CrossMark

Phonology **34** (2017) 299–324. © Cambridge University Press 2017
doi:10.1017/S095267571700015X

*Learning opacity in Stratal Maximum Entropy Grammar**

Aleksei Nazarov
Harvard University

Joe Pater
University of Massachusetts Amherst

Opaque phonological patterns are sometimes claimed to be difficult to learn; specific hypotheses have been advanced about the relative difficulty of particular kinds of opaque processes (Kiparsky 1971, 1973), and the kind of data that is helpful in learning an opaque pattern (Kiparsky 2000). In this paper, we present a computationally implemented learning theory for one grammatical theory of opacity, a Maximum Entropy version of Stratal OT (Bermúdez-Otero 1999, Kiparsky 2000), and test it on simplified versions of opaque French tense–lax vowel alternations and the opaque interaction of diphthong raising and flapping in Canadian English. We find that the difficulty of opacity can be influenced by evidence for stratal affiliation: the Canadian English case is easier if the learner encounters application of raising outside the flapping context, or non-application of raising between words (e.g. *life* with [ʌɪ]; *lie for* with [aɪ]).

1 Introduction

(Kiparsky 1971, 1973) draws attention to cases of historical change that suggest that at least some opaque phonological process interactions are difficult to learn. In his work on Stratal Optimality Theory (Stratal OT), he further claims that independent evidence about the stratal affiliation of a process is helpful in learning an opaque interaction (Kiparsky 2000). In this paper, we develop a computationally implemented learner for a weighted-constraint version of Stratal OT, and show that for an initial set of test cases, opaque patterns are indeed generally more

* E-mail: ANAZAROV@FAS.HARVARD.EDU, PATER@LINGUIST.UMASS.EDU.
We would like to thank Ricardo Bermúdez-Otero, Paul Boersma, Jeroen Breteler, Ivy Hauser, Jeff Heinz, Coral Hughto, Gaja Jarosz, Marc van Oostendorp, Olivier Rizzolo, Klaas Seinhorst and Robert Staubs, as well as audiences at the 21st Manchester Phonology Meeting, the University of Massachusetts Amherst and the University of Amsterdam for their insightful feedback on this paper and for stimulating discussion. We also thank the editors of this volume and two anonymous reviewers for their very helpful and useful comments. We are grateful to the National Science Foundation for supporting this work through grants BCS-0813829 and BCS-1424077. All errors are ours.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

difficult than their transparent counterparts, and that information about stratal affiliation does make learning of an opaque pattern easier.

Our results support the viability of Stratal OT as a theory of opaque interactions (Bermúdez-Otero 1999, 2003, Kiparsky 2000; see below). However, it is not clear whether a stratal setup is the only source of opaque interactions – in fact, some cases of opacity may stem from other factors (McCarthy 2007). A broader assessment of alternative theories of opacity will not be attempted in this paper; see Jarosz (2016) for work on learning opacity in the Serial Markedness Reduction framework (Jarosz 2014).

In the Stratal OT approach to opacity (see especially Bermúdez-Otero 1999, 2003), grammars are chained together to produce outputs that would not be possible within a single level. For example, in the classic case of the opaque interaction between diphthong raising and flapping in Canadian English, illustrated in (1), *mitre* can be produced with raising in (word-level) Grammar 1 and then flapping in (phrase-level) Grammar 2. (Here and throughout, we enclose the output of the word-level grammar in ‖ ‖.) As we will see below, the final candidate cannot be produced by a single OT grammar with the standard set of constraints assumed for this case (though see Pater 2014).

(1) /maɪtɚ/ → (Grammar 1) ‖mʌɪtɚ‖ → (Grammar 2) [mʌɪɾɚ]

This chaining of grammars results in a hidden structure problem (Tesar & Smolensky 2000), insofar as only the output of the final grammar, not any earlier ones, is available in the learning data. Bermúdez-Otero (2003) develops a revision to the constraint-demotion algorithm that is specifically tailored to the problem of hidden structure in Stratal OT. In this paper, we instead apply a general approach to learning hidden structure in probabilistic OT to the specific case of chained grammars.

The variant of probabilistic OT adopted here is Maximum Entropy (MaxEnt) Grammar (Goldwater & Johnson 2003, Hayes & Wilson 2008). The MaxEnt formalism has a broad basis in applied mathematics, including connectionism (Smolensky 1986); see Johnson (2013) for an overview, as well as Goldwater & Johnson (2003). MaxEnt defines the probability distribution over a candidate set in a conveniently direct way: candidate probabilities are proportional to the exponential of the weighted sum of violations.

Candidate probabilities are illustrated in the tableau in (2), in which weights are indicated under constraint names (which are simplified versions of the ones used for the case study in §3.2). The weighted sum appears in the column headed '$\mathcal{H}$', for harmony; MaxEnt is a probabilistic version of Harmonic Grammar (Smolensky & Legendre 2006; see also the overview in Pater 2016). The exponentiation of $\mathcal{H}$ appears under $e^{\mathcal{H}}$, and the resulting probability (obtained by normalisation of $e^{\mathcal{H}}$ to sum to 1 across all candidates) under p.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

(2)

/maɪtɚ/	*aɪt 7	IDENT(C) 5	IDENT(V) 1	*VTV 0	$\mathcal{H}$	$e^{\mathcal{H}}$	p
a. ∥maɪtɚ∥	−1			−1	−7	0.001	0.002
b. ∥maɪɾɚ∥		−1			−5	0.007	0.018
c. ∥mʌɪtɚ∥			−1	−1	−1	0.368	0.973
d. ∥mʌɪɾɚ∥		−1	−1		−6	0.002	0.007

The *aɪt constraint penalises the unraised diphthong [aɪ] before a voiceless consonant, and IDENT(V) penalises raising to [ʌɪ]. With *aɪt having greater weight than IDENT(V), raising before a voiceless stop, as in (2c), has greater probability than the faithful (2a). *VTV penalises an alveolar stop between vowels, and IDENT(C) penalises the change from stop to flap. With IDENT(C) weighted higher than *VTV, the probability on flapping candidates (2b) and (2d) is lowered. Candidate (d) is the correct output in Canadian English, but it has a proper superset of the violations of (b), since flapping also satisfies *aɪt. No weighting of the constraints will give [mʌɪɾɚ] the highest probability in the tableau. In this respect, the situation is the same as in single-level OT, which cannot make (d) optimal. In MaxEnt, (d) can tie with (b); when the constraints they violate have zero weight, they are assigned equal probability (see Pater 2016 for further discussion and references on harmonic bounding and probabilistic weighted-constraint theories). In Table VI in §3.2, we will see that [mʌɪɾɚ] can emerge from a second chained grammar with highest probability.

Our way of dealing with the hidden structure problem posed by Stratal OT is an extension of the approach to Harmonic Serialism in Staubs & Pater (2016). The probability that the grammar assigns to an output form is the sum of the probabilities of the derivations that lead to it. The probability of a derivation is the product of the probabilities of each of its component steps. The probability of each step is determined by the weights of the constraints that apply at that level (word or phrase; see §2.2 for more details). We use a batch-learning approach, in which the objective is to minimise the divergence between the predictions of the grammar and the learning data, subject to a regularisation term that prefers lower weights. This is a standard approach to learning MaxEnt grammars (see e.g. Goldwater & Johnson 2003, Hayes & Wilson 2008).

Jarosz (2006, 2015) shows how this approach of considering every possible hidden structure in generating an overt candidate can be applied to phonology in her work on expectation maximisation with Optimality Theory. Here we follow Eisenstat (2009), Pater *et al*. (2012), Johnson *et al*. (2015) and Staubs & Pater (2016) in adopting a weighted-constraint variant of this general approach.

When there is hidden structure, the learner is not guaranteed to converge on the global optimum, i.e. a set of weights that best satisfy the

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

objective function (the solution space is not guaranteed to be convex). Instead it may only find a local optimum, a set of weights that is the best within the space that the learner can explore. We provide examples of such local optima in Table III in §2.3 and Tables IX and X in §3.3.

A simple and standard way of finding a global optimum in such cases is to try different starting positions for the learner (initialisations), in the hope that one or more will allow the learner to find the global optimum. In the limit, as long as the space from which we are sampling initialisations contains points from which there is a path towards an optimal solution, this will allow the learner to find a global optimum. In our test cases, the learner always found a successful solution at least once for every dataset. A successful solution is defined here as a grammar that grants more than 0.50 probability to the overt forms that have highest probability in the learning data. Our measure of 'ease of learning' for the various patterns we study is the probability that a learner will be successful over a set of random initialisations.

In the next section we provide more details about our grammatical and learning theories in the context of the first test case, a relatively simple example of opaque tense–lax vowel alternations in French. In §3 we turn to the opaque interaction of diphthong raising and flapping in Canadian English, and the effect of supplying evidence to the learner of phrasal non-raising. §4 provides an exploration of the mechanisms that lead to incorrect learning of (predominantly) opaque patterns. Finally, overall conclusions and directions for further research can be found in §5.

2 Tensing/laxing in Southern French

2.1 Data and analysis

As reported by Moreux (1985, 2006), Rizzolo (2002) and Eychenne (2014), certain dialects of Southern French have a synchronic process which tenses mid vowels in open syllables, while laxing them in closed ones (as well as open syllables followed by a schwa syllable; we will disregard this complication, and refer the reader to Selkirk 1978, Moreux 1985 and Rizzolo 2002 for potential explanations of this). This is conventionally called the *loi de position* ('law of position'), and is illustrated in (3) (examples from Rizzolo 2002 and Eychenne 2014).

(3) a.	/sel/	→ [sɛl]	*sel*	'salt'	/se/	→ [se]	*sait*	'knows'
b.	/pøʁ/	→ [pœʁ]	*peur*	'fear'	/pøʁ-ø/	→ [pø.ʁø]	*peureux*	'fearful'
c.	/poʁ/	→ [pɔʁ]	*pore*	'pore'	/poʁ-ø/	→ [po.ʁø]	*poreux*	'porous'

Tensing/laxing is made opaque (is counterbled) by resyllabification across word boundaries, at least in some cases, as shown in (4b), from

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

Rizzolo (2002: 51), where the lax vowel ‖œ‖ is retained in the surface open syllable [pœ].[1]

(4) *Opaque interaction between resyllabification and tensing/laxing*
   a. /kɑ̃pøʁ/ → [kɑ̃.pœʁ]
      campeur 'camper'
   b. /kɑ̃pøʁ ɑ̃ʁaʒe/ → ‖kɑ̃.pœʁ # ɑ̃.ʁa.ʒe‖ → [kɑ̃.pœ.ʁɑ̃.ʁa.ʒe]
      campeur enragé 'enraged camper'

In the simulation reported here, we investigated whether our learner could deal with this simple case of opacity. This also serves as a simple case to illustrate in more detail the functioning of the learner. The data that we offered the learner consisted of exceptionless *loi de position* as well as exceptionless opacity through resyllabification, as in (5).

(5) *Opaque interaction, as in Southern French*
   a. /set a/ → (‖sɛt. # a‖ →) [sɛ.ta] *cette A* 'this (letter) A'
   b. /se ta/ → (‖se. # ta‖ →) [se.ta] *c'est 'ta'* 'it is 'ta''

The constraint set that we used for this subset of real French was maximally simple, and is given in (6). Correct syllabification was taken for granted, in order to keep the constraint and candidate sets as small as possible, so that the interactions yielding opacity are easy to see.

(6) a. *[−tense]/open
      Assign a violation mark for every [−tense] segment in an open syllable.
   b. *[+tense]/closed
      Assign a violation mark for every [+tense] segment in a closed syllable.
   c. IDENT(V)
      Assign a violation mark for every vowel that is not identical to its input correspondent.

The original formulation of Stratal OT (Bermúdez-Otero 1999, Kiparsky 2000) allows for three levels of derivation: a stem level, a word level and a phrase level, which all have the same constraints, but different rankings or weightings. For this problem, we will only consider two levels – a word-level and a phrase-level grammar. The word-level grammar evaluates each word individually, without regard to its neighbours in the phrase. By contrast, the phrase-level grammar, which operates on the output of the word-level grammar, does evaluate entire phrases together.

[1] Rizzolo (2002: 51) reports that this effect is not unexceptional, at least in the dialect he describes. In our case study, we will consider an idealised version of this pattern, in which the opaque interaction is unexceptional.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

With this set-up, then, opaque interaction is obtained when word-level *[+tense]/closed and *[−tense]/open have high weight and word-level IDENT(V) has low weight, while the opposite holds at the phrase level. This corresponds to activity of the *loi de position* at the word level, and its inactivity at the phrase level.

This latter scenario is the only option to derive this opaque interaction. As can be seen in tableau (8) below, phrase-level markedness disprefers surface [sɛ.ta], so that the mapping from /e/ to [ɛ] cannot be derived at the phrase level. At the same time, word-level markedness prefers the word-level output that does map /e/ to [ɛ], ‖sɛt#a‖, because the first vowel is in a closed syllable at that level, since the phrasal context is invisible. This can be seen in tableau (7a).

Activity of *loi de position* at the word level can be expressed with the weights in (7) (found in successful runs of the learner described in §2.2).

(7) a.

/set#a/	*[−tns]/open 6.24	*[+tns]/closed 6.24	IDENT(V) 0	$\mathcal{H}$	$e^{\mathcal{H}}$	p
i. ‖set#a‖		−1		−6.24	0.002	0.00
ii. ‖sɛt#a‖			−1	0	1	1.00
b. /se#ta/						
i. ‖se#ta‖				0	1	1.00
ii. ‖sɛ#ta‖	−1		−1	−6.24	0.002	0.00

At the phrase level, giving high weight to IDENT(V) and zero weight to both markedness constraints results in a probability of 1.00 for all faithful mappings. This is illustrated in (9) for the phrase-level derivation that takes place if the word-level form output is ‖sɛt.#a‖: phrase-level outputs retain the [ɛ] that was created by closed-syllable tensing at the word level with a probability of 1.00.

(8)

/sɛt#a/	*[−tns]/open 0	*[+tns]/closed 0	IDENT(V) 6.93	$\mathcal{H}$	$e^{\mathcal{H}}$	p
a. ‖se.ta‖			−1	−6.93	0.001	0.00
b. ‖sɛ.ta‖	−1			0	1	1.00

The single-stratum mappings shown in (7) and (8) are assembled into derivational paths that lead from an underlying representation (UR) to a surface representation (SR). The probability of a single derivational path is obtained by multiplying the probabilities of every step in the path, as illustrated in (9). This is because every derivational step in Stratal OT is by definition independent of earlier or later derivational steps (Bermúdez-Otero 1999, Kiparsky 2000).[2]

[2] See also Odden (2011) on the Markov property traditionally ascribed to ordered phonological rules.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

(9) *Probabilities of derivational paths*

a. $p(/\text{set}\#\text{a}/ \rightarrow \|\text{set}\#\text{a}\| \rightarrow [\text{se.ta}]) =$
$p(/\text{set}\#\text{a}/ \rightarrow \|\text{set}\#\text{a}\|) \times p(\|\text{set}\#\text{a}\| \rightarrow [\text{se.ta}]) = 0.00 \times 1.00 = 0.00$

b. $p(/\text{set}\#\text{a}/ \rightarrow \|\text{sɛt}\#\text{a}\| \rightarrow [\text{sɛ.ta}]) =$
$p(/\text{set}\#\text{a}/ \rightarrow \|\text{sɛt}\#\text{a}\|) \times p(\|\text{sɛt}\#\text{a}\| \rightarrow [\text{sɛ.ta}]) = 1.00 \times 1.00 = 1.00$

The expected probability of a surface form (more accurately, a UR/SR mapping) is the sum of the probabilities of all derivational paths that lead from the input to that surface form, as in Table I.[3] This table shows that, given the weights in (7) and (8), the desired output candidates (cf. (5)) are generated with a probability of 1.00.

underlying	derivational path	*p*	surface	*p*
/set#a/	/set#a/ → ‖set#a‖ → [se.ta]	0.00	[se.ta]	0.00
	/set#a/ → ‖sɛt#a‖ → [se.ta]	0.00		
	/set#a/ → ‖set#a‖ → [sɛ.ta]	0.00	[sɛ.ta]	1.00
	/set#a/ → ‖sɛt#a‖ → [sɛ.ta]	1.00		
/se#ta/	/se#ta/ → ‖se#ta‖ → [se.ta]	1.00	[se.ta]	1.00
	/se#ta/ → ‖sɛ#ta‖ → [se.ta]	0.00		
	/se#ta/ → ‖se#ta‖ → [sɛ.ta]	0.00	[sɛ.ta]	0.00
	/set#a/ → ‖sɛ#ta‖ → [sɛ.ta]	0.00		

Table I
Computation of expected probabilities of UR/SR mappings by summing over all derivational paths.

This serial approach does not have look-ahead capabilities: the phrase-level grammar cannot influence which mappings will be preferred at the word level, in accordance with Bermúdez-Otero (1999). A closely related model that does have look-ahead capabilities is explored by Boersma & van Leussen (to appear), whose predictions for the same dataset would also be interesting to explore in future work.

Now that we have explained and illustrated the generation of probabilities over UR/SR mappings using the successful weights in (7) and (8), in §2.2 we will explain the structure of the learner that arrived at these weights. In §2.3 we will show how often this learner arrives at this successful set of weights within a sample of 100 runs, and what happens within that same sample when the learner does not find this analysis.

[3] Since there are several derivational paths per surface form, but only one surface form per derivational path, UR/SR mappings are bundles of non-overlapping sets of derivational paths. This means that, in computing expected probabilities of UR/SR mappings, each derivational path's probability is counted once, so that UR/SR probabilities are guaranteed to add up to 1.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

2.2 Learning

Staubs (2014b) provides an implementation in R (R Core Team 2013) of a variant of the approach to learning serial grammars discussed in Staubs & Pater (2016). We modified it minimally to allow different violations of the same constraint at different derivational stages, which is sometimes necessary in a stratal framework.

For instance, the constraint *[+tense]/closed in the Southern French simulation will be violated at the word level by a sequence [set#a], because even though [ta] could form a separate syllable, the word boundary between [t] and [a] makes it necessary to syllabify [t] as a coda. However, the same constraint remains unviolated at the phrase level for the same sequence, since the grammar can now look beyond word boundaries and evaluate entire phrases, so that [ta] does form a syllable: [se.ta].

Expected probabilities of UR/SR mappings are computed given sets of candidates, violation vectors and constraint weights, as described and illustrated in §2.1. In addition, the learner is also given observed probabilities of UR/SR mappings. Since we are working with categorical data, all UR/SR mappings found in the data were given a probability of 1.00, and all others a probability of 0.00. The learner then minimises the Kullback-Leibler (KL) divergence (Kullback & Leibler 1951) of the expected probabilities from the observed probabilities, which is a measure of how closely the grammar has been able to fit the data; see (10).

(10) *KL divergence* (D_{KL})
For all UR/SR mappings $\mathrm{UR} \rightarrow \mathrm{SR}_{i=1}$, …, $\mathrm{UR} \rightarrow \mathrm{SR}_{i=k}$:

$$D_{KL}(P_{obs} \| P_{exp}) = \sum_{i=1}^{k} p_{obs}(UR \rightarrow SR_i) \times \log\left(\frac{p_{obs}(UR \rightarrow SR_i)}{p_{exp}(UR \rightarrow SR_i)}\right)$$

Minimisation of this objective function is achieved by the L-BFGS-B method of optimisation (Byrd *et al*. 1995). This is a pseudo-Newtonian batch-optimisation method, which uses approximations of the first-order and second-order derivatives of the objective function to take incremental steps towards an optimum. The 'optim' implementation of this method in R was used, specifying a minimum of zero on all constraint weights.

To place an upper bound on the objective function, an L2 regularisation term (Hoerl & Kennard 1970) with $\mu = 0$ and $\sigma^2 = 10{,}000$ was added. This regularisation term penalises weights as they increase, keeping the optimal solution finite, and driving a constraint's weight down to zero whenever it is not helpful in the analysis.

Summarising, our algorithm calculates probabilities of derivations in Stratal MaxEnt, sums over them to get the expected probabilities of overt forms and fits the observed distribution of UR/SR mappings by minimising KL divergence.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

2.3 Results

The learner described in §2.2 was run 100 times on the French *loi de position* data, with every run starting from random starting weights for each constraint, drawn with replacement from a uniform distribution over [0, 10]. The results are given in Table II.

dataset	learned successfully
/set#a/ → [sɛ.ta] /se#ta/ → [se.ta]	96

Table II
Results for 100 runs for the Southern French dataset.

Thus opaque *loi de position* was learned successfully for an overwhelming majority of start weights. However, there is still a slight chance of incorrect learning, which we defined as a probability of 0.50 or less on at least one surface form that has a probability of 1.00 in the learning data (see also §1). The four unsuccessful runs yielded grammars which had free variation between tense [e] and lax [ɛ], as in Table III.

/set#a/	*p*	/se#ta/	*p*
[se.ta]	0.50	[se.ta]	0.50
[sɛ.ta]	0.50	[sɛ.ta]	0.50

Table III
UR/SR mapping probabilities for the local optimum.

This is a local optimum that arises when all constraint weights are set to zero. KL divergence from the actual data (our measure of error) is smaller for these weights than for any other weights within the weight space explored by the learner.

As will be explained in §4.1, an initially incorrect weighting at the word level leads to a tendency for the weight of phrase-level faithfulness to be lowered. At the same time, when phrase-level IDENT(V) reaches zero weight, any effect of closed-syllable laxing from the word level cannot be transmitted to the final output. Once this happens, 0.50 probability for all candidates is the best possible fit to the data, and the all-zero weight version of that solution minimises the penalty on the regularisation term, which prefers smaller weights.

It could be argued that this weight set is not a local optimum in the strict sense. There is a locally available solution that is better than all-zero weights: assigning minimal weight ($0 + \varepsilon$ for any positive ε) to word-level *[−tense]/open and *[+tense]/closed, as well as phrase-level IDENT(V), while keeping all other constraints at 0, will lower the objective function

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

with respect to all-zero weights, as this will nudge the probabilities of both /set#a/ → [sɛ.ta] and /se#ta/ → [se.ta] slightly above 0.50. However, for the reasons outlined above, the learner cannot see this possibility when it retreats to all-zero weights.

3 Raising and flapping in Canadian English

3.1 Data

Our second case study is a classic case of opacity, attested in Canadian English (Joos 1942, Idsardi 2000 and references therein, Pater 2014): the interaction between Canadian raising and flapping. The low-nucleus diphthongs [aɪ aʊ] raise to [ʌɪ ʌʊ] before voiceless consonants, as in (11a), and the coronal oral stops /t d/ become a voiced flap [ɾ] in a variety of contexts (De Jong 2011) – among others, between two vowels if the first is stressed and the second is not, as in (11b).

The two processes display a counterbleeding interaction: the fact that flapping cancels the voicelessness of underlying /t/ does not prevent that /t/ from triggering raising. This is illustrated in (11c).

(11) a. /laɪf/ → [lʌɪf] *life* cf. /laɪ/ → [laɪ] *lie*
b. /kʌt-ɚ/ → [kʌɾɚ] *cutter*
c. /maɪtɚ/ → [mʌɪɾɚ] *mitre* /saɪdɚ/ → [saɪɾɚ] *cider*

In addition, raising is restricted to the word domain, as illustrated in (12a), while (b) shows there is no evidence of such a restriction for flapping.

(12) a. /laɪ # fɔɹ/ → [laɪ fɔɹ] *lie for* *[lʌɪ fɔɹ] cf. [lʌɪf]
b. /laɪ # tu/ → [laɪ ɾə] *lie to* *[lʌɪ ɾə] [mʌɪɾɚ]

This fits with a stratal analysis in which raising applies only at the word level, and flapping only at the phrase level, as in (13).

(13) *Sketch of stratal analysis of flapping and raising*

UR	/maɪtɚ/	
word level	‖mʌɪtɚ‖	*raising only*
phrase level	[mʌɪɾɚ]	*flapping only*

The transparent counterpart to this pattern, claimed to be spoken by some Canadian English speakers (Joos 1942), is a language which also has both raising and flapping, as in (11) and (12), but in which the application of flapping blocks the application of raising, as in (14), since the flap [ɾ] is not voiceless. However, the existence of this transparent dialect is disputed in later literature (Kaye 1990). In our simulations, we use this (perhaps hypothetical) transparent dialect as a non-opaque baseline against which to compare the mainstream, opaque dialect of Canadian English.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

(14) /maɪtɚ/ → [maɪrɚ]

In the next subsection, we investigate the learnability of the opaque and the transparent versions of the Canadian English data. We consider the contrast between opacity and transparency, and between various possible datasets.

3.2 Simulation setup

Canadian English provides at least two pieces of independent evidence regarding the stratal affiliation of the opaque raising process: evidence for its application outside the flapping context, as in (11a), and evidence for its word-boundedness, as in (12a). Based on Kiparsky (2000), we predict the availability of such evidence to make the opaque interaction more learnable. To investigate this, we considered the four datasets in Table IV, each of which contains a subset of the relevant mappings observed in Canadian English, and provides various degrees of independent evidence for the stratal affiliation of the opaque process.

	mitre–cider	*mitre–cider–life*	*mitre–cider–lie for*	*mitre–cider–life–lie for*
opaque	/maɪtɚ/→[mʌɪrɚ] /saɪdɚ/→[saɪrɚ]	/maɪtɚ/→[mʌɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪf/→[lʌɪf]	/maɪtɚ/→[mʌɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪ#fɔɹ/→[laɪ fɔɹ]	/maɪtɚ/→[mʌɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪf/→[lʌɪf] /laɪ#fɔɹ/→[laɪ fɔɹ]
transparent	/maɪtɚ/→[maɪrɚ] /saɪdɚ/→[saɪrɚ]	/maɪtɚ/→[maɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪf/→[lʌɪf]	/maɪtɚ/→[maɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪ#fɔɹ/→[laɪ fɔɹ]	/maɪtɚ/→[maɪrɚ] /saɪdɚ/→[saɪrɚ] /laɪf/→[lʌɪf] /laɪ#fɔɹ/→[laɪ fɔɹ]

Table IV
Opaque datasets for Canadian English.

Mitre–cider provides no independent evidence regarding the opaque process – its synchronic activity and stratal affiliation must be inferred from the opaque interaction. *Mitre–cider–life* provides evidence for the opaque process applying outside the interaction, since the [f] in *life* is not subject to flapping: raising must be synchronically active at some stratum. *Mitre–cider–lie for* provides evidence for the non-application of raising at the phrase level, since *lie for* does not undergo raising, despite having [f] after the diphthong across a word boundary. Finally, *mitre–cider–life–lie for* combines both pieces of evidence. These datasets were investigated with both an opaque and a transparent interaction between raising and flapping.

Our learner had access to the four constraints in (15). As in the Southern French case, only two derivational levels were used: a word level and a phrase level.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

(15) a. IDENT[low]
Assign a violation mark for raising or lowering a diphthong.

b. IDENT[son]
Assign a violation mark for any underlying consonant whose [sonorant] specification is not identical to that of its output correspondent. This constraint penalises the transition from /t d/ to [ɾ], and from /ɾ/ to [t d].

c. *ÝTV
Assign a violation mark for an alveolar stop [t d] between a stressed and a following unstressed vowel.

d. *aɪ/__[−voice] (*aɪ/[−voice])
Assign a violation mark for a non-raised diphthong before a voiceless consonant.

Whenever *aɪ/[−voice] is sufficiently higher-weighted than IDENT[low], the raising process will be active. The flapping process is active when the weight of *ÝTV is sufficiently greater than that of IDENT[son].[4] The opaque interaction is captured if there is raising but no flapping at the word level, and flapping but no raising at the phrase level (see §1 on the impossibility of capturing the interaction within one stratum). This is reflected in the weights found at successful learning trials for *mitre–cider–life–lie for*. At the word level, *aɪ/[−voice] is weighted far above IDENT[low], and *ÝTV far below IDENT[son]; the opposite is true of the phrase level, as shown in Table V.

The word-level tableau in (16) shows that the weights above ensure raising but no flapping at the word level.

(16)

/maɪtɚ/	ID[son] 10.44	ID[low] 5.02	*ÝTV 0	*aɪ/[−vce] 11.13	$\mathcal{H}$	$e^{\mathcal{H}}$	p
a. ∥maɪtɚ∥			−1	−1	−11.13	0.00001	0.00
b. ∥maɪɾɚ∥	−1				−10.44	0.00003	0.00
c. ∥mʌɪtɚ∥		−1	−1		−5.02	0.007	0.99
d. ∥mʌɪɾɚ∥	−1	−1			−15.46	0.00000	0.00

The candidates considered at the phrase level are the same as those at the word level.[5] As can be seen in the phrase-level tableaux in (17), it is essential for the generation of these probabilities that /maɪtɚ/ exhibit raising at

[4] We assume that the transition from /t d/ to [ɾ] entails a change in [sonorant].

[5] We did not consider derivational paths with the changes /t/ → ∥ɾ∥ → [d] or /d/ → ∥ɾ∥ → [t]: the phrase-level candidate set did not include [m{a,ʌ}ɪdɚ] or [s{a,ʌ}ɪtɚ], so that every tableau had four candidates. However, we are confident that the presence of these candidates would not have significantly changed our results: as shown in (17b), candidates that change ∥ɾ∥ to [t] or [d] at the phrase level incur extra violations of IDENT[son] and *ÝTV, while not yielding any improvement on the other two constraints.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

word level				phrase level			
Id[son]	Id[low]	*V́TV	*aɪ[−vce]	Id[son]	Id[low]	*V́TV	*aɪ[−vce]
10.44	5.02	0	11.13	0	6.81	6.12	0

Table V
Sample successful weights for opaque *mitre–cider–life–lie for*.

the word level, since the phrase level must preserve diphthongs faithfully, otherwise /maɪtɚ/ will not surface with both raising and flapping (i.e. as attested [mʌɪɾɚ]).

(17) a.

/mʌɪtɚ/	Id[son] 0	Id[low] 6.81	*V́TV 6.12	*aɪ/[−vce] 0	$\mathcal{H}$	$e^{\mathcal{H}}$	p
i. [maɪtɚ]		−1	−1	−1	−12.93	0.000	0.00
ii. [maɪɾɚ]	−1	−1			−6.81	0.001	0.00
iii. [mʌɪtɚ]			−1		−6.12	0.002	0.00
iv. [mʌɪɾɚ]	−1				0	1	1.00

b.

/maɪɾɚ/							
i. [maɪtɚ]	−1		−1	−1	−6.12	0.002	0.00
ii. [maɪɾɚ]					0	1	1.00
iii. [mʌɪtɚ]	−1	−1	−1		−12.93	0.000	0.00
iv. [mʌɪɾɚ]		−1			−6.81	0.001	0.00

The surface-form probabilities resulting from the weights in Table V match the actual opaque Canadian English data extremely closely, as shown in Table VI.

/maɪtɚ/	p	/saɪdɚ/	p	/laɪf/	p	/laɪ#fɔɹ/	p
[maɪtɚ]	0.00	[saɪdɚ]	0.00	[laɪf]	0.00	[laɪ fɔɹ]	0.99
[maɪɾɚ]	0.00	[saɪɾɚ]	0.99				
[mʌɪtɚ]	0.00	[sʌɪdɚ]	0.00	[lʌɪf]	1.00	[lʌɪ fɔɹ]	0.01
[mʌɪɾɚ]	0.99	[sʌɪɾɚ]	0.00				

Table VI
Surface form probabilities generated by Table V.

As opposed to the opaque interaction, the transparent interaction does not need to order flapping after raising. In fact, for transparent *mitre–cider* and *mitre–cider–lie for*, raising need not be represented in the grammar at all, because the raised diphthong is absent from the data – see the lack of weight on *aɪ/[−voice] at both levels in Tables VIIa and b. For transparent *mitre–cider–life*, raising can be represented either at the

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

	dataset (transparent)	word level				phrase level			
		ID [son]	ID [low]	*ÝTV	*aɪ [−vce]	ID [son]	ID [low]	*ÝTV	*aɪ [−vce]
a.	*mitre–cider*	0	6.78	0	0	0	6.78	6.78	0
b.	*mitre–cider–lie for*	0	7.15	0	0	0	7.15	6.75	0
c.	*mitre–cider–life* (var. 1)	0	6.41	0	0	0	5.72	5.73	11.45
d.	*mitre–cider–life* (var. 2)	0	6.07	0	11.74	0	6.77	6.36	0
e.	*mitre–cider–life–lie for*	0	6.34	0	11.98	0	7.03	6.34	0

Table VII

Sample weights for successful runs of various transparent datasets.

word level, by giving high weight to word-level *aɪ/[−voice] and zero weight to word-level IDENT[son], as in (c) – or at the phrase level, by giving high weight to *ÝTV and *aɪ/[−voice] at the phrase level, as in (d). Transparent *mitre–cider–life–lie for*, however, does require raising to take place at the word level, because the non-raised diphthong in /laɪ#fɔɹ/ → [laɪ fɔɹ] precludes raising at the phrase level. The weights found for this dataset are very similar to the second set of weights for *mitre–cider–life*, as shown in (e).

Thus the opaque interaction of raising and flapping requires word-level raising and phrase-level flapping. The transparent interaction, however, only requires flapping at the word level, while raising can be represented at either level.

Finally, the addition of both *life* and *lie for* to the transparent dataset leads to the necessity of representing raising at the word level, even though this is not otherwise required for the transparent interaction.

3.3 Results

Simulations were run as described in §2.2, except that L2 regularisation was made stronger by setting σ^2 to 9,000 instead of 10,000, to prevent the learner from considering constraint weights that tend towards infinity for the opaque *mitre–cider* dataset.

All four datasets were examined with both opaque and transparent interaction between raising and flapping. The same 100 sets of initialisations drawn with replacement from a uniform distribution over [0, 10] were used for all eight datasets. The results are given in Table VIII.

As can be seen, independent evidence regarding the opaque process yields a clear increase in learnability for the opaque cases. Evidence that

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

dataset	learned successfully	
	opaque	transparent
mitre–cider	51	100
mitre–cider–life	61	99
mitre–cider–lie for	87	100
mitre–cider–life–lie for	92	93

Table VIII
Results for 100 runs for the Canadian English dataset.

the opaque process is word-bounded (*lie for*) has a stronger effect than evidence for independent activity of the opaque process (*life*).

For all transparent datasets except *mitre–cider–life–lie for*, performance is (almost) at ceiling. For *mitre–cider–life–lie for*, however, the opaque and transparent versions are learned at a near equal rate.

Whenever (opaque or transparent) Canadian English was not learned successfully (i.e. at least one attested SR was given a probability of 0.50 or less by the grammar), the learner ended up in one of the four local optima summarised in Table IX.

optimum	occurs in:	inputs	outputs
I	opaque *mitre–cider* *mitre–cider–lie for* *mitre–cider–life–lie for*	/maɪtɚ/ /saɪdɚ/ (/laɪf/) (/laɪ#fɔɹ/)	[mʌɪɾɚ] ~ [maɪɾɚ] [sʌɪɾɚ] ~ [saɪɾɚ] ([laɪf] ~ [lʌɪf]) ([laɪ fɔɹ] ~ [lʌɪ fɔɹ])
II	opaque *mitre–cider–lie for* *mitre–cider–life–lie for*	/maɪtɚ/ /saɪdɚ/ (/laɪf/) /laɪ#fɔɹ/	[mʌɪɾɚ] 0.67 ~ [maɪɾɚ] [sʌɪɾɚ] 0.67 ~ [saɪɾɚ] ([lʌɪf]) [laɪ fɔɹ] 0.67 ~ [lʌɪ fɔɹ]
III	opaque *mitre–cider–life* transparent *mitre–cider–life*	/maɪtɚ/ /saɪdɚ/ /laɪf/	[mʌɪɾɚ] ~ [maɪɾɚ] [sʌɪɾɚ] ~ [saɪɾɚ] [lʌɪf]
IV	transparent *mitre–cider–life–lie for*	/maɪtɚ/ /saɪdɚ/ /laɪf/ /laɪ#fɔɹ/	[maɪɾɚ] [saɪɾɚ] [laɪf] ~ [lʌɪf] [laɪ fɔɹ] ~ [lʌɪ fɔɹ]

Table IX
Local optima observed for the various Canadian English datasets. Parentheses indicate words that are not shared between all datasets for which this local optimum is attested.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

For each underlying representation the table lists the surface representations that have a probability of more than 0.00 in that local optimum. The symbol '~' will be used as a shorthand for 0.50 probability on both surface representations shown, unless indicated otherwise. Weights generating each local optimum are given in Table X.

	word level				phrase level			
	ID [son]	ID [low]	*V́TV	*aɪ [−vce]	ID [son]	ID [low]	*V́TV	*aɪ [−vce]
local optimum I	0	0	0	0	0	0	7.75	0
local optimum II	0	6.60	0	0	0	5.91	5.92	5.90
local optimum III	0	0.69	6.14	6.54	6.51	5.86	0.04	0
local optimum IV	0	0	0	0	0	0	7.01	6.60

Table X
Sample weights for local optima.

These four local optima have in common that they try to represent both processes with minimal appeal to the interaction between levels. For instance, local optima I and IV are attempts to represent the data without appealing to underlying representations, by setting all faithfulness constraints to zero.

Local optimum II does not represent the raising process at the word level, as would be necessary for any dataset involving *life* with a raised diphthong and *lie for* with a non-raised diphthong (see §3.2). This means that the vowel contrast between *life* and *lie for* has to be modelled as within-word variation.

Finally, local optimum III is a consequence of representing both raising and flapping at the word level. Since applying raising and flapping at the same level leads to lack of raising in *mitre*, variation between raising and non-raising is created by lowering the weight of word-level IDENT[low].

As in the case of Southern French, local optima I and IV are not local optima in the narrow sense: if the weights of word-level IDENT[son], IDENT[low] (in the case of local optimum I) and *aɪ/[−voice] and phrase-level IDENT[low] are increased by even a little, this will decrease the value of the objective function. It is this 'bottleneck effect' (see §4.1) that leads the learner to land in this state. However, local optima II and III are true local optima, in the sense that any neighbouring values will increase the value of the objective function.

As a final note, none of these outcomes represent opaque analyses. Local optima II and III resemble opaque analyses, since they both assign a significant amount of probability to [mʌɪɾɚ], but they crucially ignore the difference between underlying /t/ in *mitre* and underlying /d/ in

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

cider, and allow unmotivated raising in both words (because the constraints that regulate raising either have a weight of 0 or are tied with competing constraints).

3.4 Summary

We have found that independent evidence about the stratal affiliation of an opaque pattern can significantly improve the learnability of the opaque interaction – especially the addition of evidence that the process does not apply at the phrase level. Furthermore, the presence of both *lie* and *lie for* makes the transparent and opaque interactions equally learnable, and in fact destroys the transparent interaction's learning advantage.

Whenever the languages are not learned successfully, this is because either phrase-level faithfulness is given zero weight, making it impossible to transfer information from the word level to the phrase level, or raising and flapping are represented at the same level when they need to be represented at different levels. We will now turn to a discussion of the 'bottleneck effect' (see §3.3), an obstacle posed by the current hidden structure learning problem that leads the learner away from the global optimum, as well as ways in which evidence for stratal affiliation helps the learner overcome this obstacle.

4 Difficulties in learning hidden structure

4.1 Cross-level dependencies

The relative difficulty of learning opaque interactions found in our results seems to stem from the fact that the effectiveness of the weighting on each level depends on the weights on the other level. Specifically, high weight on phrase-level faithfulness is only effective when word-level constraints are weighted appropriately, while the result of word-level weighting can only be transmitted to the surface representation when phrase-level faithfulness has a high enough weight.

We will show here how the learner fails to reach the global optimum if it finds appropriate weights only for word-level constraints or only for phrase-level faithfulness, and we will show that this scenario is more likely to occur in opaque cases than in transparent cases. Furthermore, we will show how the types of evidence for stratal affiliation of the opaque process offered by the Canadian English datasets increase the likelihood that the learner will find the global optimum.

When the learner has not found weights that generate a desirable distribution at the word level, it gets closer to its objective by lowering the weights of phrase-level faithfulness, rather than raising them. For instance, consider the weighting in Table XI, which leads to a local optimum for opaque *mitre–cider*.

At the word level, this weighting gives a non-raised diphthong in *mitre* highest probability, because high-weighted *ÝTV and low-weighted

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

word level				phrase level			
Id[son]	Id[low]	*ÝTV	*aɪ[−vce]	Id[son]	Id[low]	*ÝTV	*aɪ[−vce]
1	7	3	1	0	6.28	6.29	0

Table XI
Sample initialisation leading to local optimum for *mitre–cider*.

*aɪ/[−voice] leads to flapping, which blocks diphthong raising, as illustrated for the word-level tableau in (18).

(18)

/maɪtɚ/	Id[son] 1	Id[low] 7	*ÝTV 3	*aɪ/[−vce] 1	$\mathcal{H}$	$e^{\mathcal{H}}$	p
a. ∥maɪtɚ∥			−1	−1	−4	0.01	0.05
b. ∥maɪɾɚ∥	−1				−1	0.37	0.95
c. ∥mʌɪtɚ∥		−1	−1		−10	0.00	0.00
d. ∥mʌɪɾɚ∥	−1	−1			−8	0.00	0.00

At the same time, as shown in (19), it displays flapping and lack of raising on the phrase level, as desired for the opaque interaction (cf. §3.2).

(19)

	∥mʌɪtɚ∥	Id[son] 0	Id[low] 6.28	*ÝTV 6.29	*aɪ/[−vce] 0	$\mathcal{H}$	$e^{\mathcal{H}}$	p
a.	i. [maɪtɚ]		−1	−1	−1	−12.57	0.000	0.00
	ii. [maɪɾɚ]	−1	−1			−6.28	0.002	0.00
	iii. [mʌɪtɚ]			−1		−6.29	0.002	0.00
	iv. [mʌɪɾɚ]	−1				0	1	1.00
b.	∥maɪɾɚ∥							
	i. [maɪtɚ]	−1		−1	−1	−6.29	0.002	0.00
	ii. [maɪɾɚ]					0	1	1.00
	iii. [mʌɪtɚ]	−1	−1	−1		−12.57	0.000	0.00
	iv. [mʌɪɾɚ]		−1			−6.28	0.002	0.00

In cases like this, KL divergence decreases (and fit to the data increases) as the weight of phrase-level Ident[low] approaches zero. As shown in (10) above, KL divergence is computed based on observed and expected probabilities. Table XII shows the observed probabilities for opaque *mitre–cider*, and Table XIII shows expected probabilities for various weights of phrase-level Ident[low] and the resulting KL divergence.

If the learner has assigned zero weight to phrase-level faithfulness, moving towards appropriate weights at the word level does not lower KL divergence. This is illustrated in Table XIV. Word-level and phrase-level expected probabilities are shown only for *mitre*, but KL

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

/maɪtɚ/				/saɪdɚ/			
[maɪtɚ]	[maɪɾɚ]	[mʌɪtɚ]	[mʌɪɾɚ]	[saɪdɚ]	[saɪɾɚ]	[sʌɪdɚ]	[sʌɪɾɚ]
0	0	0	1	0	1	0	0

Table XII
Observed distribution for opaque *mitre–cider*.

weights (phrase level)				/maɪtɚ/				/saɪdɚ/				D_{KL}
Id [son]	Id [low]	*ÝTV	*aɪ	[aɪt]	[aɪɾ]	[ʌɪt]	[ʌɪɾ]	[aɪt]	[aɪɾ]	[ʌɪt]	[ʌɪɾ]	
0	**6.28**	6.29	0	0.00	1.00	0.00	**0.00**	0.00	**1.00**	0.00	0.00	5.87
0	**1**	6.29	0	0.00	0.72	0.00	**0.27**	0.00	**0.72**	0.00	0.27	1.63
0	**0**	6.29	0	0.00	0.50	0.00	**0.50**	0.00	**0.50**	0.00	0.50	1.39

Table XIII
KL divergence for grammars with word-level weights as in Table XI and various weights of phrase-level Ident[low].

divergence is computed for both *mitre* and *cider*. Phrase-level constraint weights are as given in Table XI, except for Ident[low], whose weight is set to zero.

weights					word-level outputs for /maɪtɚ/				surface representations				D_{KL}
Id [son] Wd	Id [low] Wd	*ÝTV Wd	*aɪ [−vce] Wd	Id [low] Phr	aɪt	aɪɾ	ʌɪt	ʌɪɾ	aɪt	aɪɾ	ʌɪt	ʌɪɾ	
1	7	3	1	**0**	.05	.95	**.00**	.00	.00	.50	.00	.50	1.39
4	3	2	4	**0**	.09	.64	**.23**	.03	.00	.50	.00	.50	1.39
6	3	1	6	**0**	.04	.11	**.84**	.01	.00	.50	.00	.50	1.39
7	3	0	7	**0**	.02	.02	**.96**	.00	.00	.50	.00	.50	1.39

Table XIV
KL divergence for *mitre–cider* when phrase-level Ident[low] has zero weight.

As can be seen in Table XIV, raising the weights of word-level Ident[low] and *ÝTV and lowering those of word-level Ident[son] and *aɪ/[−voice] leads to a desirable result at the word level: raising and no flapping (cf. (17)). However, this information is not factored into the distribution over UR/SR mappings if phrase-level Ident[low] has a weight of 0.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

At the same time, phrase-level IDENT[low] has a motivation to decrease its weight until word-level outputs with a raised diphthong gain a cumulative probability of at least 0.50. The lower the weight of IDENT[low], the closer the winning candidate will be to having a probability of 0.50, since setting IDENT[low] to zero means that the phrase level will assign equal probability to raised and non-raised diphthongs. As long as the current word-level weights are such that the winning candidate has less than 0.50 probability, this is an improvement. In the particular case discussed here, IDENT[low] will be motivated to have non-zero weight only when the raised diphthong is more likely than the non-raised diphthong at the word level, so that the probability of [mʌɪɾɚ] increases as the weight of IDENT[low] increases.

This is what yields the aforementioned bottleneck effect: if the current word-level weights lead to lower probability on the attested candidate compared to all-zero weights at the word level, then phrase-level faithfulness might be set to zero before a more desirable set of word-level weights is found.[6] Thus word-level information needs to travel through the bottleneck of the phrase-level faithfulness constraints in order to have an effect on the distribution over UR/SR mappings. The bottleneck effect is not limited to just this dataset, but applies to any dataset with dependence between levels, including the Southern French case in §2.3 above and other Canadian English datasets (see §4.2 below).

4.2 Advantage from evidence for stratal affiliation

As we showed in §3.3, the addition of various kinds of evidence for the stratal affiliation of the opaque process of raising dramatically increases the likelihood that the opaque interaction will be learned.

Adding *life* to the opaque interaction means that raising at either the word level or the phrase level will be rewarded, regardless of the weighting of the flapping constraint *ÝTV. This means that increasing the weight of word level *aɪ/[−voice] leads to a sharper drop in KL divergence for opaque *mitre–cider–life* than for opaque *mitre–cider–lie for*, as shown in Table XV.[7]

In this way, the presence of *life* in the dataset means a sharp increase in the gradient of the word-level constraint *aɪ/[−voice] in a position in the weight space that would otherwise have a low gradient for that constraint. We have not studied how the exact numerical increase in the gradient of this constraint, which depends on the relative frequency of data points

[6] More generally, the effect happens when the current word-level weights produce a distribution over surface candidates that has a higher KL divergence from the actual distribution than a uniform distribution (i.e. when the word-level grammar makes the predictions for surface forms go in the opposite direction of the actual data).

[7] However, lowering the weight of phrase-level IDENT[low] is also rewarded more strongly compared to opaque *mitre–cider*, so that the bottleneck effect becomes stronger with the addition of *life*. This is probably why adding *life* only had a modest effect on learnability.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

weights (word level)				weights (phrase level)				D_{KL}	D_{KL}
Id [son]	Id [low]	*ÝTV	*aɪ [−vce]	Id [son]	Id [low]	*ÝTV	*aɪ [−vce]	*mitre–cider*	*mitre–cider–life*
1	7	3	1	0	6.28	6.29	0	5.87	11.31
1	7	3	7	0	6.28	6.29	0	5.85	6.54
1	7	3	14	0	6.29	6.29	0	5.85	5.85

Table XV

Correlation between the weight of word-level *aɪ/[−vce] and KL divergence for *mitre–cider* and *mitre–cider–lie for*.

like *life* in the total corpus (we only considered a minimal corpus here), relates to learnability rates – this is a matter for future research. However, we can confidently say that the categorical presence of raising outside its interaction with flapping has a positive influence on the learnability of the interaction.

Adding *lie for* to the opaque interaction, as in Table XVI, means that high weight on phrase-level Ident[low] is penalised less strongly when the word-level weights do not generate the desirable candidates with high probability.[8] This is because the word *lie* itself contains neither a flapping nor a raising context, so that flapping and raising constraints do not interact with the identity of the diphthong at the word level. It is not necessary to determine the mutual weighting of word-level *ÝTV and Ident[son] or *aɪ/[−voice] and Ident[low] to give the desirable word-level output for *lie for* more than 0.50 probability – non-zero weight on word-level Ident[low] is sufficient.

weights (word level)				weights (phrase level)				D_{KL}	D_{KL}
Id [son]	Id [low]	*ÝTV	*aɪ [−vce]	Id [son]	Id [low]	*ÝTV	*aɪ [−vce]	*mitre–cider*	*mitre–cider–lie for*
1	7	3	1	0	6.28	6.29	0	5.87	5.87
1	7	3	7	0	1	6.29	0	1.63	1.94
1	7	3	14	0	0	6.29	0	1.39	2.08

Table XVI

Correlation between the weight of phrase-level Ident[low] and KL divergence for *mitre–cider* and *mitre–cider–lie for*.

When both *life* and *lie for* are present in the dataset, the grammar must represent the raising process at the word level in order to generate lack of raising in *lie for* (because word-level *aɪ/[−voice] cannot see that *lie for* has

[8] Another effect of adding *lie for* is that high weight on phrase-level *aɪ/[−voice] is penalised more strongly, because this constraint prefers a raised diphthong in *lie for*.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

/aɪ/ before /f/, while phrase-level *aɪ/[−voice] can). As shown in §3.2, this introduces an additional dependency between word and phrase levels for the transparent datasets: while the transparent interaction between flapping and raising does not require raising to apply before flapping, the combination of *life* and *lie for* does.

This creates a bottleneck effect for the transparent dataset, which explains the decrease in learnability for transparent *mitre–cider–life–lie for*. The increase in learnability for opaque *mitre–cider–life–lie for*, on the other hand, can be seen as the cumulative effect of *life* and *lie for* on the opaque interaction, as reviewed above.

5 Concluding remarks

We have presented here an approach to learning opaque and transparent interactions in a MaxEnt version of Stratal OT. Our goal was to investigate whether the general setup of Stratal OT – chained parallel OT grammars with independent rankings or weightings of constraints – predicts learnability differences between opaque and transparent process interactions, and also whether evidence of a process's stratal affiliation makes it easier to learn opaque interactions.

Our first case study was opaque tensing/laxing in Southern French. We found that it was learned at a high rate of accuracy, but the solution space has a local optimum – one where the grammar does not represent the phonological process at all, which yields free variation in the data.

We then looked at the opaque interaction between diphthong raising and flapping in Canadian English (Joos 1942, Idsardi 2000 and references therein). The opaque raising process also applies in contexts where flapping is irrelevant, and it does not apply across word boundaries – both of which constitute evidence for the stratal affiliation of raising.

We found that, without additional evidence for stratal affiliation, the opaque interaction was learned at a rate of about 50%, while its transparent counterpart was learned at ceiling. However, addition of this additional evidence improved the learnability of the opaque interaction to a maximum of 92%, while the learnability of the transparent interaction decreased to a similar rate. This confirms Kiparsky's prediction: evidence of the stratal affiliation of raising does improve its learnability when it is opaque. We also found that the advantage of transparent over opaque interactions is relative: the dataset with full evidence for stratal affiliation did not produce a learnability difference between the two.

Our explanation for the observed learning difficulties in this framework has to do with a bottleneck effect in the transmission of information from earlier derivational levels through faithfulness constraints. This effect makes it more difficult to find the global optimum when the learner starts with word-level weights that predict the wrong surface form. However, as shown in §4.2, this effect is mitigated for opaque Canadian English by information about an opaque process's stratal affiliation,

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

because this information either boosts desirable word-level weightings, punishes undesirable phrase-level weightings or diminishes the bottleneck effect in general.

The two cases that we considered differ in their learnability rate. The Southern French dataset contains evidence that the opaque process (tensing/laxing) is word-bounded, making it analogous to the opaque *mitre–cider–lie for* dataset in the Canadian English case study. While Southern French was learned at a rate of 96%, *mitre–cider–lie for* was only learned at a rate of 87%. A possible explanation is that the Southern French case study did not include constraints on syllable structure, while the process that makes tensing/laxing opaque is resyllabification. We predict that the learnability of Southern French would be slightly reduced if such constraints were introduced into the simulation.

We have limited ourselves here to two case studies and one learning implementation. Other approaches to learning Stratal OT are possible too: probabilistic ranked constraint learning with expectation maximisation (Jarosz 2006, 2015, 2016), Noisy Harmonic Grammar (Coetzee & Pater 2011) or Stochastic OT (Boersma 1998). These might differ in their particular learning strategies, the local optima they find and the distribution over outputs generated at local optima. Nonetheless, the mechanisms responsible for the learnability differences that we found are quite general. Interdependence between phrase-level faithfulness and appropriate weighting at the word level poses a challenge for finding a grammar that generates the learning data. For this reason, we predict that other learning approaches will find results similar to ours (an early confirmation of this can be found in Jarosz 2016). However, more work is needed that explores the learnability of these and other opaque interactions in other learning frameworks, including the non-serial multi-level framework in Boersma & van Leussen (to appear). Further work is also needed to examine how other cases of opacity behave in our framework. In particular, more complex interactions that involve more constraints and/or more derivational levels would be essential to test the predictions of a general advantage for transparency, and of improved learning of opacity given evidence of stratal affiliation. In addition, as pointed out in §4.2, it is important to understand how prevalent evidence for stratal affiliation must be in the data to ensure successful learning.

Finally, further research would need to be done before one could draw conclusions about the relationship of our findings to attested patterns of language learning, language change and language typology. In naturalistic language learning, we are unaware of any evidence that learners have biases toward transparent interactions, or that they find some cases of opacity easier than others. Such evidence may be difficult to come by outside the laboratory, since confounds are generally controlled for when the learning of different patterns is compared. There is some evidence that transparent interactions are easier in artificial language learning (Kim 2012), but this line of research is only beginning. There is also evidence that opacity

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

can be innovated by children (e.g. Dinnsen & Farris-Trimble 2008); whether or not this is a challenge to our model or other related theories would require a detailed consideration of the constraints involved and the steps on the learning path. To apply our model to historical change would require adding the analogue of rule loss, that is the possibility of mislearning an opaque interaction as a contrast; our constraint set cannot generate this possibility. Finally, to draw conclusions about how these results fit the relative degree of typological attestation of opaque and transparent interactions, we would need not only a model that connects learning to typological attestation (see e.g. Staubs 2014a), but also empirical research to determine whether or not opacity is underattested.

REFERENCES

Bermúdez-Otero, Ricardo (1999). *Constraint interaction in language change: quantity in English and Germanic*. PhD dissertation, University of Manchester & University of Santiago de Compostela.

Bermúdez-Otero, Ricardo (2003). The acquisition of phonological opacity. In Spenader *et al*. (2003). 25–36.

Boersma, Paul (1998). *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. PhD dissertation, University of Amsterdam.

Boersma, Paul & Jan-Willem van Leussen (to appear). Efficient evaluation and learning in multilevel parallel constraint grammars. *LI* **48**.

Byrd, Richard H., Peihuang Lu, Jorge Nocedal & Ciyou Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**. 1190–1208.

Coetzee, Andries W. & Joe Pater (2011). The place of variation in phonological theory. In Goldsmith *et al*. (2011). 401–434.

De Jong, Kenneth J. (2011). Flapping in American English. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.) *The Blackwell companion to phonology*. Malden, Mass.: Wiley-Blackwell. 2711–2729.

Dinnsen, Daniel A. & Ashley W. Farris-Trimble (2008). An opacity-tolerant conspiracy in phonological acquisition. *Indiana University Working Papers in Linguistics* **6**. 99–118.

Eisenstat, Sarah (2009). *Learning underlying forms with MaxEnt*. MA thesis, Brown University.

Eychenne, Lucien (2014). Schwa and the *loi de position* in Southern French. *Journal of French Language Studies* **24**. 223–253.

Goldsmith, John A., Jason Riggle & Alan C. L. Yu (eds.) (2011). *The handbook of phonological theory*. 2nd edn. Malden, Mass.: Wiley-Blackwell.

Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Spenader *et al*. (2003). 111–120.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**. 379–440.

Hoerl, Arthur E. & Robert W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**. 55–67.

Idsardi, William J. (2000). Clarifying opacity. *The Linguistic Review* **17**. 337–350.

Jarosz, Gaja (2006). *Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory*. PhD dissertation, Johns Hopkins University.

Jarosz, Gaja (2014). Serial markedness reduction. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.) *Proceedings of the 2013 Meeting on*

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

Phonology. Available (May 2017) at http://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/40.

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

Jarosz, Gaja (2016). Learning opaque and transparent interactions in Harmonic Serialism. In Gunnar Ólafur Hansson, Ashley Farris-Trimble, Kevin McMullin & Douglas Pulleyblank (eds.) *Proceedings of the 2015 Annual Meeting on Phonology*. Available (May 2017) at http://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/3671.

Johnson, Mark (2013). A gentle introduction to maximum entropy, log-linear, exponential, logistic, harmonic, Boltzmann, Markov Random Fields, Conditional Random Fields, etc., models. Slides of paper presented to the Macquarie University Machine Learning Reading Group. Available (May 2017) at http://web.science.mq.edu.au/~mjohnson/papers/Johnson12IntroMaxEnt.pdf.

Johnson, Mark, Joe Pater, Robert Staubs & Emmanuel Dupoux (2015). Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 303–313.

Joos, Martin (1942). A phonological dilemma in Canadian English. *Lg* **18**. 141–144.

Kaye, Jonathan (1990). What ever happened to Dialect B? In Joan Mascaró & Marina Nespor (eds.) *Grammar in progress: GLOW essays for Henk van Riemsdijk*. Dordrecht: Foris. 259–263.

Kim, Yun Jung (2012). Do learners prefer transparent rule ordering? An artificial language learning study. *CLS* **48:1**. 375–386.

Kiparsky, Paul (1971). Historical linguistics. In William Orr Dingwall (ed.) *A survey of linguistic science*. College Park: University of Maryland Linguistics Program. 576–642.

Kiparsky, Paul (1973). Abstractness, opacity, and global rules. In Osamu Fujimura (ed.) *Three dimensions in linguistic theory*. Tokyo: TEC. 57–86.

Kiparsky, Paul (2000). Opacity and cyclicity. *The Linguistic Review* **17**. 351–365.

Kullback, S. & R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**. 79–86.

McCarthy, John J. (2007). *Hidden generalizations: phonological opacity in Optimality Theory*. Sheffield & Bristol, Conn.: Equinox.

McCarthy, John J. & Joe Pater (eds.) (2016). *Harmonic Grammar and Harmonic Serialism*. London: Equinox.

Moreux, Bernard (1985). La 'Loi de Position' en français du Midi. I: Synchronie (Béarn). *Cahiers de Grammaire* **9**. 45–138.

Moreux, Bernard (2006). Les voyelles moyennes en français du Midi: une tentative de synthèse en 1985. *Cahiers de Grammaire* **30**. 307–317.

Odden, David (2011). Rules v. constraints. In Goldsmith *et al*. (2011). 1–39.

Pater, Joe (2014). Canadian raising with language-specific weighted constraints. *Lg* **90**. 230–240.

Pater, Joe (2016). Universal Grammar with weighted constraints. In McCarthy & Pater (2016). 1–46.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith (2012). Learning probabilities over underlying representations. In *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology*. Montreal: Association for Computational Linguistics. 62–71. Available (May 2017) at www.aclweb.org/anthology/W12-2308.

R Core Team (2013). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press

Rizzolo, Olivier (2002). *Du leurre phonétique des voyelles moyennes en français et du divorce entre licenciement et licenciement pour gouverner*. PhD dissertation, University of Nice-Sophia Antipolis.
Selkirk, Elisabeth (1978). The French foot: on the status of 'mute' e. *Studies in French Linguistics* **1:2**. 141–150.
Smolensky, Paul (1986). Information processing in dynamical systems: foundations of Harmony Theory. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group (eds.) *Parallel Distributed Processing: explorations in the micro-structure of cognition*. Vol. 1: *Foundations*. Cambridge, Mass.: MIT Press. 194–281.
Smolensky, Paul & Géraldine Legendre (eds.) (2006). *The harmonic mind: from neural computation to optimality-theoretic grammar*. 2 vols. Cambridge, Mass.: MIT Press.
Spenader, Jennifer, Anders Eriksson & Östen Dahl (eds.) (2003). *Variation within Optimality Theory: Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory'*. Stockholm: Department of Linguistics, Stockholm University.
Staubs, Robert (2014a). *Computational modeling of learning biases in stress typology*. PhD dissertation, University of Massachusetts Amherst.
Staubs, Robert (2014b). *Stratal MaxEnt Solver*. Software package. Available (July 2017) at http://www.linguist.robertstaubs.org/HGR/serialMaxEnt.zip.
Staubs, Robert & Joe Pater (2016). Learning serial constraint-based grammars. In McCarthy & Pater (2016). 369–388.
Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.

https://doi.org/10.1017/S095267571700015X Published online by Cambridge University Press