A corpus study and comparative analysis of formal learning proposals of Korean sound-symbolic vowel harmony

Darrell Larsen and Jeffrey Heinz
dlarsen,heinz@udel.edu

University of Delaware

University of Tokyo
July 26, 2010

Acknowledgments

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

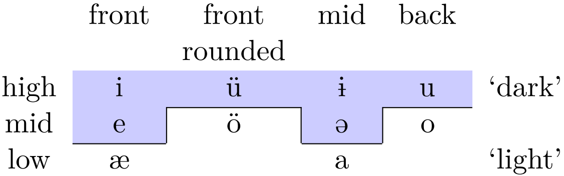
Main Goals of Presentation

1. Provide quantitative support for vowel harmony in sound-symbolic forms in Korean
2. Establish that [u] behaves like neutral vowels [i] and [ɨ] (Cho 1994), and to a lesser extent, [ü] (Lee 1984)
3. Compare specific learning proposals:
   3.1 tier-based bigram learners (Hayes and Wilson 2008, Goldsmith and Riggle, to appear)
   3.2 strictly piecewise learners (Heinz 2010, in press, Heinz and Rogers 2010)

Main Goals of Presentation

1. Provide quantitative support for vowel harmony in sound-symbolic forms in Korean
2. Establish that [u] behaves like neutral vowels [i] and [ɨ] (Cho 1994), and to a lesser extent, [ü] (Lee 1984)
3. Compare specific learning proposals:
   3.1 tier-based bigram learners (Hayes and Wilson 2008, Goldsmith and Riggle, to appear)
   3.2 strictly piecewise learners (Heinz 2010, in press, Heinz and Rogers 2010)

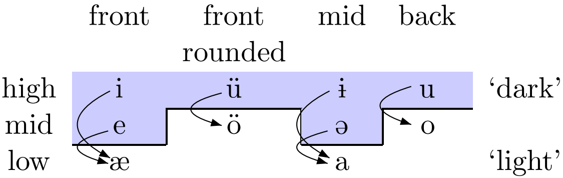
Vowel harmony in sound-symbolic morphemes



	front	front rounded	mid	back	
high	i	ü	ɨ	u	‘dark’
mid	e	ö	ə	o	
low	æ		a		‘light’

- Vowels [i] and [ɨ] are ‘dark’ in initial syllables, transparent in noninitial syllables (Kim-Renaud 1976, Cho 1994, inter alia)
- Notation: #V denotes vowels in initial syllbles

Light-dark pairs (Kim-Renaud 1976)



	front	front rounded	mid	back	
high	i	ü	ɨ	u	‘dark’
mid	e	ö	ə	o	
low	æ		a		‘light’

Connotations in sound-symbolic words

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

- Examples

‘dark’	[p^{h}uŋdəŋ]	‘splash’ (e.g. person falling into water)
‘light’	[p^{h}oŋdaŋ]	‘splash’ (e.g. a small stone falling into water)
‘dark’	[pənccək]	‘sparkling, twinkling’ (e.g. flash of light)
‘light’	[panccak]	‘sparkling, twinkling’ (e.g. stars)

Questions for Corpus Study

1. Is VH robust within sound-symbolic reduplicant morphemes phonotactically?
2. Are the vowels [i] and [ɨ] transparent in noninitial syllables and ‘dark’ in initial syllables?
3. Do [u] and [ü] behave as transparent vowels in noninitial syllables?
4. Do [u] and [ü] behave as ‘dark’ vowels in initial syllables?

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.
- At most one token of each sound-symbolic form was taken.
- Only reduplicating morphemes were selected for ease of extraction and to minimize possibility of non-sound symbolic words from entering
- Only morphemes of 2 or 3 syllables (pre-reduplication) were used.
- Morphemes containing diphthongs not traditionally discussed in VH literature were excluded (e.g. [wa] 와 . . .)
- Total of 4,006 such sound-symbolic morphemes 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. of 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?

ㄱ_# #_	[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				

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

If so:

1. should allow D, N vowels to follow
2. should not allow L vowels to follow

Do neutral vowels behave as 'dark' vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D. . .	466	42
#L. . .	31	1030
#N. . .	231	33
#u. . .	285	30

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D. . .	466	42
#L. . .	31	1030
#N. . .	231	33
#u. . .	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D. . . and #N. . .	p=0.07919	

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D…	466	42
#L…	31	1030
#N…	231	33
#u…	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D…and #N…	p=0.07919	
#D…and #u…	p=0.622	

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D…	466	42
#L…	31	1030
#N…	231	33
#u…	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D…and #N…	p=0.07919	
#D…and #u…	p=0.622	
#N…and #u…	p=0.3118	

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D. . .	466	42
#L. . .	31	1030
#N. . .	231	33
#u. . .	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D. . . and #N. . .	p=0.07919	
#D. . . and #u. . .	p=0.622	
#N. . . and #u. . .	p=0.3118	
#L. . . and #D. . .	p<2.2 × 10e-16	*

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D. . .	466	42
#L. . .	31	1030
#N. . .	231	33
#u. . .	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D. . . and #N. . .	p=0.07919	
#D. . . and #u. . .	p=0.622	
#N. . . and #u. . .	p=0.3118	
#L. . . and #D. . .	p<2.2 × 10e-16	*
#L. . . and #N. . .	p<2.2 × 10e-16	*

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

Do neutral vowels behave as ‘dark’ vowels in initial position? Does [u]?

Morpheme frequency of which vowel types in initial syllables precede dark or light vowels.

	D	L
#D. . .	466	42
#L. . .	31	1030
#N. . .	231	33
#u. . .	285	30

Pairwise comparison of rows with chi-square tests (df=1)

#D. . . and #N. . .	p=0.07919	
#D. . . and #u. . .	p=0.622	
#N. . . and #u. . .	p=0.3118	
#L. . . and #D. . .	$p<2.2 \times 10e\text{-}16$	*
#L. . . and #N. . .	$p<2.2 \times 10e\text{-}16$	*
#L. . . and #u. . .	$p<2.2 \times 10e\text{-}16$	*

- #L differs significantly from #D,#N, and #u, but #D,#N, and #u do not differ significantly from each other.

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

If so

1. should appear after both L and D vowels
2. in 3-syllable words, should allow harmony to pass over it

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
. . . D	982	31
. . . L	106	1059
. . . N	850	756
. . . u	474	270

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
…D	982	31
…L	106	1059
…N	850	756
…u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

…D and …L	$p<2.2 \times 10e\text{-}16$	*

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
…D	982	31
…L	106	1059
…N	850	756
…u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

…D and …L	p<2.2 × 10e-16	*
…D and …N	p<2.2 × 10e-16	*

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
...D	982	31
...L	106	1059
...N	850	756
...u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

...D and ...L	$p<2.2 \times 10e\text{-}16$	*
...D and ...N	$p<2.2 \times 10e\text{-}16$	*
...D and ...u	$p<2.2 \times 10e\text{-}16$	*

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
…D	982	31
…L	106	1059
…N	850	756
…u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

…D and …L	p<2.2 × 10e-16	*
…D and …N	p<2.2 × 10e-16	*
…D and …u	p<2.2 × 10e-16	*
…L and …N	p<2.2 × 10e-16	*

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
…D	982	31
…L	106	1059
…N	850	756
…u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

…D and …L	p<2.2 × 10e-16	*
…D and …N	p<2.2 × 10e-16	*
…D and …u	p<2.2 × 10e-16	*
…L and …N	p<2.2 × 10e-16	*
…L and …u	p<2.2 × 10e-16	*

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

Morpheme frequency of which vowel types follow dark or light vowels.

	D	L
…D	982	31
…L	106	1059
…N	850	756
…u	474	270

Pairwise comparison of rows with chi-square tests (df=1)

…D and …L	$p<2.2 \times 10e\text{-}16$	*
…D and …N	$p<2.2 \times 10e\text{-}16$	*
…D and …u	$p<2.2 \times 10e\text{-}16$	*
…L and …N	$p<2.2 \times 10e\text{-}16$	*
…L and …u	$p<2.2 \times 10e\text{-}16$	*
…N and …u	$p<1.187 \times 10e\text{-}6$	*

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

	D	L
...D	982	31
...L	106	1059
...N	850	756
...u	474	270

Morpheme frequency of which vowel types follow dark or light vowels.

- In non-initial syllables N and [u] are significantly different than D and L vowels.
- N and [u] are significantly different from each other probably because noninitial [u] is preceded by L vowels about half as much as noninitial N vowels are.
- Probably due to historical reasons: the dark vowels, including [u], only occurred with neutral and dark vowels; the change appears to be a modern development (see, e.g., Lee 1984, Cho 1994).
- The evidence is consistent with the claim that [u] is much more like N than either D or L.

Q3.2: Does [u] allow harmony to pass over it?

Morpheme frequency of which vowel types occur between harmonizing vowels.

	D_D	L_L
D	121	1
L	0	153
N	160	138
u	77	46

Pairwise comparison of rows with chi-square tests (df=1)

D and L	$p<2.2 \times 10e\text{-}16$	*
D and N	$p<2.2 \times 10e\text{-}16$	*
D and u	$p<2.2 \times 10e\text{-}16$	*
L and N	$p<2.2 \times 10e\text{-}16$	*
L and u	$p<2.2 \times 10e\text{-}16$	*
N and u	$p<1.187 \times 10e\text{-}6$	*

- As before, the evidence is consistent with the claim that [u] is much more like N than either D or L.

Status of [ü]

- The remaining [+high] vowel appears to behave like [u] as well.
- Caution: only limited data (46/4,006 forms contain [ü]). Chi-square tests are unreliable.

	D	L
#ü. . .	13	1
#u. . .	285	30
. . . ü	7	3
. . . u	474	270
	D_D	L_L
ü	2	1
u	77	46

Status of [ü]

- None of the chi-square tests reveal a significant difference between
  1. the vowel [ü] as compared to [u] and N in any of the cases considered
  2. the vowel [ü] in initial syllables and D vowels in initial syllables
- The chi-square tests do reveal significant differences between
  1. noninitial [ü] and noninitial D and L vowels.
  2. initial [ü] and initial L vowels
- Although the results of these tests may be inaccurate, they are 100% consistent with the hypothesis that [ü] is a neutral vowel, which is transparent noninitially and a dark vowel-like word-initially (and inconsistent with the hypothesis that it is a dark vowel everywhere)

Another reason to think [ü] is neutral

1. Only one form of the type LDL, but the D is [ü].
2. No forms of type DLD.
3. Only two LLD forms and in both cases the D vowel is [ü].
4. No forms of type DDL.

⇒ If [ü] is actually neutral, these exceptions are explained.

Part 1: Conclusions

1. VH is strongly attested in Korean sound-symbolic reduplicant base morphemes
2. Vowels [i, ɨ, u] behave as ‘dark’ vowels in initial syllables, but as transparent vowels in noninitial syllables, supporting Cho (1994)
3. The vowel [ü]’s behavior is more consistent with neutrality than with darkness, supporting Lee’s (1984) suggestion (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)
- Strictly Piecewise learners (Rogers et al. 2009, Heinz in press, Heinz 2010, Heinz and Rogers, 2010)

Part 2: Outline

- Introduce main concepts behind learners
- Explain categorical versions in the context of an idealized VH pattern in Korean sound-symbolic forms (i.e. no exceptionless) and their predictions (qualitative evaluation0
- Show results of probabilistic versions trained on the actual corpus (quantitave evaluation)

Tier-based bigram model

A phonological relationship exists between adjacent members on a tier

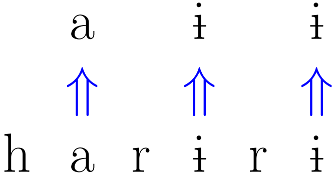
h a r ɨ r ɨ

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

(Goldsmith 1976, Hayes and Wilson 2008, Goldsmith and Riggle to appear, inter alia)

Tier-based bigram model

A phonological relationship exists between adjacent members on a tier



a ɨ ɨ

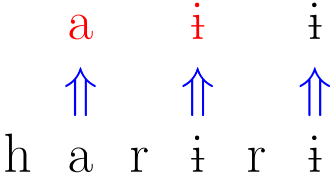
h a r ɨ r ɨ

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

(Goldsmith 1976, Hayes and Wilson 2008, Goldsmith and Riggle to appear, inter alia)

Tier-based bigram model

A phonological relationship exists between adjacent members on a tier



a ɨ ɨ

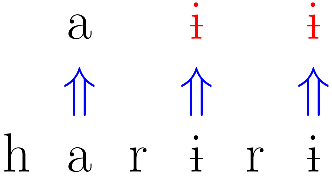
h a r ɨ r ɨ

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

(Goldsmith 1976, Hayes and Wilson 2008, Goldsmith and Riggle to appear, inter alia)

Tier-based bigram model

A phonological relationship exists between adjacent members on a tier



a ɨ ɨ

h a r ɨ r ɨ

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

(Goldsmith 1976, Hayes and Wilson 2008, Goldsmith and Riggle to appear, inter alia)

Strictly 2-Piecewise model

A phonological relationship exists between every two sufficiently similar sounds.

h a r ɨ r ɨ

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

(Hansson 2001, Rose and Walker 2004, Heinz in press)

Strictly 2-Piecewise model

A phonological relationship exists between every two sufficiently similar sounds.

h a r ɨ r ɨ

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

(Hansson 2001, Rose and Walker 2004, Heinz in press)

Strictly 2-Piecewise model

A phonological relationship exists between every two sufficiently similar sounds.

h a r ɨ r ɨ

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

(Hansson 2001, Rose and Walker 2004, Heinz in press)

Strictly 2-Piecewise model

A phonological relationship exists between every two sufficiently similar sounds.

h a r ɨ r ɨ

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

(Hansson 2001, Rose and Walker 2004, Heinz in press)

Strictly 2-Piecewise model

A phonological relationship exists between every two sufficiently similar sounds.
Radical interpretation: any two sounds are sufficiently similar

h a r ɨ r ɨ

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

(Hansson 2001, Rose and Walker 2004, Heinz in press)

Strictly 2-Local (Bigram) Learner

Categorical Version (Garcia et al. 1990, Heinz 2007):

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# }

Grammars of Strictly 2-Local (Bigram) Learners

$$G = \left\{ \begin{array}{ccccc} \#\text{D} & \text{DD} & & \text{DN} & \text{D}\# \\ \#\text{L} & & \text{LL} & \text{LN} & \text{L}\# \\ \#\text{N} & \text{ND} & \text{NL} & \text{NN} & \text{N}\# \end{array} \right\}$$

- 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 Model (Strictly 2-Local distributions)

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

Word	Prob(word)
LNL	0.003611
DND	0.006353
LND	0.007325
DNL	0.003132
NDD	0.001942
NLL	0.001178

Strictly 2-Piecewise Learner (Precedence Learner)

Categorical Version (Heinz 2007, 2010, in press):

time	**word**	**2-long subsequences**	**grammar**
0			∅
1	CNCDCD	{C…N, C…C, C…D, N…C, N…D, D…C, D…D}	{C…N, C…C, C…D, N…C, N…D, D…C, D…D }
2	CLCNCLC	{C…L, C…C, L…C, L…N, L…L, N…C, N…L}	{C…N, C…C, C…D, N…C, N…D, D…C, D…D, C…L, L…C, L…N, L…L, N…L }
3	DCCDCNC	{D…C, D…D, D…N, C…D, C…C, C…N, N…C}	{C…N, C…C, C…D, N…C, N…D, D…C, D…D, C…L, L…C, L…N, L…L, N…L, D…N }

Grammars of the Strictly 2-Piecewise Learners

$$G = \left\{ \begin{array}{ccccc} \text{D}\ldots\text{D} & & \text{D}\ldots\text{N} & \text{D}\ldots\text{C} & \text{D}\ldots\# \\ & \text{L}\ldots\text{L} & \text{L}\ldots\text{N} & \text{L}\ldots\text{C} & \text{L}\ldots\# \\ \text{N}\ldots\text{D} & \text{N}\ldots\text{L} & \text{N}\ldots\text{N} & \text{N}\ldots\text{C} & \text{N}\ldots\# \\ \text{C}\ldots\text{D} & \text{C}\ldots\text{L} & \text{C}\ldots\text{N} & \text{C}\ldots\text{C} & \text{C}\ldots\# \\ \#\ldots\text{D} & \#\ldots\text{L} & \#\ldots\text{N} & \#\ldots\text{C} & \#\ldots\# \end{array} \right\}$$

- Allows harmony to spread without a vowel tier:

  D…x…D
  L…x…L

- Disallows disharmonious sequences with transparent vowel intervening:

  *D…N…L (because *D…L)
  *L…N…D (because *L…D)

- Fails to distinguish between initial and noninitial N:

#LNL	*#NLL
L…N, N…L, L…L	N…L, L…L

Learning Strictly 2-Piecewise Distributions

- A trained probabilistic SP2 learner (Heinz & Rogers 2010) learns the transparency of noninitial N vowels, and to some extent, the behavior of initial-syllable N vowels.

Word	Prob(word)
LNL	0.002893
DND	0.004357
LND	0.000142
DNL	0.000255
NDD	0.001867
NLL	0.000657

Quantitative Comparison

Using the trained SP2 and SL2 probability distributions, we calculated the expected number of each word type.

word type	actual	SP2	SL2
DD	455	502.5	473.4
DL	47	56.5	10.8
DN	637	563.6	237.5
…			

Then we could compute the correllation (Spearman's r) between the expected number and the actual number:

	SP2	SL2	# of words
All	0.95	0.55	4006
Disyllabic words	0.97	0.87	3020
Trisyllabic words	0.47	0.31	986

SL2 distributions and SP2 distributions have the same number of parameters!

Part 2: Conclusion

1. Qualititave and quantitative analysis shows the SP2 learner outperforms the SL2 learner.
2. The tier-based SL2 learner fails to learn what the SP2 learner is able to learn: the transparency of noninitial N vowels.
3. The SL2 learner also fails to learn the bifunctionality of 'neutral' vowels in Korean VH
4. The probabilistic SP2 learner is able to learn the bifunctionality to some degree because the corpus contains fewer N… L subequences than N… D ones. This difference in degree is not distinguishable by the cateogorical version.

Potential solution for tier-based bigram learner

N vowels project to harmony tier only 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 SP2 learners

Treat vowels in initial syllables differently

- The learner realizes N1...L is bad but N2...L is OK (since N1 is the initial vowel and N2 is the non-initial one).
- Sounds at word boundaries typically have special properties (Beckman 1997, Endress 2009)
- But the learner also learns D1 and D2 behave the same, etc. Seems to be missing the right generalization.

Summary

1. Vowel harmony in sound-symbolic forms in Korean is robustly attested in the phonotactics.
2. The vowels [u] and [ü] behave like the neutral vowels [i, ɨ].
3. Strictly 2-Piecewise learners are better suited to capture vowel harmony over transparent vowels than tier-based bigram learners; however, both learners fail to satisfactorily capture the bifunctionality of N vowels in Korean.