value of 0.090. The preferences of the full constraint set for both underlying and epenthetic schwa are shown in (36). With this constraint set any of the three frameworks can match the data in an individual cell of the table in Table 14 to arbitrary precision, and they can also get the general pattern of cumulative constraint interaction. The question is how closely they can fit the overall pattern.

(36) Difference vectors for constraint scores: negative values favor schwa deletion, positive differences favor schwa realization

	NoSchwa	*CCC	*Clash	Max	Dep	*Cluster
la terre s**e** vend	–1	+1	+1	+1	0	0
la terre s**e** vend bien	–1	+1	0	+1	0	0
le vin s**e** vend	–1	0	+1	+1	0	+1
le vin s**e** vend bien	–1	0	0	+1	0	+1
mets ta vest**e** rouge	–1	+1	+1	0	–1	0
mets ta vest**e** marron	–1	+1	0	0	–1	0
la bott**e** jaune	–1	0	+1	0	–1	+1
la bott**e** chinoise	–1	0	0	0	–1	+1

We first present a MaxEnt model whose weights were obtained by using a batch learner (Staubs 2011) that incorporates an optimization algorithm that finds weights that minimize the difference between the training data and the model predictions, in terms of Kullback–Leibler divergence (Kullback & Leibler 1951). This is an implementation of the same general approach to MaxEnt grammar and learning that is presented in Goldwater & Johnson (2003) and Wilson (2006) (as well as Hayes & Wilson 2008, though their model defines a probability distribution over all possible words, rather than over a set of candidates for a given UR). The optimization algorithm was L-BFGS-B (Byrd et al. 1995) as implemented in R. The weights were constrained to be above zero, and a Gaussian prior with variance 100,000 was imposed (the prior seemed to have no effect, as a weaker prior did not change the solution).

Table 15 shows the predicted probabilities for realized schwa in each of the eight environments, as well as the difference between the experimental probabilities and predicted probabilities (positive values indicate that the predicted value is too high, negative too low). The sum of absolute differences for this MaxEnt model with respect to the empirical data is 0.253 (mean over contexts = 0.032; we present SSE and K-L divergence in the summary table at the end of this section).

The constraint weights producing these probabilities are shown in the Table 16. As mentioned in the previous section, the probabilities result from the formula $e^n/(1 + e^n)$, where *n* is the weighted sum of difference scores. For the *la botte chinoise* type of epenthetic schwa (C_σσ́), whose probability is 0.109, the weighted sum is the negative of the weights of NoSchwa and Dep, plus the weight of *Cluster: –1.015 + –1.084 + 0 = –2.099. The corresponding underlying schwa type, *le vin se vend* (C_σσ́), differs in the absence of the negative contribution of Dep, and presence of the positive contribution of Max, thus leading to a higher baseline probability for underlying schwa in Table 15.

The contribution of the high weighted *CCC is seen in the probability differences between the columns in Table 15 while contribution of the somewhat lower weighted *Clash is seen in the probability differences between rows.[10] As discussed in the previous section, the function relating weight differences to probability differences is a sigmoid centered at 0.50 probability. Therefore, the highest possible contribution of a weight difference is when the midpoint between the probability where the constraint doesn't apply and the probability where it does apply is 0.50. Thus, the greatest contribution of the *CCC constraint is in the penultimate epenthetic context, where it yields a probability increase of 0.608 (0.775–0.167), and the midpoint is closest to 0.50 (0.471). This is in line with the empirical differences, where this context has the highest difference between preceding singleton and cluster. One might think that to get a greater

Table 15: MaxEnt's predicted probabilities after batch training, errors in parentheses.

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.633 (–0.015)	0.967 (0.029)
	_σσ́	0.514 (–0.048)	0.948 (0.034)
Epenthetic schwa	_σ́	0.167 (0.045)	0.775 (–0.058)
	_σσ́	0.109 (0.019)	0.678 (–0.005)

Table 16: MaxEnt constraint weights after batch training.

*CCC	2.845
Dep	1.084
Max	1.069
NoSchwa	1.015
*Clash	0.490
*Cluster	0.000

[10] Because *FF is relevant only in the singleton contexts, the effective value of *CCC is diminished by the weight of *Cluster, but *Cluster here has a zero weight.

difference between C_σ́ and CC_σ́ for epenthetic schwa than for underlying schwa one would need a separate constraint, but in fact, this follows in the MaxEnt model from the difference in the baseline probability value in each case. Since the baseline probability for underlying case is the singleton probability of 0.633, the MaxEnt model is predicted to yield a smaller probability increase after the cluster. It is worth noting, though, that the MaxEnt model winds up producing a slightly smaller difference between the columns than in the empirical data for epenthetic schwa, and a slightly larger difference for underlying schwa.

*Clash has its greatest effect on probability differences in the realization of underlying schwa in the C_ environment (0.633–0.514 = 0.119), again because the midpoint is the closest to 0.50. This fits the empirical data in terms of producing a greater effect for *Clash in the singleton than in the cluster environment for underlying schwa, and also in terms of producing a greater effect for *Clash in singletons for underlying schwa than epenthetic schwa. One subtle mismatch with the empirical data is that the greatest effect for *Clash is in fact in the cluster environment for epenthesis (the rightmost column). The MaxEnt model cannot match this because the baseline in that case is further away from 0.50.

To obtain fitted models for Stochastic OT and Noisy HG, we must use on-line learners; no batch approaches are available because it is computationally costly to calculate or estimate model predicted probabilities in those frameworks. In on-line learning, the learner receives a single piece of data at each learning step and uses the grammar to generate a prediction just for that datum, updating the constraint values if the learning datum and the prediction mismatch. Conveniently, it is possible to conduct on-line learning in a nearly identical way across the three frameworks. For MaxEnt, the on-line method is referred to as Stochastic Gradient Ascent (Jäger 2007), and in applying it to Noisy HG, Boersma and Pater (2016) call it the Harmonic Grammar Gradual Learning Algorithm (HG-GLA). The weights are updated by the difference in violation vectors between the learner's prediction and the learning datum, scaled by a learning rate, or plasticity. In Stochastic OT's GLA, constraints preferring the correct learning datum are promoted by the plasticity amount, and those preferring the learner's own incorrect prediction are demoted. When the differences between the candidate vectors are always zero or one, as in our examples (see 34), the HG-GLA and the OT-GLA are identical.

The learning simulations were conducted in Praat (Boersma & Weenink 2017). Constraints were given an initial value of 2, and the plasticity was set to 0.1. The learners received 100,000 samples from the target distributions. These distributions were the experimental results in Section 3, with equal probability given to each off the 8 contexts. The learner then received 3 more sets of 100,000 samples of data, with the plasticity set at 0.01, 0.001 and 0.0001 respectively. This training regime is based on the Praat defaults, but with an initial weight value of 2 rather than 10 so as to get comparable results across the frameworks, and with a correspondingly lower initial plasticity. The noise for Stochastic OT and Noisy HG was set at 0.2, rather than the Praat default of 2, because of the lower initial weight and plasticity. We conducted 20 runs for each model.

The MaxEnt model trained on-line predicts distributions very similar to those of the model trained in a batch fashion. Table 17 shows the results from the model that provides the closest fit to the data, with a sum of absolute differences of 0.240 (mean over contexts = 0.030). The 20 runs had an average summed absolute difference of 0.256 (mean 0.032), with a maximum of 0.269 (mean 0.034).

The weights producing that distribution, shown in Table 18, are somewhat different from those for the batch model, but we again have a relatively high weight for *CCC, and a relatively low weight for *Clash.

The predictions of the best fitting Stochastic OT model are shown in Table 19. The sum of absolute differences with respect to the empirical data is higher than the best MaxEnt model, 0.299 (mean 0.037). The average sum of absolute differences over 20 runs was 0.330 (mean 0.041), and the maximum was 0.381 (mean 0.048). The distributions of these error measures for the MaxEnt models and the Stochastic OT models are non-overlapping: the worst fitting of the 20 on-line MaxEnt models had less error than the best fitting of the Stochastic OT models. We'll show shortly that this holds for other ways of measuring error as well.

Like the MaxEnt models, the Stochastic OT predictions get the general pattern of cumulative constraint interactions, and the individual fits are sometimes even somewhat better. The bulk of the error is in the rightmost column epenthetic schwa: the values of the two rows are too close together with respect to the empirical data, which means the effect of *Clash in the cumulative interaction with *CCC is too weak. In the empirical data, the effect of *Clash is superlinear: there is a 0.032 difference in the C_ context, and a 0.150 difference in the CC_ context. As discussed in the last section, Stochastic OT produces cumulative interactions that are predictably sublinear in probability space, here leading to a gross mismatch with the empirical data, which show a 0.060 difference in the C_ context, and 0.013 in CC_.

Table 17: Proportions of schwa for MaxEnt after on-line training.

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.637 (–0.011)	0.968 (0.030)
	_σσ́	0.518 (–0.043)	0.950 (0.036)
Epenthetic schwa	_σ́	0.166 (0.043)	0.778 (–0.056)
	_σσ́	0.109 (0.019)	0.682 (–0.001)

Table 18: MaxEnt constraint weights after on-line training.

*CCC	3.532
NoSchwa	1.798
Max	1.184
Dep	0.982
*Cluster	0.670
*Clash	0.502

Table 19: Proportions of schwa of the best fitting Stochastic OT model.

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.648 (0.000)	0.914 (–0.025)
	_σσ́	0.567 (0.005)	0.907 (–0.006)
Epenthetic schwa	_σ́	0.169 (0.047)	0.778 (–0.064)
	_σσ́	0.109 (0.019)	0.682 (0.073)

The Stochastic OT constraint values producing this distribution are shown in Table 20. In contrast with the MaxEnt values, the Stochastic OT constraint values are much closer together. This is because variation, and the consequent cumulativity, requires constraints to be relatively close in value so that their ranking will vary across samples from the noise distribution. Nonetheless, we see the same general pattern of *CCC having a higher value than *Clash.

The final set of predictions is those of the best fitting Noisy HG model, shown in Table 21. The sum of absolute differences with respect to the empirical data is comparable to the best Stochastic OT model, 0.295 (mean error 0.037). The average over 20 runs was also similar, 0.327 (mean error 0.041), as was the maximum, 0.381 (mean error 0.0486). The distribution of error over the eight contexts was somewhat different; the best fitting model is again typical.

The Noisy HG model succeeds in getting a greater spread than Stochastic OT between CC_σ́ and CC_σσ́ for epenthesis. In this respect mimicking MaxEnt, and approaching the empirical spread. In doing this, though, it also creates a greater spread between the values in the C_ column than is motivated by the empirical data. Here Noisy HG is producing a slightly sublinear pattern: the effect of *Clash on the probability is a 0.088 difference on its own (penultimate column), and 0.081 in conjunction with *CCC (rightmost). In this respect, it is intermediate between the superlinear pattern of MaxEnt, and the highly sublinear pattern of Stochastic OT. Noisy HG patterns like MaxEnt in giving both CC_σ́ and CC_σσ́ for underlying schwa too much probability of realized schwa, and both contexts of C_ too little; these models are not quite fitting the extent to which *CCC has a greater effect in the Epenthetic contexts.

The weights producing the Noisy HG distribution are given in Table 22. As in MaxEnt, the additive nature of constraint interaction in this weighted constraint model allows constraints with even small weights to have an effect on the outcome. Again, the greater effect of *CCC than *Clash seen in the probability distributions is reflected in the weights, even allowing for the effect of *Cluster in singleton contexts.

Our comparisons of models' fit to the empirical data have thus far been made in terms of differences in raw probability. There are other ways of measuring fit, and one might wonder whether the outcome is different using other metrics. In Table 23, we provide the mean, best

Table 20: Stochastic OT constraint values.

*CCC	2.402
Dep	2.144
Max	2.097
NoSchwa	2.047
*Clash	1.977
*Cluster	1.551

Table 21: Proportion of schwa of the best-fitting Noisy HG model.

	Following context	Preceding context	
		C_	CC_
Underlying schwa	_σ́	0.634 (−0.014)	0.977 (0.038)
	_σσ́	0.527 (−0.035)	0.963 (0.050)
Epenthetic schwa	_σ́	0.195 (0.072)	0.766 (−0.067)
	_σσ́	0.107 (0.016)	0.690 (0.002)

Table 22: Noisy HG constraint weights.

*CCC	2.299
NoSchwa	1.955
*Cluster	1.746
Max	0.211
Dep	0.166
*Clash	0.034

Table 23: Error for each model.

	Absolute Error			Sum of Squared Error			K-L Divergence		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
Stochastic OT	0.330	0.299	0.381	0.043	0.037	0.052	0.086	0.064	0.112
NoisyHG	0.327	0.295	0.371	0.035	0.031	0.045	0.035	0.034	0.037
MaxEnt	0.256	0.240	0.269	0.021	0.019	0.023	0.020	0.020	0.021

and worst fits for each model in terms of sum of squared error and Kullback-Lieber divergence, and also repeat the absolute error values reported in the text.[11] In all cases, MaxEnt had consistently lower error than the other models. When error is measured in terms of SSE or K-L Divergence, the Noisy HG values are lower than those for Stochastic OT, and the MaxEnt *vs.* Stochastic OT difference is enhanced, because the error in the Stochastic OT predictions is concentrated in just two of the contexts (the _CC column for epenthetic schwa).

In sum, all three models – MaxEnt, Noisy HG and Stochastic OT – were able to capture the general pattern of cumulative constraint interaction seen in the empirical data, and provided reasonable fits to the attested values. The MaxEnt model did slightly better than the other models, anxd in comparison to Stochastic OT, at least some of that success is attributable to its ability to produce superlinear cumulativity in probability space.

5 Predictions and directions for future work

Since the predictions of our generative models are only as trustworthy as the data they're trained on, we've taken many steps to model the simplest, most controlled data set possible — collecting a lot of judgment data for a relatively small set of contexts. This is necessary because the realization of schwa is conditioned by a multitude of factors, naturally occuring data are very noisy, and accurately estimating probabilities requires many tokens. As a result, using corpus data makes it difficult to isolate the fine-grained differences in the predictions of the models.

Although we've looked at the interaction of just three factors that condition the realization of schwa (type of boundary, stress, and number of preceding segments), we expect the same types of constraint interaction regardless of the phonological constraints under consideration. A richer model would take into account factors that we controlled for and mentioned in passing, such as the sonority profile of the consononant cluster, the number of preceding syllables, h-aspiré, and individual differences between speakers. Future work will determine whether our present findings scale up when more factors are considered in light of naturally occuring speech.

[11] Absolute error was calculated with respect to the probability of schwa in each context. Sum of squared error and K-L divergence were calculated over the probability of each of schwa and no-schwa. K-L divergence is formulated to be calculated over entire probability distributions. If SSE were calculated over just probability of schwa, the value would be half of that reported, and if absolute error were calculated for both schwa and no-schwa, it would double.

6 Conclusion

In this paper, we described and modeled the interaction of two phonological factors that condition French schwa alternations: schwa is more likely after two consonants than one (the cluster factor) and in the penultimate syllable than elsewhere (the stress factor). Each of these factors has been identified in the literature on French schwa, but their interaction in probability space hasn't been previously described or formalized. Using data from a judgment study, we showed that both factors play a role in schwa epenthesis and deletion, including in contexts where the stress factor has previously been described as having no effect. We then provided a characterization of patterns of cumulative interaction as sub- through super-linear, showing that Stochastic OT is limited to sublinear cumulativity. Because superlinearity is attested in our experimental data, Stochastic OT fared less well in fitting the data than the weighted constraint probabilistic models Noisy HG and MaxEnt, with MaxEnt yielding the best fit to the data. These results add to a growing body of work showing that weighted constraints provide a better fit to probabilistic natural language data than ranked constraints, particularly when it comes to cumulativity.

Abbreviations

HG = Harmonic Grammar, MaxEnt = Maximum Entropy Grammar, OT = Optimality Theory

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** List of experimental items. DOI: https://doi.org/10.5334/gjgl.583.s1
- **Appendix B.** Participant background. DOI: https://doi.org/10.5334/gjgl.583.s2

Acknowledgements

We are grateful to three anonymous reviewers for their comments, which greatly improved the quality of the manuscript. We are also grateful for feedback and comments from the audiences at the 46th Linguistic Symposium on Romance Languages, Stonybrook University, as well as discussions with François Dell, Edward Flemming, Bruce Hayes, and Kie Zuraw.

Funding Information

Joe Pater's work on this project was supported by NSF grants 1424077 and 1650957 to the University of Massachusetts Amherst.

Competing Interests

The authors have no competing interests to declare.

References

Anttila, Arto. 1997. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout & W. Leo Wetzels (eds.), *Variation, change, and phonological Theory*, 35–68. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/cilt.146.04ant

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bayles, Andrew, Aaron Kaplan & Abby Kaplan. 2016. Inter- and intra-speaker variation in French schwa. *Glossa: A Journal of General Linguistics* 1(1). DOI: https://doi.org/10.5334/gjgl.54

Benor, Sarah & Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82(2). 233–278. DOI: https://doi.org/10.1353/lan.2006.0077

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21. 43–58.

Boersma, Paul. 2007. Some listener-oriented accounts of h-aspiré in French. *Lingua* 117(12). 1989–2054. DOI: https://doi.org/10.1016/j.lingua.2006.11.004

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1). 45–86. DOI: https://doi.org/10.1162/002438901554586

Boersma, Paul & David Weenink. 2017. Praat: doing phonetics by computer (Version 6.0.29).

Boersma, Paul & Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Bürki, Audrey, Mirjam Ernestus, Cédric Gendrot, Cécile Fougeron & Ulrich Hans Frauenfelder. 2011. What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The Journal of the Acoustical Society of America* 130(6). 3980–3991. DOI: https://doi.org/10.1121/1.3658386

Byrd, Richard H., Peihuang Lu, Jorge Nocedal & Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16(5). 1190–1208. DOI: https://doi.org/10.1137/0916069

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Coetzee, Andries W. & Joe Pater. 2011. The place of variation in phonological theory. In John A. Goldsmith, Jason Riggle & Alan C. Yu (eds.), *Handbook of Phonological theory*, 2nd ed., 401–434. Cambridge, MA: Blackwell. DOI: https://doi.org/10.1002/9781444343069.ch13

Côté, Marie-Hélène. 2000. *Consonant cluster phonotactics: A perceptual approach*. MIT dissertation.

Côté, Marie-Hélène. 2007. Rhythmic constraints on the distribution of schwa in French. In José Camacho, Nydia Flores-Ferrán, Liliana Sánchez, Viviana Déprez & María José Cabrera (eds.), *Romance linguistics 2006: selected papers from the 36th Linguistic Symposium on Romance Languages (LSRL)*, 81–95. Amsterdam & Philadelphia: Benjamins.

Côté, Marie-Hélène & Geoffrey Stewart Morrison. 2007. The nature of the schwa-zero alternation in French clitics: experimental and non-experimental evidence. *Journal of French Language Studies* 17. 159–186. DOI: https://doi.org/10.1017/S0959269507002827

Delattre, Pierre. 1939. Accent de mot et accent de groupe. *The French Review* 13(2). 141–146.

Dell, François. 1985. *Les règles et les sons*. Paris: Hermann.

Drummond, Alex. 2013. Ibex farm. *Online Server*. http://spellout.net/ibexfarm.

Durand, Jacques, Catherine Slater & Hilary Wise. 1987. Observations on schwa in southern French. *Linguistics* 25(5). 983–1004. DOI: https://doi.org/10.1515/ling.1987.25.5.983

Fougeron, Cécile, Cedric Gendrot & Audrey Bürki. 2007. On the acoustic characteristics of French schwa. In Jürgen Trouvain & William J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 942–944. Saarbrücken: Saarland University.

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

Grammont, Maurice. 1914. *Traité pratique de prononciation française*. Paris: Delagrave.

Guy, Gregory R. 1997. Violable is variable: Optimality Theory and linguistic variation. *Language Variation and Change* 9. 333–347. DOI: https://doi.org/10.1017/S0954394500001952

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440. DOI: https://doi.org/10.1162/ling.2008.39.3.379

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan*, 467–479. Stanford: CSLI.

Jäger, Gerhard & Anette Rosenbach. 2006. The winner takes it all – almost. Cumulativity in grammatical variation. *Linguistics* 44(5). 937–971. DOI: https://doi.org/10.1515/LING.2006.031

Jarosz, Gaja. 2015. Expectation driven learning of phonology. Ms. University of Massachusetts Amherst.

Jetchev, Georgi. 1999. Schwa or "ghost" vowels in French: a harmonic phonology account. *Italian Journal of Linguistics* 11(2). 231–272.

Jun, Sun-Ah & Cécile Fougeron. 2000. A phonological model of French intonation. *Intonation*, 209–242. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-011-4317-2_10

Kaplan, Aaron. 2016. Local optionality with partial orders. *Phonology* 33(2). 285–324. DOI: https://doi.org/10.1017/S0952675716000130

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86. DOI: https://doi.org/10.1214/aoms/1177729694

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 715–762. DOI: https://doi.org/10.2307/412333

Léon, Pierre R. 1966. Apparition, maintien et chute du "e" caduc. *La Linguistique* 2. 111–122.

Lyche, Chantal & Jacques Durand. 1996. Testing government phonology ou pourquoi le choix du schwa. *Current Trends in Phonology: Models and Methods* 2. 443–471.

Mazzola, Michael L. 1991. Stress clash and segment deletion. In Christiane Laeufer & Terrell Morgan (eds.), *Theoretical analyses in Romance linguistics: Selected papers from the nineteenth Linguistic Symposium on Romance Languages (LSRL XIX)*, 81–96. Amsterdam: John Benjamins.

Mazzola, Michael L. 2014. Schwa at the phonology/syntax interface. In Marie-Hélène Côté & Éric Mathieu (eds.), *Variation within and across Romance Languages: Selected papers from the 41st Linguistic Symposium on Romance Languages (LSRL)*, Ottawa 5–7 May 2011, 101–118. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/cilt.333.08maz

McCarthy, John J. & Alan Prince. 1995. Faithfulness and Reduplicative Identity. In Jill Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.), *University of Massachusetts Occasional Papers in Linguistics 18*, 249–384. Amherst, MA: GLSA Publications.

McPherson, Laura & Bruce Hayes. 2016. Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology* 33(1). 125–167. DOI: https://doi.org/10.1017/S0952675716000051

Morin, Yves-Charles. 1974. Règles phonologiques à domaine indéterminé: Chute du cheva en français. *Cahiers de Linguistique de l'Université Du Québec* 4. 69–88. DOI: https://doi.org/10.7202/800029ar

Pater, Joe. 2016. Universal grammar with weighted constraints. *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Pizzo, Presley. 2015. *Investigating Properties of Phonotactic Knowledge Through Web-Based Experimentation*. Amherst, MA: University of Massachusetts dissertation.

Post, Brechtje Maria Bowine. 2000. *Tonal and phrasal structures in French intonation* Vol. 34. The Hague: Thesus.

Prince, Alan & Paul Smolensky. 2004. *Optimality Theory: constraint interaction in generative grammar*. Malden, MA: Blackwell. DOI: https://doi.org/10.1002/9780470759400

Racine, Isabelle. 2008. *Les effets de l'effacement du schwa sur la production et la perception de la parole en français*. Geneva: University of Geneva dissertation.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Smolensky, Paul & Géraldine Legendre. 2006. *The harmonic mind: from neural computation to Optimality-Theoretic grammar*. Cambridge, MA: MIT Press.

Staubs, Robert. 2011. *HG in R (hgR). Software package*. Amherst, MA: University of Massachusetts Amherst. Software available at http://blogs.umass.edu/hgr/hg-in-r.

Tranel, Bernard. 1981. *Concreteness in Generative Phonology: Evidence from French*. Berkeley: University of California Press.

Tranel, Bernard. 1987. *The Sounds of French*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511620645

Verluyten, S. Paul. 1982. *Recherches sur la prosodie et la métrique du français*. Antwerp: Universiteit Antwerpen dissertation.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30(5). 945–982. DOI: https://doi.org/10.1207/s15516709cog0000_89

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language* 93(3). 497–548. DOI: https://doi.org/10.1353/lan.2017.0035

How to cite this article: Smith, Brian W. and Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5(1): 24. 1–33. DOI: https://doi.org/10.5334/gjgl.583

Submitted: 05 December 2017 **Accepted:** 24 October 2019 **Published:** 05 March 2020

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS