Ms, Department of Linguistics, University of Massachusetts, Amherst, 2005. Has minor corrections from published version, can be cited as in the Proceedings of the 29th Boston University Conference on Language Development, edited by Alejna Brugos, Manuella R. Clark-Cotton, and Seungwan Ha, Cascadilla Press, Somerville, MA. 2005. 482-492.

Learning a Stratified Grammar

Joe Pater
University of Massachusetts, Amherst

Introduction

Phonological processes and structures are often limited to a particular set of a language's words, such as loanwords, Latinate words in English, or Yamato words in Japanese. Typically, the etymologically older, or "core" set of words is more restricted in the structures that it permits, and is (hence) subject to more processes. To capture such restrictions in Optimality Theory, Itô and Mester (1999) propose stratified grammars of the form in (1), where *M* stands for a Markedness constraint, and *Faith* stands for a faithfulness constraint. *Faith-L1* is a lexically specific version of that faithfulness constraint (see also Pater 2000 for other arguments for lexically specific constraints).

(1) **A schematic stratified grammar**
FAITH-L1 >> M >> FAITH

The ranking of M over FAITH forces words to satisfy the markedness constraint by violating the faithfulness constraint. FAITH-L1 applies to the set of words in a language that violate M (i.e. loanwords, or other non-core words). Its rank above M allows just those words to bear the marked structure.

Stratified grammars pose special learnability problems. The simple schematic example in (1) raises the most basic one: how do lexically specific constraints get constructed, and indexed to the proper set of words? In cases where no morpho-phonological alternations exist, this problem has been argued to be irresolvable (Ota 2004). In this paper, I address this issue by proposing that learners are initially conservative, in that when they encounter a word that requires an adjustment to the grammar, they first assume that this adjustment is specific that word. More formally, in terms of Tesar and Smolensky (1998) *et seq.*, when Error-Driven Constraint Demotion produces a Mark-Data pair, faithfulness constraints preferring the winner are indexed to the lexical item in question. When this proposal is incorporated into Prince and Tesar's (2004) Biased Constraint Demotion Algorithm, it automatically yields an answer to the

* Portions of this work were also presented at the University of Victoria and NYU. Thanks to the participants there and at the BUCLD, as well as Adam Albright, Andries Coetzee, John McCarthy and Bruce Tesar, for comments and discussion.

further problem of how lexically specific faithfulness constraints are correctly interspersed into a hierarchy of markedness constraints (Itô and Mester 1999). In applying stratified grammars to lexically gradient phonotactics, this paper also suggests a way of expressing one kind of gradient well-formedness in Optimality Theory (cf. Frisch, Pierrehumbert and Broe 2004).

1. A Stratified Grammar

In this section, I provide a simple hypothetical example of a language with a stratified grammar, which will form the basis of the learnability discussion. The constraints that I use are defined in (2).

(2)

NoCoda	Syllables end in vowels (*C]$_\sigma$)
*Complex	No consonant clusters
Max	Input segments have output correspondents ('No deletion'; McCarthy and Prince 1999)

The hypothetical language generally lacks both codas and clusters; most of the words are made up only of CV syllables. The lack of codas and clusters is a purely static generalization; there are no alternations. There is, however, a small set of words with codas (two words, [bat] and [net]), and an even smaller set with clusters (one word, [pla]).[1]

(3) **Hypothetical stratified grammar**

Grammar: Max-L1 >> *Complex >> Max-L2 >> NoCoda >> Max

Lexicon: /pla$_{L1}$/ /bat$_{L2}$/ /net$_{L2}$/ /pa/ /ti/ /la/ /me/ /go/ /ra/

Max-L1 and Max-L2 are lexically specific versions of Max. The ranking of *Complex and NoCoda above Max enforces the general absence of clusters and codas. Since there are no alternations, the choice of this particular faithfulness constraint is arbitrary; clusters and codas could equally be avoided through epenthesis. The exceptional words /bat/ and /net/ are able to violate NoCoda because they indexed to Max-L2, which ranks above the markedness constraint. Because it is indexed to Max-L1, /pla/ is able to violate *Complex.

Without lexically indexed constraints and stratification, the grammar would have to be as in (4). This grammar is the same as for a language that freely allows codas and clusters.

---

1. Stratification is usually discussed, in Itô and Mester (1999) and elsewhere, in etymological, not in quantitative terms. Often, these will correlate: loanwords are typically relatively rare. The learnability account presented here is dependent on a quantitative interpretation of stratification; see the end of section 2 for relevant discussion.

(4) **Hypothetical grammar without stratification**
MAX >> *COMPLEX, NOCODA

It has in fact been suggested that the grammar in (4) would be the correct one for a language like our hypothetical one, which lacks any alternations that would require MAX to be dominated by *COMPLEX and NOCODA (Inkelas, Orgun and Zoll 1997, Rice 1997). However, this hypothetical case demonstrates the unhappy consequence of this position: that even a single exception (like [pla]) forces an otherwise general pattern (like the absence of clusters) to be completely unexpressed in the grammar.

The need to distinguish between fully acceptable and exceptional patterns is underlined by the ability of native speakers to distinguish between them. Coetzee (2004) examined English native speakers' judgments and lexical decision times on nonce words of the shape [sC$_1$VC$_2$], where C$_1$ and C$_2$ are identical stops, [p], [t], or [k]. The results show that the subjects ranked the words as in (5), where where '>' indicates higher grammaticality rating, and slower lexical decision time.

(5) stVt > skVk > spVp

This ranking appears to correlate with the extent to which the structures are attested in existing words.[2] Homorganic oral stops in this context are completely unattested when they are labials, are somewhat more common when they are velars (especially if intervening nasals and liquids are included), and are quite common when they are coronals (see e.g. Davis 1991, Lamontagne 1993). Examples appear in (6).

(6)

sTVT	e.g. stud, stood, stead, stat, state, staid…
sKVK	e.g. skeg, skag, skunk, skank, skulk
sPVP	e.g. ?

2. Coetzee (2004) treats both sKVK and sPVP as unattested, which they are if one considers only words exactly like his stimuli, which have voiceless stops without intervening nasals or liquids. His account of the difference between them in his results rests on a universal ordering of *sPVP >> *sKVK. While this ordering holds of the Germanic sCVC cases he examines, it does not appear to hold more broadly of OCP-Place, as Berg, Coetzee and Pater (2005) show for Muna.

The stratified grammar for this case appears in (7), using an undifferentiated faithfulness constraint, and markedness constraints that directly penalize each of the three structures.

(7) *sPVP >> FAITH-L1 >> *sKVK >> FAITH >> *sTVT

The ranking of *sPVP above both faithfulness constraints encodes the unacceptability of sPVP forms. The ranking of both faithfulness constraints above *sTVT allows sTVT words to surface faithfully. For sKVK words like *skag*, one might say that their marginal status is encoded by the presence of the lexical exception feature (i.e. its indexation to FAITH-L1). However, this does not deal with the nonce word judgments, nor does it deal with further gradience, in which two exceptional structures differ in their degree of integration into the language, as in our hypothetical language.[3]

Gradient well-formedness judgments can be modeled using a stratified grammar by calculating the probability of a form surfacing faithfully, given all possible lexical indexations (cf. Anttila 1997 on variation). Returning to our hypothetical language, a CVC nonce form would surface faithfully if indexed to either MAX-L1 or MAX-L2, but not if left unindexed, in which case it would lose its coda. That is, it surfaces faithfully in 2/3 indexations. A CCV nonce form would surface faithfully only if indexed to MAX-L1, or in 1/3 indexations. A CV nonce form would surface faithfully with any indexation. In this way, the grammar in (3) ranks these forms as in (8):

(8) **Ranking of word types by hypothetical stratified grammar**
CV > CVC > CCV

Alternatively, one could use the relative rank of the markedness constraints in the grammar to directly assess the relative well-formedness of the word types (see Coetzee 2004). Either way, stratified grammars can be used to account for speakers' knowledge of gradient well-formedness that arises from the relative degree of attestation of a structure in the lexicon. As Frisch *et al.* (2004) point out, this sort of gradience has yet to be dealt with in Optimality Theory (see also

3. English may provide a real example, if sCVC words with heterorganic stops are judged as more acceptable than sTVT words, so that sTVT and sKVK are "exceptions" of different acceptability (see Lamontagne 1993: 267 for evidence that sTVT words are underrepresented). The OCP-Place facts of Arabic (Frisch *et al.* 2004) and Muna (Berg *et al.* 2005) likely do as well, though the relevant judgment data are lacking. Several other studies show native speaker sensitivity to multiple degrees of acceptability, but these involve variation and/or alternations (e.g. Zuraw 2000, Boersma and Hayes 2001).

Berkley 2000; cf. Anttila 1997 and Boersma and Hayes 2001 on variation, and Zuraw 2000 on gradience in morpho-phonological alternations).

This simple hypothetical example shows how stratification allows a grammar to encode intermediate degrees of well-formedness, which motivates the use of stratification even in cases that do not involve alternation (see also Itô and Mester 2001).[4] Now we turn to how such a grammar can be learned without evidence of alternation.

2. Learning a Stratified Grammar

In Error-Driven Constraint Demotion (Tesar and Smolensky 1998), the learners' current grammar is used to parse the data it encounters. If the grammar yields an output that does not correspond to the learning data, the constraints are reranked using the Constraint Demotion Algorithm. Unmodified, this approach would yield an unstratified grammar like that in (4) for our hypothetical language. If there were a corresponding alternation that applied to core but not peripheral forms, then inconsistency detection (Tesar 1998, Prince 2002) could be applied to trigger the creation of a lexically specific constraint.[5] But for purely phonotactic stratification, an unstratified grammar would successfully parse all of the forms.

This is a special case of the subset problem for phonotactic learning discussed by Smolensky (1996), Hayes (2004) and Prince and Tesar (2004). These papers posit learning biases in which markedness constraints are ranked over faithfulness constraints, so that learners are not trapped in the superset grammar with faithfulness over markedness. The twist here is that the language does in fact provide evidence for the F >> M ranking, but only in a limited set of words. To deal with this special case, I elaborate on Prince and Tesar's (2004) Biased Constraint Demotion (BCD) Algorithm (cf. Itô and Mester 1999, who elaborate on Smolensky 1996).

---

4. The main argument that Itô and Mester (2001) provide for stratification is its ability to account for "impossible nativizations", which are implicational relationships between exceptional structures. Taking our hypothetical language as an example, given an input form with a coda and a cluster, if the output preserves the cluster, it must also preserve the coda, since MAX-L1 dominates both *COMPLEX and NOCODA. However, most real cases of this type involve separate faithfulness constraints interacting with each of the markedness constraints. For these, it is unfortunately unclear why indexing a word for the higher ranked faithfulness constraint should entail an indexation to the lower one as well (cf. Itô and Mester 1999).

5. See Pater (2004) for some development of this approach; see also Ota (2004) for discussion of Japanese postnasal voicing in these terms.

When Error-Driven Constraint Demotion detects an error, it creates a Mark-Data pair that provides the information used for constraint demotion. The learning datum is called the Winner, and the output of the learner's grammar is called the Loser. Constraints assign a 'W' when they prefer the Winner, and an 'L' when they prefer the Loser. I adopt Tesar's (1998) proposal that Mark-Data pairs are retained for further learning, forming a set that Tesar and Prince (2004) term a support. The main elaboration that I propose is that when Mark-Data pairs are formed, faithfulness constraints that prefer the winner are indexed to the lexical item in question.

BCD iteratively places constraints in strata according to the following steps, which favor high-ranking markedness constraints, and hence a restrictive grammar:

(9) i. Identify constraints that prefer no losers
ii. If any of these are markedness constraints, install them in the current stratum, and return to step i.
iii. If there are no available markedness constraints, install faithfulness constraints that prefer winners, and return to step i.
iv. If there are no faithfulness constraints that prefer winners, install those that prefer no losers, and return to step i.

To illustrate how this modified BCD algorithm creates a stratified grammar, I use the constraint set, and hypothetical language introduced in the last section:

(10) Words: [pa] [ti] [la] [me] [go] [ra] [mat] [net] [fle]
Constraints: NOCODA *COMPLEX MAX

Before any data are presented to the algorithm, it creates the following grammar, with the markedness constraints ranked above the faithfulness constraint:

(11) NOCODA, *COMPLEX >> MAX

If this grammar is used to parse any of the non-CV words, it will yield an error. For example, given [mat], the grammar yields [ma]. A Mark-Data pair is then created, with an indexed faithfulness constraint:

(12)

Input	W ~ L	NOCODA	*COMP	MAX-L1
mat_{L1}	mat ~ ma	L		W

BCD will now produce the following:

(13) *COMPLEX >> MAX-L1 >> NOCODA >> MAX

Since the indexed constraint applies only to /mat/, the grammar in (13) would also produce an error upon encountering [net] (and [fle]). All of the words with marked structures will lead to errors and the creation of Mark-Data pairs, while the unmarked CV words will never produce errors or Mark-Data pairs. The full support tableau for this language will thus be as in (14).

(14)

Input	W ~ L	NO CODA	*COMP	MAX-L1	MAX-L2	MAX-L3
mat_{L1}	mat ~ ma	L		W		
net_{L2}	net ~ ne	L			W	
fle_{L3}	fle ~ fe		L			W

Applying BCD to this support entails choosing amongst three lexically specific faithfulness constraints. Prince and Tesar (2004: 267) propose that the choice amongst faithfulness constraints is made by identifying ones that "free up" markedness constraints for ranking:

(15) "**Smallest effective F sets.** When placing faithfulness constraints into the hierarchy, place the *smallest set* of F constraints that *frees up some markedness constraint*."

To free up a markedness constraint means to eliminate all its L marks. A Mark-Data pair is eliminated when a constraint that prefers its Winner is installed; the installation of that constraint guarantees that the Winner will be optimal in the resulting grammar. Installing MAX-L3 will eliminate the Mark-Data pair for /fle/, and will free up *COMPLEX. To free up NOCODA, both MAX-L1 and MAX-L2 must be installed. Therefore, the smallest effective F set is {MAX-L3}. When this constraint is installed, the Mark-Data pair for /fle/ is eliminated, indicated by removing the row from the support tableau in (16):

(16) Grammar: MAX-L3 >>

Input	W ~ L	NO CODA	*COMP	MAX-L1	MAX-L2	MAX-L3
mat_{L1}	mat ~ ma	L		W		
net_{L2}	net ~ ne	L			W	

*COMPLEX is then installed due to the markedness bias:

(17) MAX-L3 >> *COMPLEX >>

This does not eliminate any further Mark-Data pairs, since *COMPLEX prefers no Winners, so the support remains as in (16).

Before proceeding further, it is worth noting the important role of the "Smallest effective F sets" clause to the success of BCD in creating a stratified grammar. Part of the goal is to have markedness constraints ranked according to how often they are violated in the language, with higher rank correlating with fewer violations (see also Boersma 1998 in a different context). A lexically specific faithfulness constraint is created for each occurrence of a word with a violation of a markedness constraint. A markedness constraint that is violated rarely will create few faithfulness constraints. This will be a "small effective F set", and it, followed directly by this markedness constraint, will be installed before a markedness constraint that is violated more often, since its effective F set will be larger. In the present example, *COMPLEX is only violated once, and so its associated effective F set consists only of MAX-L3. NOCODA is violated twice, so its effective F set consists of MAX-L1 and MAX-L2.

Once we have installed *COMPLEX, MAX-L1 and MAX-L2 will be installed together to free up NOCODA. This eliminates all of the Mark-Data pairs from further consideration, so the support tableau is now empty.

(18) MAX-L3 >> *COMPLEX >> MAX-L1, MAX-L2

With no more data to account for, NOCODA will be installed next due to the markedness bias, and then the general MAX constraint:

(19) MAX-L3 >> *COMPLEX >> MAX-L1, MAX-L2 >> NOCODA >> MAX

Lexically specific constraints can be collapsed as follows:

(20) Merge instantiations of any constraint that occupy the same stratum

This produces the desired grammar, and lexicon:

(21) Grammar: MAX-L1 >> *COMPLEX >> MAX-L2 >> NOCODA >> MAX
Lexicon: /fle$_{L1}$/ /mat$_{L2}$/ /net$_{L2}$/ /pa/ /ti/ /la/ /me/ /go/ /ra/

Further collapse of lexically specific constraints is necessary to produce a non-stratified grammar when a structure is well attested in a language. This can be accomplished by imposing a maximum size on the set of words targeted by a lexically specific constraint:

(22) If the number of words indexed by a constraint is greater than *x*, remove indexation, and delete any lower ranked instantiation of the constraint

Once indexation has been removed, the learner will also stop making errors, and creating Mark-Data pairs and lexically specific constraints. For example, if we assumed that $x = 1$,[6] the step in (22) would result in (23) for our hypothetical language.

(23) Grammar: MAX-L1 >> *COMPLEX >> MAX >> NOCODA
Lexicon: /fle$_{L1}$/ /mat/ /net/ /pa/ /ti/ /la/ /me/ /go/ /ra/

If a learner with this grammar encountered another word with a coda, it would parse it faithfully, and no Mark-Data pair would be created.

The arbitrary diacritics that lexical indexation creates may also be eliminated in favor of grammatical categories. For example, Smith's (1997) cross-linguistic study shows that nouns often contain structures that are banned in verbs. Learners of these languages might start by treating these structures as arbitrary exceptions, and then identify the category Noun as the property uniting the indexed items, which would replace the arbitrary diacritic (e.g. 'MAX-L1' → 'MAX-NOUN'). Since faithfulness constraints targeting categories do not require lexical markings, these may be more robust than arbitrary exceptions, and less prone to regularization (see also Anttila 2002 for evidence of an arbitrary pattern being grammaticalized). Similarly, strata like those discussed Itô and Mester (1999), which tend to show a clustering of properties, might use just one (arbitrary) lexical diacritic across a number of faithfulness constraints, which could also increase robustness.

3. Conclusions

A stratified grammar, consisting of lexically indexed faithfulness constraints interspersed between markedness constraints, provides a means of expressing intermediate grades of well-formedness in Optimality Theory. In the absence of alternations, such stratified grammars raise three learnability problems:

(24) i. How does a learner create lexically specific constraints for exceptions to phonotactics?
ii. How do the markedness constraints get in the right order?
iii. How do the faithfulness constraints get interspersed correctly?

Itô and Mester (1999) propose a solution to the second two of these problems, but it requires that the learner encounter the data in a particular order, and it does

6. It is of course impossible to know exactly what value *x* should have, though 1 is clearly too low for real cases.

not address the first problem. In this paper, I have proposed that the first problem can be solved by lexically indexing faithfulness constraints during the creation of Mark-Data pairs. With this addition, Biased Constraint Demotion takes care of the rest.

References

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R van Hout and W. L. Wetzels (eds.) *Variation, Change and Phonological Theory*. Amsterdam, John Benjamins.

Anttila, Arto. 2002. Morphologically Conditioned Phonological Alternations. *Natural Language and Linguistic Theory* 20. 1-42.

Berkley, Deborah. 2000. *Gradient Obligatory Contour Principle Effects*. Ph.D. Dissertation, Northwestern University.

Berg, René van den, Andries Coetzee and Joe Pater. 2005. Lexically Gradient Phonotactics in Muna and Optimality Theory. Ms, University of Massachusetts, Amherst.

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

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32. 45-86.

Coetzee, Andries. 2004. *What It Means To Be A Loser: Non-Optimal Candidates In Optimality Theory*. Ph.D. Dissertation, University of Massachusetts, Amherst.

Davis, Stuart. 1991. Coronals and the phonotactics of non-adjacent coronals in English. In C. Paradis, J.-F. Prunet (eds.) *The Special Status of Coronals: Internal and External Evidence*. San Diego: Academic Press. 49-60.

Frisch, Stefan, Janet Pierrehumbert and Michael Broe. 2004. Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory* 22. 179-228.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the Early Stages. In R. Kager, J. Pater and W. Zonneveld (eds.), *Constraints in Phonological Acquisition*. Cambridge University Press.

Inkelas, Sharon, Cemil Orhan Orgun and Cheryl Zoll. 1997. Implications of lexical exceptions for the nature of grammar. In I. Roca (ed.) *Constraints and derivations in phonology*. Oxford: Clarendon Press. 393-418.

Itô, Junko and Armin Mester. 1999. The Phonological Lexicon. In N. Tsujimura (ed.) *Handbook of Japanese Linguistics*. Oxford: Blackwell. 62-100.

Itô, Junko and Armin Mester. 2001. Covert generalizations in Optimality Theory: the role of stratal faithfulness constraints. *Studies in Phonetics, Phonology, and Morphology* 7. 273-299.

Lamontagne, Greg. 1993. *Syllabification and Consonant Cooccurrence Conditions*. Ph.D. Dissertation, University of Massachusetts, Amherst.

McCarthy, John and Alan Prince. 1999. Faithfulness and identity in prosodic morphology. In R. Kager, H. van der Hulst and W. Zonneveld (eds.) *The Prosody-Morphology Interface*. Cambridge: Cambridge University Press. 218-309.

Ota, Mits. 2004 The learnability of the stratified phonological lexicon. *Journal of Japanese Linguistics* 20, 4.

Pater, Joe. 2000. Nonuniformity in English stress: the role of ranked and lexically specific constraints. *Phonology* 17. 237-274.

Pater, Joe. 2004. Exceptions in Optimality Theory: Typology and Learnability. Presented at the Conference on Redefining Elicitation: Novel Data in Phonological Theory, New York University (http://people.umass.edu/pater/exceptions.pdf)

Prince, Alan. 2002. Entailed Ranking Arguments. Ms., Rutgers University.

Prince, Alan, and Bruce Tesar. 2004. Learning Phonotactic Distributions.In R. Kager, J. Pater and W. Zonneveld (eds.) *Constraints in Phonological Acquisition*. CUP. 245-291.

Rice, Keren. 1997. Japanese NC clusters and the redundancy of postnasal voicing. *Linguistic Inquiry* 28. 541-551.

Smith, Jennifer. 1997. Noun faithfulness: On the privileged behavior of nouns in phonology. Ms., University of Massachusetts.

Smolensky, Paul. 1996. The Initial State and Richness of the Base in Optimality Theory. Technical Report JHU-CogSci-96-4, Cognitive Science Department, Johns Hopkins University.

Tesar, Bruce. 1998. Using the mutual inconsistency of structural descriptions to overcome ambiguity in language learning. In P. Tamanji and K. Kusumoto (eds.) *Proceedings of the North East Linguistic Society* 28. Amherst, MA: GLSA, University of Massachusetts. 469-483.

Tesar, Bruce, and Alan Prince. 2004. Using Phonotactics to Learn Phonological Alternations. In *The Proceedings of CLS 39, Vol. II: The Panels.*

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology.* Ph.D. Dissertation, UCLA.