ISSN 1695-6885 (in press); 2014-9719 (online) Catalan Journal of Linguistics 15, 2016 **53-66**

Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints*

Claire Moore-Cantwell
Yale University
clairemoorecantwell@gmail.com

Joe Pater
University of Massachusetts Amherst
pater@linguist.umass.edu

Received: February 25, 2016
Accepted: June 22, 2016

Abstract

The number of exceptions to a phonological generalization appears to gradiently affect its productivity. Generalizations with relatively few exceptions are relatively productive, as measured in tendencies to regularization, as well as in nonce word productions and other psycholinguistic tasks. Gradient productivity has been previously modeled with probabilistic grammars, including Maximum Entropy Grammar, but they often fail to capture the fixed pronunciations of the existing words in a language, as opposed to nonce words. Lexically specific constraints allow existing words to be produced faithfully, while permitting variation in novel words that are not subject to those constraints. When each word has its own lexically specific version of a constraint, an inverse correlation between the number of exceptions and the degree of productivity is straightforwardly predicted.

Keywords: exceptions; variation; computational phonology; Maximum Entropy Grammar; indexed constraints

Resum. *Excepcionalitat gradual i gramàtica de màxima entropia amb restriccions especificades lèxicament*

El nombre d'excepcions a una generalització fonològica sembla que afecta de forma gradual la seva productivitat. Les generalitzacions amb relativament poques excepcions són bastant productives, per les mesures en tendències a la regularització i per les produccions de mots sense sentit i altres tasques psicolingüístiques. La productivitat gradual s'ha modelat prèviament amb gramàtiques probabilístiques, incloent-hi la gramàtica de màxima entropia, però sovint no aconsegueixen recollir les pronunciacions fixes de paraules existents en una llengua, contràriament al que passa amb les paraules sense sentit. Les restriccions especificades lèxicament permeten produir els mots existents de manera fidel i al mateix temps permeten variació en mots nous, que

* This material is based on work supported by NSF Graduate Research Fellowship DGE-0907995 to C.M.C., NSF grant BCS-424077 to the University of Massachusetts Amherst, and a City of Paris Research in Paris fellowship to J.P. We thank the participants in the OCP Exceptionality Workshop, and members of the UMass phonological community, especially Robert Staubs, for helpful discussion.

no estan subjectes a aquestes restriccions. Quan cada mot té la seva pròpia versió d'una restricció especificada lèxicament es prediu directament una correlació inversa entre el nombre d'excepcions i el grau de productivitat.

Paraules clau: excepcions; variació; fonologia computacional; gramàtica de màxima entropia; restriccions indexades

Table of Contents

1. The problem: gradient exceptionality

The most general form of the problem that we address in this paper is the inadequacy of a two-way distinction between a regular / rule-governed / general phonological pattern and an exceptional / lexical / minor one. This two-way theoretical distinction is inadequate because it does not match the observed data. Phonological patterns across languages display a continuum of productivity, or conversely, of exceptionality. This is demonstrated at length in Hayes' (2008: ch. 9) textbook chapter, which offers a number of examples at various points of this continuum.

We discuss two empirical examples of gradient exceptionality in this paper. In this first section, we use the example of gradient exceptionality in lexical stress placement to elaborate on the problem, and in Section 2 we use it to exemplify our solution. In section 3, we present computational modeling results for another case, that of voicing alternations in Dutch (Ernestus and Baayen 2003).

1.1. Gradient exceptionality in lexical stress placement

The problem of gradient exceptionality was perhaps first pointed out by Fidelholtz (1979: 58):

> It appears to be a problem for linguistic theory that there is nothing in the formal description of Polish stress which would indicate that Polish is a 'penultimate-stress' language, as compared with the similar rules in English, which is essentially a free-stress language.

When the penultimate syllable is light, stress falls on the penultimate syllable of some English and Polish words (e.g. English *banána*, Polish *spokójny* 'quiet'), and on the antepenult on others (e.g. English *Cánada*, Polish *fizyka* 'physics'). In English, both patterns are well-attested (Pater 1994), and each word's pronunciation is stable; there is apparently no regularization to either penultimate or antepenultimate stress. In Polish, there are very few antepenultimately stressed words – Peperkamp *et al*. (2010) note that about 0.1% of the vocabulary has exceptional stress, which also includes final stress. Antepenultimately stressed words tend to

be borrowings and or learned words, and there is frequent regularization to penultimate stress.

Besides frequency of regularization, another piece of evidence that the difference in lexical statistics correlates with a difference in productivity comes from Peperkamp *et al*. (2010) psycholinguistic research with native speakers of languages with varying degrees of exceptionality. They use a nonce word memory task to determine how well participants are able to encode stress differences. Participants hear a sequence of 5 bisyllables that differ only in the placement of stress. The task is to report which syllable was stressed in each of the words. Performance on this task is compared with performance on a similar task with a segmental contrast. Peperkamp *et al.* (2010) find evidence of three-way grouping of participants by language background. Spanish participants, whose language has the highest degree of exceptionality (closest to English), perform as well on the stress contrast as on the segmental contrast. French, Finnish and Hungarian participants, whose languages have predictable stress, perform much worse on the stress contrast. Polish participants' level of performance is in between the two other groups.

The problem in linguistic theory that Fidelholtz is alluding to in the cited passage is that the number of exceptions, few or many, does not affect the grammatical status of a pattern in a standard generative grammar. The formal descriptions of both English and Polish, as well as Spanish, would require lexical marking of one of the patterns, with the other generated by rule. This is true of the SPE formalism of Fidelholtz's time, of metrical rule approaches, and of OT accounts with lexically specific constraints, posited independently by Kraska-Szlenk (1995) for Polish and Pater (2000) for English stress. Standard generative accounts can generate the existing words of the languages, but they do not account for the varying degrees of productivity of the patterns.

1.2. Previous generative approaches to gradient productivity

Gradient productivity has been the subject of considerable research in the recent generative literature, especially in the modeling of nonce word judgments or productions. Zuraw (2000), Ernestus and Baayen (2003), Hayes and Wilson (2008) and Hayes, Zuraw, Siptár and Londe (2009) and others attack the problem using stochastic grammars, either Stochastic OT (Boersma 1998) or Maximum Entropy Grammar (MaxEnt: Goldwater and Johnson 2003; referred to as a log-linear model in Ernestus and Baayen). As a simple example of this approach, the tableaux in (1) and (2) show the activity of two MaxEnt grammars, one generating penultimate stress with 0.95 probability, and another choosing between penultimate and antepenultimate stress at chance. We consider only left-headed binary feet, and have all unstressed vowels as schwa, as in English. The constraints choosing the position of the foot are Align-R, which demands a foot at the right edge of the word, and Nonfinality, which demands that the final syllable be unfooted. Violations are indicated by negative integers, and the weights of the two constraints are shown beneath their names. The column labeled *H* shows the weighted sum of violations, or Harmony (Smolensky and Legendre 2006). The probability of each candidate

is shown in the last column; it is proportional to the exponential of its *Harmony*. When Align-R has higher weight, as in (1), penultimate stress has higher probability. When the two constraints have equal weight, as in (2), penultimate and antepenultimate stress have equal probability.

(1) *A grammar choosing penultimate stress 95% of the time*

/bætækæ/	ALIGN-R 4	NONFIN 1	*H*	*p*
bə(tækə)		–1	–1	0.95
(bætə)kə	–1		–4	0.05

(2) *A grammar choosing penultimate stress 50% of the time*

/bætækæ/	ALIGN-R 2	NONFIN 2	*H*	*p*
bə(tækə)	–1		–2	0.50
(bætə)kə		–1	–2	0.50

These tableaux would be appropriate for a case in which multiple pronunciations of a single word are generated by a single speaker's grammar. For the production of a nonce word by a native speaker of English, a grammar of this type would be appropriate, since for a given word of this shape either penultimate or antepenultimate stress would be assigned with about equal probability (Moore-Cantwell 2015). For existing words in the English lexicon, however, this would not be the correct analysis (or as we will shortly claim, it's not the complete analysis). Words with final schwa in English, like *Cánada* and *banána* do have close to a 50/50 split in penultimate *vs*. antepenultimate stress (Moore-Cantwell 2015), but each of the words is produced in a single way. In other words, English stress displays what's traditionally called exceptionality, or what we might more neutrally call lexically conditioned application. An unelaborated MaxEnt grammar of this type is instead appropriate for what's traditionally called variation (Goldwater and Johnson 2003), as displayed in English t/d-deletion (though it's worth noting that variation is often, if not always, lexically conditioned; Coetzee and Pater 2011).

The challenge in modeling gradient exceptionality is to account for how it can affect responses in nonce word production and other psycholinguistic tasks, while at the same time allowing existing real words to be produced in a non-variable way. This challenge has long been recognized (Zuraw 2000), but it is sometimes not addressed. For example, Hayes *et al*. (2009) provide a MaxEnt grammar for Hungarian vowel harmony that is trained on the lexicon, and that generates patterns that are compared with nonce word productions. They do not confront the problem that the grammar would generate the wrong outcome for many real Hungarian words (see also Ernestus and Baayen's 2003 similar Stochastic OT and MaxEnt modeling of Dutch voicing alternations). To deal with gradient

exceptionality, Zuraw (2000) develops a model in which morphologically complex words are stored both as wholes and in decomposed forms, with the grammar choosing between the two based on the constraint ranking. This might be applied to the Dutch or Hungarian examples, but it does not seem to be applicable when the gradient pattern is over underived words, as in the stress cases we have been discussing here. Conversely, Hayes and Wilson (2008) develop a MaxEnt model of phonotactics that defines a probability distribution over the space of possible words, and which generates well-formedness scores for underived nonce forms. The Hayes and Wilson (2008) model is not applicable to alternations (though see the extension in Allen and Becker 2015 and Gouskova and Becker to appear).

To meet this challenge, we propose to combine MaxEnt grammar with lexically specific constraints, which are able to encode lexical conditioning (see Pater 2010 for an overview and comparison with alternatives, such as co-grammars). Real words have stable pronunciations because their associated lexically specific constraints have sufficiently high weight; nonce words have no associated constraints. Lexically conditioned constraints have in fact been suggested as a means of coping with gradient exceptionality in a deterministic version of OT (Pater 2005, Becker, Nevins and Ketrez 2011), but this requires a special grammar-external calculation over the lexicon to get the influence of the degree of regularity.

2. The proposal: Maximum Entropy Grammar with Lexically Specific Constraints

We assume that the grammar is composed of general constraints, and lexically specific versions of (some of) them, indexed to particular items. Continuing with the example from the last section, we show in (3) how lexically indexed constraints can stabilize the pronunciation of individual lexical items. As we saw in (2) above, equally weighted Align-R and Nonfin on their own generate equal probability for penultimate and antepenultimate stress. The tableau in (3) adds an indexed Align-R constraint for *banana*, whose weight leads to near-fixed penultimate stress for that word, and an indexed Nonfin constraint for *Canada*, whose weight leads to near-fixed antepenultimate stress. A correct grammar would presumably make the probability of misstressing either of these words vanishingly low. Scaling these weights (multiplying them by a constant) would bring the probability of the correct form arbitrarily close to 1.

(3)

	Align-R-*i* 5	Nonfin-*j* 5	Align-R 2	Nonfin 2	*H*	*p*
→ bə(nǽnə)$_i$				–1	–2	0.99
(bǽnə)nə$_i$	–1		–1		–7	0.01
kə(nǽdə)$_j$		–1		–1	–7	0.01
→ (kǽnə)də$_j$			–1		–2	0.99

There is a range of possibilities for how the set of lexically specific constraints is composed. In Pater (2010), for example, it is proposed that lexically specific constraints are posited only to resolve inconsistency. For example, if we had just Align-R and Nonfin in our constraint set, and no other constraints (such as faithfulness) to distinguish *Canada* from *banana*, then adding one of the indexed constraints would be necessary, and sufficient, to generate the correct pronunciations. The proposal in Pater (2010) does not deal with gradient productivity. To do so, we assume that constraints have lexically specific instantiations for every lexical item. Because we have separate constraints for every lexical item, the number of items displaying one or another pattern can affect both the weight of the general constraints, and of their lexically specific versions. When a pattern is relatively general across the lexicon, the general constraint winds up with a relatively high weight. In this case there are relatively few exceptions, and the lexically specific constraints encoding them need to have relatively high weight for the words to be pronounced in a fixed fashion. On the other hand, when a pattern is relatively rare, the general constraint winds up with a relatively low weight.

Exactly how lexically specific constraints are induced is a topic we leave for further research, as it will require considering cases more realistic than the small illustrations we present here (for example, to determine whether unworkably large constraint sets are created under some assumptions). One general question in previous research on lexically indexed constraints is whether markedness constraints, faithfulness constraints, or both, can be indexed (see Pater 2010 for references and discussion). Choices in this regard will not only be important in terms of keeping the size of the constraint set manageable, but it will also effect how exactly learners generalize. An attractive possibility is that the lexically indexed constraints are simply the constraints encoding the phonological features of the word, that is, that they are realizational constraints in the sense of Xu (2007), but a full exploration of the consequences of abandoning underlying representations and faithfulness is of course beyond the scope of this short paper.

As an illustration of our proposal, we consider a set of simple stress 'languages' with varying degrees of regularity. Stress can fall in one of two positions, and there is a general constraint preferring each (e.g. Align-L and Align-R). There are 100 words, and thus 100 lexically specific versions of Align-L, and 100 of Align-R. The languages have stress in one position in 100 to 50 words (fully regular to fully lexical), making 51 languages.

Our learning algorithm is batch Gradient Descent. It is similar to the Stochastic OT / HG Gradual Learning Algorithm (GLA; Boersma 1998; Boersma and Pater 2016), except that each epoch is a presentation of the entire dataset (see Pater and Staubs 2013 and Moreton, Pater, and Pertsova 2015). We used batch Gradient Descent for convenience: a single run closely approximates the averaging of results of multiple runs of the regular on-line GLA (the batch algorithm doesn't have a stochastic component, and thus gives a single outcome for a single starting point). We ran it for 1000 epochs, with a learning rate of 0.1, and starting constraint weights of zero.

The resulting grammars give probability near 1 for the correct pronunciations of the words it was trained on. To generate probabilities for nonce word productions,

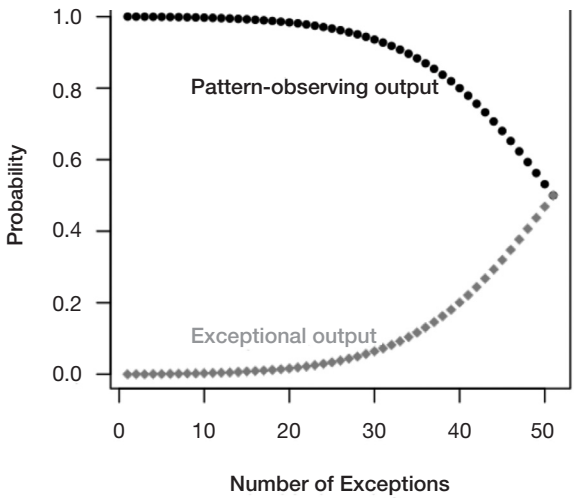

Figure 1. Probabilities assigned by grammars with only the general constraint weights, learned for grammars with 0-50 exceptions.

we simply remove the lexically indexed constraints. The resulting probabilities are shown in Figure 1.

When there are no exceptions, the grammar without the lexically specific constraints assigns stress following the general pattern with probability 1.0, and when the number of forms following each pattern is equal, nonce words are predicted to be assigned to one or the other pattern with equal probability. In between, the grammar assigns a range of probabilities, thus allowing for the modeling of gradient productivity. Interestingly, the curve is not linear: adding an exception to a language with relatively few exceptions is predicted to have less of an effect on productivity than adding an exception to a language that has relatively many. We do not know of any current data that bear in this prediction.

Our claim that gradient productivity exists across the stress patterns of the world's languages was not based on nonce word data. In section 1.1, we introduced two pieces of evidence that the stress pattern of Polish, while admitting exceptions, is more productive than that of English or Spanish: exceptions are more often regularized, and speakers have more difficulty encoding lexical stress in an experimental task. In terms of our model, these differences between Polish on the one hand, and English and Spanish on the other, can be understood as follows. In Polish, the general constraint(s) demanding penultimate stress would have relatively high weight compared with English or Spanish. For a word with antepenultimate stress to be encoded faithfully, the relevant lexically specific constraint would itself need to have relatively high weight. Thus, encoding lexical stress would require more experience for a Polish learner (assuming gradual learning), and absent that experience, regularization or unfaithful encoding would be predicted to be more common.

3. Case study: Dutch voicing alternations

In this section, we turn to a case of gradient productivity within a single language, for which we have existing nonce word data to model. For this detailed case study, we analyze the classic case of Dutch voicing alternations presented by Ernestus and Baayen (2003). In Dutch, word-internal obstruents contrast in voicing, as can be seen in the minimal pair [vɛrʋɛidɛn] ('to widen') *vs*. [vɛrʋɛitɛn] ('to reproach'). However, obstruents devoice word-finally, leading to neutralization: [vɛrʋɛit] is a homophone, meaning either widen or reproach. Final neutralization is exceptionless in Dutch, but Ernestus and Baayen (2003) show that there are gradient phonological generalizations about the likelihood that a word-final voiceless obstruent will surface as voiced under suffixation – that is, in terms of the standard analysis, about whether it is underlyingly voiced.

In the lexicon of Dutch, both the place and manner of the word-final obstruent affects the likelihood that a word will undergo alternation, with the obstruent becoming voiced word-internally. For example, as shown in Table 1, about 70% of words with word-final labial fricatives (f/v) alternate, while only 9% of words with word-final labial stops alternate. Data from Ernestus and Baayen's (2003) search of the CELEX corpus are given under the column headed "% voiced in lexicon".

Also shown in Table 1 are the percentages from Ernestus and Baayen's production experiment. They found that participants could 'guess' whether or not a form should alternate based on its neutralized form, and their guesses followed the statistics of the lexicon. A participant would be presented with a nonce word, say *kuuf*, and would be asked for the past tense form of the word, formed by adding *-te* for final voiceless obstruents and *-de* for final voiced obstruents. A response of *kuufte* indicates that the participant has interpreted the final f of *kuuf* as underlyingly voiceless, while a response of *kuuvde* indicates that the participant has interpreted the f as voiced. As Table 1 indicates, for each place and manner category, participants roughly matched the probability of voicing in the lexicon. Participants exhibited an overall preference against voicing which is not seen in the lexicon (likely because they were given the devoiced form as a prompt), but otherwise, they roughly match the probability of voicing in the lexicon for each type of word-final obstruent.

Table 1. Percent of word-final voiceless consonants that become voiced intervocalically under derivation, in Ernestus and Baayen's 2003 CELEX corpus search and experimental results

	% voiced in lexicon	% voiced in experiment
p/b	9%	4%
t/d	25%	9%
s/z	33%	23%
f/v	70%	49%
x/ɣ	97%	80%

This is a case where participants 'frequency match', closely copying in nonce forms the distribution in the lexicon of their language. Some generalizations reported by Ernestus and Baayen are nearly exceptionless, while others have many exceptions. Our model represents the generalizations with fewer exceptions via higher weights on the general constraints, lower weights on the lexically specific constraints of observers, and very high weights on lexically specific constraints of violators. Generalizations with large numbers of exceptions are represented with lower weights on the general constraints, and similar weights on lexically specific constraints for observers and for violators.

As even the title of Ernestus and Baayen's paper implies ("Predicting the unpredictable"), the standard analysis in which the voicing of the derived form is simply stored as part of the Underlying Representation of the underived form does not predict that speakers should show awareness of the generalizations about which types of consonant alternate. An alternative is to treat the pattern as lexically conditioned intervocalic voicing (see also Becker, Kitrez and Nevins 2011 on Turkish), as in the following adaptation of Ernestus and Baayen's (pg. 20) analysis, which uses a subset of their constraints. We used seven general constraints, including two very general constraints demanding that all intervocalic obstruents be voiced or voiceless. We note from the outset that these constraints may well be different from those one would use in a standard OT analysis, and that they gloss over details of the Dutch system, such as the effects of final vowel length, and additionally do not represent the morphological and prosodic environments relevant to the voicing alternation. We use these constraints for ease of comparison with Ernestus and Baayen's analysis, and because they are sufficient to illustrate our approach.

(4) *The most general constraints*

*VTV Assign a violation mark to an intervocalic voiceless consonant

*VDV Assign a violation mark to an intervocalic voiced consonant

Each lexical item is associated with a lexically specific version of one of these constraints – words with a voiced obstruent, like [vɛrʊɛidɛn], are assigned a copy of *VTV, while words with a voiceless obstruent ([vɛrʊɛitɛn]) are assigned a copy of *VDV. For convenience, we omitted lexically specific constraints that would be violated in the correct output (these would simply receive a weight of zero in the learned grammar).

(5) $*VTV_{WIDEN}$ Assign a violation mark to an intervocalic voiceless consonant when its underlying correspondent is contained in /vɛrʊɛid/ 'widen'

$*VDV_{REPROACH}$ Assign a violation mark to an intervocalic voiced consonant when its underlying correspondent contained in /vɛrʊɛit/ 'reproach'

Additionally, we use the following five place-specific constraints, which are general in the sense that they do not pertain to particular lexical items. They militate against voicing for the obstruent types that tend to be voiceless intervocalically (p/b, t/d and s/z), and militate against voicelessness for the obstruent types that tend to be voiced intervocalically (f/v, x/ɣ).

(6) *Place and manner constraints*

*P[+voice]: Assign a violation mark to an intervocalic voiced labial stop

*T[+voice]: Assign a violation mark to an intervocalic voiced coronal stop

*S[+voice]: Assign a violation mark to an intervocalic voiced sibilant

*F[-voice]: Assign a violation mark to an intervocalic voiceless labial fricative

*X[-voice]: Assign a violation mark to an intervocalic voiceless velar fricative

Note also that our training data consist of the pairs of derivationally related words from Ernestus and Baayen's corpus search, and that they report that the statistics for obstruents at other locations than root-final differ (pg. 7). A fuller simulation may well need to specify the constraints in (5) and (6) to root-final position.

In this simulation, we did not allow all general markedness constraints to be indexed to specific lexical items; rather we only allowed the most general markedness constraints *VTV and *VDV to be lexically indexed. This choice was made primarily to save on processing time (with 1700 lexical items, allowing all seven constraints to be indexed to any relevant lexical item would result in a possible 6800 constraints). As discussed in Section 2, more investigation is needed into whether all constraints should be available for indexation, or if not, on what basis constraints are chosen for indexation.

Like the above stress simulation, this simulation used Batch Gradient Descent, where a single learning epoch constitutes an update over the entire set of training data, which match the corpus data from Ernestus and Baayen (2003). The algorithm was run for 10,000 epochs with a learning rate of 0.01. Constraint weights were regularized using a Gaussian prior with a mean of zero and a variance of 100.[1]

After 10,000 epochs, the predicted percentages of voicing on all real words in the training data was very close to either 100% or 0%, depending on whether that word was voiced or voiceless in the lexicon. Table 2 shows the percentage of voicing predicted by our simulation for each type of input obstruent for 'wug' words. These are generated from the weights of the constraints, without any lexically specific constraints. Again, our assumption is that novel forms do not have associated lexically specific constraints (or if they do, that they have very low weight). The overall pattern matches the place and manner distinctions shown in the lexical data and in the experiment. The trends are generally more extreme in the predictions of our model than in either the lexicon or in the experimental data; the percent voicing is closer to zero for the voiceless trend cases, and closer to 100 for voiced. Patterns

1. The parameter settings can affect the outcome, especially the setting of the regularization term. The need to tune the parameters to match the experimental data is a potential weakness of this approach.

Table 2. Percentage of intervocalically voiced forms, in the training data, in the outcome of our learning simulation, and Ernestus and Baayen's experiment

		Training Data			
		% voiced		% voiced	%voiced
	Trend	Lexicon	no. forms	Experiment (EB)	Simulation
p/b	voiceless	9%	230	4%	**1%**
t/d	voiceless	25%	719	9%	**9%**
s/z	voiceless	33%	451	23%	**18%**
f/v	voiced	70%	166	49%	**84%**
x/ɣ	voiced	97%	131	80%	**99%**

in experimental data of this type tend to show more randomness than in lexica. To adjust for this, Hayes et al. (2009) use a temperature parameter in their MaxEnt model, and a similar approach could be adopted here.

Figure 2 shows the model's behavior on items with a velar fricative, which in the lexicon are voiced nearly exceptionlessly. On the left, the weights of the four types of constraints relevant to such a lexical item are plotted over the course of the learning period. Epochs are on the x axis, and weights on the y axis. The general constraint *VTV has a very low weight throughout the entire learning process (except for a 'burn-in' period over the first approximately 300 epochs). Because it applies to all types of word in the training data in the same way, it does much less work than the place/manner specific constraints. The constraint demanding voicing on velar fricatives *X[-voice] gets a relatively high weight, and in particular much higher than the weight of the lexically-specific *VTV constraints associated with lexical items which follow the trend. The weights of the lexically specific *VDV constraints, on the 3% of lexical items which violate the trend are very high - they must overcome *X[-voice] in evaluation for the exceptional words to be pronounced correctly.

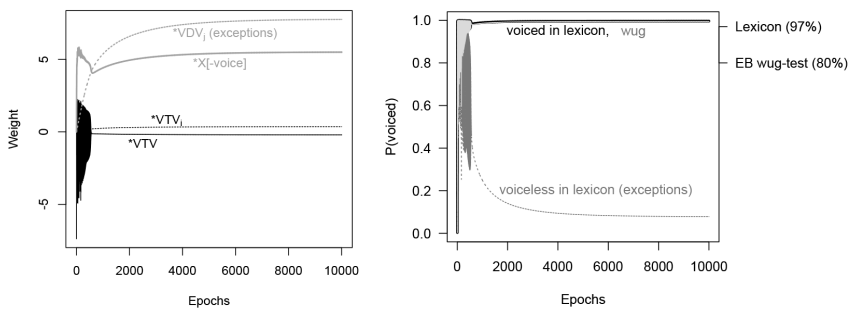

Figure 2. Velar fricatives – relevant constraint weights (left) and predicted probabilities of different word types (right) over the course of the simulation.

The right-hand side of Figure 2 shows the probability of voicing on words that are voiced in the lexicon (black) and words that are voiceless in the lexicon (grey), both of which have associated lexically specific constraints. The black line is also the predicted probability for nonce words – the data from the nonce words exactly matches the probabilities of the words that are voiced in the lexicon. Words that follow the trend toward voicing are learned early and at the end of learning are pronounced correctly 100% of the time. However, the exceptional voiceless words are learned much slower, and even at the end of learning still have some non-zero probability of error. Nonce words are also predicted to voice velar fricatives 100% of the time, in contrast to the 80% voicing on these items observed by Ernestus and Baayen – this is one instance of the general "trend exaggeration" seen in the output of our model discussed with respect to Table 2 above.

Figure 3 shows the weights and probabilities associated with forms with [s/z], which have a lower lexical probability of intervocalic voicing than [x/ɣ], and fall in the middle of the range for different place/manner combinations in Table 2. As for [x/ɣ], the general *VTV constraint gets zero weight. The constraint *S[+voice] gets some weight, but this time both lexically specific constraints of observers (*VDV$_i$ in this case) and of violators (*VTV$_j$) get higher weights than the general markedness constraint. The probability of correct pronunciation of an existing lexical item observing the generalization is the same as the correct pronunciation existing lexical item violating the generalization - close to 1.0. However, the probability of voicing on a nonce word, unspecified for voicing and lacking a lexically specific constraint, is 0.33. This is close to the value observed by Ernestus and Baayen for these forms, 23% voiced.

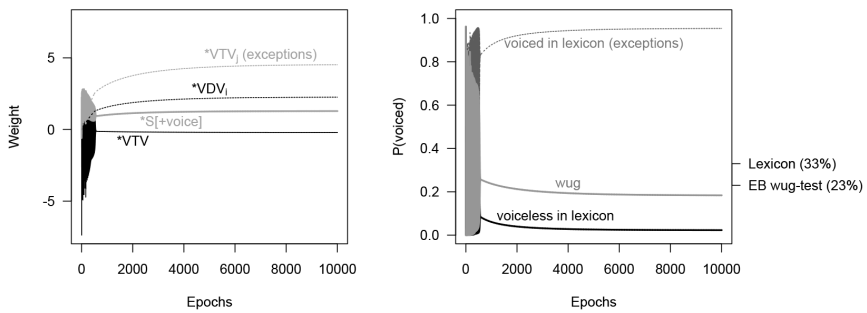

Figure 3. Coronal fricatives – relevant constraint weights (left) and predicted probabilities of different word types (right) over the course of the simulation.

4. Conclusions

Recent work in generative linguistics has begun to address the problem of gradient productivity by adopting probabilistic grammars. Our proposal seeks to address an outstanding problem in the modeling of lexically gradient patterns: that although the

proposed grammar models successfully model the gradience seen in experimental results such as nonce word productions, they often do not deal with the fixed pronunciations seen in individual lexical items. We have shown that by incorporating lexically specific constraints into a Maximum Entropy model, both gradient productivity and fixed pronunciation of individual lexical items can be successfully modeled. Moreover, the influence of the lexicon on the weights of the constraints encoding the general patterns comes from the basic operation of the learning algorithm with the proposed constraint set, rather than through any special calculation over the lexicon. We have illustrated our proposal using a toy case of differences of lexical frequency of stress patterns across languages, as well as an attested case of gradience in nonce word production from Dutch voicing alternations. Much remains to be done in developing this model and in comparing it to alternatives, such as MaxEnt models that operate over sub-lexicons (Allen and Becker 2015, Gouskova and Becker to appear, Moore-Cantwell and Staubs 2014), and various forms of analogical model (see Ernestus and Baayen 2003 and Moore-Cantwell 2015, as well as the Appendix of Moreton 2015 on formal connections between analogical and MaxEnt models). We find the results thus far with our relatively simple model encouraging, and present this as a potentially useful further step in addressing a longstanding problem.

References

Allen, Blake & Becker, Michael. 2015. Learning alternations from surface forms with sublexical phonology. Unpublished manuscript, University of British Columbia and Stony Brook University. Available as lingbuzz/002503

Becker, Michael & Gouskova, Maria. To appear. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry.* Available as lingbuzz/001622.

Becker, Michael, Ketrez, Nihan & Nevins, Andrew. 2011. The Surfeit of the Stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87(1): 84-125. <http://dx.doi.org/10.1353/lan.2011.0016>

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

Boersma, Paul & Pater, Joe. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy & Joe Pater (eds.). *Harmonic Grammar and Harmonic Serialism*, 389-434. London: Equinox Press.

Ernestus, Mirjam & Baayen, Harald. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79: 5-38. <http://dx.doi.org/10.1353/lan.2003.0076>

Fidelholtz, James. 1979. Stress in Polish - with some comparisons to English stress. *Poznań Studies in Contemporary Linguistics* 9: 47-61. Available at <http://wa.amu.edu.pl/psicl/files/9/04_Fidelholtz.pdf>.

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

Hayes, Bruce. 2008. *Introductory Phonology*. Malden, MA: Wiley-Blackwell.

Hayes, Bruce & Wilson, Colin. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379-440. <http://dx.doi.org/10.1162/ling.2008.39.3.379>

Hayes, Bruce, Zuraw, Kie, Siptár, Peter & Londe, Zsuzsa. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85: 822-863. <http://dx.doi.org/10.1353/lan.0.0169>

Kraska-Szlenk, Iwona. 1995. *The phonology of stress in Polish*. Ph.D. dissertation, University of Illinois, Urbana-Champaign.

Moreton, Elliott, Pater, Joe & Pertsova, Katya. 2015. Phonological concept learning. *Cognitive Science.* <http://dx.doi.org/10.1111/cogs.12319>

Moore-Cantwell, Claire. 2015. The phonological grammar is probabilistic: New evidence pitting abstract representation against analogy. Unpublished manuscript, Yale University. Available at: <http://blogs.ubc.ca/amp2015/files/2015/09/Moore-Cantwell.pdf>.

Moore-Cantwell, Claire & Staubs, Robert. 2014. Modeling morphological subgeneralizations. In Kingston, John, Moore-Cantwell, Claire, Pater, Joe & Staubs, Robert (eds.). *Proceedings of the 2013 meeting on phonology*, Linguistic Society of America, Washington DC. <http://dx.doi.org/10.3765/amp.v1i1.42>

Pater, Joe. 1994. Against the underlying specification of an 'exceptional' English stress pattern. *Toronto Working Papers in Linguistics* 13: 95-121. Available at <http://twpl.library.utoronto.ca/index.php/twpl/article/view/6336/3324>.

Pater, Joe. 2000. Non-uniformity in English stress: the role of ranked and lexically specific constraints. *Phonology* 17: 237-274. <http://dx.doi.org/10.1017/S0952675700003900>.

Pater, Joe. 2005. Learning a stratified grammar. In Brugos, Alejna, Clark-Cotton, Manuella R. & Ha, Seungwan (eds.). *Proceedings of the 29th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. 482-492.

Pater, Joe. 2010. Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution. In Steve Parker (ed.). *Phonological Argumentation: Essays on Evidence and Motivation*, 123-154. London: Equinox.

Pater Joe & Staubs, Robert. 2013. Modeling learning trajectories with batch gradient descent. Paper presented October 27th to the Northeast Computational Phonology Circle, MIT. <http://people.umass.edu/pater/pater-staubs-gradient-descent-2013.pdf>.

Peperkamp, Sharon, Vendelin, Inga & Dupoux, Emmanuel. 2010. Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics* 38: 422-430. <http://dx.doi.org/10.1016/j.wocn.2010.04.001>

Xu, Zheng. 2007. *Inflectional morphology in Optimality Theory*. PhD dissertation, Stony Brook University.

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*. PhD dissertation, University of California, Los Angeles.