Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints

Claire Moore-Cantwell and Joe Pater, University of Massachusetts Amherst

This paper shows that the combination of lexically specific constraints and Maximum Entropy Grammar yields a novel approach to a long-standing problem in phonological theory: the gradience of exceptionality. The problem of gradient exceptionality was perhaps first noted by Fidelholtz (1979: 58):

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

When the penultimate syllable is light, stress falls on the penultimate syllable of some English and Polish words (e.g. English *banána*), and on the antepenult on others (e.g. *Cánada*). In English, both patterns are well-attested (Pater 1994), and each word's pronunciation is stable. In Polish, on the other hand, there are very few antepenultimately stressed words (about 0.1% of the vocabulary according to Peperkamp *et al*. 2010), and they tend to be borrowings, with frequent regularization to penultimate stress. Peperkamp and colleagues provide psycholinguistic evidence of a difference between the two languages: English speakers are much better than Polish at recalling the placement of stress on a sequence of novel words. The formal descriptions of both languages require lexical marking of one of the patterns (in English it is hard to say which one), but the number of exceptions – few or many – does not affect the grammatical status of the pattern at all in a standard generative grammar. This is true of the SPE formalism of Fidelholtz's time, of metrical rule approaches, and of OT accounts with lexically specific constraints, posited independently by Kraska-Szlenk (1995) for Polish and Pater (1995 [2000]) for English stress.

When lexically specific constraints are incorporated into a Maximum Entropy Grammar framework (MaxEnt: Goldwater and Johnson 2003), the outcome of learning does yield a grammatical difference between a language like Polish and one like English. MaxEnt differs from standard OT in having weighted rather than ranked constraints, and in defining a probability distribution over the members of a candidate set. Lexically specific constraints are clones of general constraints that apply to single lexical items. The tableaux in (1) illustrate the basic workings of MaxEnt and lexically specific constraints. We consider only candidates with a trochaic foot in final or non-final position. The general constraints conflict in wanting the foot to be final (Align-R) or not (Nonfinality). Lexically specific Align-R-*i* applies only to *banana*. Hand-chosen weights are given beneath the constraint names. This is a language with a general pattern of antepenultimate stress (Nonfinality > Align-R), and where *banana* is an exception. The column headed *H* shows the weighted sum of violations, and the probability (*p*) of each candidate in a tableau is proportional to exp(*H*). The probabilities of the correct stress patterns on both *banána* and antepenultimately stressed *Cánada* approach 1 (and could be made arbitrarily close to it by scaling the weights).

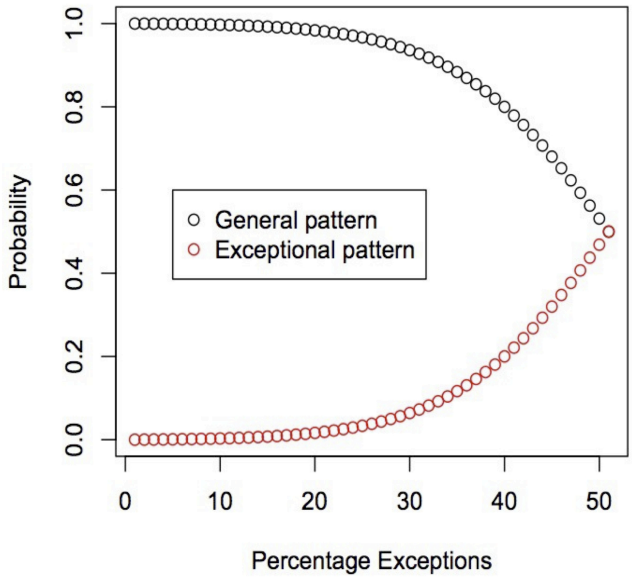
(1)

	Align-R-*i* 10	Nonfin 6	Align-R 1	*H*	*p*
ba(nána)$_i$		–1		–6	0.99
(bána)na$_i$	–1		–1	–11	0.01
Ca(náda)		–1		–6	0.01
(Cána)da			–1	–1	0.99

Given basic assumptions about how MaxEnt grammars are learned, the weights of the general constraints (and of the lexically specific ones) will vary as a function of the number of lexically specific constraints. As the number of exceptions increases, the strength of the encoding of a pattern in the general constraints decreases. This is illustrated in the result of a series of 51 learning simulations in which there were two possible locations for stress, and 100 words, with a general stress pattern with between zero and 50 exceptions. We posited a general constraint for each stress pattern, and two lexically specific constraints for each word - one favoring the general pattern and one favoring the exceptional pattern. Constraint weights started at zero, and were updated over 1000 epochs of Gradient Descent, with a learning rate of 0.1. At the end of learning, each word had probability of correct stress placement approaching 1, in all of the simulations. The grammatical encoding of the generalization can be seen in the weights of the general constraints, and in the resultant probabilities granted to stress in each position in the absence of lexical constraints. These probabilities are shown in Figure 1, and can be interpreted as the probability of each pattern being applied to a nonce word.

The probability of the general pattern ranges from 1.0 for an exception-less pattern, to 0.5 when stress is distributed equally across the two positions. This gradient model captures the difference between Polish and English: when the language has relatively few exceptions, the general pattern will apply not only to totally new forms, but whenever the speaker forgets a form – leading to regularization, especially of low-frequency items. However, the more exceptions the language has, the less regularization will obtain since the probability of a speaker choosing the exceptional form on a novel or forgotten word increases.

Figure 1: Probability of stress on a novel word (no lexically specific constraints) in each position (black = general pattern, red = exceptional stress)



The gradient strength of encoding of a general pattern is also compatible with further results from Peperkamp *et al.* (2010), who show that speakers of languages with fully predictable stress do even worse than Polish speakers in encoding novel stress patterns, and that Spanish speakers, whose language has a degree of exceptionality between Polish and English, also perform at a level intermediate between those two groups. The more strongly encoded a pattern is in terms of the weights of the general constraints, the higher the weights must be on exceptional items' lexically specific constraints in order to faithfully encode those exceptions. Since learning proceeds gradually, higher weights on lexically specific constraints would take longer to learn.

This result builds on several recent demonstrations of the usefulness of MaxEnt grammars for capturing lexically gradient patterns. Our model differs from Hayes and Wilson (2008) in being applicable to alternations as well as phonotactics, and from Hayes, Zuraw, Siptár and Londe (2009) in being able to encode word-specificity of a pattern. To show the breadth of coverage of this model, we will also present results of simulations on the learning of lexically specific alternations of differing degrees of strength in a single language (Dutch voicing: Ernestus and Baayen 2003 *et seq.*).