A computational model of phonotactic acquisition

Predictictability of exceptional patterns in Hungarian

Ildikó Emese Szabó
ildi.szabo@nyu.edu

New York University

OCP—Workshop on Variation in Phonology
13th of January, 2016

Word-final vowel length in Hungarian

	Front		Back	
	Rounded	Unrounded	Rounded	Unrounded
High	y yː	i iː	u uː	
Mid	ø øː		o oː	
Low		ɛ eː	ɒ	aː

Different phonotactic restrictions for different vowels (Nádasdy and Siptár, 1994; Siptár and Törkenczy, 2000; Törkenczy, 2006; Mády and Reichel, 2007).

- High rounded vowels: have to be long (preferably)
  - Standard Hungarian——*Lexical exceptions*
  - Budapest Colloquial Hungarian——*Free variation*
- Mid vowels: have to be long
  - *No exceptions*
- Low vowels: have to be short
  - *Lexical exceptions*

Outline

Phonotactic exceptionality: Previous approaches

Mostly treated as uninteresting and random

- Especially phonotactic exceptions

Treatments of other variation

- Rule-based
  - Categorical lexical exceptions (Chomsky and Halle, 1968 et seq.)
  - Variable rules (Labov, 1969; Cedergren and Sankoff, 1974):
    - Never proposed for phonotactics
    - No limitations to factors
- Constraint-based Grammars
  - Lexically-indexed constraints and co-phonologies (Anttila, 2002)
  - Added random variation (like Noisy Harmonic Grammar—Boersma and Pater, pear; Boersma and Weenink, 2010):
    - No connection between the two

Exceptionality: Exceptionality as variation

Two types of variation (Hayes et al., 2009; Rebrus and Törkenczy, 2013, 2015):

- Lexical variation: the appearance of the alternants is linked to specific lexical items.

  An example from phonotactics:
  [eː] is banned word-finally, but

matiné	[mɒtineː]	∼	*[mɒtinɛ]	'matinee'
lé	[leː]	∼	*[lɛ]	'juice'
kér-né	[keːrneː]	∼	*[keːrnɛ]	'he/she/it would ask for it'

  are possible words.

- Vacillation: Several forms are possible for a given word form (even as intraspeaker variation).

  An example from phonotactics:
  In BCH, supposedly only [yː] is allowed word-finally, but we find both [y] and [yː].

Main claims

1. The two types of phonotactic exceptions are conditioned by the same factors (perceivability and functional load of the contrast).
2. The two types of exceptions have distinct configurations.

However:
This model does not predict *exact* amounts of variations nor *definite* appearance of exceptions.

Contrast Limitation Model (CLM):Theoretical background

Lexicon (the input of the learner):

- A collection of tokens sorted into word forms

Functional load

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

Where A and B are sounds, N(A) is the number of tokens of A, and N(A;B) is the number of tokens of words containing A, but if that A is replaced by B, it would still yield a possible word.

- Significant differences between high, mid and low vowels (light, intermediate and heavy functional load, respectively)—Halácsy et al. (2004).

Perception

- Often-used contrasts are more prominent in perception (Liberman et al., 1957; Studdert-Kennedy et al., 1970)
- Functional load expands the perceptual space (Feldman and Griffiths, 2007)

Phonotactic strength

- Non-categorical preference of one member of a pair of sounds in a given environment
- Examples:

AC > BC	(*BC)
A# > B#	(*B#)

CLM: The Algorithm

1. Generate the lexicon—two sounds (*A and B*) with randomly generated set of token based on productional data (F1-, F2-data from Mády and Reichel (2007))
2. *n*-many iterations (here, 5000) in each of which two new tokens are generated: one based on A, one based on B (both with random noise)
    1. For each new token point, a *certainty value* is calculated based on:
        1. The *distance* of the token from the average of A and of the average of B
        2. The functional load of A and B ($ln(L_{CONTRAST})$)
    2. If this certainty value meets the threshold value *t*, the token is sorted into either A or B (into the set whose average is closer).
    3. If the certainty value of the point is below a threshold value *t*, the token is sorted into not A or B but set C.
    4. After *q*-many iterations, a random element is deleted from each set (forgetting).

CLM: Results I.

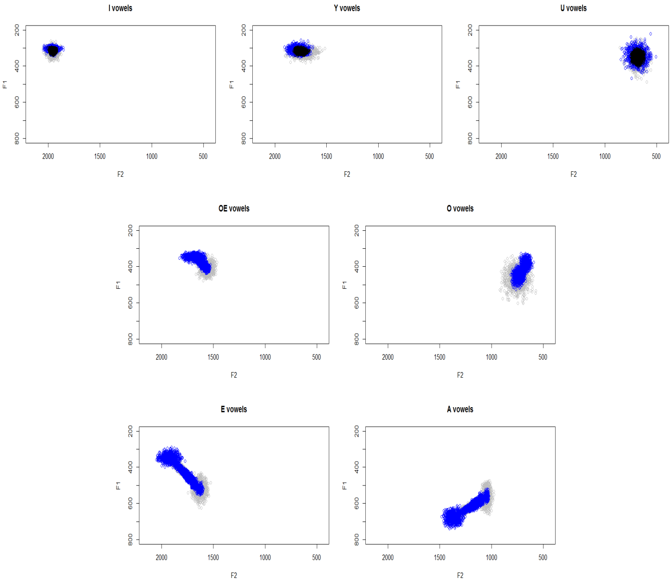
Vowels		Short	Long	Undecided
U	(u – uː)	470.44	473.02	8907.46
Y	(y – yː)	1511.86	1376.72	6962.42
I	(i – iː)	916.2	795.88	8138.8
O	(o – oː)	2393.77	2235.68	5230.56
OE	(ø – øː)	2606.21	2610.99	4648.3
A	(ɒ – aː)	4951.6	4948.4	0.0
E	(ɛ – eː)	4939.59	4960.41	0.0

Set threshold: -3.7

Vowels		Short	Long	Undecided
U	(u – uː)	899.41	898.32	8058.41
Y	(y – yː)	1768.1	1636.13	6447.06
I	(i – iː)	1167.34	1002.3	7681.32
O	(o – oː)	5039.81	4860.19	0.0
OE	(ø – øː)	4939.42	4960.58	0.0
A	(ɒ – aː)	4944.55	4955.45	0.0
E	(ɛ – eː)	4938.24	4961.76	0.0

Set threshold: -4.7

CLM: Results II.



Vacillation

With CLM, there is one consistent generalization for Hungarian:

$$N(undecided_{HIGH}) \geq N(undecided_{MID}) \geq N(undecided_{LOW})$$

Conditions of vacillation:

- Light functional load
- Low perceivability
- Can be simulated as categorizational failure

Lexical exceptions

Two cases:

- Low vowels
- High rounded vowels (in Standard Hungarian)

These are cases where

- Salient categories are close to each other

  [eː] and [i(ː)]
  [u] and [uː], [y] and [yː]
- Variant trading for the sake of phonetic consistency (unlabeled input will sometimes be mislabeled to preserve the categories—Blevins and Wedel (2009))
- In this case, the two categories have different phonotactic distributions

As a result, generalizations on the sounds' distributions (phonotactic restrictions) will be less categorical.

Conclusions

1. The two types of phonotactic exceptions are conditioned by the same factors
   - Exceptionality conditioned by phonetics—influenced by distribution: modular approaches cannot be maintained
2. The two types of exceptions have distinct configurations
   - Vacillation
     - ★ Light functional load
     - ★ Low perceivability
     - ★ Categorizational failure
   - Lexical exceptions
     - ★ Salient categories close to each other
     - ★ Different phonotactic distribution
     - ★ Regular categorizational mistakes (variant trading)

Topics for future research

- Generalizing the algorithm to consonants and to other languages
- Generalizability of the patterns to other areas of phonology
- Phonology vs. phonetics (distribution-sensitive, language-specific perception)
- Vacillation and lexical variation from a diachronic perspective—implicational relationship

Thank you!

References

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

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

Boersma, P. and Pater, J. (to appear). Convergence properties of a gradual learning algorithm for harmonic grammar. In McCarthy, J. and Pater, J., editors, *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Boersma, P. and Weenink, D. (2010). {P} raat: doing phonetics by computer. Computer program.

Cedergren, H. J. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row Publishers, New York.

Feldman, N. and Griffiths, T. L. (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., and Trón, V. (2004). Creating open language resources for hungarian. In *Proceedings of Language Resources and Evaluation Conference*, pages 203–210.

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

Labov, W. (1969). Contraction, deletion, and inherent variability of the english copula. *Language*, 45:715–762.

Liberman, A. M., Safford Harris, K., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–68.

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

References (cont.)

Nádasdy, A. and Siptár, P. (1994). A magánhangzók. In Ferenc, K., editor, *Strukturális magyar nyelvtan*, volume 2. Fonológia, pages 42–182. Budapest: Akadémiai Kiadó.

Rebrus, P. and Törkenczy, M. (2015). Monotonicity and the limits of disharmony. In *LSA Proceedings of the 2014 Annual Meetings on Phonology*. Linguistic Society of America.

Rebrus, P. and Törkenczy, M. (22nd October 2013). Magánhangzó–diszharmónia (vowel disharmony). In *Talk at RIL HAS, Budapest, Hungary.*

Siptár, P. and Törkenczy, M. (2000). *The Phonology of Hungarian*. The Phonology of the World's languages. Oxford University Press, New York.

Studdert-Kennedy, M., Liberman, A., Harris, K., and Cooper, F. (1970). Motor theory of speech perception: a reply to lane's critical review. *Psychological Review*, 77:234–249.

Törkenczy, M. (2006). *The Phonotactics of Hungarian*. Dsc dissertation, Hungarian Academy of Sciences, Budapest.

Functional load: Cumulative Contrast Load

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

- Token-sensitive: as opposed to type-based
- Observed over possible: not biased by the frequency of the given sound
- Pair-wise comparison: not sensitive to the bias of one member of the sound-pair being more frequent than the other one

Vowels	Cumulative Contrast Load	By percentage
u – uː	0.009191649	0.9192%
y – yː	0.004174471	0.4174%
i – iː	0.003240602	0.3241%
o – oː	0.020582427	2.0582%
ø – øː	0.022096762	2.2097%
ɒ – aː	0.05981946	5.9819%
ɛ – eː	0.128850935	12.8851%

Data from SzóSzablya Hungarian Webcorpus, Halácsy et al. (2004)