Representational Strength Theory: Combining lexical idiosyncrasy and probabilistic grammar

Claire Moore-Cantwell

Abstract

This paper presents a novel theory of lexical exceptionality within a probabilistic, constraint-based grammar. In Representational Strength Theory (RST), traditional faithfulness constraints are replaced by Phonological Form Constraints (PFC's), which encode lexicalized properties of a word. A series of simulations demonstrate that the weights of PFC's can be learned alongside the weights of markedness constraints so that probability-matching behavior is predicted on novel words, but real words are accurately represented, whether they are exceptions to the language's grammar or not. Because RST can model the learning process for the lexicon and the grammar at the same time, it makes specific predictions about the grammar-lexicon relationship. Two predictions are explored here: (1) that high-frequency exceptions should be more stable than low-frequency ones, and (2) that features which are entirely predictable from the grammar should not be stored on individual words.

1 Introduction

This paper presents Representational Strength Theory, a theory of exceptionality within a constraint-based grammatical framework. As work in phonology continues with probabilistic grammars like Maximum Entropy (MaxEnt) grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008), more and more progress is made in understanding the complex, detailed, and probabilistic knowledge that speakers have about their language. In particular, systems that were once thought of as an absolute, categorical pattern with some

memorized exceptions, are now reinterpreted as probabilistic patterns where individual words of the language are sometimes high-probability, and sometimes low probability. This reinterpretation allows us a new look into the relationship between the grammar and the lexicon. How, and why do speakers know probabilistic generalizations over words whose details all need to be memorized anyway? The lexicon of a language is traditionally assumed to be as simplified as possible, leaving out details which the grammar can predict. Does this assumption hold up when the grammar is not always deterministic? Finally, we do not learn all the words of our language perfectly. Low frequency words in particular are subject to misremembering and regularization. How does imperfect lexical knowledge interact with the grammar during language use and learning?

Representational Strength Theory (RST) provides a foothold into these questions by making knowledge of the various properties of lexical items explicitly learnable and subject to various amounts of pressure from the grammar. RST begins with the MaxEnt grammar framework, and encodes the listed properties of lexical items as weighted constraints. Higher-weighted properties are generally more important, contrastive, and unpredictable, while lower-weighted properties or properties not encoded at all are generally more predictable and less contrastive. However, overspecification and underspecification are possible in this theory, and frequently arise as errors or at early stages of learning. Representing lexical items in this way allows lexical entries to be learned alongside the phonological grammar, and allows us to quantify how strongly a certain property is associated with a lexical item, and how well a lexical item has been learned in general.

This model can be used to generate empirical predictions about the interaction between the lexicon and the phonological grammar. For example, high-frequency words in some circumstances resist phonological change, and in other cases lead it (Bybee, 1985; Phillips, 2006; Bybee, 2007). The degree to which two sounds contrast in a language depends on both the grammar of the language, and how frequent the two sounds are in the language's lexicon (Hall, 2013; Hall and Hall, 2016; Scobbie and Stuart-Smith, 2008). More frequent words of a language tend to exhibit more extreme behavior in patterns with token variation (Morgan and Levy, 2016; Smith and Moore-Cantwell, 2017). Certain types of grammatical patterns only exhibit token variation, while others can exhibit type variation, deeply involving the lexicon (Zuraw, 2016).

These phenomena, and more, require an explicit theory of lexicon-phonology interaction during learning in order to be modeled, and ultimately understood. This paper will focus on just two empirical predictions - the stability of high-frequency exceptions, and the relationship between contrast and grammar - but RST can in principle serve as a jumping-off point for modeling many phenomena of this type.

This paper presents the general framework of RST, including a schema for representing lexical properties as constraints, called Phonological Form Constraints (PFC's). Because RST does not use underlying forms in the traditional sense, but instead encodes a lexical item's behavior as a series of constraints with weights, there are no Faithfulness constraints in this theory, and the grammar can be understood simply as the relative weighting of the markedness constraints in the model. Although this is a major departure from current phonological theory, PFC's and markedness constraints are sufficient to describe basic phonological processes: epenthesis, deletion, and feature change. A learning algorithm for RST is presented, containing two crucial elements: (1) a decay mechanism for PFC's, which ensures lexical representations are as simplified as possible, and (2) a method for determining, on a language specific basis, which features should be encoded on novel words. The model begins agnostic and later bootstraps, using words it already knows to determine how new words should be represented. By the end of learning, only features which are contrastive in the language are represented on new words.

Finally, a series of simulations will be presented, using stress patterns as a test case to illustrate the learning process and its outcomes. These simulations illustrate the following predictions of RST: (1) Phonological grammar is probabilistic, learned to match the distribution in the lexicon of a language as closely as possible. (2) Exceptional lexical items are more stable when they are high frequency, and exhibit a tendency to regularize when they are low-frequency. (3) Features that are not contrastive in a language are difficult for speakers to encode on novel words.

1.1 Exceptions in probabilistic phonology

Many phonological and morphological patterns in languages come with exceptions. These can be solitary exceptions like the past tense of *go* in English (*went*), or groups of exceptions such as the exceptional pasts

led, read, fed,.... Groups like this can pattern together in phonological shape, linguistic origin, or part of speech, but crucially it is impossible to predict from these things alone whether a word will be an exception or not. One way to represent this lexical idiosyncrasy is through a "rule-and-exceptions" approach, in which the general rule (here "add -d for the past tense") is used unless an irregular form is specifically listed in the lexicon. Such an approach predicts that only the rule, and not the exceptions, should affect speakers' behavior on novel words.

Even in the case of the English past tense, the "rules-and-exceptions" approach seems to oversimplify the situation. Exceptions are not entirely random, but pattern together in clusters, and exhibit productivity based on their phonological characteristics (Albright and Hayes, 2003). For a clearer example of this, consider the case of Polish stress (Fidelholtz, 1979). Polish in general has penultimate main stress, illustrated by alternations such as *język* ∼ *języka* ∼ *językámi*. However, there are a small number of highly patterned exceptions: nearly all are antepenultimate[1]. Antepenultimate exceptions are mainly borrowings (*matemátyka, státua, káliko*), but a few are native (*ógulem, szczéguly, okólika*). Not only are nearly all exceptions antepenultimate (not final, pre-antepenultimate, or any other stress), but all have a light penultimate syllable. Borrowings with a heavy penultimate syllable are instead regularized to the majority penultimate pattern. Thus, exceptions are far from random, and seem to follow their own grammar. Fidelholtz analyzes this system using a main stress rule which assigns stress to the penult, and a minor rule which assigns stress to the antepenult only when the penult is light.

Many cases of 'patterned exceptions' like this exist. In fact, they may be the typical case, with random arbitrary exceptions being somewhat unusual. A body of more recent phonological research, starting at least with Zuraw (2000), tests the productivity of patterns with exceptions like the Polish case. Overwhelmingly, researchers find that not only are the 'major rules' productive, but so are the 'minor rules'. Although nonwords would have no way to be lexically marked as exceptions, speakers seem to arbitrarily and probabilistically choose some nonwords to undergo minor rules, in proportion to how common that 'minor rule' pattern is in the lexicon of the language. Hayes et al. (2009) call this the 'Law of Frequency Matching':

[1] A few are final, and all of these are interjections: *patatáj, galóp, akurát, korékt, (h)ohó, ahá, ojéj*

(1) Law of Frequency Matching: Speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies.

This 'Law' turns out not to be precise. Often speakers match some patterns in their lexicon well, but over-represent, under-represent, or completely ignore others. However, some degree of frequency matching has been obtained across many pattern types and across many languages. Examples include Zuraw (2000); Eddington (2004); Ernestus and Baayen (2003); Albright and Hayes (2003); Zuraw (2007); Moore-Cantwell (2012); Normann-Vigil (2012); Linzen et al. (2013); Colavin (2013); Jun (2015); Zhang and Liu (2016); Bayles et al. (2016); Kim (2016); Becker et al. (2017); Garcia (2017); Kim (2017); Kumagai and Kawahara (2018); Smith and Pater (2020).

In light of this discovery that speakers often behave probabilistically on wug-tests, rather than consistently choosing a default pattern, phonologists have proposed the adoption of probabilistic models of phonological competence, such as Variable Rules (Labov, 1969), Partially Ordered Constraints (Anttila, 1997), Stochastic OT (Boersma and Hayes, 2001), Noisy Harmonic Grammar (Coetzee, 2009), and Maximum Entropy Grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008). All of these models, rather than choosing a single output form for a given input form, predict a probability distribution over two or more possible outputs. A speakers' choice of an output (at least on novel words) is a sample from that probability distribution. When the predicted probability distribution matches the distribution of forms in the lexicon, these models predict lexical frequency matching behavior.

If the Polish stress case were a case of lexical frequency matching, then speakers would choose penultimate stress most of the time on a wug-test, but they would choose antepenultimate stress (on light-penult words) a small percentage of the time. In the Polish lexicon, the percentage of exceptional stresses turns out to be very small - less than 1% of words (Peperkamp et al., 2010). So, if participants did lexical-frequency match on this pattern it might be hard to tell in a noisy experimental setting. Let us consider a more typical case: Dutch voicing alternations (Ernestus and Baayen, 2003).

In Dutch, voicing is contrastive, but word-final obstruents devoice. This is illustrated in (2).

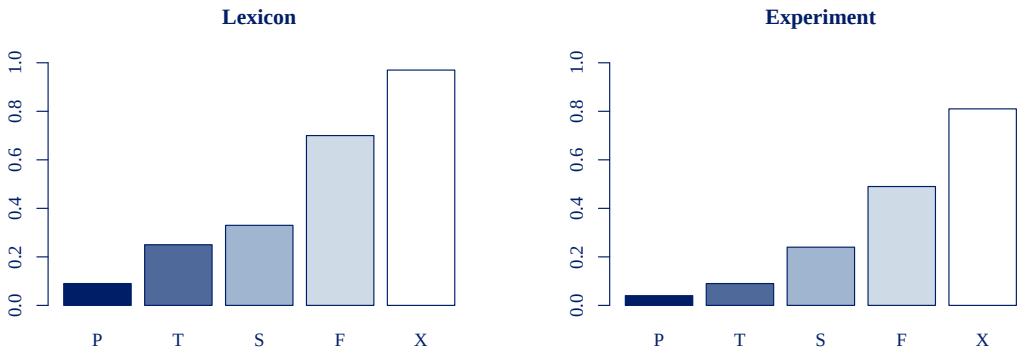
(2) a. verwij**d**en [vɛrʊɛi**d**ən] 'widen-INF'
verwij**t**en [vɛrʊɛi**t**ən] 'reproach-INF'

b. verwij**d** [vɛrʊɛi**t**] 'widen'
verwij**t** [vɛrʊɛi**t**] 'reproach'

However, not all voicing specifications are equally likely. Ernestus and Baayen find that in the lexicon of Dutch, certain word-final obstruents are very likely to be voiced (labials), while others are very unlikely (velar fricatives). Others are in the middle, like alveolar stops. They present participants in an experiment with novel verbs in the suffixless form, as in (2b.) above, and ask then to produce a suffixed form, thus choosing a voicing specification for the final obstruent of the novel word.

The rates of voicing they found using this method are shown in Figure 1 below, with the lexicon statistics. Participants do not choose the voiceless form every time, though this would make sense considering they are given a voiceless form to start with and voiceless obstruents are in all cases allowed in the language. They also do not choose the most common voicing specification in the lexicon (voiceless), or the most common voicing specification for each obstruent. Rather, they behave probabilistically, seeming to draw their responses from a distribution very similar to the distribution found in the lexicon of Dutch.

Figure 1: Data from Ernestus and Baayen (2003), Figure 1. Participants behave probabilistically on novel words, roughly matching the statistics of the lexicon.



Ernestus and Baayen consider a variety of possible models of this phenomenon, including analogical models, statistical models, and constraint-based grammar models. Though they do not come to a strong conclusion about which type of model best fits their data, subsequent work has often settled on Maximum Entropy Grammar (MaxEnt) models for similar data. Applied to the Dutch data, this would look something like Table 1.

Table 1: Maximum Entropy tableau, using constraints from Ernestus and Baayen (2003). Inputs are wug-words, and predicted probabilities match those produced by participants in the experiment.

	p	$\mathcal{H}$	*P[+voice]	*S[+voice]	*X[-voice]
dap			3.26	2.35	1.42
→ dap**ə**n	**0.96**	-3.26			
→ dab**ə**n	**0.04**	0	1		
bers					
→ bers**ə**n	**0.91**	-2.35			
→ ber**z**ən	**0.09**	0		1	
kix					
→ ki**x**ən	**0.19**	-1.42			1
→ ki**ɣ**ən	**0.81**	0			

Table 1 uses weighted constraints to predict a probability distribution over possible output forms. For each candidate, constraint weights are multiplied by the number of the candidates violations, and summed to get the Harmony ($\mathcal{H}$), which is then converted to a probability distribution using the logit function. In Table 1 weights are chosen specifically so that they will predict the probabilities observed in Ernestus and Baayen's experiment, though fitting weights in MaxEnt models in general is a complex problem, and precise matches like this are not always attainable.

A seldom-discussed consequence of using probabilistic grammatical models like MaxEnt is that they force a re-evalutation of how lexical entries are encoded. Individual words are not 'exceptions' or 'observers' of a particular pattern, but rather their behavior is more or less probable according to the grammar. Traditionally, in a "rules-and exceptions" type of framework, exceptions are special in that their lexical entries contain specifications for certain features which do not have to be stored on rule-following forms. This understanding of exceptions stems from common phonological practice that lexical entries "should contain only idiosyncratic properties of items, properties not predicted by the general rule." (Chomsky and Halle, 1968, page 12) When

the grammar is probabilistic, then the relevant features must be specified on both observers and violators of a pattern.

Table 2 illustrates this for Dutch. A "rules-and-exceptions" framework would refrain from generally listing voicing specification in the lexicon. Here, Pattern-following forms like *zoog* ([zoɣ-ən]), would not have any voicing specification listed in the lexicon for their final segments. This is illustrated in the tableau using the symbol /X/ to stand in for a velar fricative unspecified for voicing in the lexicon. Voicing would only be listed on exceptions like *poch* ([pux-ən]).

Table 2: Predictions of an account of Dutch word-final voicing with lexical simplification. Inputs are real words of Dutch. X in /zoX/ indicates an underlying segment with no voicing specification.

	p	$\mathcal{H}$	*X[-voice]	Ident-Voice
/pux/			1.42	10
→pu**x**ən	**0.999**	-1.42	1	
pu**ɣ**ən	**0.0**	-10		1
/zoX/				
→zo**x**ən	**0.19**	-1.42	1	
→zo**ɣ**ən	**0.81**	0		

Table 2 contains a markedness constraint demanding that velar fricatives be voiced, as well as a faithfulness constraint: Ident-Voice. For /pux/, the 'exception', this faithfulness constraint forces the output to match the lexical entry in voicing, but for /zoX/, the pattern-observing form, there is no underlying voicing specification to match, so neither candidate violates Ident-Voice. Instead, the grammar (in this case the sole constraint *X[-voice]) determines the outcome. The flaw here is immediately visible: the correct candidate zoɣən is the most probable, but it is not at 100%. The grammar predicts that zoxən should also surface around 20% of the time. In order for this word to surface correctly, its voicing specification must be stored in the lexicon, just like the exception. No lexical simplification is possible in this system, even though there are relatively strong grammatical preferences.

In some patterns of variation each individual word varies. A classic example is English word-final t/d deletion (Guy, 1980; Coetzee, 2009, and others). Phonological factors condition the deletion of t/d, but real words vary. 'West' is pronounced sometimes with the final t ([wɛst]), and sometimes without the t ([wɛs]). Deletion can also occur 'halfway', resulting in outputs like [wɛstˀ], [wɛsʔ], or [wɛsˀ]. Patterns like t/d deletion

do not have the problem illustrated in Table 2, since variability in outputs for real words of the language is a correct prediction. Variation like Dutch voicing alternations, where each word behaves consistently has been called 'type variation', 'lexical variation' or, as in Zuraw (2016), 'polarized variation'. Zuraw argues that polarized variation is very common, and provides simulations of diachronic change showing that it should be the norm when variation is between two distinct phonological categories, while within-word variation like English t/d deletion should be more common when outcomes fall along a phonetic continuum.

In the study of probabilistic grammar, polarized variation like the Dutch case should not be treated as a special case, but rather central to the enterprise. We need some mechanism to ensure that real words of a language are produced correctly when virtually all of them diverge from the grammar's predictions.

1.2 Models of exceptionality

Previous models of lexical storage and exceptionality within a probabilistic framework fall into roughly two categories: *Lexical Listing*, and *Multiple Grammars*. The most straightforward example of a *Lexical Listing* approach is simply having the relevant feature specified in the UR of every form. In the Dutch case above, the voicing of every word-final obstruent would be stored, not just the exceptional ones. More consequentially, morphologically complex forms would be listed whole in the lexicon, even when they are not exceptions (Zuraw, 2000, 2010). Other types of lexical listing are more complicated: Abstract UR's of various types can indirectly determine a word's behavior in different morphological contexts. Abstract UR's could include phonemes that never surface (Hayes, 2008, chapter 12), floating features or tones (Bermúdez-Otero, 2012; Tebay and Zimmermann, 2020; Trommer, 2020) or additional structure such as segment activity (Smolensky and Goldrick, 2016; Zimmermann, 2019).

While the lexical listing approach often produces satisfying analyses to individual problems, it is not a general theory of exceptions in probabilistic grammar, largely because it does not account for Frequency Matching behavior. If all obstruents in Dutch must be underlyingly specified for voicing, there is nothing in particular to be gained for the learner by acquiring the probabilistic grammar. A high-weighted faithfulness constraint like in Table 2 will be both necessary and sufficient to achieve a perfect match with the lexicon.

The markedness constraints, like *X[-voice] are unnecessary.

Multiple Grammars models come in two main varieties: Constraint Indexation, and Cophonologies. Constraint Indexation models (Pater, 2000, 2010; Becker, 2009; Becker et al., 2011) use 'cloned' copies of markedness or faithfulness constraints, which are then indexed to specific lexical items, only assigning a violation when that lexical item appears in the input. Cophonologies approaches (Anttila, 2002; Inkelas et al., 1997; Sande et al., 2020) work similarly, but lexical items are indexed to particular constraint *rankings*, instead of to cloned copies of constraints. An illustration of how constraint indexation might work in OT is given in Table 3.

Table 3: Illustration of the indexed-constraints approach, using Dutch word-final voicing alternations

/pux/$_j$	IDENT-VOICE$_j$	*X[-voice]	IDENT-VOICE
→ puxən		1	
puɣən	1!		1
/zox/			
zoxən		1!	
→ zoɣən			

In general, constraint indexation or cophonologies by themselves cannot account for Frequency Matching behavior. Becker et al. (2011) propose an extra mechanism whereby novel words are assigned to a certain indexed version of a constraint probabilistically, based on how many existing words are indexed to that constraint. This mechanism does predict Frequency Matching, but adds an extra step to the phonological computation process. Moore-Cantwell and Pater (2016) incorporate indexed constraints into MaxEnt Grammar, which does predict Frequency Matching, however in order to predict the correct outcome for real words, this model requires massive redundancy in the cloned constraints themselves.

A model which is slightly different from constraint indexation and cophonologies, but bears mentioning here, is Scaled MaxEnt. Versions of this model have been proposed in Linzen et al. (2013); Shih (2018); Coetzee and Kawahara (2013); Coetzee and Pater (2009), and the model is developed in detail in Zymet (2018, 2019). Scaled MaxEnt uses lexically specific scaling factors, which enable lexical items to add or subtract from the weights of general constraints. Thus, Frequency Matching is predicted by the 'default'

weights of the constraints, and the behavior of individual lexical items is a function of the default weights and a lexical item's scaling factors. Zymet (2019); Shih (2018) demonstrate that scaling factors can be learned as random effects in a heirarchical mixed-effects regression model, where the main effects are general constraint weights. If scaling factors are fit this way, then Frequency Matching obtains, and individual lexical items' behavior can be fit as well.

2 Representational Strength Theory

2.1 Phonological Form Constraints

The crucial mechanism of Representational Strength Theory (RST) is the use of Phonological Form Constraints (PFC's). PFC's constitute the phonological portion of a word's lexical entry, and encode speakers' knowledge of the sounds of a word, their arrangement, and sometimes even how they vary in different contexts. Each PFC has a weight, which can be any positive real number. They compete with weighted markedness constraints to determine the correct output for a given form. These constraints do the job of both UR's and faithfulness constraints in typical constraint-based models, and as a result RST has neither UR's nor Faithfulness. A PFC's definition encodes a feature's value and position within a lexical item - the job of a UR, and its weight encodes how important it is that that feature surface correctly - the job of a faithfulness constraint. Rather than using Faithfulness constraints to determine which features of a lexical item are the most important to preserve, RST allows the lexical entry *itself* to compete with markedness constraints.

2.1.1 Overview

Table 4 illustrates the activity of PFC's in the case of Dutch word-final obstruent voicing. The tableau contains three inputs, the real words *poch* and *zoog*, and a novel word, labelled Wug. Each of the real words come with a PFC. These are Poch-$\mathbb{S}_3$-[-voice] and Zoog-$\mathbb{S}_3$-[+voice], which assign violations as follows:

(3) a. Poch-$\mathbb{S}_3$-[-voice] Assign a violation to any candidate output for input Poch which does not have

a value of [-voice] on its third segment.

b. Zoog-$\mathbb{S}_3$-[+voice] Assign a violation to any candidate output for input Zoog which does not have a value of [+voice] on its third segment.

These constraints are irrelevant for candidates which are not derived from the input noted in the constraint. To indicate this, cells of the tableau have been greyed out where each constraint does not apply. Note that there is no corresponding constraint for the novel word, since it does not have a lexical entry.

Table 4: Tableau illustrating the interaction of PFC's and markedness in two Dutch words, and one nonword

	p	$\mathcal{H}$	*X[-voice] 1.42	Poch-$\mathbb{S}_3$-[-voi] 6.5	Zoog-$\mathbb{S}_3$-[+voi] 3.4
Poch+Inf					
→ puxən	.99	-1.42	1		
puɣən	.01	-6.5		1	
Zoog+Inf					
zoxən	.01	-4.82	1		1
→ zoɣən	.99	0			
Wug+Inf					
→ kixən	.19	-1.42	1		
→ kiɣən	.81	0			

The constraint *X[-voice], as in previous tableaux, is a general markedness constraint penalizing all candidates with a voiceless velar frictive, regardless of their input. Its weight of 1.42 (same as in Table 2) predicts the correct distribution of voicing and voicelessness on wug-words, illustrated by Wug + Inf. The weight of Poch-$\mathbb{S}_3$-[-voice] is relatively high, while the weight of Zoog-$\mathbb{S}_3$-[+voice] is relatively low. This is because *poch* is an exception to the overall generalization that velar fricatives are voiced, while *zoog* obeys this generalization. The weight of Zoog-$\mathbb{S}_3$-[+voice] cannot be zero, however, or the tableau would predict that speakers devoice the fricative in *zoog* 19% of the time, in line with the wug-word. For features of a word that are completely predictable, meaning that the grammar predicts the correct outcome 100% of the time, no PFC would be needed at all.

The PFC's given in Table 4 illustrate the basic concept: PFC's both define a feature of the lexical item, and demand 'faithfulness' to it. Because these constraints replace the UR, which would encode the position

Table 5: Word-final devoicing in Dutch: though the PFC for *zoog* demands voicing, the markedness constraint overcomes that preference

	p	$\mathcal{H}$	$*\left[\begin{smallmatrix}+\text{voice}\\-\text{sonorant}\end{smallmatrix}\right]\#$ 10.6	*X[-voice] 1.42	Poch-$\mathbb{S}_3$-[-voi] 6.5	Zoog-$\mathbb{S}_3$-[+voi] 3.4
Poch						
→ pu**x**	.99	-1.42		1		
pu**ɣ**	.01	-17.1	1		1	
Zoog						
→ zo**x**	.99	-4.82		1		1
zo**ɣ**	.01	-10.6	1			
Wug						
→ ki**x**	.99	-1.42		1		
ki**ɣ**	.01	-10.6	1			

of the feature, the PFC must also include a statement of the feature's position. How exactly positions should be defined is a complex and unresolved issue of this theory, but detailed discussion is included in Section 2.1.4.

Like faithfulness constraints, PFC's can be out-weighted by markedness constraints, resulting in a surface form which does not adhere to all demands of PFC's. This would be analogous to a surface form which does not match the UR.

In Table 5, the constraint $*\left[\begin{smallmatrix}+\text{voice}\\-\text{sonorant}\end{smallmatrix}\right]\#$ assigns a violation to all word-final voiced obstruents. Its weight is high enough to overcome the preferences of both *X[-voice] and the PFC Zoog-$\mathbb{S}_3$-[+voi].

2.1.2 Conflicting PFC's

Unlike in standard constraint-based models, in RST a single lexical entry may have multiple competing representations. This aspect of the theory is similar to approaches using UR-constraints (Pater et al., 2012; Boersma, 2001), in which a single lexical entry may have multiple distinct UR's, and candidate outputs differ in both their surface form and which UR they were derived from. UR constraints demand that a specific UR be used, and are violated when it is not. In RST, lexical entries may simply have competing PFC's, which make opposing demands on the surface form of a particular lexeme. Such an analysis would be appropriate in cases where a lexical item behaves differently in different environments, but in ways that are not completely predictable from the grammar. Situations like this include phonologically conditioned allomorphy (Carstairs, 1988; Kager, 2008), and lexically specific variation (Hayes et al., 2009; Pater and Smith, 2011; Zuraw, 2016), where words exhibit idiosyncratic rates of undergoing a certain

variable process.

To illustrate how this works in RST, consider the case of stress clash in English: two stresses in a row (a 'clash') are tolerated in a wide range of environments, including within words, but sometimes stress clash across words is repaired by moving a stress leftwards (the 'Rhythm Rule', Prince, 1983). This rule shifts primary stress onto a secondary stress earlier in the word, so that *thìrtéen* becomes *thírtèen mén*, and *àchromátic* becomes *áchromàtic léns*. Not all words undergo stress shift, however. *sùperstítious* does not become **súperstìtious mán*, and *sìncére* does not become *síncère létter*[2]

In RST, words like *thirteen* and *achromatic* would have two competing stress specifications. The first, with main stress further right (*thìrtéen, àchromátic*) would be represented using PFC's with a higher weight, while the second, with main stress further left (*thírtèen*, *áchromàtic*) would be represented using PFC's with a slightly lower weight. In isolation, the versions with the stress further right will surface, while in clash contexts the *Clash constraint can gang with the alternative, lower-weighted PFC's to make the versions with main stress farther to the left preferable. This is illustrated here for the word *thirteen*. The word has two competing PFC's, one demanding main stress on the first syllable, and the other demanding main stress on the second syllable. PFC's determining the segmental content of the word are not shown.

(4) a. Thirteen - σ_2 - stress: (10) Assign a violation to every candidate output for the lexical entry Thirteen which does not have main stress on the **second** syllable

b. Thirteen - σ_1 - stress: (6) Assign a violation to every candidate output for the lexical entry Thirteen which does not have main stress on the **first** syllable

Table 6 demonstrates what happens to the word *thirteen* in different contexts. In isolation, the word surfaces correctly with final main stress. Even though there are competing PFC's for this word's stress pattern, the higher weight of Thirteen-σ_2-stress compared to Thirteen-σ_1-stress means that the correct pronounciation surfaces 98% of the time. On the other hand, when the word is in a clash context, the *Clash constraint conspires with the lower-weighted Thirteen-σ_1-stress to predict a high probability on the stress-shifted *thírtèen mén*. The weights were chosen here to predict a probability of about 85% for stress shift, since Grabe and Warren (1995) found that the Rhythm Rule applies with about this frequency in a production experiment.

[2]Bolinger (1981) suggests that which words do and do not undergo the Rhythm Rule may vary from person to person. These examples reflect the author's own judgments.

Table 6: Tableau illustrating competing PFC's for *thirteen*

Thirteen		p	$\mathcal{H}$	*Clash	Thirteen-σ_2-stress	Thirteen-σ_1-stress	Sincere-σ_2-stress
				6	10	6	12
a. →	thìrtéen	**0.98**	-6			1	
b.	thírtèen	0.02	-10		1		
Thirteen Men							
c. →	thìrtéen mén	**0.12**	-12	1		1	
d. →	thírtèen mén	**0.88**	-10		1		
Sincere							
e. →	sìncére	**1**	0				
f.	síncère	0	-12				1
Sincere Men							
g. →	sìncére mén	**0.99**	-6	1			
h.	síncère mén	0.01	-12				1

The *Clash constraint must be sufficiently low-weighted that it cannot overcome the single stress PFC for the word *sincere*. Thus, when *thirteen* exhibits stress shift this is a gang effect between the relatively low-weighted *Clash and *thirteen*'s lower-weighted PFC. In this example, the idiosyncrasy of *thirteen* vs. *sincere* is encoded in their PFC's, not as propensity to undergo or resist stress shift per se, but rather as ambiguity, or conflict, about the lexical entry of the word itself. The grammar can then take advantage of this conflict to create a more unmarked surface structure.

2.1.3 PFC's as a substitute for Faithfulness

Thus far we have seen examples of PFC's demanding a specific feature value, and sometimes being overruled by a markedness constraint. In these examples, PFC's essentially do the work of Ident constraints. However, since PFC's are meant to replace all faithfulness constraints, they must not only do the work of the Ident family of constraints, but of Max and Dep as well. To this end, a brief illustration of deletion and epenthesis is in order. How PFC's could militate against deletion is reasonably clear. A PFC demands that a certain feature surface in a certain position - so all that is needed is to stipulate that if the position itself does not exist, then the PFC is automatically violated.

The case of epenthesis is somewhat more difficult. Since each PFC demands that one specific feature should surface, there is no unified representation of the entire word that encodes information about what does not occur. A candidate with an epenthetic segment will not violate any PFC's, and since there are no Dep constraints, the only

Table 7: Tableau illustrating how PFC's prevent epenthesis or deletion

		p	$\mathcal{H}$	NoCoda	*Coronal	*Nasal	*Sonorant	*Structure	Bean-$\mathbb{S}_3$-[+nasal]
Bee				2	2	2	2	1	15
a. →	bi	**0.99**	-2					2	
b.	bin	0.00	-11	1	1	1	1	3	
Bean									
c.	bi	0.00	-17					2	1
d. →	bin	**0.99**	-11	1	1	1	1	3	

thing preventing rampant epenthesis would be markedness constraints. Within Optimality Theory, *which* segments get epenthesized is governed by low-ranked markedness constraints (Lombardi, 2002). For example, a glottal stop might get epenthesized instead of another consonant because most consonants violate some markedness constraints like *Velar, *Coronal, *Nasal, etc. These same general markedness constraints will prevent most epenthesis in RST. All that is required is that they be weighted low enough to be easily overcome by PFC's when necessary, and weighted high enough to prevent unmotivated epenthesis. These general markedness constraints will be sufficient so long as every possible epenthetic segment violates some markedness constraint. If the analyst prefers to view certain segments as totally unmarked, such as glottal stop or schwa, then the inclusion of a generalized *Structure constraint would be necessary (Prince and Smolensky, 1993/2004, pg 25). *Structure assigns a violation for every output segment. In RST, it would effectively prevent any segment from surfacing if it did not have a PFC or markedness constraint supporting it.

The prevention of both epenthesis and deletion is demonstrated in Table 7, using the minimal pair *bee*/*bean*, which differ only in the presence or absence of a segment - the n. Note that while only one PFC is shown, demanding that there be a [+nasal] feature in the third segment for Bean, in a full analysis there would also be PFC's demanding that this segment be coronal, that it be sonorant, that it be voiced, etc. No PFC's are shown for Bee because it does not have any that adjudicate between the two candidates shown. Its PFC's would demand that the vowel be high, that the onset be voiced, and so on.

In Table 7, epenthesis of a spurious n is prevented in the word Bee, because a series of markedness constraints disprefer it, and no PFC demands that the n be present. Meanwhile, in Bean, the n does surface even though it incurs violations of several markedness constraints. Deletion of the n is prevented by the high-weighted PFC Bean-*[+nasal]-cod-*σ_1.

2.1.4 Defining Positions

PFC's must refer to the position within the word of the feature they are specifying. This section will sketch out some potential different approaches to representing position and linear order with PFC's. A system of precedence relations is ultimately proposed here, but this is an area of RST that requires further exploration.

In the examples above, linear order and position were determined by PFC's by simply specifying in each PFC a segment at a specific linear location in the output candidate. I will call this the "simple-linear-order approach". Directly specifying the order of segment that a PFC applies to has one obvious problem: epenthesis and deletion would completely disrupt the activity of all PFC's. Consider vowel-initial words of English, such as Apple: A candidate surface form with a glottal stop at the beginning, [ʔæpl], would incur violations of *all* PFC's. Since the first segment is a glottal stop, PFC's demanding that $\mathbb{S}_1$ be a vowel, low, and front would be violated. Since the second segment is now æ, PFC's demanding that $\mathbb{S}_2$ be labial, voiceless, and a stop, would be violated. Finally, since the third segment is a p, the PFC's demanding that $\mathbb{S}_3$ be l would be violated, and the l, now the fourth segment of the word, would essentially be treated as epenthetic. Similar issues would arise in candidates with deletion.

A compelling alternative approach would be to tie position definitions to prosodic structure. Rather than specifying first-segment, second-segment, third-segment, etc., this approach would specify a segment's prosodic role, beginning with its role in the syllable. The PFC Bean-$\mathbb{S}_3$-[+nasal], from Table 7 would be replaced by a PFC demanding a [+nasal] feature in coda position: Bean-Coda-[+nasal]. This strategy, which I will call the "Prosodic-structure-as-ordering approach" is somewhat intuitive, and relates cleanly to the "filler-and-role" structure used by Gradient Symbolic models (Smolensky, 1990; Smolensky et al., 2014), which like RST allow elements of a lexical entry to vary in their activity. It has the additional benefit that for the most part segment order would never need to be directly specified, but rather would fall out from each segment's prosodic role: onsets precede nucleii, codas follow. This could in principle carry all the way up the prosodic heirarchy. For example, syllables are often parsed into feet, in which one syllable is strong. In trochaic feet the strong syllable is always first, and the weak syllable is second. In iambic feet, the opposite is always true. Then, if a segment is epenthesized or deleted, the PFC's referring to segments around it would still apply correctly. Adding the glottal stop to the beginning of [ʔæpl], for example, would not affect the PFC's making demands about the content of the two nucleii, and about the onset of the second syllable.

In this approach, there would of course be at least a little bit of 'leftover' ordering that still must be specified,

Table 8: Contrastive syllabification in the prosodic-structure-as-ordering approach

		p	$\mathcal{H}$	Onset	NoCoda	Tree-Ons-σ_2-[+coronal]	Sky-Cod-σ_1-[+coronal]
Tree				5	2	10	15
a. →	pa.ta	**0.99**	0				
b.	pat.a	0.00	-17	1	1	1	
Sky							
c.	pa.ta	0.00	-15				1
d. →	pat.a	**0.99**	-7	1	1		

such as complex onsets, codas, or nucleii, unparsed syllables, and elements of the same prominence level at higher levels of structure. The main downfall of the prosodic-structure-as-ordering approach, however, is that it requires that syllable structure be specified in the lexical entry. This in turn predicts (1) that resyllabification should be very difficult, and (2) that syllable structure should in general be contrastive the way order of segments generally is. Both of these predictions run counter to what is actually observed in the world's languages: in general resyllabification occurs liberally across word boundaries, or to accommodate epenthesis or deletion of segments. On the other hand, syllabification has not been observed to be contrastive in any language.

Table 8 illustrates a hypothetical language in which the word for 'tree' is pronounced [pa.ta] with the t in onset position, but the word for 'sky' is pronounced [pat.a] with the t in coda position. Although markedness constraints Onset and NoCoda militate against the [pat.a] pronounciation, the PFC for Sky is able to overcome this pressure. Not only is this language representable with PFC's like those in the tableaux, which specify a segment's position as its role in the syllable, but languages with resyllabification across word or morpheme boundaries would actually be quite difficult to represent. Resyllabifying a coda as an onset, for example, would be the same as deleting a segment and then epenthesizing a new segment identical to the one deleted, but with a different syllabic role.

A good system for representing positions must be able to accommodate epenthesis and deletion, and be flexible with syllable structure. A few other desiderata are in order, summarized below in (5). Patterns in which a segment is partially preserved from input to output, such as in cases of vowel coalescence or nasal substitution (np → m), must be representable. In order for a coalescence candidate to win, preservation of a segment's features even while the segment's root node is deleted must still satisfy some PFC. As will be demonstrated below, this is accomplished by separating PFC's demanding the *existence* of a feature from PFC's demanding a specific location for that feature. Metathesis also must be representable, in the sense that it must be treated as a single operation. That is, it must

violate one or just a few PFC's - those pertaining to segment order - and not violate 'spurious' PFC's. It should not violate the same set of PFC's as would deletion plus epenthesis of one of the segments. Finally, PFC's must be able to represent features that are not necessarily affiliated with a single segment, like stress. Under most analyses, stress is a property of syllables, so ideally a PFC would be able to demand stress on a particular syllable, or demand that certain syllables be parsed into a foot.

These goals for a system of representing positions are summarized below:

(5) Goals:

a. Epenthesis and deletion of segments must not disrupt the behavior of other segments in the word

b. Resyllabification must be easy

c. Metathesis must be treated as one, or relatively few, PFC violations, *not* as deletion plus epenthesis

d. Features must be able to move from one segment to another while still satisfying some PFC's

e. Phenomena, like stress, whose host is not a segment but a higher-level structure must be representable without having to refer to a specific segment.

The "precedence-structure-of-segments approach" accomplishes these goals using precedence structure rather than linear order of segments. Instead of defining position using prosodic roles, or by defining a strict linear order, this approach uses abstract 'segment' objects, and defines precedence relationships between them. PFC's establish the existence of a segment object (e.g. $\mathbb{S}_i$) either by demanding a feature assignment to it (b. below), or by implicating it in a precedence relationship (c. below). Finally, features would generally come with 'existence' PFC's, as in a. below, which demand that a particular feature specification be included in a word's output without defining a location for that feature specification.

(6) Types of PFC's in the "precedence-structure-of-segments approach":

a. LEXEME- $\exists$ [αfeature]: Assign a violation to every candidate output for LEXEME which does not contain a specification of [αfeature].

b. LEXEME-[αfeature]-$\mathbb{S}_i$: Assign a violation to every candidate output for LEXEME in which $\mathbb{S}_i$ does not have feature value [αfeature].

c. LEXEME-$\mathbb{S}_i \prec \mathbb{S}_k$: Assign a violation to every candidate output for LEXEME in which $\mathbb{S}_i$ does not precede

$\mathbb{S}_k$

This system would generally accomplish the goals in (5): Epenthesis and deletion would not disrupt precedence structure of a lexical item. Metathesis would consist of two $\mathbb{S}$'s appearing in incorrect precedence relations to each other, but correct precedence relation to the rest of the string. Only the precedence PFC's would be violated - PFC's like in (6) b. would remain unviolated. The inclusion of PFC's like (6) a. which demand the existence of a feature without linking it to a particular $\mathbb{S}$, would allow for phenomena like coalescence, in which a segment is deleted but not all of its features are. Coalescence would violate PFC's like (6) b. but not (6) c. Finally, not every PFC is required to refer to $\mathbb{S}$ objects. Stress PFC's, for example, could still refer to syllables. Because all PFC's refer to output structure, such stress PFC's would refer to output syllables.

2.2 Richness of the Base and phonological contrast

Removing general faithfulness constraints from a constraint-based system, as RST does, has consequences beyond the representation of particular phonological processes. In this section, I discuss how both Richness of the Base and the representation of phonological contrast in the grammar must be re-imagined within RST. Instead of being the consequence of constraint rankings between markedness and faithfulness, these properties are instead a consequence of the learning process in RST, which will be briefly outlined in this section.

Richness of the Base is the phonological maxim that a grammar should always produce an output that is allowed in the language, regardless of input (Prince and Smolensky, 1993/2004). Richness of the Base effectively ensures that grammatical models correctly describe productive generalizations that speakers have in their heads, and do not behave strangely on novel or unusual inputs. This is generally achieved in a constraint-based analysis via relative ranking or weighting of markedness and faithfulness. If a constraint like $*\left[\begin{smallmatrix}+\text{voice}\\-\text{sonorant}\end{smallmatrix}\right]\#$, from Table 5, is ranked (or sufficiently highly weighted) over IDENT-[voice], then all inputs will surface in a way that obeys the grammar. It will be impossible to construct an input that will surface with a voiced final obstruent. In RST, the grammar does not contain faithfulness constraints, and PFC's, which do the job of faithfulness, are part of the input. There is no configuration of the grammar, then, that will always produce word-final devoicing no matter what the input.

Table 9 illustrates the problem, using the invented input ZOOG'. This input resists word-final devoicing because it has a very high-weighted PFC. Since there is in principle no upper bound on PFC weights, possible inputs can be

Table 9: An exceptional input, Zoog', fails to undergo word-final devoicing.

	p	$\mathcal{H}$	*[+voice −sonorant]# 10.6	*X[-voice] 1.42	Zoog'-$\mathbb{S}_3$-[+voi] 100
Zoog'					
zo**x**	0.0	-101.42		1	1
→ zo**ɣ**	1	-10.6	1		
Wug					
→ ki**x**	.99	-1.42		1	
ki**ɣ**	.01	-10.6	1		

constructed with arbitrarily high PFC weights to overcome any grammar. The learning algorithm proposed below offers instead a practical upper-limit on PFC weights. PFC's are induced when needed, at a particular starting weight, and they decay over time. Because of this decay, most PFC's will not achieve a weight higher than the starting weight specified in the model. Only lexical items with which the learner has a great deal of experience, and evidence that they should have a high-weighted PFC, will achieve a high enough weight to overcome strong grammatical pressures. These will generally be only high frequency exceptions.

In constraint-based models, contrast is typically encoded as a particular relationship between markedness and faithfulness: if there is a voicing contrast in the language, Ident-[voice] will outrank *[voice], for example. In RST, each word will have its own weight on the PFC or PFC's that determine voicing on its consonants. Therefore, there is no unified representation of contrast. The learning algorithm presented below places contrast in the learning algorithm instead. The induction of PFC's is mitigated by a process which bootstraps off of existing words of the language. PFC's specifying a certain feature (say, voicing) are more likely to be induced if many words of the language already specify that feature, and are unlikely to be induced if very few or no words already specify that feature. Because of the decay process for PFC's, features that are completely predictable by the grammar tend not to be specified on many words, which then makes them more difficult to induce on new words. So, in practice non-contrastive features are not stored on new words while contrastive features are.

3 Learning RST with the Gradient Lexicon and phonology Learner

The Gradient Lexicon and Phonology Learner (GLaPL) has as its goals (1) to learn appropriate weights for all markedness constraints, correctly predicting speakers' choices on nonce-words, and (2) to induce appropriate PFC's across the lexicon and learn weights on them that correctly predict the behavior of each individual word.

Figure 2: Illustration of the Gradual learning Algorithm, adapted from Boersma and Hayes (2001), pages 52-53. α=0.01

Learning step *i*

	C1	C2	C3	C4	C5	C6	C7
/underlying form/	13	11	9.6	4	2.12	1	0.2
a. *observed*	1→	1→	1			1→	
b. *predicted*			1	←1			←1

Constraint weights at step *i+1*:

C1	C2	C3	C4	C5	C6	C7
12.99	10.99	9.6	4.01	2.12	0.99	0.21

3.1 Structure of the algorithm

GLaPL takes as its starting point the Perceptron update rule (Rosenblatt, 1958), used in the gradual learning algorithm (GLA) (Boersma and Hayes, 2001). This is an error-driven learner, meaning that it learns only when it makes an error, or a mismatch between its own predictions and reality. In a single learning instance, a surface form and an underlying form are presented to the learner as a pair. Perceptron then predicts its own surface form from the underlying form using its current constraint weights. If the predicted surface form matches the correct (observed) surface form, then there is no error, and the constraint weights stay the same. However, if the predicted and observed surface forms do not match, then the constraint weights are updated by a small amount. Constraints which prefer the incorrect predicted form are demoted a little, and constraints which prefer the correct observed form are promoted a little. This is illustrated below in Figure 2. The amount by which constraints are updated is a parameter of the model, typically called α.

In Figure 2, constraint C1 is violated by the observed form but not by the predicted form, therefore it prefers the incorrect predicted form, and is demoted a little. Constraints C2 and C6 are also demoted a little for the same reason. C4 and C7 prefer the correct observed form, and are therefore promoted a little. Constraints C3 and C5 are not updated because neither distinguishes between the observed and the predicted form. C3 is violated by both, while C5 is violated by neither. The value for α is set to 0.01 in Figure 2, so that is how much each constraint changes by, whether its weight is moved up or down. This number is a fairly typical value for α, but because it is such a small number, constraints weight change only very slowly. The constraint weights shown in step *i+1* of Figure 2 will make essentially the same predictions as the constraint weights from step *i*. This learning process is repeated a large number of times - perhaps thousands or millions - on different input-output pairs. The learner will typically see a

single input-output pair many times, perhaps in proportion to its lexical frequency. There is no distinct end state to the learning process, so learning can continue arbitrarily long. However, constraint weights tend to stabilize at some point during learning. When this happens depends on the size of the learning data and the complexity of the grammar being learned.

Boersma and Hayes (2001) developed the GLA to learn Stochastic OT, a model in which constraints have real-number 'ranking values' which get translated into a strict ranking at evaluation, but the algorithm can also be used to learn straight Harmonic Grammar, Noisy Harmonic Grammar, and, relevant for the present paper, MaxEnt Grammar. When using the GLA to learn MaxEnt Grammar, one wrinke arises, which is that MaxEnt never predicts a single winner, but rather a probability distribution over candidates. However, for simple updating as described here, a single predicted form is required. This can be achieved simply by sampling a single 'predicted' form from the predicted probability distribution.

GLaPL uses the same basic update rule as the GLA, but it must update both markedness constraints and PFC's. Its update also includes an induction mechanism: when an appropriate PFC does not exist, it must be added to a word's lexical entry. Finally, GLaPL includes a decay mechanism for PFC's. Their weight gradually decreases to zero over time if they are not used. This decay is independent of the error-driven update process, and proceeds even when no errors are made.

For the simulations reported below, a different α is used for markedness constraints than for PFC's: 0.01 for markedness but 0.1 for the PFC's. The need for these different α values can be understood theoretically in the following way: updates of markedness constraints should be relatively conservative, since with the perceptron update rule, *every* form can potentially cause an update. In a probabilistic system, where different forms could push constraint weights in different directions, a high α could cause markedness constraint weights to be unstable, bouncing back and forth and never converging. On the other hand, PFC weights only ever pertain to one lexical item, and often don't even compete with each other. A large α in this case is advantageous because it allows the learner to learn an individual lexical item quickly, but poses relatively little danger of instability.

3.2 Inducing and decaying PFC's

In GLaPL, PFC's are induced only when motivated by an error, and they decay throughout the learning process. Decay is determined by a constant decay rate, by which all PFC weights are decremented at each learning iteration.

If a PFC's weight decays to zero, that PFC is removed from the model entirely. Decay is the only aspect of the learner that is not tied to errors, so it proceeds even when no errors are made. Because of this, it enforces sparsity in PFC's, so that unnecessary PFC's decay to zero and are never re-induced.

PFC's are induced by comparing the observed candidate to the predicted candidate. Where they differ, a PFC can be induced for each feature that differs between the observed and predicted candidates, demanding the value which is present in the predicted candidate. In the simulations presented below, this process is somewhat simplified in that only one PFC is ever induced per error. In a more complex learning problem, the difference between the observed and predicted candidate could potentially be large, requiring the induction of several PFC's in a single learning step.

Which potential PFC's ought to be induced between any observed-predicted pair is language-specific, and is guided by the phonemic contrasts of the language. Suppose for a certain word, the current stage of learning predicts [tak], while the observed form is [t:g]. In a language where vowel length is phonemic and obstruent voicing is phonemic, two PFC's should be induced for this form. However, suppose a language requires that vowels be long before voiced obstruents and short elsewhere. In this case, the learner should only induce a PFC for the voicing of the obstruent, and not for the length of the vowel. The vowel length would presumably be predictable based on the markedness constraints present in the system.

The proposed mechanism for determining which PFC's to induce is as follows. This process is designed to bootstrap knowledge about what features are contrastive (that is, what features need to be stored) from earlier stages of learning. The learner proceeds as follows:

1. At the beginning of learning, induce every possible PFC. (That is, make no commitments about what features are contrastive)
2. Once the learner's lexicon is sufficiently large, begin *sampling*:
    - For each potential PFC, referring to feature F, which distinguishes OBSERVED from PREDICTED:
    - Randomly select n entries from the existing lexicon
    - For each entry, check whether any of its PFC's refer to feature F
    - If any do, then induce the PFC for the current learning datum. If none do, then do not induce a PFC.

At the beginning of learning, the learner remains agnostic about what kinds of PFC's to induce, but later it can

essentially guess whether inducing a new PFC is a good idea based on what it has already experienced. As learning proceeds, this guess will get better. Early in learning, most features will be available to the learner to create new PFC's from, but as the weights of the markedness constraints develop, some PFC's - the ones specifying properties of lexical items which the grammar also determines - will become unnecessary and decay away to zero.

The sample size, n, of lexical entries used to decide whether to induce a new PFC is a parameter of the learner that can be set to different values. When it is low, the learner is extremely conservative about inducing PFC's and will often fail to induce them when they are needed. This could still result in eventual correct learning, but it would take several exposures to a given word to get that word right. When n is large, the learner is more liberal in inducing PFC's, and will get individual words right faster, but will take longer to converge on the simplest lexicon with only necessary PFC's.

Another parameter of the model is the starting weight of the PFC's themselves. When PFC's are induced, they must have some non-zero weight in order to (a) actually help in getting the word right, and (b) not immediately decay away. The starting weight and the decay rate together determine the minimum time a PFC will stick around in the model. If the starting weight is 10, and the decay rate is 1, then all PFC's will decay away to zero in just ten iterations. This would be the same if the starting weight were 1 and the decay rate 0.1. On the other hand, if the starting weight is 10 but the decay rate is 10^{-10}, then it will take a hundred trillion iterations for a PFC to decay to zero with no interruptions.

One way to think about an RST model is in terms of how much words' behavior is encoded in the lexicon, vs. how much is encoded in the grammar (the weight of markedness constraints). A longer PFC decay time will lead to model fits with more burden on the lexicon, while a shorter decay time will lead to more burden on the grammar. In the limit, if PFC's have an extremely long decay time, and high starting weight, the grammar will not be learned at all for a very long time, or perhaps it may never be learned if the α value for PFC's is also high. On the other hand, if PFC's decay to zero very quickly, then the weights of markedness constraints will be learned quickly and well, but individual lexical items will struggle to be produced correctly by the grammar.

4 Stress pattern simulations

In this section, a series of simulations will be presented illustrating three major predictions of RST. First, like most MaxEnt learners, RST with the Gradient Lexicon and Phonology Learner predicts frequency matching as described

in (1). For novel inputs, the grammar will predict probabilities over output form types which match the frequencies of those types in the training data. In the simulations presented here, this match will be almost perfect. However, if a pattern exists in the data which the model's markedness constraints cannot adequately capture, the match will be correspondingly worse.

Second, exceptional lexical items will be more stable when they are higher frequency, and will tend to exhibit regularization at lower frequencies. That is to say, their PFC's will maintain a relatively high weight when the form is higher frequency, and may fail to maintain any PFC at all for the exceptionful feature at lower frequencies. A low-frequency exception will struggle to maintain its exceptional status during learning.

Finally, these simulations will illustrate how some features emerge as contrastive and some as non-contrastive within a given language. The PFC induction process is subject to the bootstrapping process described in section 3.2 above, so that when no existing words of a language have PFC's specifying a certain feature, then no PFC will be induced for that feature on a novel word either. A feature that is specified on many words of the language, however, will be easily carried over to new words as well. The bootstrapping process also predicts intermediate degrees of contrast (a feature is *sometimes* specified on a novel word), which interestingly does appear to have psychological reality. Peperkamp et al. (2010) illustrate this in a series of experiments, and their data will also provide a convenient foundation for the first two points described above as well.

Peperkamp et al. (2010) examine the phenomenon of 'stress deafness' in several languages. Stress deafness is observed in speakers of languages with completely predictable stress, and refers to those speakers' general inability to learn the stress patterns of new words, at least before they have had training in a language with contrastive stress. The term 'stress deafness' is a bit misleading, however, since the difficulty seems to lie not in acutally perceiving the difference in stress between two items (say, númi and numí, examples from the paper), but rather in remembering which stress pattern occurred when. Participants are exposed to sequences of stresses, say *númi numí númi númi*, and are asked to reproduce the order in which they heard the different stresses - for this example a correct answer would be something like 1, 2, 1, 1. Speakers of French, Hungarian, and Finnish, all with completely predictable stress, have difficulty with this task, while speakers of English and Spanish, languages with contrastive stress, do much better. Interestingly, they also test speakers of Polish, whose stress pattern is overwhelmingly penultimate stress, but which has occasional exceptions (as discussed in the Introduction above). Polish speakers exhibit an intermediate level of 'stress deafness', performing better at the task than French, Finnish, or Hungarian speakers, but worse than English

Table 10: Toy data used in the simulation

	Regular penultimate stress (French/Finnish/Hungarian)			5 exceptions (Polish)			100 exceptions (Spanish/English)		
	σ	$\sigma\sigma$	$\sigma\sigma\sigma$	σ	$\sigma\sigma$	$\sigma\sigma\sigma$	σ	$\sigma\sigma$	$\sigma\sigma\sigma$
Final	1000	0	0	1000	0	0	1000	0	0
Penultimate	N/A	1000	1000	N/A	1000	995	N/A	1000	900
Antepenultimate	N/A	N/A	**0**	N/A	N/A	**5**	N/A	N/A	**100**

or Spanish speakers. One way to interpret their results is to say the stress is noncontrastive in French, Finnish, and Hungarian, totally contrastive in English and Spanish, and partially contrastive in Polish. Polish speakers will sometimes succeed in encoding stress on a novel word (in the parlance of RST, they will sometimes induce a PFC for stress), but not always.

The following simulation uses toy stress data meant to simulate these three classes of language. In every dataset, penultimate stress is the default pattern, but the datasets vary in how many exceptions to that default they exhibit. Toy data is used rather than actual language data, so that we can see how the behavior of the model is affected by number of exceptions alone, holding the actual stress pattern of the language constant, as well as the distribution of word sizes. Each toy language consists of 3000 words, with 1000 each monosyllables, disyllables, and trisyllables. Exceptions are always trisyllables with Antepenultimate stress, similar to what is found in Polish. Including just one kind of exception is a simplification as well, since languages often exhibit multiple types of stress exceptions, which can even be processed differently (Domahs et al., 2008, 2013). The toy data is shown in Table 10.

The words of these toy languages were distributed in frequency according to a zipfian distribution. Each toy word was randomly assigned a frequency, but 'exceptions' - trisyllables with Antepenultimate stress - were spread across the frequency spectrum so that there were always high-frequency exceptions and low-frequency exceptions. The distribution of frequencies for the three datasets is shown in Figure 3.

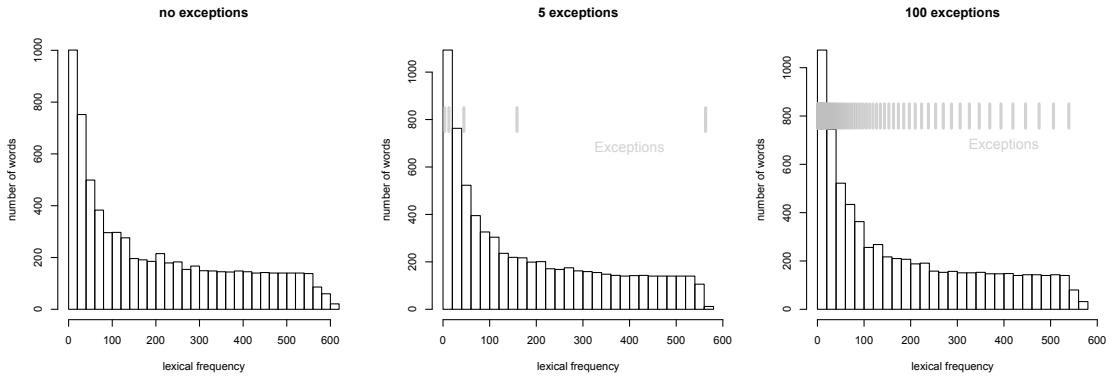
The markedness constraints used in the simulation are typical stress constraints, defined in (7).

(7) Markedness Constraints for penultimate stress

a. Align-Ft-R: Assign a violation for every syllable intervening between the right edge of the word and a foot.

b. Align-Ft-L: Assign a violation for every syllable intervening between the left edge of the word and a foot.

Figure 3: Histograms illustrating the frequency distributions in the toy data. Grey bars indicate the lexical frequencies of exceptional forms.



c. NonFinality-Foot: Assign a violation if the final syllable is parsed into a foot.

d. Iamb: Assign a violation to every foot not stressed on the rightmost syllable.

e. Trochee: Assign a violation to every foot not stressed on the leftmost syllable.

f. HaveStress: Assign a violation to any candidate with no stressed syllable.

A language with exceptionless penultimate stress will have a high weight on Align-Ft-R and Trochee, creating right-aligned trochees, as in *pa(táka)*. Since monosyllables are too small to host either a trochee or an iamb, the constraint HaveStress forces them to be stressed anyway. Since all the training data has stress on monosyllables, we can expect this constraint to always have a high weight. Exceptions are antepenultimately stressed trisyllables, such as *(páta)ka* which observe Align-Ft-L, NonFinality-Foot, and still observe Trochee. When there are no antepenultmate exceptions, we expect Align-Ft-L and NonFinality-Foot to have low weight, but when there are many exceptions, those two constraints should have a small amount of weight - though still not as much as the constraints defining the majority pattern. Since no forms in any of the training data observe Iamb, we expect it it always have a very low weight.

All simulations were run for a total of 1.5 million iterations, with the following parameter settings:

(8) Parameter settings for penultimate stress simulations

a. Learning rate (α): 0.01

b. PFC learning rate (β): 0.1

c. Sample size of words for determing whether to induce a new PFC (n): 10

d. PFC starting weight: 10

e. Decay rate: 0.0001

Finally, each language was learned with these same parameter settings 100 times, in order to assess the stability of the results.

4.1 Results

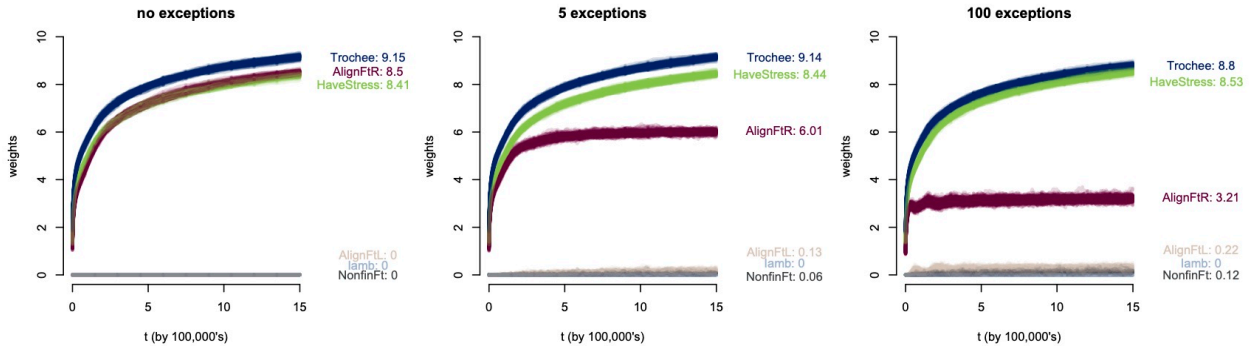
As expected, the model produced probability matching to the data it was given. Table 11 shows the predicted probabilites for different stress patterns for each length of word, for representative runs of the model. As can be seen, the weights of the markedness constraints (illustrated in Figure 4) predict virtually 100% penultimate stress when 100% of the data has penultimate stress. In the '5 exceptions' language, where 0.5% of the trisyllables take antepenultimate stress, the markedness constraints predict antepenultimate stress 0.4% of the time - very close of 0.5%. When 10% of trisyllables take antepenultimate stress, this is again exactly what the markedness constraints predict. This result is relatively unsurprising - nearly every learning algorithm for MaxEnt predicts this matching when appropriate constraints are included. However, it illustrates that the inclusion of PFC's in the model does not affect the probability matching behavior. The model does not, for example, fail to learn any weight on the antepenultimate-preferring constraints Align-Ft-L and NonFinality-Ft just because it is given a mechanism (PFC's) to override the grammar's preferences for specific words.

Table 11: Predicted probabilites across stress patterns for each toy language

	no exceptions	5 exceptions (0.5%)	100 exceptions (10%)
σ́	.999	.999	.999
σ	0.0	0.0	0.0
σ́σ	.999	.999	.999
σσ́	0.0	0.0	0.0
σ́σσ	**0.0**	**0.004**	**0.09**
σσ́σ	**0.999**	**0.996**	**0.91**
σσσ́	0.0	0.0	0.0

Learning was remarkably stable across different runs. Figure 4 illustrates the progression of the weights of each markedness constraint across learning iterations, with the 100 runs of each language graphed on top of each other.

Figure 4: The development of weights of markedness constraints across learning iterations. 100 runs of each language are plotted on top of each other, and average final weights across all runs are labelled at the end of each plot.



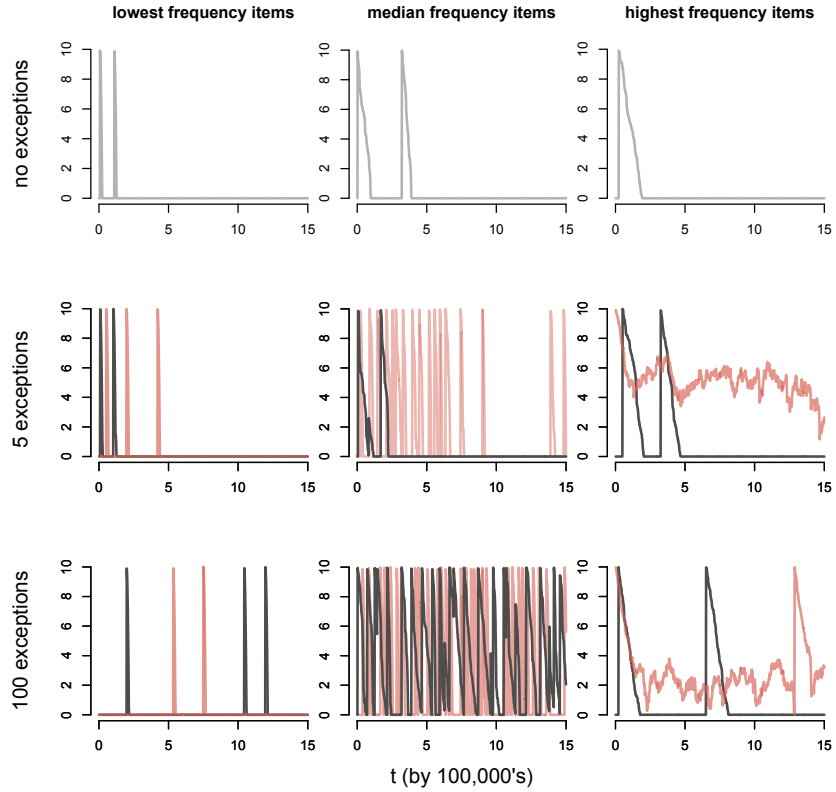
Each run is a thin line, and Figure 4 shows all runs plotted on top of each other, so the variation from run to run can be viewed. Average weights at the end of learning are plotted at the end of each curve.

For all three languages, TROCHEE and HaveStress acquired a high weight, which stabilized at around 250,000 learning iterations, and after that continued to slowly rise throughout learning. Since no observed outputs violate these constraints, there is no upper bound on their weights. The constraint which varies from language to language the most is ALIGN-FT-R, which gains an average weight of 8.5 for the language with no exceptions, 6.01 for the language with 5 exceptions, and 3.21 for the language with 100 exceptions. This is the primary constraint which demands penultimate stress rather than antepenultimate. Its weight is mirrored by that of ALIGN-FT-L and NONFINALITY-FOOT, both of which prefer antepenultimately stressed outputs, and which both get a tiny amount of weight in the 5 exceptions case, and a little bit more weight in the 100 exceptions case. ALIGN-FT-R rises swiftly at the beginning of learning, and then remains incredibly stable throughout the rest of learning. The majority pattern of penultimate stress keeps the weight of this constraint as high as possible, but the algorithm's continued exposure to a small number of exceptional antepenultimate stressed forms creates an upper limit on the constraint's weight. The resulting weight is finely tuned to produce a probability distribution over output forms which mirrors the distribution in the input - thus the model exhibits probability matching behavior.

To understand how PFC's behave in the three different types of languages, consider Figure 5. These graphs illustrate the behavior of individual PFC's in one representative run of the learner for each of the three toy languages. For each language, PFC's for penultimately stressed words, in grey, are compared to PFC's for antepenultimately stressed exceptions, in red. Note that the 'no exceptions' language has no antepenultimately stressed words, so only

PFC's for penultimately stressed words are shown in that case. The highest PFC weight shown here is 10, the starting weight for PFC's in these model runs. Each PFC's weight starts at 10 when it is first induced, and drops down almost immediately. In some cases, the weight drops all the way to zero very quickly, resulting in a thin spike on the graph. In other cases the decay is more shallow, resulting in a fatter peak. Most PFC's shown here are induced multiple times throughout learning, indicated by multiple sharp peaks on the graph. The PFC in the upper left panel, for example, is induced, decays to zero almost immediately, and then is induced and decays to zero a second time, all before 250,000 learning iterations.

Figure 5: Individual PFC's from representative runs of the three toy languages, illustrating the difference between high frequency and low frequency forms. Red lines are exceptional antepenultimate items while grey lines are grammar-observing penultimate items. y axis is PFC weight.



The graph also divides words up by frequency, with PFC's for the lowest frequency items on the right, median-frequency items in the middle, and highest-frequency items on the right. Frequency affects the behavior of both exceptional and non-exceptional items. This is most clearly visible in the top row of the figure, for the language

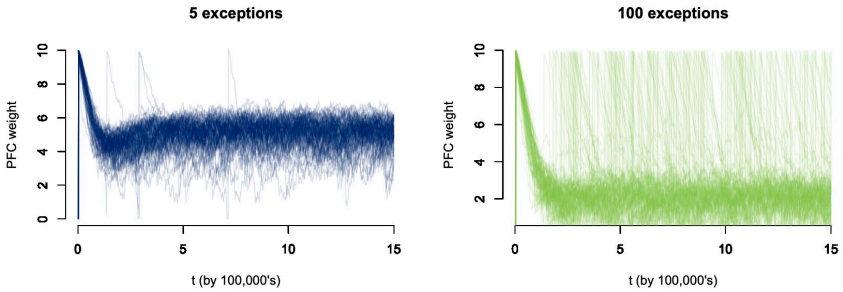
with no exceptions. The PFC shown for the lowest-frequency lexical item is induced and then immediately decays to zero, twice. The PFC for the median frequency item is also induced on two separate occasions, but decays to zero somewhat more slowly, resulting in a shallower slope on the graph from 10 down to zero after the constraint is induced. The PFC for the highest frequency item, in the top right corner of Figure 5, decays to zero just like the others but does so much more slowly, over the course of around 200,000 iterations - rather than 80,000-100,000 for the median frequency word, and less than 20,000 for the low frequency word. This same frequency effect can be seen in the grey lines indicating PFC weights for penultimately stressed words, in all three languages. Weights for low-frequency words decay very quickly, and weights for high-frequency words decay much more slowly.

The second and third rows of the figure show a pattern: PFC's for low- and mid-frequency items are unstable. In the lowest-frequency items, neither the stress PFC for the exceptional (in red) nor the pattern-observing form (in grey) stick around very long during learning. This means that essentially, the lowest frequency forms are predicted to behave exactly according to the grammar's predictions - if exceptional, they are regularized. The median-frequency forms illustrate an interesting difference between the 5-exceptions case and the 100-exceptions case. In the 5-exceptions case, the exception struggles to be correctly represented, its PFC being induced and forgotten over and over throughout learning. However, the median frequency pattern-observer exhibits a PFC for stress only early in learning. Once the grammar is converged on, at about 600,000 iterations according to the progression of markedness weights in Figure 4, that grammar predicts the correct outcome for this pattern-observing form 99.5% of the time. This high accuracy means that errors are only very rarely made on this form, which means the PFC is not re-induced. In the 100-exceptions case however, both the exception and the pattern-observer have PFC's reinduced repeatedly throughout learning. The correct grammar for this language will predict the wrong output for the pattern-observing form a full 10% of the time, leading to a fair number of errors, and therefore reinduction of the PFC throughout learning.

The highest frequency forms for the 5-exceptions and the 100-exceptions case exhibit relative stability in the PFC's of high-frequency exceptions, though high-frequency pattern-observers still have their stress patterns forgotten and re-learned throughout. On the whole, this figure illustrates the general principle among these simulations that the high-frequency exceptions exhibit stability while low-frequency exceptions regularize.

The point that high-frequency exceptions exhibit relative stability is further illustrated in Figure 6. In this figure, the PFC weight for the highest-frequency exception in each of the 100 runs is illustrated for both the 5-exceptions case and the 100-exceptions case. In both there is a great deal more variation across runs than what was seen for the

Figure 6: PFC weights on the highest-frequency exception over 100 runs for the 5-exceptions language and the 100-exceptions language.

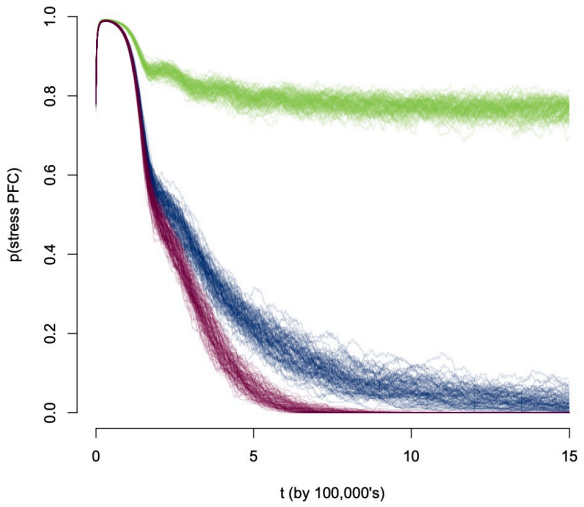


constraint weights, but in both cases, the PFC for high-frequency exceptions persists throughout learning, and only on the occasional run does it decay away and have to be re-induced. Interestingly, the PFC weight in the 100 exceptions case seems to stabilize at a lower weight, and in fact be generally less stable than the one in the 5-exceptions case, possibly indicating the degree of exceptionality: When the grammar already predicts a 10% chance on the exceptional outcome, then a lower weight is necessary on the PFC to add to that to predict near 100% correct behavior. When the grammar only predicts the exceptional outcome 0.5% of the time, then a higher weight is necessary on the PFC.

Finally, let us examine the probability for each of these languages that a PFC determining stress will be induced on a novel word. Recall that this probability is the probability that the model will find, in its random sample of n words, a word already in the lexicon of the language which has a PFC referring to stress. If none are found, it will not induce a new stress PFC, but if at least one is found it will. Figure 7 graphs this probability for each run of the toy datasets: No exceptions at the bottom, in magenta, 5 exceptions in the middle in blue, and 100 exceptions at the top, in green. For the no-exceptions case, this probability falls to zero in every run by about halfway through learning. In this language, stress PFC's are induced at the beginning of learning, but as soon as the grammar is learned, they decay to zero and are never re-induced. PFC's are only induced when the model makes an error, and this model never makes stress errors once it has learned the stress pattern - because that pattern is completely predictable.

The 100-exceptions case, in green, illustrates what happens when there is a large number of exceptions. Even though the grammar of this language stabilizes about halfway through learning (as can be seen in the progression of the constraint weights in Figure 4), the number of stress PFC's remains high, and therefore the probability that a new stress PFC will be induced remains high - though not at 100%. It is useful to remember for this language that only trisyllables have exceptions - monosyllables and disyllables still don't need stress to be specified on them.

Figure 7: Probability of inducing a new stress PFC on error, plotted for each run of the toy datasets. The top bundle, in green, is the 100-exceptions case, the middle bundle, in blue is the 5-exceptions case, and the bottom bundle, in magenta is the no exceptions case.

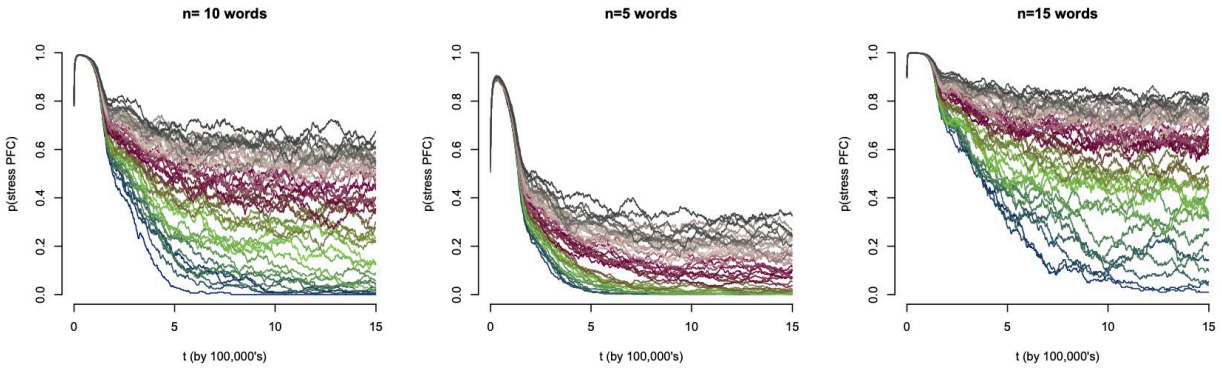


A language with a higher number of exceptions, or with exceptions more spread out across word types, would likely exhibit a 100% or near 100% rate of new stress PFC induction.

Finally, the 5-exceptions case, in blue, mirrors what Peperkamp et al. (2010) found for speakers of a language with very few exceptions: There is an intermediate ability to induce new stress PFC's. Interestingly, the predicted rate of new stress PFC's is variable from run to run, sometimes stabilizing at zero, and sometimes stabilizing just above zero, up to 10%. If these runs accurately model Polish speakers' learning process, we might expect some speakers to exhibit French/Finnish/Hungarian-like stress deafness, while others to exhibit a greater ability to encode stress, though never as great as English/Spanish speakers. Where all of these lines fall depends acutely on the setting of the model parameter which determines how many words are examined for existing PFC's before making a determination for a new word. In the runs presented here, 10 words were sampled. However, Figure 8 illustrates the effect of different settings of this parameter.

Figure 8 illustrates the behavior of the model across dataset with different numbers of exceptions, starting at the bottom of each graph with a dataset like those above, but with only one exception. The highest line in each graph is a dataset with 50 exceptions. The different panels illustrate the development of the probability of inducing a new stress constraint across different settings of the n parameter. When n is smaller, the probability of inducing a new

Figure 8: Effect of number of exceptions on probability of inducing a new PFC. Each line represents a single run of one particular dataset, with numbers of exceptions ranging from 1 (the lowest line) to 50 (the highest)



stress PFC descends very quickly and even for 50 exceptions, stabilizes at a relatively low value, around 40%. When n is higher, the probability descends more slowly, and stabilizes at a wider range of values, depending on how many exceptions are present. For the case with only one exception, the probability still bottoms out close to zero, though it takes longer to do so. Under reasonable settings of the n parameter, the basic facts of contrast on novel words hold: When no words in the lexicon contrast in stress, stress will be very difficult to represent on novel words. If there are even 5 exceptions, it will be slightly easier, at least for some speakers. The more exceptions there are, the easier stress will be to represent on novel words.

5 Conclusion

This paper presented Representational Strength Theory, a theory based on Maximum Entropy Grammar, which uses Phonological Form Constraints instead of Faithfulness to enforce individual lexical item's preferences. By using PFC's, which have weights that are learned along the weights of markedness constraints, the acquisition of the lexicon and the grammar can be modeled together, as demonstrated in the simulations presented in the final section. These simulations illustrate that RST can learn probabilistic phonology alongside the real words of a language. In fact, the model makes two significant predictions: (1) Exceptional lexical items are more stable when they are high frequency than when they are low frequency. This prediction is empirically correct, as has been pointed out as early as Bybee (1985). (2) Features that are not contrastive in a language should be difficult for speakers to encode, and furthermore, contrastiveness is a continuum. Languages in which the grammar perfectly predicts some feature can be said to be

completely noncontrastive for that feature, but languages in which that feature is semi-predictable may exhibit an intermediate level of contrast. This prediction lines up well with results in Peperkamp et al. (2010).

PFC's can in principle do all the work of an underlying form plus faithfulness constraints, though more research is necessary to fully understand the different predictions of the two models. The ordering issue discussed in section 2.1.4 is also of importance, especially as it relates to the induction of novel PFC's during learning. The simulations presented in this paper simplified many things, including ordering relationships.

Finally, RST and the GLaPL learning algorithm allow a new look into the relationship between the grammar and the lexicon, bearing on interesting open questions such as: How are high-frequency words represented differently than low-frequency words? How does the lexicon of a language support or hinder the acquisition of certain probabilistic patterns? How, and why do speakers know probabilistic generalizations over words whose details all need to be memorized anyway? When grammatical generalizations are learned, how much and what kind of redundancy do lexical items actually exhibit?

References

Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161, 2003. doi: 10.1016/S0010-0277(03)00146-X.

Arto Anttila. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, editors, *Variation, Change, and Phonological Theory*, pages 35–68. John Benjamins, Amsterdam, 1997.

Arto Anttila. Morphologically conditioned phonological alternations. *Natural Language & Linguistic Theory*, 20(1): 1–42, 2002.

Andrew Bayles, Aaron Kaplan, and Abby Kaplan. Inter- and intra- speaker variation in French schwa. *Glossa*, 1(1): 19, 2016.

Michael Becker. *Phonological trends in the lexicon: the role of constraints*. PhD thesis, University of Massachusetts Amherst, 2009.

Michael Becker, Nihan Ketrez, and Andrew Nevins. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 87(1):84–125, March 2011. doi: 10.1353/lan.2011.0016.

Michael Becker, Lauren Eby Clemens, and Andrew Nevins. Generalizations of French and Portuguese plural alternations and initial syllable protection. *Natural Language and Linguistic Theory*, 35:299–345, 2017.

Ricardo Bermúdez-Otero. The architecture of grammar and the division of labour in exponence. In Jochen Trommer, editor, *The morphology and phonology of exponence: the state of the art*, pages 8–83. Oxford University Press, Oxford, 2012.

Paul Boersma. Phonology-semantics interaction in OT, and its acquisition. In Kirchner et al, editor, *Papers in Experimental and Theoretical linguistics*, volume 6. University of Alberta, Edmonton, 2001.

Paul Boersma and Bruce Hayes. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86, 2001. doi: 10.1162/002438901554586.

Dwight Bolinger. Two kinds of vowels, two kinds of rhythm. Indiana University Linguistics Club publications, Bloomington, IN, 1981.

Joan Bybee. *Morphology: A study of the relation between meaning and form.* John Benjamins, 1985.

Joan Bybee. *Frequency of use and the organisation of language.* Oxford University Press, Oxford, 2007.

Andrew Carstairs. Some implications of phonologically conditioned suppletion. *Yearbook of Morphology*, 1:67–94, 1988.

Noam Chomsky and Morris Halle. *The sound pattern of English.* Harper and Row, New York, Evanston, and London, 1968.

Andries W. Coetzee. An integrated grammatical/non-grammatical model of phonological variation. In Young-Se Kang, Jong-Yurl Yoon, Hyunkyung Yoo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Kyoung-Ae Kim, and Hye-Kyung Kang, editors, *Current Issues in Linguistic Interfaces.*, volume 2, pages 267–294. Hankook-munhwasa, Seoul, 2009.

Andries W. Coetzee and Shigeto Kawahara. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31:47–89, 2013.

Andries W. Coetzee and Joe Pater. The place of variation in phonological theory. In John A. Goldsmith, Jason Riggle, and Alan C. Yu, editors, *Handbook of Phonological theory*, pages 401–434. Wiley-Blackwell, 2 edition, 2009. doi: 10.1002/9781444343069.ch13.

Rebecca Irene Victoria Colavin. *Phonotactic Probability in Amharic: a psycholinguistic and computational investigation*. PhD thesis, UCSD, 2013.

Ulrike Domahs, Richard Wiese, Ina Bornkessel-Schlesewsky, and Matthias Schlesewsky. The processing of German word stress: evidence for the prosodic heirarchy. *Phonology*, 25:1–36, 2008.

Ulrike Domahs, Safiye Genc, Johannes Knaus, Richard Wiese, and Barış Kabak. Processing (un-)predictable word stress: Erp evidence from turkish. *Language and Cognitive Processes*, 28(3):335–354, 2013.

David Eddington. Issues in modeling language processing analogically. *Lingua*, 114:849–871, 2004.

Mirjam Ernestus and Harald Baayen. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79(1):5–38, 2003. doi: 10.1353/lan.2003.0076.

James Fidelholtz. Stress in Polish - with some comparisons to English stress. *Poznań studies in contemporary linguistics*, 9:47–61, 1979.

Guilherme D. Garcia. Weight gradience and stress in Portuguese. *Phonology*, 34(1):41–79, 2017.

Sharon Goldwater and Mark Johnson. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120, 2003.

Esther Grabe and Paul Warren. Stress shift: do speakers do it or do listeners hear it. *Connell & Arvaniti (1995)*, pages 95–110, 1995.

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

Daniel Currie Hall and Kathleen Currie Hall. Marginal contrasts and the contrastivist hypothesis. *Glossa*, 1(1):1–23, 2016.

Kathleen Currie Hall. A typology of intermediate phonological relationships. *The Linguistic Review*, 30(2):215–275, 2013.

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

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

Bruce Hayes, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85(4):822–863, December 2009. doi: 10.1353/lan.0.0169.

Sharon Inkelas, Orhan Orgun, and Cheryl Zoll. The implications of lexical exceptions for the nature of grammar. In Iggy Roca, editor, *Derivations and constraints in grammar*, chapter 13, pages 393–418. Clarendon paperbacks, Oxford, 1997.

Jongho Jun. Korean n-insertion: a mismatch between data and learning. *Phonology*, 32:417–458, 2015.

René Kager. Lexical irregularity and the typology of contrast. *The nature of the word: studies in honor of Paul Kiparsky*, pages 397–432, 2008.

Hyoju Kim. Contextual distribution of English loanword word-initial tensification in Korean. *Studies in Phonetics, Phonology and Morphology*, 22(2):245–288, 2016.

Seoyoung Kim. Phonological trends in Seoul Korean compound tensification. In *Proceedings of AMP 2016*, 2017.

Gakuji Kumagai and Shigeto Kawahara. Stochastic phonological knowledge and word formation in Japanese. *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, 153:57–83, 2018.

William Labov. contraction, deletion, and inherent variability of the English copula. *Language*, 45(4):715–762, 1969.

Tal Linzen, Sofya Kasyanenko, and Maria Gouskova. Lexical and phonological variation in Russian prepositions. *Phonology*, 30:453–515, 2013.

Linda Lombardi. Coronal epenthesis and markedness. *Phonology*, 19(2):219–251, 2002.

Claire Moore-Cantwell. Over-and under-generalization in morphological learning. In Stefan Keine and Shayne Sloggett, editors, *The proceedings of the 42nd meeting of the North East Linguistics society*, volume 2, pages 41–54, 2012.

Claire Moore-Cantwell and Joe Pater. Gradient exceptionality in maximum entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics*, 15:53–66, 2016. URL `http://revistes.uab.cat/catJL/article/view/v15-moore-cantwell-pater/183-pdf-en`.

Emily Morgan and Roger Levy. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:382–402, 2016. doi: http://dx.doi.org/10.1016/j.cognition.2016.09.011.

Ingrid Normann-Vigil. Accounting for variation of diminutive formation in Porteño Spanish. *Mester*, 41(1):99–122, 2012.

Joe Pater. Non-uniformity in english secondary stress: the role of ranked and lexically specific constraints. *Phonology*, 17(2):237–274, 2000. doi: 10.1017/S0952675700003900.

Joe Pater. Morpheme-specific phonology: constraint indexation and the inconsistency resolution. In Steve Parker, editor, *Phonological argumentation: essays on evidence and motivation*, pages 123–154. Equinox, 2010.

Joe Pater and Brian Smith. Le 'e' en français: élision, épenthèse, les deux, ni l'in ni l'autre? In *Phonologie du français contemporain: Contraintes, variation, prosodie*, December 2011.

Joe Pater, Robert Staubs, Karen Jesney, and Brian Smith. Learning probabilities over underlying representations. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology (SIGMORPHON2012)*, pages 62–71, Montréal, Canada, June 2012. Association of computational Linguistics.

Sharon Peperkamp, Inga Vendelin, and Emmanuel Dupoux. Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, 38:422–430, 2010.

Betty Phillips. *Word frequency and lexical diffusion*. Palgrave Macmillan, Basingstoke, 2006.

Alan Prince. Relating to the grid. *Linguistic Inquiry*, 14:19–100, 1983.

Alan Prince and Paul Smolensky. *Optimality Theory: constraint interaction in generative grammar*. Blackwell, Malden, MA, 1993/2004.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Hannah Sande, Peter Jenks, and Sharon Inkelas. Cophonologies by ph(r)ase. *Natural Language & Linguistic Theory*, 38:1211–1261, 2020.

James M Scobbie and Jane Stuart-Smith. Quasi-phonemic contrast and the fuzzy inventory: Examples from scottish english. *Contrast in phonology: Theory, perception, acquisition*, pages 87–114, 2008.

Stephanie Shih. Learning lexical classes from variable phonology. In *Selected Papers from Asian Junior Linguists Conference*, volume 2, pages 1–15. ICUWPL, 2018.

Brian W. Smith and Claire Moore-Cantwell. Emergent idiosyncrasy in english comparatives. In Andrew Lamont and Katie Tetzloff, editors, *NELS 47: Proceedings of the 47th meeting of the North East linguistic society*, pages 127–140, Amherst, MA, 2017. Graduate Linguistic Student Association.

Brian W. Smith and Joe Pater. French schwa and gradient cumulativity. *Glossa*, 5(1):24, 2020.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216, 1990.

Paul Smolensky and Matthew Goldrick. Gradient symbolic representations in the grammar: The case of French liason. *Rutgers Optimality Archive*, 1552, 2016.

Paul Smolensky, Matt Goldrick, and D. Mathis. Optimization and quantization in gradient symbol systems: a framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, 38(6):1102–1138, 2014.

S oren Eggert Tebay and Eva Zimmermann. Exceptionality in assamese vowel harmony: A phonological account. *Glossa*, 5(1):102, 2020.

Jochen Trommer. The subsegmental structure of German plural allomorphy. *Natural Language and Linguistic Theory*, 2020. doi: https://doi.org/10.1007/s11049-020-09479-7.

Jie Zhang and Jiang Liu. The productivity of variable disyllabic tone sandhi in Tianjin Chinese. *Journal of East Asian linguistics*, 25:1–35, 2016.

Eva Zimmermann. Gradient symbolic representations and the typology of ghost segments. In Katherine Hout, Anna Mai, Adam McCollum, Sharon Rose, and Matthew Zaslansky, editors, *Proceedings of AMP 2018*, 2019.

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

Kie Zuraw. The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language*, 83(2):277–316, 2007.

Kie Zuraw. A model of lexical variation and the grammar with application to tagalog nasal substitution. *Natural Language and Linguistic Theory*, 28(2):417–472, 2010.

Kie Zuraw. Polarized variation. *Catalan Journal of Linguistics*, 15:145–171, 2016.

Jesse Zymet. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning.* PhD thesis, UCLA, 2018.

Jesse Zymet. Learning a frequency-matching grammar together with lexical idiosyncrasy: Maxent versus hierarchical regression. In *Proceedings of the Annual Meetings on Phonology*, volume 7, 2019.