# Modeling lexical and stochastic exceptions in phonotactics: a categorizational problem

*Ildikó Emese Szabó—New York University*

**Problem:** Some generalizations are exceptionless, but others are just tendencies. In this paper I propose a learning model sensitive to how much work a contrast does (its functional load) that captures the patterns observed in Hungarian vowel length contrasts. Thereby I also argue that lexical and stochastic exceptions are closely linked, and are results of different malfunctions in categorization: a systematic failure to categorize and mislabeling in crucial environments.

Functional load is a measure of how often a contrast is used to distinguish words from each other. For Hungarian vowel length pairs, it is lightest for high ones (/u-uː/;/y-yː/ and /i-iː/), intermediate for mid (/o-oː/;/ø-øː/), and heaviest for low vowels (/ɔ-aː/;/ɛ-eː/). Perceivability patterns the same way: low distinguishability for short-long high vowels, intermediate for mid, and low for low vowels (confusability peaks for /u-uː/;/y-yː/;/i-iː/, and /iː-eː/ Mády, 2010).

The length of word-final vowels is subject to distributional restrictions that vary in strength (Siptár & Törkenczy, 2000; Mády & Reichel, 2007; Mády, 2010). High rounded vowels must be long (*/u#/, */y#/), but any word-final high vowel can be short in Budapest Colloquial Hungarian (BCH). In Standard Hungarian (SH) the restriction has lexical exceptions—such as [fɔlu] 'village' or [aːru] 'goods'. The mid vowel restriction requiring them to be long word-finally is exceptionless. Phonologically low vowels must be short (i.e. only /ɔ, ɛ/ are permitted, /aː, eː/ are not), but they also exhibit lexical exceptionality—i.e. certain lexical items violate this restriction: e.g. [kaːveː] 'coffee', [ørøkːeː] 'forever', [burʒoaː] 'bourgeois'. Neither amount nor type of exceptions seem to line up with the perceivability and functional load scales, and hence exceptionality may seem random. Note that in SH low and high vowels both show lexical exceptions, but them forming a natural class excluding mid vowels would be difficult to motivate.

**Model:** The following model represents the two types of exceptionality for phonotactic restrictions as two distinct categorizational patterns. It uses two forms of input: acoustic properties and raw frequency for functional load. Acoustic data (e.g. vowel formants) are given for the two sounds between which the phonotactic restriction expresses a preference (here: A and B). Based on means and standard deviations, the learner generates its starting set of tokens for both sounds ($set_A$, $set_B$). The second form of input is a corpus, where for each two tokens, the learner has information about whether they are instances of the same word or not. The learner then counts the functional load of the A–B contrast ($L_{FUNC}$(A;B)) N(A;B) is the number of A tokens that if replaced with B, the resulting word would still occur in the corpus. N(A) is the total frequency of A in the corpus.

$$L_{FUNC}(A;B) = \frac{N(A;B) + N(B;A)}{N(A) + N(B)}$$

Consider the following toy corpus S ={pig, pig, peg, pug, peg, gap, pip, pup, pun}. If we are looking for the $L_{FUNC}$ of 'i' and 'u' (A is 'i', B is 'u'), N(A;B) will be 2 ('pig' and 'pug form a minimal pair, and 'pig' occurs twice in the corpus), N(B;A) will be 1 ('pig' and 'pug form a minimal pair, and 'pug' occurs once), N(A) will be 3 (because the corpus contains 'pig' twice and 'pip' once, 3 'i's in total) and N(B) will be 3 as well (because the corpus contains 'pug', 'pup' and 'pun', all three of those once). Therefore $L_{FUNC}$ =(2+1)/(3+3)=0.5

In each iteration, the learner generates data points from both starting sets based on properties of already existing points in $set_A$ and $set_B$ but with some noise added. It will then sort these points into one of three sets, $set_A$, $set_B$, or $set_C$ for undecided. It calculates a certainty score based on two factors: how big the difference between the point's distance from *set*$_A$ and *set*$_B$ is (how obvious it is, which sound the point is a token for) and how big the functional load of the contrast is (how important it is for the learner is to categorize the token). Essentially, the certainty score is an equivalent of perception—with distinctions being sharpened by functional load (Feldman and Griffiths, 2007). If the score reaches a specified threshold, the point is sorted into either *set*$_A$ or *set*$_B$—whichever it is closest to. Otherwise, the point ends up in *set*$_C$.

A set of simulations using different thresholds reveals typical patterns for lexical and stochastic exceptionality. Because of functional load directly contributing to the certainty score, light functional load leads to a larger *set*$_C$—i.e. the two categories are less salient. This results in a systematic failure to categorize: a model of stochastic exceptionality. Lexical variation (occasional suspension of the restriction), however, requires two salient categories where at least one is frequently misperceived as a sound with a different distribution. In other words, lexical exceptions in phonotactics stem from categorizational mistakes.

**Results:** The two inputs were the SzóSzablya Webcorpus (Halácsy et al., 2004) that has 589 million tokens and 7.2 million types as a corpus and Mády and Reichel (2007)'s acoustic data. Functional load indeed patterned with vowel height: the length distinction for high vowels had the lightest functional load (0.35%), mid vowels were intermediate (2.3%) and low vowels had the heaviest functional load (10.1%). 100 simulations were run for each Hungarian length-pair. With 5 000 iterations each, each simulation categorized 10 000 new points (5 000 generated based on both *set*$_A$ and *set*$_B$). With any threshold, in all simulations, the scale shown below in (1) was observed. In (1), $|C_{HIGH}|$ stands for the number of high vowel points left "undecided" (*set*$_C$). This matches the degree of stochastic variation in BCH—it shows up strongly for high vowels, mid vowels start to exhibit it while low vowels are stable.

(1) $|C_{HIGH}| \geq |C_{MID}| \geq |C_{LOW}|$

As for lexical exceptions, the low vowel /eː/ and the high vowel /iː/ have different distributions (/i/ can appear word-finally), but are often mixed up (17.3% and 8.1% confusability: Mády, 2010). In the model, it was borne out as some word-final tokens being falsely categorized as /eː/'s—keeping the categories distinct and consistent. This makes the restriction on word-final /eː/'s less strong, allowing for exceptions. In the case of /u/-/uː/ and /y/-/yː/ the accidental mislabeling happens within the length pair. Even if the categories are successfully distinguished (as in SH), restrictions on these pairs have lexical exceptions, as they are frequently confused with each other.

**Conclusions**: Above, I proposed a model that succeeds at identifying the two types of phonotactic exceptions with different categorizational malfunction patterns. Through the link of functional load, acoustics and perception, stochastic exceptions can be modeled as a systematic failure to categorize and lexical exceptions as results of categorizational mistakes. This is a first step of establishing either a more general relationship between stochastic and lexical patterns or a potential implicational link between them from a sound change perspective.

**References**: **FELDMAN, N & T. L. Griffiths.** 2007. A rational account of the perceptual magnet effect. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. **HALÁCSY P., Kornai A., Németh L., Rung A., Szakadát I. & Trón V.** 2004. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference* 203-210. **MÁDY K.** 2010. Shortening of long high vowels in Hungarian: a perceptual loss? In *Proceedings of Sociophonetics at the crossroads of speech variation, processing and communication, Pisa*. **MÁDY, K. & U.D. Reichel.** 2007. Quantity distinction in the Hungarian vowel system—just theory or also reality? In *Proceedings of the 15th ICPhS, Saarbrücken*. **SIPTÁR P. & Törkenczy M.** (2000). *The Phonology of Hungarian.* The Phonology of the World's languages. Oxford University Press, New York.