Comparing Learning Models for Korean Sound-symbolic Vowel Harmony

Darrell Larsen and Jeffrey Heinz

March 20, 2010

PLC 34

Main Goals of Presentation

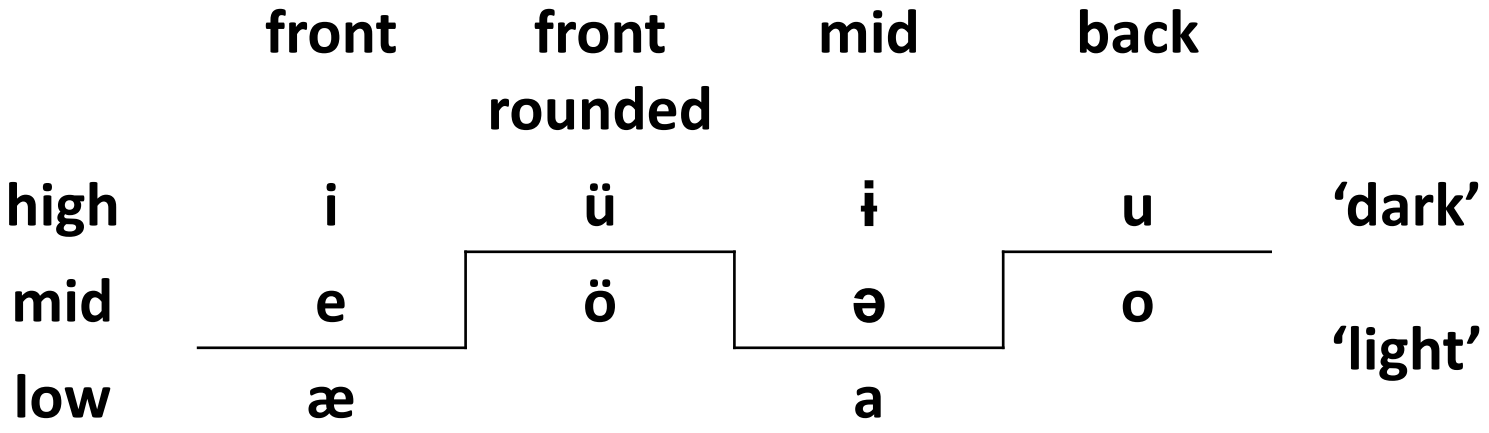
1. Provide quantitative support for vowel harmony in sound-symbolic forms in Korean

2. Establish that [u] behaves like transparent vowels [i] and [ɨ] (Cho, 1994), and to a lesser extent, [ü]

3. Pinpoint challenges for specific learning proposals (tier-based bigram and precedence)

Sound-symbolic Harmony

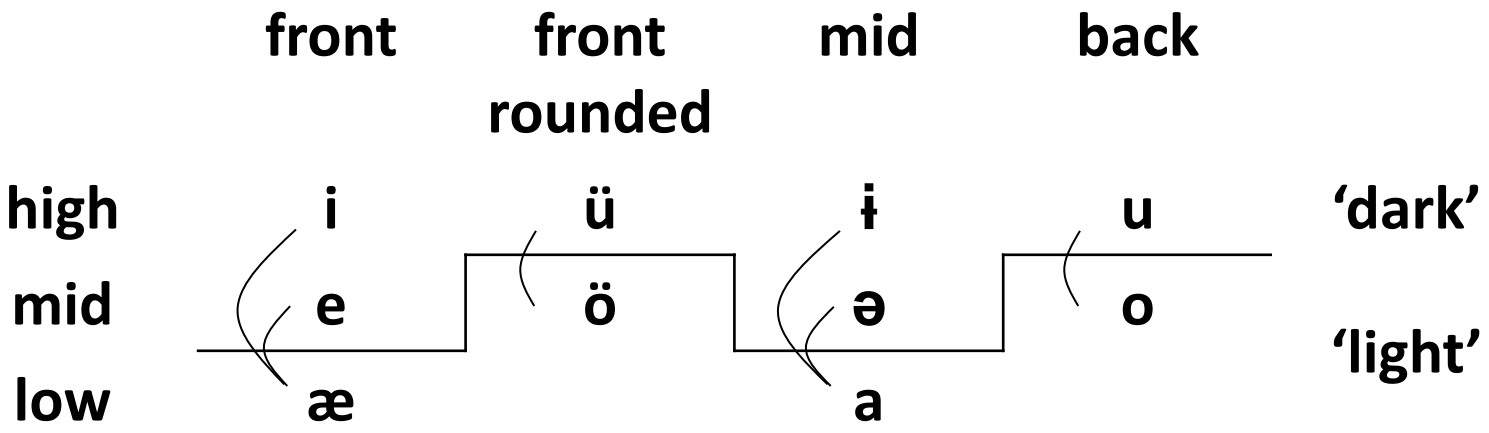
- Vowel harmony in sound-symbolic morphemes



- [i] and [ɨ] are 'dark' in initial position, transparent in noninitial position (Kim-Renaud 1976, Cho 1994, inter alia)

Sound-symbolic Harmony

- light-dark pairs (Kim-Renaud 1976)



Sound-symbolic Harmony

Connotations in sound-symbolic words

‘light’	brightness, lightness, sharpness, quickness, smallness, thinness
‘dark’	darkness, heaviness, dullness, slowness, deepness, thickness

Examples

‘dark’ vowels	[pʰuŋdəŋ]	‘splash’ (e.g. person falling into water)
‘light’ vowels	[pʰoŋdaŋ]	‘splash’ (e.g. a small stone falling into water)
‘dark’ vowels	[pənccək]	‘sparkling, twinkling’ (e.g. flash of light)
‘light’ vowels	[panccak]	‘sparkling, twinkling’ (e.g. stars)

Questions for Corpus Study

1. Is VH robust within sound-symbolic reduplicant morphemes phonotactically?

2. Do transparent vowels and [u] behave as ‘dark’ vowels in initial position?

3. Does [u] behave as a transparent vowel in noninitial position?

4. Does [ü] also behave as a neutral vowel?

About the Corpus

- Designed to aid the National Institute of the Korean Language's development of 'The Great Standard Korean Dictionary' (표준국어대사전) http://www.hangeul.pe.kr/symbol/words.htm
- Original corpus contains 29,000 entries of sound-symbolic words.
- Many are variants built on same underlying sound-symbolic form.
- Only one token of each sound-symbolic form was taken
- For ease of extraction, and to minimize possibility of non-sound symbolic words from entering, only reduplicants were selected
- Only reduplicants of 2 or 3 syllables (pre-reduplication) were used.
- Reduplicants containing diphthongs not traditionally discussed in VH literature were excluded (e.g. [wa] 와…)
- Total of 4,006 such sound-symbolic reduplicants were found.

Types of Reduplication

1) reduplication of one-syllable forms

sal-sal ‘gently, softly; slowly’

2) reduplication of two-syllable forms

curəŋ-curəŋ ‘in clusters’ (e.g. grapes hanging ~)

3) reduplication of three-syllable forms

harɨrɨ-harɨrɨ ‘thin and soft texture’ (e.g. paper, cloth)

4) reduplication of first syllable onto second, and of third syllable onto fourth

c^{h}ikc^{h}ikp^{h}okp^{h}ok ‘chugga chugga’ (e.g. train)

Q1) Is vowel harmony robust in sound-symbolic reduplicants?

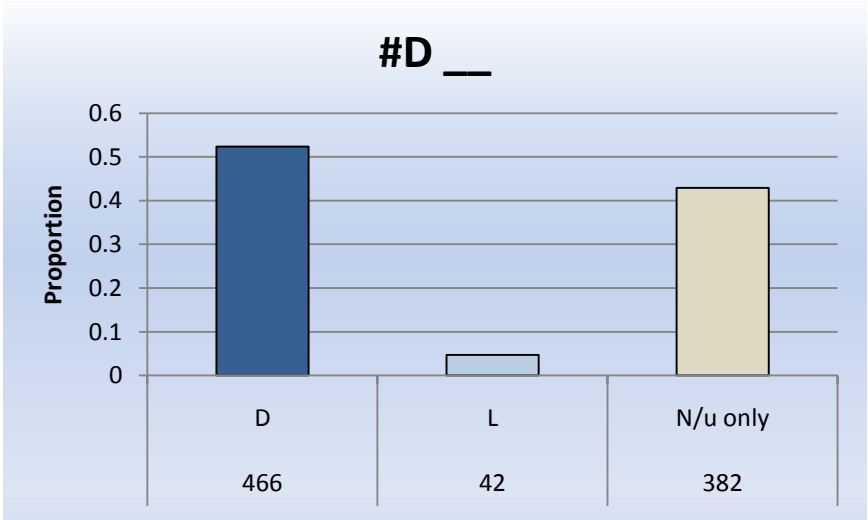
- Out of 4,006 morphemes, only 3.4% contain both L and D vowels. This is when counting initial N vowels as D.

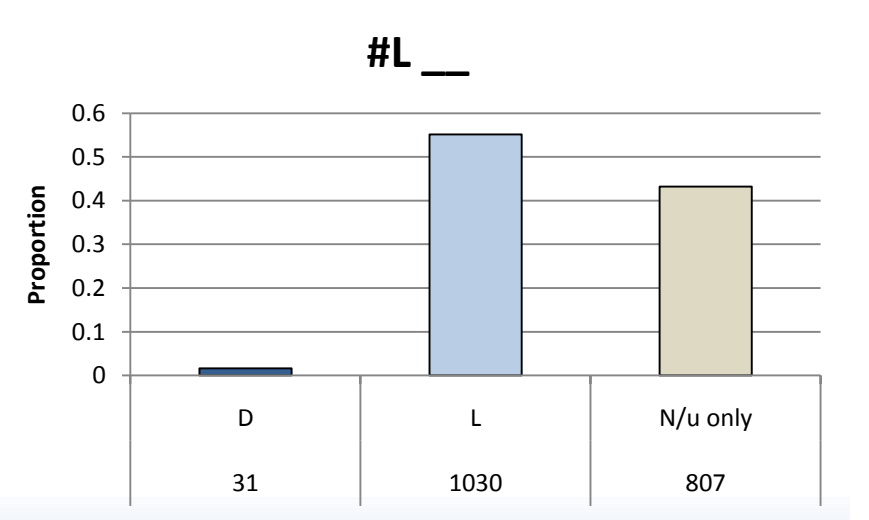
¬ #__ / #__	L [a] 아 (925)	L [o] 오 (223)	L [æ] 애 (33)	L [ö] 외 (0)	D [ə] 어 (973)	D [e] 에 (1937)	D [ü] 위 (10)
L [a] 아 (952)					16	3	3
L [o] 오 (605)					3	3	
L [æ] 애 (281)					3		
L [ö] 외 (27)							
D [ə] 어 (769)	31						
D [e] 에 (85)	10						
D [ü] 위 (36)	2						
D [u] 우 (647)	28		2				
D [i] 이 (378)	21	3					
D [ɨ] 으 (226)	7		1				

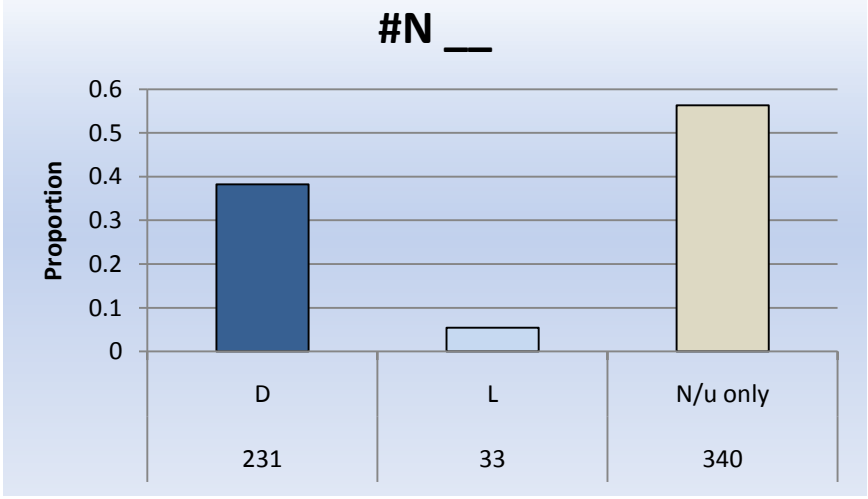
Q2) Do neutral vowels behave as ‘dark’ vowels in initial position?

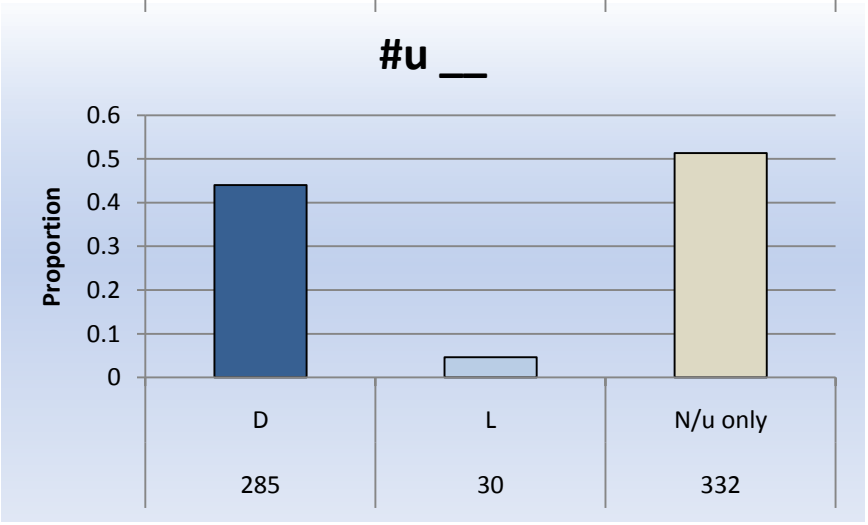
- If so:
  i. should allow D, N vowels to follow
  ii. should not allow L vowels to follow

Q2) Do neutral vowels behave as ‘dark’ vowels in initial position?





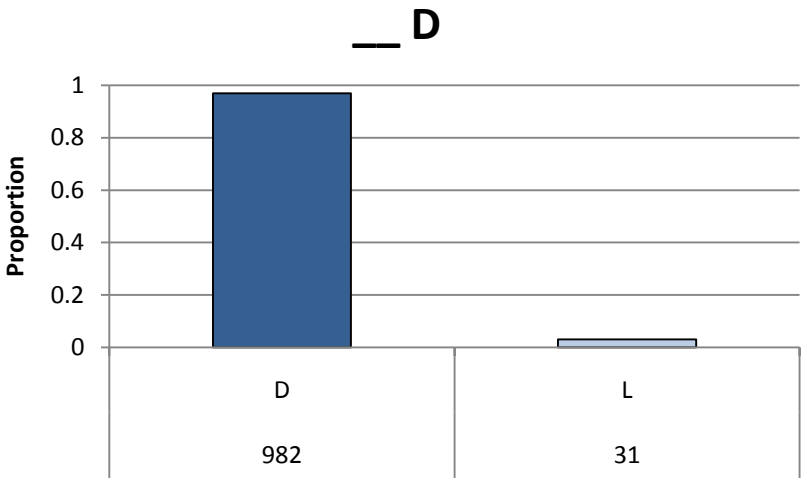


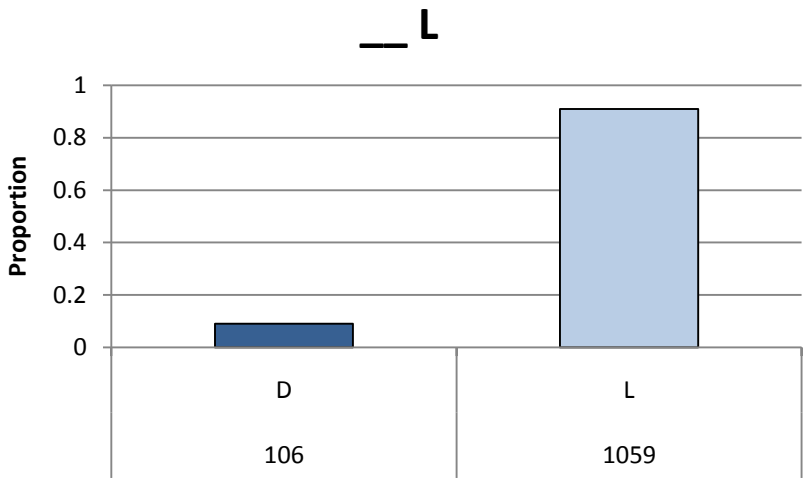


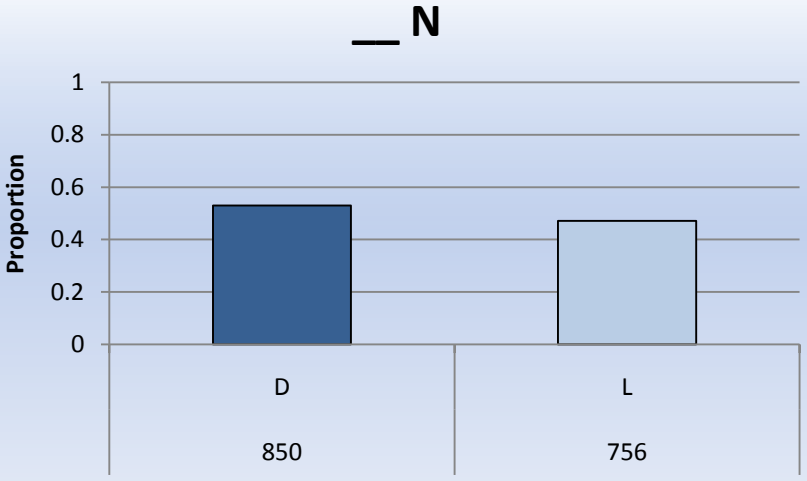
Q3) Does [u] behave as a neutral vowel in noninitial position?

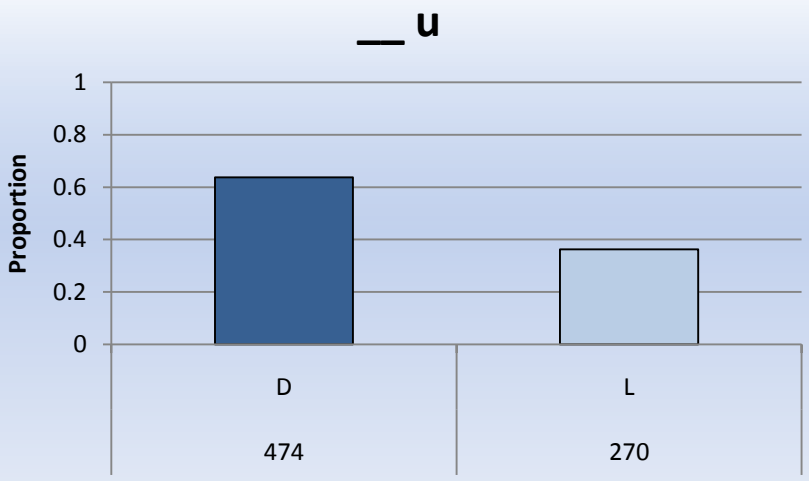
- If so:
  i. should appear after both L and D vowels
  ii. in 3-syllable words, should allow harmony to pass over it

Q3i) Does noninitial [u] appear after both D and L?

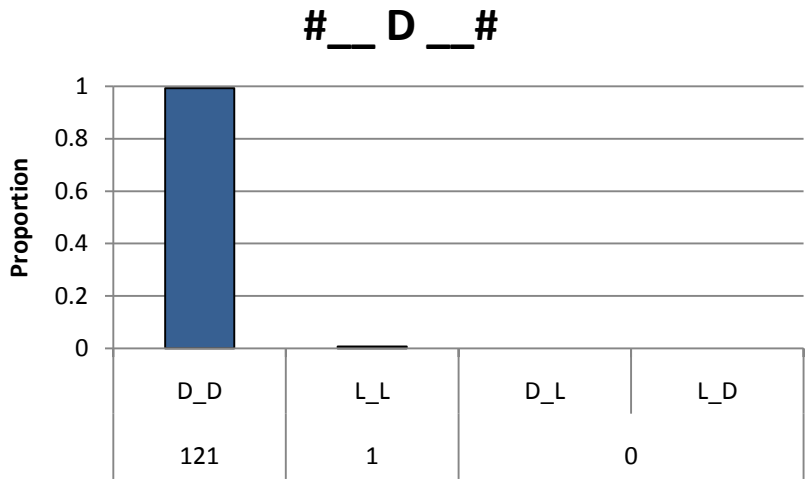


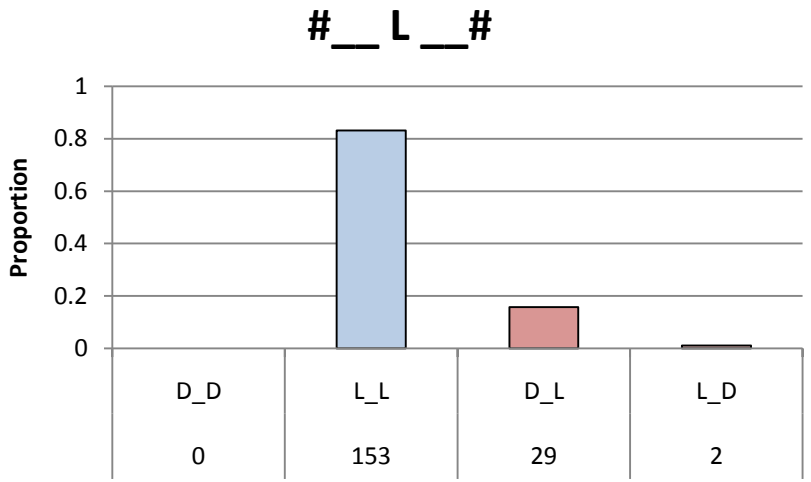


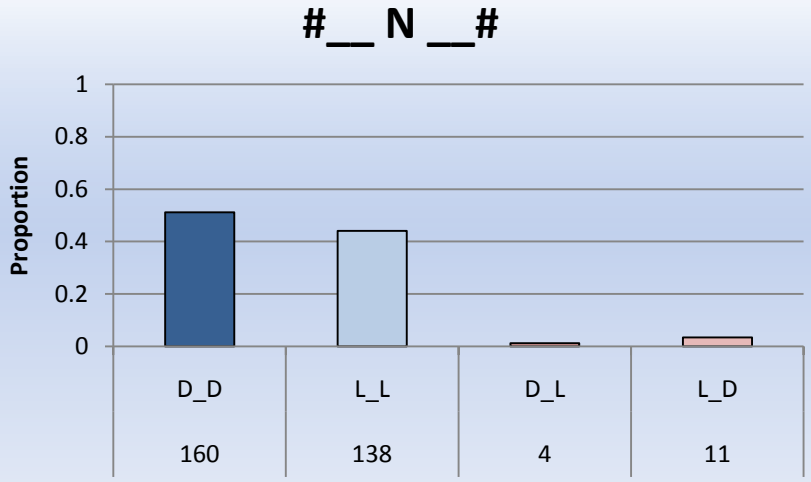


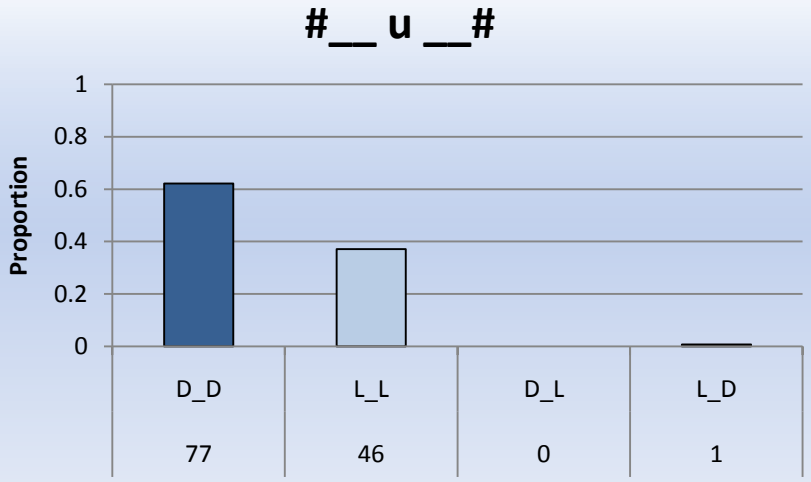


Q3ii) Does [u] allow harmony to pass over it?







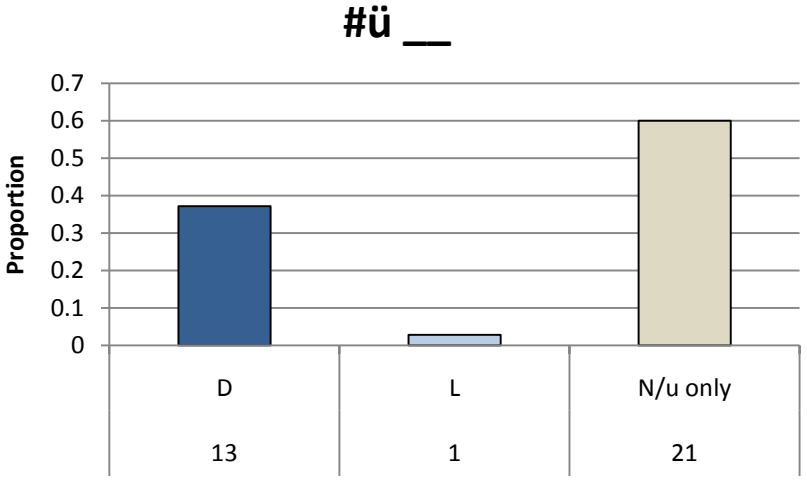


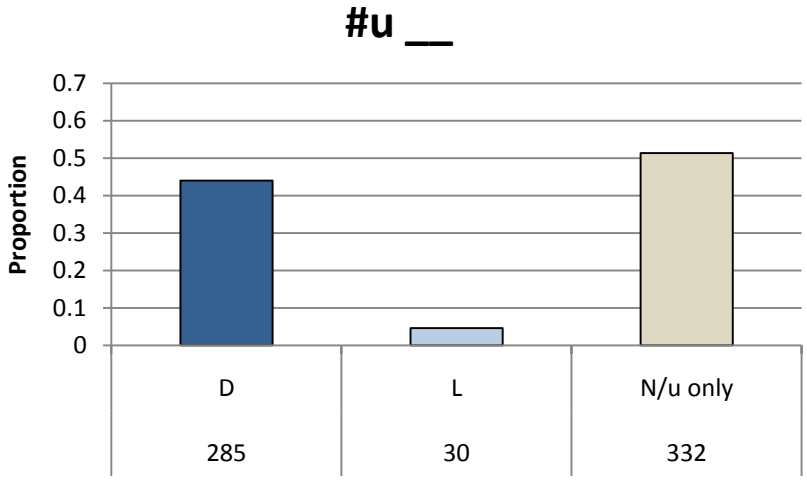
Status of [ü]

- The remaining [+high] vowel appears to behave like [u] as well.
- Only limited data (46/4,006 forms contain [ü])

Status of [ü]

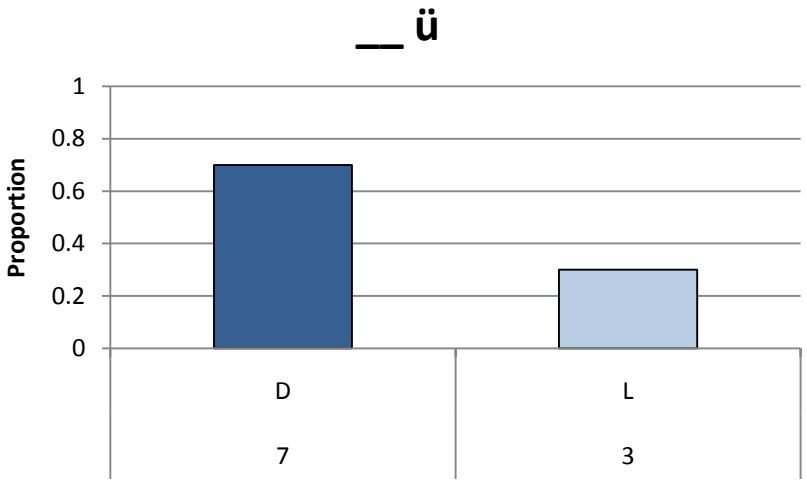
- In initial position:

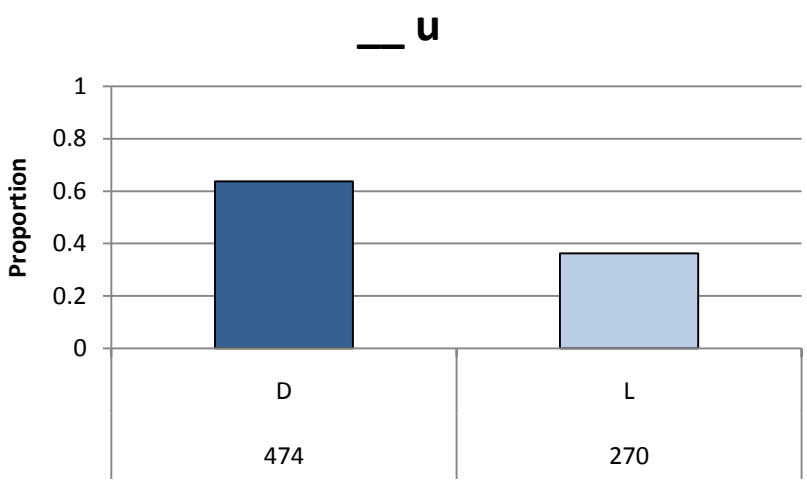




Status of [ü]

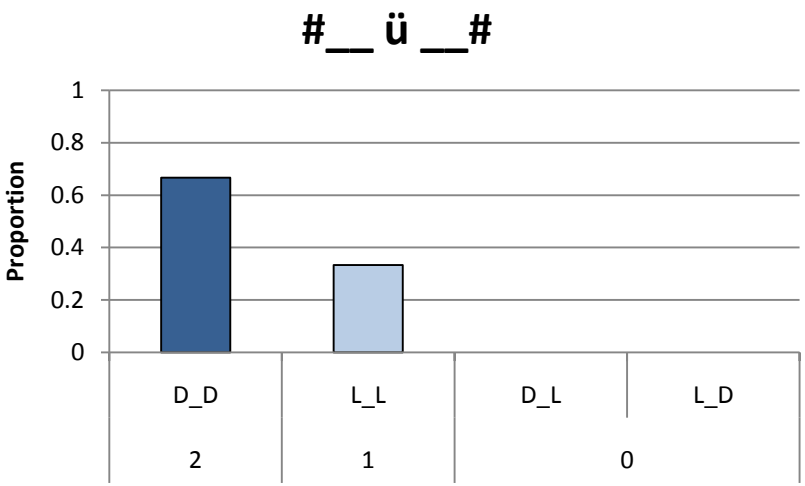
- In noninitial position:

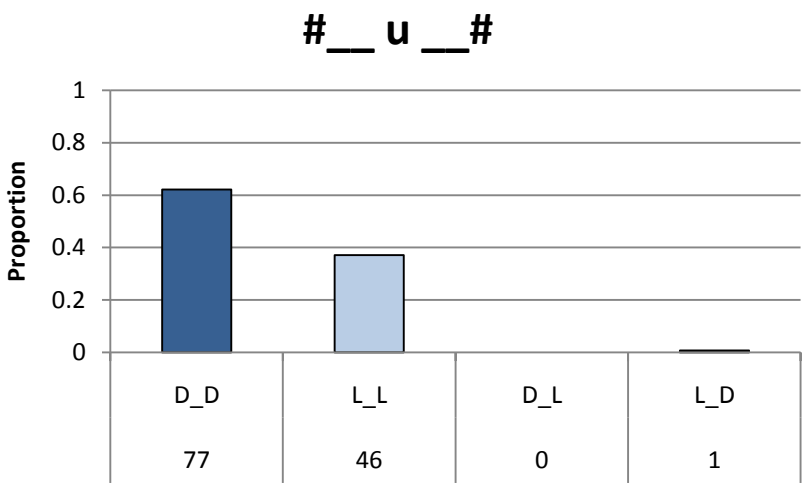




Status of [ü]

- Medial position – can VH pass over [ü]?





Part 1: Conclusions

1. VH is strongly attested in Korean sound-symbolic reduplicant base morphemes

2. [u] behaves like transparent vowels in noninitial position, supporting Cho (1994)

3. [i, ɨ, u] behave as ‘dark’ vowels in initial syllables, but as transparent vowels in noninitial syllables

4. [ü] appears to behave like [u], though more data is needed

Part 2: Learning Models

- Can existing learning models account for the behavior of neutral vowels?

- Previous approaches to long-distance learning:
  - Bigram learner applied over a vowel tier (Hayes and Wilson (2008), Goldsmith and Xanthos (2009), Goldsmith and Riggle (to appear))

  - Precedence learner (Rogers et al. (2009), Heinz to appear, Heinz and Rogers, under review)

Bigram Learner

- Categorical version:

Time	Word	Bigrams	Grammar
0			∅
1	NDD	{#N, ND, DD, D#}	{ **#N, ND, DD, D#** }
2	LNL	{#L, LN, NL, L#)	{ #N, ND, DD, D#, **#L, LN, NL, L#** }
3	DDN	{#D, DD, DN, N#}	{ #N, ND, DD, D#, #L, LN, NL, L#, **#D, DN, N#** }

Bigram Learner

Grammar_{VH} =

#D	DD		DN	D#
#L		LL	LN	L#
#N	ND	NL	NN	N#

- fails to capture vowel harmony over transparent vowels

  allows: *LND *DNL

  LN+ND DN+NL

- fails to distinguish between initial and noninitial N

  allows: #LNL *#NLL

  #L+LN+NL #N+NL + LL

Bigram Learner

- A trained probabilistic bigram learner (Jurafsky & Martin, 2008) also fails to make the right distinctions:

Word	Prob(word)
L N L	0.003611
D N D	0.006353
L N D	0.007325
D N L	0.003132
N D D	0.001942
N L L	0.001178

Precedence Learner

- Categorical version (Heinz 2007, to appear):

Time	Word	Precedence Relations	Grammar
0			∅
1	NDD	{#…N, #…D, N…D, D…D, D…#, N…#}	{ **#…N, #…D, N…D, D…D, D…#, N…#** }
2	LNL	{#…L, #…N, L…N, N…L, L…L, L…#, N…#)	{ #…N, #…D, N…D, D…D, D…#, **#…L, L…N, N…L, L…L, N…#, L…#** }
3	DDN	{#…D, #…N, D…D, D…N, D…#, N…#}	{ #…N, #…D, N…D, D…D, D…#, #…L, L…N, N…L, L…L, N…#, L…#, **D…N** }

Precedence Learner

$$\text{Grammar}_{VH} = \left\{ \begin{array}{lllll} \#...D & D...D & & D...N & D...\# \\ \#...L & & L...L & L...N & L...\# \\ \#...N & N...D & N...L & N...N & N...\# \end{array} \right\}$$

- allows harmony to spread without a vowel tier

  D…x…D L…x…L

- and disallows disharmonious sequences with transparent vowel intervening

  *D…N…L *L…N…D

- but fails to distinguish between initial and noninitial N

  #LNL *#NLL

  L…N, N…L, L…L N…L, L…L

Precedence Learner

- A trained precedence learner (Heinz & Rogers, under review) learns the transparency of noninitial N vowels, but not the behavior of initial-syllable N vowels.

Word	Prob(word)
L N L	0.002893
D N D	0.004357
L N D	0.000142
D N L	0.000255
N D D	0.001867
N L L	0.000657

Part 2: Conclusion

- The tier-based bigram learner fails to learn what the precedence learner is able to learn: the transparency of noninitial N vowels.

- Neither the bigram learner nor the precedence learner can account for bi-functionality of ‘neutral’ vowels in Korean VH

Potential solution for tier-based bigram learner

- N vowels only project to harmony tier if initial
- Captures transparency for noninitial N vowels because they are not on the tier
- Captures behavior of initial N because it learns that NL sequences are absent on the tier

- But… How do you learn which vowels are N?

Potential solution for precedence learner

- Treat initial vowels differently
- The learner realizes N_1…L is bad but N_2…L is OK.
- Sounds at word boundaries frequently behave differently (Endress 2009)
- But the learner also learns D_1 and D_2 behave the same, etc. Seems to be missing the right generalization.

Conclusions

1. Vowel harmony in sound-symbolic forms in Korean is robust in the phonotactics.

2. [u] behaves like the transparent vowels [i, ɨ]; [ü] appears to as well.

3. A precedence learner is better suited to capture vowel harmony over transparent vowels than a tier-based bigram learner; however, both learners fail to capture the bifunctionality of N vowels in Korean.

Acknowledgments

- Phonology and Phonetics Lab Group at the University of Delaware, Karthik Durvasula, Bill Idsardi, James Rogers

Thank you for listening!

Questions?

References

Cho, Mi-Hui. (1994). “Vowel Harmony in Korean: A Grounded Phonology Approach.” Ph.D. dissertation. Indiana University.

Goldsmith, John and Jason Riggle. (to appear). “Information theoretic approaches to phonological structure: the case of Finnish vowel harmony.” In *Natural Language and Linguistic Theory*.

Goldsmith, John and Aris Xanthos. (2009). “Learning Phonological Categories.” In *Language*, vol. 85, no. 1, pp. 4-38.

Hayes, Bruce and Colin Wilson. (2008). “A Maximum Entropy Model of Phonotactics and Phonotactic Learning.” In *Linguistic Inquiry*, vol. 39, no. 3, pp. 379-440.

Heinz, Jeffrey. (to appear). “Learning Long Distance Phonotactics.” Manuscript. University of Delaware. *Linguistic Inquiry*.

Heinz, Jeffrey. (2007). “Inductive Learning of Phonotactic Patterns.” Ph.D. dissertation. University of California - Los Angeles.

Heinz, Jeffery and J. Rogers (under review). “Estimating Strictly Piecewise Distributions.”

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.

Kim-Renaud, Young-Key. (1976). “Semantic Features in Phonology: Evidence from Vowel Harmony in Korean.” In *Chicago Linguistic Society* Vol. 12, pp. 397-412.

Rogers J., Heinz J., Bailey G., Visscher M., Wellcome D., Edlefsen M., and Wibel S. (to appear). “On Languages Piecewise Testable in the Strict Sense.” In *Proceedings of the 11th Meeting of the Association of Mathematics of Language*.