Szabo, Ildiko Emese

A computational model of phonotactic acquisition

Predictability of exceptional patterns in Hungarian

While the topic of phonotactic generalizations is frequently examined and revisited in the literature, their exceptions are much more rarely dealt with. In the following, we present a typological classification of phonotactic exceptions and argue for a frequency- and perception-sensitive model of phonotactic acquisition. Its ability to predict types and relative extent of exceptional patterns will be illustrated and tested on Hungarian word-final vowel length data. Finally, we will explore the consequences of the model for both phonotactic representations and language change.

1 The data

Hungarian vowels of different height and length pattern differently in word final position. Word-final high rounded and mid vowels must be long (i.e. /uː, yː, oː, øː/ are permitted, /u, y, o, ø/ are not), whereas low vowels must be short (i.e. only /ɒ, ɛ/ are permitted, /aː, eː/ are not).

We find two different types of phonotactic exceptions—lexical exceptionality and vacillation[1]. For low vowels we find lexical exceptions (i.e. words that always end in long low vowels, such as *kávé* /kaːveː/ 'coffee', *örökké* /ørøkːeː/ 'forever', *burzsoá* /burʒoaː/ 'bourgeois'). Mid vowels seem to show very little exceptionality in production, but in perception the length distinction is fading Mády (2010). High vowels exhibit vacillation (vowels at the end of any word may be pronounced short) for speakers of Budapest Colloquial Hungarian (BCH, spoken mostly by younger, innovative speakers), but show lexical exceptions (such as *falu* /fɒlu/ 'village' or *áru* /aːru/ 'goods') for speakers of Standard Hungarian (SH, more typical of older, more conservative speakers).[2]

2 Modeling

2.1 The basic model—Blevins and Wedel (2009)

Blevins and Wedel (2009) propose an exemplar model employing a random walk algorithm. The relationship of two sounds can be modeled as two touching sets of tokens, A and B. Extreme values for A (i.e. variants of A that lie past the boundary between the two sets, and that are hence closer to set B than A) get sorted as instances of the other set, B. To model inconsistency in production they also add random noise.

2.2 Phonotactic implementation

First, we need to enrich their model with a measure of categorizational certainty. Tokens are not only assigned category labels, but also a value reflecting the model's certainty about its decision which correlates to its closeness to the boundary. Tokens that have a value below a certain threshold, will not be sorted into either category, but will remain unsorted. Unsorted tokens can therefore potentially form an intersections or overlaps at the category boundaries.

Learners are much more perceptually sensitive to contrasts that they have been exposed to frequently—i.e. functional load expands the perceptual space (Feldman and Griffiths, 2007)—a factor which is reflected in this model by a modifying categorizational certainty by functional-load. In order to avoid biases that can come from the unequal frequencies of the two sounds, we propose the measure of cumulative contrast load, which is calculated as given below. L denotes contrast load, A and B the two sounds, N(A-B)

---

[1] Borrowing the variational nomenclature of Hayes et al. (2009).

[2] For further description cf. Siptár and Törkenczy (2000); Mády and Reichel (2007); Mády (2010).

the instances of sound A in words that have a minimal pair contrasted by only sounds A and B in one position and N(A) the instances of sound A in the data set.

$$L_{Contrast}(A,B) = \frac{N(A-B)+N(B-A)}{N(A)+N(B)}$$

In Hungarian we find the same division for vowels with respect to cumulative length-contrast load as for their heights: high vowels have the lightest functional load (0.35%), mid vowels form the middle of this scale (2.3%) and low vowels displaying the heaviest load of length contrast (10.1%).[3] Applying its *-log* value as a decrease on categorizational certainty allows us to capture the generalization, that heavy functional load inhibits loss of contrast (Wedel et al., 2013). Language-specific perceptual data (Mády and Reichel, 2007) is implemented as a directed noise factor.

2.3 Resulting patterns

Our results show three types of configurations: disjoint, touching (with regularly occurring variant trading) and overlapping. These correspond exactly to the patterns we find in the linguistic data. In the case of disjoint sets we find no variation (e.g. for /o/ and /oː/). For touching sets with salient category boundaries, variation will only occur if the two sounds have different distributions (i.e. when different phonotactic restrictions apply to them). This is what we find for /i/ and /eː/ in Hungarian or for /u/ and /uː/ for speakers of SH. It is in the overlapping cases that we find vacillation, where categorization is not salient enough to consistently maintain a restriction (e.g. for /u/ and /uː/ for speakers of CBH).

3 Conclusion and future research

The proposed model reliably predicts the two different types of phonotactic exceptions that we find in linguistic data: lexical variation and vacillation. It also calls for a revision of how phonotactic restrictions are conceptualized: as an expression of a preference for one sound over another in a particular environment, rather than a constraint about a single sound formalized in absolute terms. By establishing a clear correspondence between frequency- and perception-based simulations and exceptional patterns attested in human languages, this model allows us to better address the relationship of phonology and phonetics. Our discussion also raises further questions about language change and its suggestion that lexical exceptionality (represented by touching sets) can be viewed as a precursor of potential vacillation and neutralization invites further investigation.

References

J. Blevins and A. Wedel. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica*, 26(2):143–183, 2009.

N. Feldman and T. L. Griffiths. A rational account of the perceptual magnet effect. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*, 2007.

P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón. Creating open language resources for hungarian. In *Proceedings of Language Resources and Evaluation Conference*, pages 203–210, 2004.

B. Hayes, K. Zuraw, P. Siptár, and Z. Cziráky Londe. Natural and unnatural constraints in hungarian vowel harmony. *Language*, 85:822–863, 2009.

K. Mády. Shortening of long high vowels in hungarian: a perceptual loss? In *Proceedings of Sociophonetics at the crossroads of speech variation, processing and communication, Pisa.*, 2010.

K. Mády and U. D. Reichel. Quantity distinction in the hungarian vowel system-just theory or also reality? In *Proceedings of the 15th ICPhS, Saarbrcken*, 2007.

P. Siptár and M. Törkenczy. *The Phonology of Hungarian.* The Phonology of the Worlds languages. Oxford University Press, New York, 2000.

A. Wedel, A. Kaplan, and S. Jackson. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186, 2013.

---

[3] Szószablya, Hungarian Webcorpus (Halácsy et al., 2004)