

Predicting innovative alternations in Korean verb paradigms

Adam Albright
MIT
albright@mit.edu

Yoonjung Kang
University of Toronto Scarborough
kang@utsc.utoronto.ca

1. Introduction

(1) A few of the challenges faced by a child learning to inflect Korean predicates

- Parse words into morphemes, and determine their meanings
- Discover phonologically predictable alternations (satisfy general phonotactics)
 - Intersonorant voicing of lax stops, post-obstruent tensification

se-da ‘count’	~	mæk-t’a ‘eat’	
---------------	---	---------------	--
 - r ~ l allophony

cær-ə ‘limp’	~	cəl-da ‘limp’	(*l/_V, *r/_C)
--------------	---	---------------	----------------
 - Coda neutralizations

təp ^h -ə ‘cover’	~	təp-t’a ‘cover’	(*C ^h /_C)
kyək ^h -ə ‘experience’	~	kyək-t’a ‘experience’	(*C’/_C)
pəs-ə ‘take off’	~	pət-t’a ‘peel off’	(*s/_C)
 - Cluster simplification¹

əps’-ə ‘lack’	~	əp-t’a ‘lack’	(*C/C_C)
nəlb-ə ‘wide’	~	nəl-t’a ‘be wide’	
 - Vowel deletion to resolve hiatus

c ^h ir-ə ‘pay for’	~	c ^h iri-da ‘pay for’	(*i/_+V)
əps’-imyən ‘lack (cond.)’	~	nolla-myən ‘be surprised (cond.)’	(*i/V+_)
- Discover phonologically conditioned allomorphy: unpredictable alternations, guided by phonotactics
 - Post-vocalic vs. post-consonantal allomorphs avoid CCC clusters

ka-mnida ‘go (defer.)’	~	ip-s’imnida ‘put on (defer.)’	
------------------------	---	-------------------------------	--
- Discover alternations not motivated by phonotactics
 - Aspiration and tensification after vowels and sonorants

co-a ‘be good’	~	co-t ^h a ‘be good’	(V-da is expected)
ci-ə ‘compose’	~	ci-t’a ‘compose’	(V-da is expected)
fir-ə ‘disliked’	~	fil-t ^h a ‘disliked’	(l-da is phon. legal)
man-ə ‘be many’	~	man-t ^h a ‘be many’	(n-da is phon. legal)
fɪn-ə ‘put on’	~	fɪn-t’a ‘put on’	(n-da is phon. legal)

- Other lexically restricted segmental alternations

tow-a ‘help’	~	top-t’a ‘help’	(p ~ w: ‘p’-irreg.)
fir-ə ‘load’	~	fɪt-t’a ‘load’	(l ~ t: ‘t’-irreg.)
hill-ə ‘flow’	~	hiri-da ‘flow’	(ll ~ ri: ‘h’-irreg.)

(2) Learning and encoding alternations: a traditional view

- Learners compare surface variants, extract unpredictable information, combine into an underlying form containing all unpredictable material—e.g.:

əps’-ə	~	əp-t’a	→ /əps/
fir-ə	~	fɪl-t ^h a	→ /silh/

- In cases where the alternation is not predictable, separate allomorphs are listed
- Prediction: if learner has incomplete data, underlying form of a morpheme may contain a subset of the necessary information, based on whatever surface forms happen to be available

fir-ə	~	???	→ /sil/
???	~	fɪl-t ^h a	→ /silh/

- This could lead the learner to project forms “incorrectly” (relative to target lg.)
 - /sil+ta/ → *[fɪl-da] (cf. adult [fɪl-t^ha]) (*’= innovative form)
- Such errors show reanalysis: UR of verb has been inferred on basis of -ə form

(3) Many possible errors

- In principle, reanalyses could be based on various forms, depending on which inflected forms happen to be known

Available data	Inferred UR	Predicted error	Cf. unseen target form
fir-ə ~	???	→ /sil/	*fɪl-da
???	~ əp-t’a	→ /əp/	*əb-ə

- Reanalysis errors could result in many different patterns being extended

a. Actual forms		[fir-ə]	[fɪt-t’a] ‘load’
b. Possible reanalysis based on [firə]		/sil-ə/	→ *fɪl-da
		/silə-ə/	→ *fɪrə-da
		/sili-ə/	→ *fɪri-da
		/silh-ə/	→ *fɪl-t ^h a
		/silʔ-ə/	→ *fɪl-t’a
c. Possible reanalyses based on [fɪt-t’a]		→ *fɪd-ə	/sit-ta/
		→ *fɪt ^h -ə	/sit ^h -ta/
		→ *fɪs-ə	/sis-ta/
		→ *fɪc-ə	/sic-ta/
		→ *fɪc ^h -ə	/sic ^h -ta/
		→ *fɪ-ə	/siʔ-ta/

¹ As Cho (1999) documents, simplification of /lC/ clusters is not enforced categorically in inflected verbal forms.

- (4) A striking observation
- Kang (2006): preliminary survey of innovations across different dialect and acquisition studies
 - In fact, many innovative forms based on reanalyses can be observed in historical change (as seen in dialect comparison), and in child errors
 - However, they are overwhelmingly asymmetrical
 - Based primarily on stem variant that occurs before vowel-initial suffixes (Kim 2001), and in particular, before suffixes that start with -ə/-a (Kang 2006) (“A-suffixes”)
 - That is, reanalyses like those in (3b) are well attested:

[ʃit-tʰa replaced by forms like { *ʃil-ta, *ʃil-tʰa, *ʃil-tʰa, *ʃiri-da }
 - By contrast, reanalyses like those in (3c), in which the -ə/-a form is changed on the basis of a pre-consonantal form are vanishingly rare²
 - Similar asymmetries seen in child errors in other languages (Spanish: Clahsen, Aveledo, and Roca 2002; German: Clahsen, Prüfert, and Eisenbeiß 2002)
 - The puzzle: why do learners seem to focus exclusively on particular forms when deciding the (morpho-)phonological behavior of words?
- (5) The single surface base hypothesis: asymmetric innovations reveal asymmetric paradigm structure (Albright 2002, 2008)
- One inflected form is designated as privileged base form
 - The base form is the same for all lexical items of a given category
 - Base forms serve as input to grammar of morphological and phonological rules/constraints, used to project remaining forms (≈ underlying form)
 - E.g., supposing that -ə form is the base:
 - Words are stored in -ə form³
 - Grammar: morphological mapping of Xə → Xta, plus phonological adjustments such as cluster simplification, post obstruent tensification, etc.

cəɾ-ə	→	cəl-ta	‘limp’
əpsʰ-ə	→	əp-tʰa	‘be without’
əb-ə	→	əp-tʰa	‘to carry on the back’
təpʰ-ə	→	təp-tʰa	‘cover’
 - Consequence: if base form is ambiguous, grammar may be uncertain about how to project other forms
 - E.g., A-suffix [ʃirə] → -ta form [ʃilta]? [ʃiltʰa]? [ʃirida]? [ʃirəda]?
 - Forms that can't be predicted correctly by the grammar must be memorized as lexical exceptions, which block regular outcome through morphological blocking (Aronoff 1976)

² Such reanalyses are seen in nouns: *kapsʰ-i* ~ *kab-i* ‘price, NOM.’ (Kenstowicz 1996; Ko 2006)

³ Note that this need not be the citation form (which is generally the -ta form in Korean).

- If learner doesn't have sufficient data about these forms (or, if lexical access or blocking fails for some reason, these forms are open to regularization
 - Asymmetry: innovative forms = forms that the grammar projects
 - Base forms are stored as heard, so remain constant
 - Non-base forms are projected by the grammar → possibility of errors/change
- (6) Why would an A-form be the base of inflection for Korean predicates?
- A potentially relevant factor: high token frequency
 - Default form with -ə/a is very frequent in spoken language, particularly in child-directed speech (Kim and Phillips 1998, Lee et al. 2003)
 - This no doubt plays a role (more on this below), but cannot be the whole story: frequency, basis of reanalysis is not the most frequent inflected form
 - Informativeness/predictiveness (Albright 2002)
 - Faced with restriction to construct a grammar based on a single surface form, which may potentially have neutralizations (i.e., insufficient information to predict remaining forms perfectly), learners seek the surface form that exhibits as many lexical contrasts, and as few neutralizations, as possible
 - Ideal base form reveals all phonological contrasts (segments of word, tonal pattern, etc.), and all morphological contrasts (inflection class)
 - If no form is perfect, the form is chosen that allows accurate projection of as many forms of as many words as possible
- (7) Are A-forms the most informative forms in Korean?
- As noted above, no single verb form reveals all lexical contrasts in Korean
 - Vowel-initial suffixes: reveal final clusters and laryngeal contrasts, but do not reveal exceptional aspiration (“final /h/”) or tensification (“final /ʔ/”)
 - Obstruent-initial suffixes reveal exceptiona laryngeal properties, but neutralize final clusters and laryngeal contrasts
 - As Kang (2006) points out, A-suffixes induce a fair number of neutralizations,
 - Casts doubt on the idea that they are uniquely informative about how to project other inflected forms
 - We will show here that although there are many neutralizations in A-forms, they are not as serious as one might think
 - Affect comparatively few words
 - Can be predicted more often than not
 - Are offset by neutralizations caused by C-initial and i-initial suffixes
 - Claim: A-forms are actually more informative about C-forms than vice versa
 - Thus, direction of reanalysis in Korean is correctly predicted

- (8) Outline of rest of the talk
- Lay out model for learning grammars to project morphologically related surface forms from one another
 - Describe simulations showing relative accuracy of projections between various forms in Korean
 - Result: A-forms are indeed more predictive about other forms than vice versa
 - Discuss predictions of using A-forms as base of morphological projection
 - Unproblematic mappings (no innovation predicted), vs. problematic/ambiguous mappings (open to innovation)
 - Fairly good match to attested forms
 - A few discrepancies—potentially revealing with respect to Korean phonology?

2. A model for learning surface mappings between inflected forms

- (9) The task
- Given a set of pairs of morphological related surface forms, learn the morphological and phonological mappings to project one from the other
 - E.g., given $-ə/a \sim -(i)myən$ ('if') pairs, learn to form one from the other

a.	siə	→	simyən	‘sour’
b.	kiə	→	kimyən	‘crawl’
c.	cəgə	→	cəgimyən	‘write down’
d.	əps’ə	→	əps’imyən	‘be without’
e.	iruə	→	irumyən	‘create’
f.	nəə	→	nəmyən	‘hand in’
g.	cəlmə	→	cəlmimyən	‘young’
- (10) The Minimal Generalization Learner (MGL; Albright and Hayes 2002, 2003)
- Step 1: parse pairs to see what they have in common, and what has changed

a.	ə	→	myən	/	si__
b.	ə	→	myən	/	ki__
c.	ə	→	imyən	/	cəg__
d.	ə	→	imyən	/	əps’__
e.	ə	→	myən	/	iru__
f.	ə	→	myən	/	nə__
g.	ə	→	imyən	/	cəlm__

- Step 2: Compare forms sharing same change, to discover context for each

b.	ə→myən	/	si__	
c.	ə→myən	/	ki__	
=	ə→myən	/	$\left[\begin{smallmatrix} -son \\ -lab \end{smallmatrix} \right]$	i__ (after non-labial obstruents + i)
e.	ə→myən	/	iru__	
=	ə→myən	/	$\left[\begin{smallmatrix} +syl \\ +high \end{smallmatrix} \right]$	— (after high vowels)
f.	ə→myən	/	nə__	
=	ə→myən	/	[+syl]__	(after vowels)

 - Likewise, comparing [cəg], [əps’], [cəlm] yields: ə → imyən / C__
- (11) Expanding contexts with phonology
- Model described so far discovers phonological context in which each change actually occurs
 - However, this misses a generalization: relation between -myən and -imyən is not arbitrary! A unified analysis is possible:
 - Suffix is -imyən, but initial /i/ is deleted in hiatus with preceding vowel
 - In order to discover this, we need to contemplate the possibility of adding -imyən after vowels—i.e., outside of the contexts where it has been observed
 - Model “clones” mappings: try -myən in -imyən contexts and vice versa
 - Check resulting outputs to see if they contain sequences known to be illegal in the language:

ə	→	imyən	applied after V predicts	*kiimyən, *iruiimyən
ə	→	myən	applied after C predicts	*əps’myən, *cəlmmyən
 - Potential phonological mappings:

i	→	∅ / V + __
∅	→	i / CC + __C
 - An issue that arises, especially in Korean
 - Evidence that the former is workable, while the latter is not, requires evidence from suffixes other than -myən/-imyən
 - As a simplification, in simulations reported here we simply provide the model with a list of generally valid phonological mappings, which the model can make use of to expand the domain of morphological mappings
 - In current case, this means providing $i \rightarrow \emptyset / V + _$, and letting model use this to generalize $ə \rightarrow imyən$ to all contexts (including after vowels)
- (12) Ambiguity and neutralization
- Although the procedure sketched above works well to learn the general distribution of the allomorphs -myən vs. -imyən, it is not possible to predict them from the A-form in all cases

- In particular, vowel coalescence and /h/ deletion before -ə/a leads to ambiguity:

kip ^h ə	/kip ^h -ə/	~	kipɪmyən	~	kipt'a	‘deep’
səgilp ^h ə	/səgilp ^h i-ə/	~	səgilp ^h i myən	~	səgilp ^h i da	‘sad’
sə	/sə-ə/	~	səmyən	~	səda	‘stand up’
nəə	/nəh-ə/	~	nəɪmyən	~	nət ^h a	‘insert’

- General ə → imyən rule yields incorrect predictions for some forms
 - [sə] → *[simyən] (correct [səmyən] appears to require Ø → myən)
 - [nəə] → *[nəmyən] (correct [nəɪmyən] is an exception to i-deletion)
- Competition: general ə → imyən pattern alongside minor patterns

(13) Assessing the strength of competing patterns: confidence

- General ə → imyən rule works for most, but not all relevant forms
- Reliability: $\frac{\text{Number of forms for which rule works}}{\text{Number of forms where rule could potentially apply}}$ (i.e., its accuracy)
- Reliability values are adjusted downwards using lower confidence limit statistics (Mikheev 1997)
 - Rationale: rules based on relatively little data inspire lower confidence than rules based on many applicable forms

(14) The end result of learning

- Set of competing rules, of varying degrees of generality and accuracy
 - General rules, along with minor rules in cases of phonological neutralization (e.g., Ø → myən: /ə/-final verbs) and also irregularity (llə → rimyən: “li-irregular” verbs)
- When grammar is invoked, all applicable rules are tried, to derive set of candidate outputs, with confidence values—e.g., for A-form /kal-a/ ‘grind’:

-ə/a form		Projected -ta form	Confidence
[kara]	→	✓ [kalda]	0.589
		* [kalt'a]	0.331
		* [karada]	0.168
		* [karida]	0.098
		* [kalt ^h a]	0.065

(15) Overregularization

- Most often, the grammar prefers the attested form (i.e., output with highest confidence is the correct form)
- However, in case of minor patterns, the grammar may not actually prefer the actual (attested) form
- These are lexical exceptions, which must be listed
- These words are open to innovative regularization, if for some reason blocking by the lexically listed form fails
 - Word is not known, or has too low frequency to be retrieved reliably

(16) Finding an informative base form

- Given that there are exceptions due to irregularities and neutralization, a certain number of listed exceptions are inevitable
- However, it would be a desirable goal to minimize reliance on listing
 - Decreased memory burden
 - Increased chance of being correct when inflecting unknown words
- The most informative base form is the one that allows most accurate projection of other inflected forms
- Strategy: for an initial batch of data, try learning grammars that use each form to project all other forms
 - ə/a → -ta; -ə/a → -ko; -ə/a → -(i)myən; -ə/a → -(si)mnida; etc.
 - ta → -ə/a; -ta → -ko; -ta → -(i)myən; -ta → -(si)mnida; etc. (and so on)
- For each mapping in each direction, calculate the accuracy of the resulting grammar in reproducing the training data
- The *base* is the form with the highest overall accuracy

3. Testing the model on Korean verbal inflection

(17) Learning data

- Database of 952 inflected predicates in Korean orthography from the National Institute of the Korean Language,⁴ augmented with token frequency information from Sejong Corpus (Kim and Kang 2000)
- Romanized using Hcode 2.1 software (Lee 1994)
- Predictable phonological processes applied automatically by script, with results spot-checked by a native speaker (the second author)
 - Cluster reduction and coda neutralizations; post obstruent tensification; nasalization; lateralization; etc.
 - Optional processes of glide formation ([kiə] ~ [kjə:] ‘crawl’) was omitted (harmlessly, since it does not lead to ambiguities)
- “Inflected forms” = verb stem + one suffix
 - We abstract away from the fact that suffixes may combine and stack
 - What matters for present purposes (stem alternations) is solely the suffix immediately following the stem
- We focus here on the most frequent affixes, according to corpus counts based on written text, consisting of a representative set of phonological shapes
 - A-initial suffixes: -ə/a, -ədo/-ado, -ədaga/adaga
 - C-initial suffixes: -ta, -ko, -ke, -ci, -nin; -(si)mnida
 - i-initial suffixes: -il, -in, -in, -imyən

⁴ www.korean.go.kr//08_new/include/Download.jsp?path=OpenPds&sub=1&idx=28

(18) Testing the predictability of suffixes with different phonological shapes

- A-suffixes (e.g., -ə/a) vs. i-suffixes (e.g., -(i)myən) vs. C suffixes (e.g., -ta)
- As Kang (2006) notes, all forms suffer from neutralizations—e.g., a sample:

Neutralizations in A-initial forms

- /C/- vs. /Ci/- vs. /Cə/-final verb stems
- Stem-final /h/, /ʔ/ (= ‘s-irregulars’)
- ‘t’-irregulars (ʃirə ~ ʃir^hmyən ‘load’) vs. regular (yərə ~ yəlm^hmyən ‘open’)

Neutralizations in i-initial forms

- /C/- vs. /Ci/-final verb stems (kip^h-i^hmyən ‘deep’ vs. ap^hi^h-myən ‘sick’)
- /li/-irregular (hillə ~ hir^hmyən ‘flow’) vs. regular (t’arə ~ t’ar^hmyən ‘follow’)

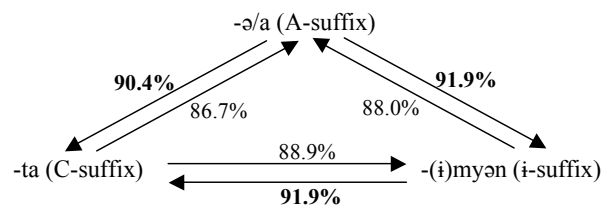
Neutralizations in obstruent-initial forms

- Stem-final clusters, laryngeal contrasts and continuants (/s/, /c/)
- ‘p’-irregulars (towa ~ topt’a ‘help’) vs. regular (copa ~ copt’a ‘be narrow’)
- ‘t’-irregulars (ʃirə ~ ʃitk’o ‘load’) vs. regulars (tada ~ tatk’o ‘close’)

- So, comparison comes down to the number of forms involved

(19) Modeling result: predictability asymmetries

- Ran model in all directions between suffixes -ə/a, -ta, and -(i)myən
- Result 1: most morphological mappings are quite accurate in both directions
 - Most difficult projection is from -ta to -ə/a, with 86.7% accuracy
 - “Multiple predictability” between surface forms (Hayes 1999)
- Asymmetries: accuracy of grammar in each direction



- Projecting from V-initial suffix to C-initial suffix is somewhat more accurate than vice versa
- Projecting C-initial form from -(i)myən (i-form) is slightly more accurate than projecting C-initial form from -ə/a (A-form) (91.9% vs. 90.4%)
- However, A-form is significantly more predictive of -(i)myən (i-form) than vice versa (91.9% vs. 88.0%)

- Result: A-form has highest average accuracy in predicting remaining forms

-ə/a	91.1%
-ta	87.8%
-(i)myən	90.0%

- Same asymmetry is seen, to a somewhat greater extent, on smaller training sets containing just the most frequent verbs (more representative of learning data for a child)
 - Smaller training sets contain same ambiguities, but they are more severe because a higher proportion of high frequency verbs are irregular
 - Upshot: usefulness of A-forms in predicting remaining forms corresponds well to observation that these forms typically act as base of reanalysis in child errors and historical change
- (20) Predictions of using A-forms as the base of morphological mappings
- Although the -ə/a form is relatively accurate, it is not perfect (8.8% errors)
 - Query: do cases where the model is erroneous or uncertain correspond to the cases where humans produce innovative forms?
- (21) Predicted error types in irregular verbs
- Errors that the model predicts, also attested in preliminary survey by Kang (2006)⁵

Type	A-form	Actual C-form	Predicted C-forms	Attested error type
p-irreg.	kiwə ‘sew’	kipt’a	*kiuda	yes (/p~w/ → /u/)
s-irreg.	na: ‘get better’	na(t)t’a	*nat ^h a	yes (/ʔ/ → /h/)
			*nada	yes (American Korean; Choi 2003)

- Predicted errors or variation, mirrored inexactly by attested innovations

Type	A-form	Actual C-form	Predicted C-forms	Attested error type
t-irreg.	murə ‘ask’	mu(t)t’a	*multa	yes, but attested change preserves tensification (*mult’a); see below
lə-irreg.	irirə ‘reach’	irida	*irilta	*iririda (with i)

- An error type that is attested, but not predicted

Type	A-form	Actual C-form	Predicted C-forms	Attested error type
li-irreg.	hillə ‘flow’	hirida	hirida	*hillida (model generates, but assigns very low confidence)

⁵ Sources: Bak (2004), H-W. Choi (2003, 2004), M.-O. Choi (1988, 1993), B.G. Kim (2003), H. Kim (2001, 2002), Park (2002, 2004), Um (1999), Yoo (2000)

(22) Predicted phonological reanalyses

- Attested error types

Type	A-form	Actual C-form	Predicted C-forms	Attested error type
ə → ∅	k ^h jə ‘turn on’	k ^h jəda	*k ^h ida	yes
i → ∅ /l_	t’ara ‘follow’	t’arida	*t’alta	yes
ə → i	sə ‘stop’	səda	*sida	yes
u → i	p ^h ə ‘dig’	p ^h uda	*p ^h ida	yes
h → ∅	k’inə ‘cut’ noa ‘let go’ irə ‘lose’	k’int ^h a not ^h a ilt ^h a	*k’int ^h a *noda *ilda	marginally attested, esp. in child errors (reanalysis to tense is more common)

- Errors that are predicted, but seem to be unattested

Type	A-form	Actual C-form	Predicted C-forms	Attested error type
a → ∅	cara ‘grow’	carada	*calda	no (though parallel to i → ∅ reanalysis)
i → ∅ /C_	kop ^h a ‘hungry’ kip’ə ‘happy’ camga ‘lock’	kop ^h ida kip’ida camgida	*kopt’a *kipt’a *camt’a	no (though parallel reanalysis does happen after /l/)

(23) A discrepancy: h → ∅ reanalysis is robustly predicted, but weakly attested

- The model frequently prefers to treat /h/-final stems as vowel or sonorant-final
- Such errors are attested in American Korean (Choi 2003), but are rarely seen elsewhere (Kang 2006, p. 194)
- Instead, attested reanalyses typically involve changes of laryngeal marking (*c’it^ha* > *c’it’a*, rather than *c’ida*)
- Kang (2006) suggests that this might be due to hypercorrection in adult innovations, but we do not have sufficient data on this point
 - Are child errors in American Korean representative of what any learner would be tempted to do, or due to some additional difference in American Korean?

(24) A systematic difference: i → ∅ after liquids, but not elsewhere

- Model predicts reanalysis to single C-final stems for a range of possible consonants: liquids, stridents, laryngeally marked obstruents, clusters
 - Liquids [t’arida] > *[t’alta]
 - Others: [kop^hida] > *[kopt’a], [camg’ida] > *[camt’a]
- However, this reanalysis seems to occur only with [r] ~ [l]
- Model’s prediction is based on assumption that Korean phonology automatically applies laryngeal neutralization, degemination, and cluster simplification:

A form		-ta form		After phonology
kop ^h a	→	kop ^h ta	→	kopt’a
camga	→	camgta	→	camt’a
hillə	→	hillta	→	hilita

- Hypothesis 1: i is preferred to break up clusters
 - Deletion vs. epenthesis Korean data provides mixed evidence about phonological repair to CC clusters (other than IC)

C-initial suffixes	Cluster simplification: /əps-ta/ → [əpt’a]
i-initial suffixes	[i] appears between stem-final C and suffix C, giving appearance of epenthesis
 - Loanword evidence: epenthesis is standard (English *picnic* → [p^hik^hinik])
 - Perhaps preference for [i] after clusters and laryngeally marked consonants reflects a form of epenthesis in this context, at least in child Korean
 - Children epenthesize in /IC/ less often than other /CC/ (Lee & Im 2004)

• Hypothesis 2: i is inserted to preserve properties of preceding C

- Preservation of [i] avoids alternations in laryngeal features, presence/absence of consonants, etc., and could be motivated by Output-Output faithfulness (Oh 2004)
 - Conjecture: r ~ l alternations are less serious than other alternations?
- Preservation of [i] helps maintain lexical contrasts (laryngeal features, clusters), and could be motivated by Paradigm Contrast constraints (Kenstowicz and Sohn, to appear)
 - Conjecture: fewer /li/ ~ /l/ contrasts than other /Ci/ ~ /C/ contrasts

• Hypothesis 3: pattern strength

- A high proportion of /Ci/-final stems have laryngeally marked C’s
- Thus, speakers hearing [...C^hə] form might be likely to infer [...C^hi/]

• We currently have no basis for deciding among these hypotheses

(25) Another systematic difference between model errors and attested innovations:

- The model occasionally makes a couple errors due to its incomplete ability to encode vowel harmony
 - [p^ha] ‘dig’ ~ *[p^hida] instead of [p^hada]
 - [wa] ‘come’ ~ *[uda] instead of [oda]
- In fact, [a] allomorph of -ə/a shows that stem vowel must be /a/, /o/
- A better ability to encode this fact would eliminate such errors

4. Base selection in the broader context of Korean inflected forms

(26) Taking stock

- Results in previous suggestion support the idea that A-forms are, in fact, the most predictive about behavior of stem with different affix shapes
 - Ranking: A-suffixes > i-suffixes > C-suffixes
- However, comparing classes of affixes in this way is an idealization; in fact, each represents many affixes, with their own segments and frequencies
- If some affix shapes are much more widely used than others, then the constant need to predict one form could make it preferable to choose a different base
- How this could happen, schematically:

In/Out	-a	-ta	-ko	-ke	Average
-a	100%	90%	90%	90%	92.5%
-ta	85%	100%	100%	100%	96.3%
-ko	85%	100%	100%	100%	96.3%
-ke	85%	100%	100%	100%	96.3%

- -a form is better at predicting C-initial forms than vice versa (95% vs. 85%)
- However, C-initial forms are perfectly mutually predictable
- Since C-initial forms outnumber A-forms, they are better *on average*

(27) In fact, this is true of Korean

- Frequency counts from the National Institute of the Korean Language⁶

A-initial	C-initial	i-initial	Other				
ə/a	57894	ta	78116	(i)n	87410	(ni)nta	22141
əsə/asə	11613	nin	60551	(i)l	30545	(si)mnida	9524
ədə/adə	2142	ko	46689	(i)myən	9832	(n)inde	4118
ədaga/adaga	1898	ke	18406	(i)myənsə	4784		
		ci	12144	(i)m	4236		

- Many frequently occurring C-initial suffixes
- If comparisons are weighted to take into account number of relevant inflected forms and their relative frequency, this could tip the balance (incorrectly) in favor of choosing a C-initial form as base
- Why does this not happen?

(28) Possibility 1: learners abstract over affix types, as in idealized simulation

- Instead of seeking the most informative affix, suppose learners seek to discover the best *affixal context*, grouping sets of affixes that behave alike with respect to phonological and morphological context
 - Take same stem allomorph, in cases of irregularity (e.g., p-irregs, li-irregs)
 - Induce same phonologically predictable alternations and neutralizations
- Could lead learners to conclude that A-forms are best, independent of freq.

- But a problem: in other known cases, frequency of individual inflected forms does seem to matter; see Albright (2008) for discussion of how frequency may influence direction of leveling in Korean noun paradigms

(29) Possibility 2 (more boring): corpus counts are not representative

- Intimate -ə/a form is highly underrepresented in written texts, while declarative -ta form is highly overrepresented
- Kim and Phillips (1998): -ə/a is actually 6.6 times more frequent than -ta in child-directed speech
 - Unfortunately, data concerns only mood markers, so not possible to determine relative frequency of other common C-initial affixes, or of other common -ə/a-initial affixes such as the past tense marker -əs’/-as’
- Safe to say that in colloquial speech (especially in child-directed speech), -ə/a forms are among the most frequent
 - Therefore, not advantageous to select a C-initial form just to be able to predict the large number of other C-initial forms

(30) Could frequency alone have explained the asymmetry?

- Given that -ə/a is so frequent, could this alone be enough to predict the fact that it is generally the source of reanalyses?
- Probably not!
 - Although -ə/a forms are quite common, other, neutralizing forms are also not infrequent in colloquial speech (-ta ‘declar.’, -n(i)n ‘progressive’, etc.)
 - Thus, it is not at all unlikely that children might hear a particular verb for the first time in a context with a following consonant
 - However, childrens’ own productions are overwhelmingly (80-100%) -ə/a forms, esp. in early stages (Kim and Phillips 1998; Lee, Lee, and Im 2003)
 - Logically, there should be words that have been heard only in the context of C-initial suffixes, but they want to produce with -ə/a → potential for innovative reanalysis
- Imposing single surface base restriction correctly blocks the model from making such reanalyses, by giving learner no means of “back-formation” to infer unknown base forms

(31) Could frequency be the reason for base selection?

- Even if we do need to assume that A-forms act as bases, do we need to assume that informativeness plays a role, or could frequency alone have done it?
- Perhaps: if Kim & Phillips counts are representative, then A-forms are indeed substantially more frequent, and could be reliably chosen as base forms
- This is unlikely to be sufficient in the long run, however...
 - Numerous other cases in which direction of reanalysis is not predicted straightforwardly by frequency (Albright 2002)
 - Historical change: analogy is often, but not always based on most frequent forms (for discussion, see Hock 1991)
 - Including cases in child morphophonology (Clahsen, Avelado, and Roca 2002; Clahsen, Prüfert, and Eisenbeiß 2002)

⁶ www.korean.go.kr//08_new/include/Download.jsp?path=OpenPds&sub=1&idx=60

5. Conclusion

- (32) Model presented here attempts to explain striking asymmetry in reanalyses in Korean verbal inflection
- Overwhelmingly based in ambiguities in A-forms, rather than other affixal contexts (Kang 2006)
 - This asymmetry is attributed the structure of the morphological grammar that Korean speakers use to project inflected forms
 - Directional mappings from A-forms to other forms
 - Direction is learned, on basis of fact that A-forms are a better basis of predicting other forms than vice versa
 - Computational modeling confirms that predictability relation is in fact true
 - Makes analysis of Korean compatible with other cases investigated so far
- (33) Open questions
- Role of frequency
 - Cooperates with phonological predictiveness in guiding base selection
 - Counts here are estimates of (child-directed) spoken language, at best, and require additional substantiation
 - Discrepancies between model predictions and attested innovations
 - In some cases, may be explained by phonological considerations not incorporated in the model; others require further investigation

References

- Albright, A. (2002). The Identification of Bases in Morphological Paradigms. Ph. D. thesis, UCLA.
- Albright, A. (2008). Explaining universal tendencies and language particulars in analogical change. In J. Good (Ed.), *Language Universals and Language Change*, pp. 144–181. Oxford University Press.
- Albright, A. and B. Hayes (2002). Modeling English past tense intuitions with minimal generalization. In *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, pp. 58–69. ACL.
- Albright, A. and B. Hayes (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119–161.
- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Bak, Suk-Hui. 2004. Ekan caykwuohwauy twu yoin ('Two factors of the restructuring of verbal stems'). *Hangul* 265. 135-169.
- Cho (1999) Intra-dialectal variation in Korean consonant cluster simplification: A stochastic approach. *Chicago Linguistic Society* 35. 43-57.
- Choi, H.-W. (2003) Paradigm leveling in American Korean. *Language Research* 39, 183–204.
- Choi, H.-W. 2004. Explaining variation in Optimality Theory: the case of L-irregular verbs. *Studies in Modern Grammar* 38. 169-194.
- Choi, Myeng-Ok. 1988. li, la-, ε(jə)-, h- pyenchiktongsauy umwunhyensangey tayhaye: pyenchiktongsalul cungsimulo ('On the phonology of irregular verbs: with special attention to li, la-, ε(jə)-, and h-irregular verbs'). *Language Research* 24, 1. 41-68.
- Choi, Myeng-Ok. 1993. Ekanuy caykwuohwawa kyochayhyenguy tanilhwa panghyang ('Restructuring of stem and direction of unification of alternants'). *Sengkoknonchong* 24. 1599-642.

- Clahsen, H., F. Avelado, and I. Roca (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language* 29, 591–622.
- Clahsen, H., P. Prüfert, S. Eisenbeiß, and J. Cholin (2002). Strong stems in the German mental lexicon: Evidence from child language acquisition and adult processing. In I. Kaufmann and B. Stiebels (Eds.), *More than Words. Festschrift for Dieter Wunderlich*, pp. 91–112. Berlin: Akademie Verlag.
- Hayes, B. (1999). Phonological restructuring in Yidj and its theoretical consequences. In B. Hermans and M. van Oostendorp (Eds.), *The Derivational Residue in Phonological Optimality Theory*, pp. 175–205. Amsterdam: John Benjamins.
- Hock, H. H. (1991). *Principles of Historical Linguistics* (2nd ed.). Mouton de Gruyter.
- Kang, Y. (2006). Neutralization and variations in Korean verbal paradigms. In *Harvard Studies in Korean Linguistics XI*, pp. 183–196. Hanshin Publishing Company.
- Kenstowicz, M. (1996). Base identity and uniform exponence: Alternatives to cyclicity. In J. Durand and B. Laks (Eds.), *Current Trends in Phonology: Models and Methods*, 363–394. University of Salford.
- Kenstowicz, M. and H. Sohn (in press) Paradigmatic Uniformity and Contrast: Korean Liquid Stems. To appear in *Phonological Studies* 2008 vol. 11, Phonological Society of Japan.
- Kim, B.-G. (2003). Pokswukicehyenguy yuhyeng (1): hyengseong yoinuy kwancemeyse ('The pattern of plural underlying forms (1): from a generative point of view'). *Cintanhakpo* 95. 165-199.
- Kim, H.-G. and B.-M. Kang (2000). Frequency analysis of Korean morpheme and word usage. Technical report, Institute of Korean Culture, Korea University, Seoul.
- Kim, H. (2001). Hwalyonghyenguy caypunsekey uyhan yongen ekan caykwuohwa: hwuum malum ekanulouy prenhwaey hanhaye ('Restructuring of verbal stems by the reanalysis of conjugated forms: with a special attention to changes into laryngeal consonant-final stems') *Kwuehak* 37.85–113.
- Kim, H. (2002). Hwalyonghyenguy caypunsekey uyhan caykwuohwawa pulmyenghwaklon ('Abduction and restructuring by reanalysis of conjugated forms'). *Language Research* 38:779-799.
- Kim, M. and C. Phillips (1998). Complex verb constructions in child Korean: Overt markers of covert functional structure. In A. Greenhill et al. (Eds.), *BUCLD22*. Somerville, MA: Cascadia Press.
- Ko, H. (2006). Base-output correspondence in Korean nominal inflection. *Journal of East Asian Linguistics* 15(3), 195–243.
- Lee, J.-Y. (1994). Hcode: Hangul code conversion program, version 2.1. <ftp://ftp.kreonet.re.kr/pub/hangul/cair-archive/code/hcode/>.
- Lee, P.-Y. and Y.-J. Im (2004). Emi hwalyong olyulu thonghay pon yuauy ene suptuk ('a study on the acquisition of language through the mistake of inflection in childhood'). *Kwuekyoyuk* 115, 65–85.
- Lee, S.-H., P.-Y. Lee, and Y.-J. Im (2003). emalemiuy suptuk kwacengey kwanhan yenkwu ('the study on the process of the acquisition of final endings: A case of Korean children under 36 months [translation as given]'). *Kwuekyoyukhakyenkwu* 18, 320–346.
- Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23(3), 405–423.
- Oh, M. (2004). Output-to-Output Correspondence in variants. In *Harvard Studies in Korean Linguistics X*, Susumo Kuno et al. (Eds.) Seoul: Hanshin Publishing, 269–289.
- Park, Sun-Woo. 2002. Hyentaykwuke 'lu' pulkyuchik hwalyongey tayhan kochal ('A study of Korean 'li' irregular predicate'). *Hanmalyenkwu* 10. 23-41.
- Park, Sun-Woo. 2004. Pulkyuchikhwalyonguy pulkyuchiksengey tayhan kemtho ('On the irregularity of Korean unsystematic conjugation'). *Chengnamemunkyooyuk* 30. 223-249.
- Um, Yongnam. 1999. Paradigmatic leveling of irregular verbs in Korean dialects: its implications for representations. In *Harvard Studies in Korean Linguistics VIII*, eds., Susumo Kuno et al. Seoul: Hanshin Publishing, 194-208.
- Yoo, Phil-Cay. 2000. Seoulpangen yongen caumekanuy hyengthayumwunlon ('Morphophonology of verb-final consonants in Seoul dialect'). *Kwuehak* 35. 35-65.