Predicting innovative alternations in Korean verb paradigms*

Adam Albright

Massachusetts Institute of Technology
Cambridge, MA 02139
albright@mit.edu

Yoonjung Kang

University of Toronto Scarborough
Scarborough, Ontario, M1C 1A4
kang@utsc.utoronto.ca

Abstract

In Korean verbal inflection, all forms in the paradigm (A-suffix, C-suffix, or ɨ-suffix forms) suffer from neutralization of some lexical contrasts, and there is no single form of the paradigm from which one can correctly predict all the other forms in the paradigm. Nevertheless, a survey of child errors and historical change show that the attested reanalyses are overwhelmingly based on ambiguities in A-forms, rather than in other affixal contexts (Kang 2006). In this study we conducted a computational modeling of learning of inflected forms of 952 Korean verbs using the Minimal Generalization Learner algorithm (MGL; Albright and Hayes 2002, 2003) to account for this striking asymmetry. The simulation result shows that A-suffix form correctly predicts the other forms in the paradigm at a higher rate than C-suffix form or ɨ-suffix form does, indicating that the A-suffix form is indeed the most informative form of the paradigm. This is in line with the previous studies showing that learners designate the most informative form as a privileged base form (Albright 2002, 2008). We compare the model's errors with attested child errors and historical changes.

Keywords: Korean, morpho-phonology, verb paradigm, analogy, acquisition, language change, computational modeling

1. Introduction

Children learning to inflect Korean verbs and adjectives faces a number of challenges simultaneously. Their primary task is to segment words in the input data into morphemes, and determine their meanings. This task is made more difficult, however, by the fact that morphemes undergo a wide variety of phonological alternations. These include fully predictable processes that satisfy general phonotactics of Korean, such as intervocalic voicing and post-obstruent tensification ([d] ~ [t'] in (1a)), allophonic [ɾ] ~ [l] alternations ((1b)), neutralization of laryngeal and manner features in codas ((1c–e)), cluster simplification ((1f–g)), and elision of [ɨ] ((1h–i)). There are also many cases of alternations caused by phonologically conditioned

allomorphy, in which the morpheme has multiple forms that are not fully predictable from one another based on regular phonological processes, but the distribution of the allomorphs is guided by phonotactic constraints ((2a)).

(1)	a.	d ~ t’	se-**da** ‘count’	mək-**t’a** ‘eat’	(*VtV, *Ct)
	b.	ɾ ~ l	**cəɾ**-ə ‘limp’	**cəl**-da ‘limp’	(*l/_V, *ɾ/_C)
	c.	C^h ~ C	**təp^h**-ə ‘cover’	**təp**-t’a ‘cover’	(*C^h/_C)
	d.	C’ ~ C	**kyək’**-ə ‘experience’	**kyək**-t’a ‘experience’	(*C’/_C)
	e.	s ~ t	**pəs**-ə ‘take off’	**pət**-t’a ‘take off’	(*s/_C)
	f.	CC ~ C	**əps’**-ə ‘lack’	**əp**-t’a ‘lack’	(*C/C_C)
	g.	CC ~ C	**nəlb**-ə ‘wide’	**nəl**-t’a ‘wide’	(*C/C_C)[i]
	h.	ɨ ~ ∅	**c^{h}iɾ**-ə ‘pay for’	**c^{h}iɾɨ**-da ‘pay for’	(*ɨV)
	i.	ɨ ~ ∅	əps’-**ɨmyən** ‘lack (cond.)’	nolla-**myən** ‘surprised (cond.)’	(*Vɨ)
(2)	a.	s’ɨ ~ ∅	ka-**mnida** ‘go (defer.)’	ip-**s’ɨmnida** ‘put on (defer.)’	(*CCC)

In addition to phonotactically motivated alternations, Korean learners are also faced with numerous irregular alternations that do not find any apparent synchronic phonotactic motivation. For example, the declarative suffix /-ta/ is expected to surface with intersonorant voicing after verb stems that end in vowels and sonorants, but it surfaces instead with either an aspirated or tense stop after many verbs ending in vowels and liquids ((3a–c)), and after all verbs ending in nasals ((3d–e)).

(3)	a.	co-a ‘be good’	~	co-**t^h**a	(V-da is expected)
	b.	ci-ə ‘compose’	~	ci-**t’**a	(V-da is expected)
	c.	ʃiɾ-ə ‘dislike’	~	ʃil-**t^h**a	(l-da is phon. legal)
	d.	man-a ‘be many’	~	man-**t^h**a	(n-da is phon. legal)
	e.	ʃin-ə ‘put on’	~	ʃin-**t’**a	(n-da is phon. legal)

Furthermore, roots may show other lexically restricted segmental alternations, such as those in (4).

(4)	a.	to**w**-a ‘help’	~	to**p**-t’a	(p ~ w: ‘p’-irreg.)
		ʃi**ɾ**-ə ‘load’	~	ʃi**t**-t’a	(**ɾ** ~ t: ‘t’-irreg.)
		h**ɨll**-ə ‘flow’	~	h**ɨɾɨ**-da	(ll ~ ɾɨ: ‘lɨ’-irreg.)

A standard view of how learners encode alternations, assumed in most work in generative phonology, is that learners compare the surface variants of each morpheme, extracting all unpredictable values and (wherever possible) incorporating them into an underlying form (UR) that contains all unpredictable values (Kenstowicz and Kisseberth 1977, chap. 1; Tesar and Prince 2007). For example, comparing the forms of the verb ‘lack’ in [**əps’**-ə] and [**əp**-t’a], the learner would establish a UR /əps/, which encodes the unpredictable presence of the cluster. Similarly, comparing the forms [ʃiɾ-ə] and [ʃil-t^ha] ‘dislike’, the learner might encode the fact that the verb unpredictably triggers aspiration on the suffix by positing the UR /silh/, with a stem-final [+aspirated] segment. The task of the learner is to arrive at a UR that is compatible with the range of attested surface variants.

A prediction of this approach is that in cases where the learner has incomplete data

about the behavior of a particular morpheme, the UR will be determined solely on the basis of whatever surface forms happen to be available. This means that the UR may contain a subset of the information needed to produce the unseen target forms. For example, if a learner had heard only [ʃiɾ-ə] but not [ʃil-t^{h}a] 'disliked', then there is no need to posit an underlying /h/ for this verb: /sil/. This provisional assumption, based on incomplete information, could lead the learner to project the declarative form ⋆[ʃil-da], which is innovative relative to the adult language. (We use here the notation '⋆' to mark innovations, which are incorrect in the adult language but correct or expected under the learner's analysis.) Such errors would reveal that a reanalysis has taken place, in this case based on the form of the stem that appears before the suffix /-ə/.

In principle, many different types of reanalysis are possible in Korean, depending on which inflected forms happen to be known. In addition to reanalyses like ⋆/sil/ based on prevocalic form [ʃiɾ-ə], it is conceivable that the learner might happen to have encountered a particular word in only preconsonantal forms: e.g., [əp-t'a] but not [əps'-ə] 'lack'. In this case, the learner would have no reason to posit a /ps/ cluster in the UR (⋆/əp/), leading to the possibility of innovations such as ⋆[əb-ə] instead of [əps'-ə]. More generally, learners operating with incomplete information are free to consider a much broader range of underlying forms than they would if all surface allomorphs were known, leading to a wide variety of possible reanalyses. These are illustrated for the irregular verb [ʃiɾ-ə] ~ [ʃit-t'a] 'load' in (5b–c).

(5)	a.	Actual forms	[ʃiɾ-ə]	[ʃit-t'a] 'load'
	b.	Possible reanalysis based on [ʃiɾə]	/sil-ə/	→ ⋆ʃil-da
			/silə-ə/	→ ⋆ʃiɾə-da
			/silɨ-ə/	→ ⋆ʃiɾɨ-da
			/silh-ə/	→ ⋆ʃil-t^{h}a
			/silʔ-ə/	→ ⋆ʃil-t'a
	c.	Possible reanalyses based on [ʃit-t'a]	→ ⋆ʃid-ə	/sit-ta/
			→ ⋆ʃit^{h}-ə	/sit^{h}-ta/
			→ ⋆ʃis-ə	/sis-ta/
			→ ⋆ʃic-ə	/sic-ta/
			→ ⋆ʃic^{h}-ə	/sic^{h}-ta/
			→ ⋆ʃi-ə	/siʔ-ta/

Given such massive surface ambiguity, it is not surprising that innovative forms are in fact widely attested in Korean. Kang (2006) surveys a variety of studies of historical change, dialect differences, and acquisition, and finds that virtually all types of irregular verbs show the effects of reanalysis. But strikingly, the attested innovations are overwhelmingly asymmetrical: they are nearly all based on the stem variant that occurs before vowel-initial suffixes (Kim 2001), and in particular, before suffixes that start with *-ə/-a* (Kang 2006) ("A-suffixes", in which the vowel quality is determined by vowel harmony). Concretely, reanalyses like those in (5b) are well-attested, while reanalyses based on pre-consonantal forms, like those in (5c), are vanishingly rare in verbs and adjectives.[ii] Such asymmetries are not unusual; in fact, they are common in historical change, and are also observed in studies of child errors in other languages (Spanish: Clahsen, Aveledo, and Roca 2002; German: Clahsen, Prüfert, and Eisenbeiß 2002). At the same time, these asymmetries are puzzling under the view that learners

establish URs based on whatever surface allomorphs are available to them, since in the case of Korean, it appears that learners focus primarily on the form of the stem before A-suffixes when deciding the (morpho-)phonological properties of words, while systematically ignoring information from other forms. The challenge is to understand why these forms would play such a privileged role in driving reanalysis in Korean.

In response to such asymmetries in historical change and child errors, Albright (2002, 2008) proposes a more restrictive model of underlying form discovery, in which learners designate a single inflected form as a privileged base form. The base form is constrained to be the same for all lexical items of a given category, and serves as the input (or underlying form) to a grammar of morphological and phonological rules (or constraints), which are used to project the remaining forms. In this model, asymmetries in innovation reflect asymmetries in paradigm structure: the base form serves as the basis of (re)analysis, while the non-base forms are projected by the grammar and are thus open to restructuring. This is referred to as the *single surface base* hypothesis.

To see the implications of this hypothesis for a language like Korean, let us assume for present that the form with the informal suffix *-ə/-a* is the base. (This assumption will be justified below.) The grammar must then operate on this form to derive other inflected forms, for example by transforming [X ə] → [X ta] to yield the declarative form, and then performing any necessary phonological adjustments such as cluster simplification, liquid allophony, intersonorant voicing, post obstruent tensification, etc.

(6)	a.	cəɾ-ə	→	cəl-da	‘limp’
	b.	əps’-ə	→	əp-t’a	‘lack’
	c.	əb-ə	→	əp-t’a	‘carry on the back’
	d.	təp^{h}-ə	→	təp-t’a	‘cover’

A consequence of this direction of mapping is that if the *-ə/-a* base form is ambiguous, the grammar may be uncertain about how to project other forms. For example, given a form like [ʃiɾə] ‘load’, should the declarative *-ta* form be [ʃilda], [ʃilt^{h}a], [ʃiɾɨda], [ʃiɾəda], (correct) [ʃi(t)t’a], or some other form? In many cases, it is impossible to know the answer based on regular grammatical mappings alone; the speaker must instead rely on memorized lexical knowledge to settle the matter. We assume that listed information about the behavior of a word supercedes whatever other output the grammar would have produced, by means of morphological blocking (Aronoff 1976). Blocking may fail, however, in cases where the learner does not have sufficient data about the correct form, or if lexical access happens to fail for some other reason. In such instances, an innovative overregularized form will be produced (Paul 1920, Marcus et al. 1992).

To summarize, the single surface base hypothesis predicts an asymmetry between base forms, which are taken as given by the grammar and must therefore remain constant, and non-basic forms, which are projected by the grammar and may therefore be overregularized if the adult form is not known or not accessed reliably. A question that must be answered is why an A-suffix such as the *-ə/-a* form would serve as the base of Korean verbal inflection. One potentially relevant factor is high token frequency. In fact, the informal *-ə/-a* form 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 in making the *-ə/-a* a likely candidate for base status, but it cannot be the whole story since

in many other documented cases, the base of reanalysis is not the most frequent form.[iii]

Albright (2002) proposes instead that the decision depends on the relative informativeness of the forms in question. The premise of this hypothesis is the following: faced with the restriction that the grammar must be based on a single surface form, which may potentially suffer from neutralizations that remove information about lexical contrasts, learners seek the surface form that exhibits as many contrasts and suffers from as few neutralizations as possible. Ideally, the base form would reveal all contrastive phonological properties (the segments of the morpheme, its tonal pattern, etc.), as well as all morphological contrasts (gender, inflection class). Unfortunately, in most languages, there is no single perfectly revealing surface form, since different inflected forms are affected by different types of neutralizations. In such cases, the learner must choose the form that has least serious neutralizations, and allows accurate projection of the inflected forms of as many words as possible.

As noted, Korean provides an excellent example of the ubiquity of neutralizations, with no single inflected form revealing all contrasts. The question that we address in this study, therefore, is whether A-suffixes such as the *-ə/-a* form are nonetheless the most informative form in Korean. This idea has some a priori plausibility, since these suffixes are vowel-initial, and therefore do not trigger neutralizations in manner and laryngeal features of stem-final consonants, or reduction of stem-final clusters—that is, they provide a phonologically advantageous "pre-vocalic" environment for the preceding stem. At the same time, some of these neutralizations have quite limited practical impact; for example, there are relatively few verb roots ending in obstruent+obstruent clusters, so it is not difficult to guess that a verb should not end in a cluster. Furthermore, although vowel-initial suffixes reveal laryngeal features and clusters, Kang (2006) points out that they also trigger neutralizations such as elision or coalescence of preceding vowels, and they fail to reveal whether the preceding root exceptionally causes aspiration or tensification of a following obstruent (see (3) above). Thus, it is an empirical question which set of neutralizations causes greater difficulty in predicting inflected forms of words.

In the remaining sections, we show that although the A-suffixes trigger various types of neutralization, they are not as serious as one might think, since they affect comparatively few words, they can be predicted more often than not, and they are more than offset by neutralizations caused by consonant-initial and ɨ-initial suffixes. The claim is that A-forms are indeed more informative about C-forms than vice versa, and that the direction of reanalysis in Korean is therefore correctly predicted by a theory that makes use of the most informative form as a base form. In order to show this, we first lay out a model for learning grammars that project morphologically related forms from one another. We then describe simulations employing this model to assess the relative accuracy of projections between various inflected forms in Korean, comparing the consequences of using A-suffixes vs. other types of affixes as the input to the grammar. The results reveal that the A-suffixes are indeed more predictive about other forms than vice versa. We discuss the predictions of a grammar that uses A-suffixed forms as the base of morphological projection, including unproblematic or unambiguous mappings (for which no innovation is expected) as well as problematic or ambiguous mappings (open to innovation). We compare these predictions to the attested range of innovations,

and find a fairly good qualitative match. A few discrepancies also emerge, however, for which we discuss possible resolutions.

2. A model for learning surface mappings between inflected forms

2.1. Predicting allomorphy based on phonological context

The model that we employ takes as its training data a set of pairs of morphologically related surface forms, and attempts to learn a grammar of morphological and phonological mappings that project one from the other. For example, suppose the learner has been given a set of pairs involving the *-ə/-a* informal suffix and the *-ɨmyən/-myən* conditional suffix, as in (7).

(7)	a.	siə	~	simyən	‘sour’
	b.	kiə	~	kimyən	‘crawl’
	c.	cəgə	~	cəgɨmyən	‘write down’
	d.	əpsʾə	~	əpsʾɨmyən	‘lack’
	e.	iɾuə	~	iɾumyən	‘create’
	f.	næə	~	næmyən	‘hand in’
	g.	cəlmə	~	cəlmɨmyən	‘young’

The model learns grammars in both directions (informal → conditional and vice versa), using the Minimal Generalization Learner algorithm (MGL; Albright and Hayes 2002, 2003). The algorithm starts by parsing each pair to see what the related forms have in common, and what has changed between forms, as in (8).

(8)	a.	ə	→	myən	/	si__
	b.	ə	→	myən	/	ki__
	c.	ə	→	ɨmyən	/	cəg__
	d.	ə	→	ɨmyən	/	əpsʾ__
	e.	ə	→	myən	/	iɾu__
	f.	ə	→	myən	/	næ__
	g.	ə	→	ɨmyən	/	cəlm__

The parse reveals that some pairs share the same change (e.g., (8a,b,e,f) all map *ə* → *myən*), while others have different changes. The model attempts to learn a grammar that predicts which change each form will take, by comparing forms that share the same change and trying to discover what phonological features they have in common. For example, comparison of the *-myən* forms in (8) yields the generalizations in (9).

(9) Iterative comparison of *ə → myən* pairs

a. ə→myən / si__
b. ə→myən / ki__
= ə→myən / $\begin{bmatrix}-\text{son}\\-\text{lab}\end{bmatrix}$ i__ (after non-labial obstruents + *i*)

e. ə→myən / iɾu__
= ə→myən / $\begin{bmatrix}+\text{syl}\\+\text{high}\end{bmatrix}$ __ (after high vowels)

f. ə→myən / næ__
= ə→myən / [+syl]__ (after vowels)

Consideration of a broader range of forms (e.g., [yəɾ-ə] ~ [yəl-myən] 'open') would show that *myən* occurs not just after vowels, but also after laterals. Likewise, comparison of the *ə → ɨmyən* pairs in (8c,d,g) yields the rule *ə → ɨmyən* / [-syl, –lat]__. Thus, the learner is able to discover that *-myən* and *-ɨmyən* occur in complementary phonological contexts.

The discovery of complementary contexts for competing affixes is useful in allowing the model to predict the conditional form of each word correctly, but in this case it is also incomplete as an analysis of Korean, since it treats the relation between *-myən* and *-ɨmyən* as completely arbitrary. This misses the generalization that a unified analysis: the suffix is /ɨmyən/, but the initial /ɨ/ of the suffix is deleted after a stem-final vowel. In order to discover this, the model must be able to consider the possibility of adding *-ɨmyən* after a vowel-final stem, even though it is never actually observed in this context. Albright and Hayes (2002) propose to accomplish this by letting the model "clone" mappings, so that *ə → myən* is tried in the contexts where *ə → ɨmyən* is known to apply, and vice versa. The resulting outputs are then checked to see whether they contain sequences that are known to be illegal in the language. For example, applying *ə → ɨmyən* after vowels yields incorrect predictions such as *[k**ɨɨ**myən] and *[iɾ**uɨ**myən], while applying *ə → myən* after consonants yields incorrect forms like *[ə**ps'm**yən] and *[cə**lmm**yən]. Comparison of the incorrect and correct forms leads the learner to consider the possibility of phonological rules of elision in suffixes (ɨ → ∅ / V + __), or epenthesis in suffixes (∅ → ɨ / CC + __C). As it turn out, when the behavior of other affixes is considered, the elision rule receives broad support, while there are /CC+C/ contexts which are repaired by deletion rather than epenthesis. Finding support for the directionality of a phonological process often requires broad consideration of a wide range of sources of data, beyond the scope of what the model can determine in considering just a limited set of morphological relations. Therefore, as a simplification, in the simulations reported here we simply provide the model with a list of generally valid phonological mappings which can be used to explain the domain of morphological rules. In the present case, this means that the model is provided with an elision rule ɨ → ∅ / V + __, which it may then use to generalize the mapping *ə → ɨmyən* to all contexts (including after vowels).

2.2. Unpredictable allomorphy

Although the procedure sketched above works well to learn the general distribution of the allomorphs *-myən* and *-ɨmyən*, it is not possible to predict the distribution of other affixes perfectly in every single case on the basis of the *-ə*/*-a* form alone. One systematic source of ambiguity is elision of stem-final *ɨ*, which creates a neutralization between consonant-final and *ɨ*-final verbs ((10a) vs. (10b)). Another difference that is not evident from the A-form is whether a stem triggers aspiration of the following consonant or not ((11b) vs. (11a)). In addition, certain lexically restricted irregular patterns are neutralized in the *-ə*/*-a* form, but are distinct in other forms ((12)).

(10)	a.	/kip^{h}/	kip^{h}ə	~	kip^{h}**ɨ**myən	~	kipt’a	‘deep’
	b.	/səkɨlp^{h}ɨ/	səgɨlp^{h}ə	~	səgɨlp^{h}**ɨ**myən	~	səgɨlp^{h}**ɨ**da	‘sad’
(11)	a.	/sə/	sə	~	sə**m**yən	~	sə**d**a	‘stand up’
	b.	/nəh/	nəə	~	nə**ɨ**myən	~	nət^{h}a	‘insert’
(12)	a.	/yəl/	yərə	~	yə**l**myən	~	yə**l**da	‘open’
	b.	/sil/$_{irreg}$	ʃiɾə	~	ʃiɾ**ɨ**myən	~	ʃi**t**t’a	‘load’

In these cases, the general rules fail. For example, the mapping of *ə* → *ɨmyən* coupled with the elison rule, ɨ → ∅ / V + __, incorrectly predicts that the conditional of [nəə] ‘insert’ should be *[nəmyən]. Similarly, the general rules incorrectly predicts that lateral-final verbs like (12a) should take [ɨmyən] in the conditional. For such cases, the mapping *ə* → *myən* is still needed as a minority pattern, existing alongside and competing with the more general (and more successful) *ə* → *ɨmyən* rule. For irregular verbs like ‘load’, even more specific rules are needed

In order to assess the competition between different patterns, the model calculates the *accuracy* or *reliability* of each rule, defined as the ratio of forms for which the rule works divided by the number of forms where the rule could potentially apply. These reliability ratios are then adjusted downwards using lower confidence limit statistics, in order to capture the fact that rules based on just a few data points tend to inspire less confidence than rules based on many applicable forms (Mikheev 1997; for details see Albright and Hayes 2002). This allows the model to estimate a *confidence* value for each rule, which determines the probability with which the model will use the rule in deriving novel outputs.

The end result of learning in this system is a grammar of competing rules of varying degrees of generality, including very general rules (such as *ə* → *ɨmyən* in any context) and very specific rules (such as *ə* → *myən* / *ɾ*__, and rules for other) each associated with a confidence score. When the grammar is invoked to produce an inflected form, all applicable rules are tried, and for each change, the rule with the highest confidence is used. This results in a set of output candidates, each given a confidence score, as shown in (13) for the A-suffixed form [kaɾa] ‘grind’. In many cases, the candidate output with the highest confidence also matches the form that is attested in the input data. In some cases, however (i.e., in the case of minority patterns) the grammar may prefer something other than the actual attested form. These cases are lexical exceptions, which must be listed and produced from memory, blocking the grammatically preferred form. Such forms are open to innovative regularization, however, if blocking fails because the word is not known or is too low frequency to be retrieved reliably.

(13) Candidate declarative *-ta* forms for informal *-ə/-a* form [kaɾa] ‘grind’

-ə/a form		Projected *–ta* form	Confidence
[kara]	→	√ [kalda]	0.589
		⋆ [kalt’a]	0.331
		⋆ [kaɾada]	0.168
		⋆ [kaɾɨda]	0.098
		⋆ [kalt^{h}a]	0.065

2.3. Selecting a base form

The example in (13) shows that the grammars that the model learns are not fully deterministic. In any language that has exceptions due to irregularities and neutralization, it is inevitable that there will be a certain number of listed exceptions. Furthermore, in languages like Korean, this is true no matter which inflected form is chosen as the starting point for morphological mappings. For example, rules starting with an A-suffix like *-ə/-a* are bound to be have difficulty predicting features like presence of stem-final *ɨ* or aspiration of a suffix-initial obstruent, while rules starting with a C-initial suffix will have difficulty predicting the laryngeal quality of stem-final obstruents or the presence of clusters. A plausible goal of the learner is to minimize reliance on listing, not only in order to decrease the burden on memory, but also (and more importantly) to increase the chance of being correct when inflecting unknown words. By comparing grammars in various directions, it is possible to assess which mappings are on average more accurate or confident in producing the correct (attested) inflected forms. Specifically, we assume that for some small initial batch of data the learner attempts to learn grammars that use each form to project all remaining forms. For each mapping in each direction, the learner then calculates the confidence with which the resulting grammar produces the form that was attested in the training data. The *base* is the form that yields grammars which are able to reproduce the training data with highest possible confidence.

3. Testing the model on Korean verbal inflection

3.1. Training data

In order to test the model, we trained it on forms drawn from a database of 952 inflected predicates (verbs and adjective) compiled by the National Institute of the Korean Language[iv] augmented with token frequency information from Sejong Corpus (Kim and Kang 2000). The inflected forms were romanized using the Hcode 2.1 software package (Lee 1994). Predictable phonological processes such as cluster reduction and coda neutralizations, post-obstruent tensification, nasalization, lateralization, and aspiration by /h/ in clusters were then applied automatically by script,[v] yielding a database of inflected forms in broad phonetic transcription. The results were spot-checked by a native speaker (the second author) to ensure that neutralizing phonological processes had been applied consistently.

In selecting input data for the morphological learner, we focus on combinations of a verb or adjective stem plus the immediately following suffix.[vi] In particular, we selected the set of affixes in (14), chosen from among the most frequent affixes (as determined by corpus counts based on written texts) to include a representative set of

phonological shapes.

(14) Affixed forms fed to the model

a. A-initial suffixes: -ə/-a, -əto/-ato, -ətaka/ataka

b. C-initial suffixes: -ta, -ko, -ke, -ci, -nɨn; -(sɨ)mnita

c. ɨ-initial suffixes: -(ɨ)l, -(ɨ)n, -(ɨ)n, -(ɨ)myən

As described above, each affix is taken as the starting point for a grammar of rules to derive the remaining inflected forms. Since the neutralizations triggered by a particular affix are primarily a function of the initial segment of the affix, many of the affixes in (14) are equivalent from the point of view of informativeness. We therefore report here just the results for three representative affixes: informal *-ə/-a* (A-suffix), declarative *-ta* (C-initial suffix) and conditional *-(ɨ)myən* (*ɨ*-initial suffix).

3.2. Comparing severity of neutralizations

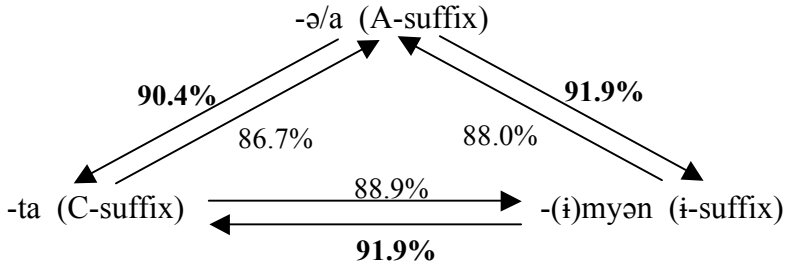
As Kang (2006) notes, all of the forms under consideration suffer from a certain degree of neutralization. The most widely-discussed neutralizations in Korean phonology are laryngeal and continuancy neutralizations among obstruents in coda position, which force all obstruents and obstruent clusters to reduce to a single unreleased stop before another obstruent. These processes are triggered by C-initial suffixes, and lead to a considerable number of neutralizations in forms like the declarative *-ta* form. In addition, consonant-initial suffixes mask several lexically irregular differences, including the difference between so-called *p*-irregular and regular /p/ verbs ((15a,c)), and the difference between *t*-irregular and regular /t/ verbs ((15d,h)). At the same time, A-suffixes such as *-ə/-a* also cause a number of neutralizations (see especially (15e–h, j–m)), as do /ɨ/-initial suffixes ((15f–m)). Thus, it is clear that all available forms suffer from numerous neutralizations.

(15)	a.	p	co**b**a	~	co**pt'**a	~	co**b**ɨmyən	'narrow'
	b.	ps	ə**ps'**ə	~	ə**pt'**a	~	ə**ps'**ɨmyən	'lack'
	c.	p-irreg	to**w**a	~	to**p**t'a	~	to**u**myən	'help'
	d.	t	ta**d**a	~	ta**(t)t'**a	~	ta**d**ɨmyən	'close'
	e.	l	yəɾə	~	yə**ld**a	~	yə**l**myən	'open'
	f.	lh	ʃiɾə	~	ʃi**lt^{h}**a	~	ʃiɾɨmyən	'dislike'
	g..	lɨ	t'aɾa	~	t'aɾɨda	~	t'aɾɨmyən	'follow'
	h.	t-irreg	ʃiɾə	~	ʃi**(t)t'**a	~	ʃiɾɨmyən	'load'
	i.	lə-irreg	iɾɨɾə	~	iɾɨda	~	iɾɨmyən	'reach'
	j.	lɨ-irreg	hɨllə	~	hɨɾɨda	~	hɨɾɨmyən	'flow'
	k.	n	ʃi**n**ə	~	ʃi**nt'**a	~	ʃi**n**ɨmyən	'put on'
	l.	nh-	ma**n**a	~	ma**nt^{h}**a	~	ma**n**ɨmyən	'many'
	m.	C	kip^{h}ə	~	kipt'a	~	kip^{h}ɨmyən	'deep'
	n.	Cɨ	ap^{h}ə	~	ap^{h}ɨda	~	ap^{h}ɨmyən	'sick'
	o.	Cə, Cu	sə	~	səda	~	səmyən	'stop'
			p^{h}ə	~	p^{h}uda	~	p^{h}umyən	'scoop'
	p.	V	kiə	~	ki**d**a	~	kimyən	'crawl'
	q.	h	c'iə		c'i**t^{h}**a	~	c'iɨmyən	'pound'
	r.	s-irreg	ciə		ci**t'**a		ciɨmyən	'compose'

For purposes of the present model, the question is how much ambiguity these neutralizations cause in practice, owing to the number of relevant words of different shapes. In order to test this, we ran the model on all pairwise mappings between the suffixes *-ə/-a*, *-ta* and *-(ɨ)myən*, and used the resulting grammars to assess the accuracy of mappings in each direction. The first result is that in spite of the impressive degree of neutralization shown schematically in (15), it is possible to construct grammars that perform most morphological mappings very accurately in all directions. The most difficult (=least accurate) mapping is from the C-initial *-ta* form to the A-initial *-ə/-a* form, which can be predicted with only 86.7 accuracy. Most of the remaining mappings can be predicted with accuracy approaching or exceeding 90%, showing a high degree of "multiple predictability" between surface forms (Hayes 1999).

Turning to asymmetries between surface forms, we find that as one might expect, it is somewhat more accurate to project from a vowel-initial suffix (ɨ or A) to a C-initial suffix than vice versa (lower left corner of (16)). Furthermore, in projecting C-initial forms, it is slightly more accurate to start with an ɨ-initial suffix (91.9%) than with an A-suffix (90.4%). This difference is offset, however, by the fact that the A-form is significantly more predictive of the ɨ-form than vice versa (91.9% vs. 88.0%). We conclude that on average, the A-initial suffix has greater accuracy in predicting the remaining forms (91.1%) than the ɨ-initial form (90.0%) or the C-initial form (87.8%).

(16) Asymmetrical mappings among suffixed forms



The simulations described here are based on the entire available lexicon of 952 lexical items. It is plausible to think, however, that a learner would wish to establish the architecture of the grammar (which form is the base, which are derived) early in the acquisition process, before an entire lexicon of data is available. It is therefore interesting to note that the same asymmetry shown in (16) is seen to an even greater extent when the model is trained on smaller data sets consisting of just the most frequent verbs, which are presumably more representative of the learning data available to a typical child. The reason is that the irregularities that lead to ambiguity tend to affect a small number of the most frequent words, which are a larger proportion of the lexicon in a smaller training sets. It therefore appears that the advantage of the A-forms in predicting the remaining forms of the paradigm would be evident not only from the unrealistically large training set employed here, but also from smaller sets of data available to children learning Korean. This result corresponds well to the observation that A-forms are also the ones that typically act as the base of reanalysis in child errors and historical change.

3.3. Predictions for innovative phonological reanalysis

According to the single surface base hypothesis, the most informative form is selected as the base form in order to avoid ambiguities when producing inflected forms. We saw in the preceding section that the *-ə/-a* form is generally quite informative in predicting the remaining inflected forms. However, it is not perfect, yielding errors on approximately 8.8% of the forms tested. This naturally leads us to wonder whether cases in which the model is erroneous or uncertain correspond to cases where humans, too, produce innovative forms.

It turns out that many of the errors predicted by the model are also attested as innovations in acquisition and language change, as discussed by Kang (2006) in a sampling of the relevant literature.[vii] One example of this is the distinction between /Cə/-, /Cu/-, and /Cɨ/-final stems (15m–o), all of which undergo elision before *-ə/-a*. For these, the model predicts reanalysis as /Cɨ/-final verbs: [sə] ~ [səda] 'stop' replaced by innovative ⋆[sɨda], and [p^{h}ə] ~ [p^{h}uda] 'scoop' replaced by ⋆[p^{h}ɨda]. Such reanalyses are in fact attested.

Another example concerns the glide in forms like [k^{h}jə] 'crawl', which in principle could correspond to either /k^{h}i-ə/ or /k^{h}jə-ə/. The attested/conservative *-ta* form of this verb is [k^{h}jəda] (i.e., based on /k^{h}jə/), but the model predicts the innovative form ⋆[k^{h}ida], which is also attested as a human innovation. Finally, *-ə/-a* forms neutralize the distinction between stems that aspirate a following consonant and the model predicts that innovations should lack aspiration: [k'ɨnə] ~ [k'ɨn**t^h**a] 'cut' replaced by ⋆[k'ɨn**d**a], [irə] ~ [ilt^{h}a] 'lose' replaced by ⋆[ilda], and so on. Such reanalyses are in fact marginally attested, particularly as child errors in American Korean (Choi 2003). However, as mentioned above, such changes are rarely seen elsewhere (Kang 2006, 194), while the more common change in such forms is in the other direction, i.e., to extend the tense or aspirated allomorphs *-t'a, -t^{h}a*.[viii] Further data is required on this point in order to determine whether the child errors in American Korean are representative of what any Korean learner would be tempted to do given reduced input data, or whether they are due to some additional difference in American Korean.

There is also ambiguity in the A-suffixed forms between /C/-final and eliding /CV/-final verbs. Among these, /C/-final verbs are more common in the lexicon, so the model predicts innovative reanalyses in the *-ta* form, particularly in replacing -C*ɨda* with ⋆-C*ta*. Such reanalyses are attested, but to a limited extent: in particular, they affect /l/-final verbs ((15e)) such as [t'arə] ~ [t'arɨda] 'follow' ((15g)), replaced by innovative C-final ⋆[t'alda]. The model also predicts parallel changes for other /Cɨ/-final items, such as ⋆[kopt'a] instead of [kop^{h}ɨda] 'hungry', ⋆[kipt'a] instead of [kip'ɨda] 'happy', and ⋆[camt'a] instead of [camgɨda] 'lock'. Interestingly, it appears that such changes are not attested after consonants other than the liquid.

There are several possible reasons why speakers may actually prefer to preserve /ɨ/ between obstruents. First, it is possible that the /ɨ/ is being employed to break up CC clusters. Usually, illegal /CC/ clusters are repaired by assimilation (normally, changing features of C_1) in Korean. However, from the point of view of a learner, data concerning /CC/ clusters may be mixed, especially in verbal inflection, since the large number of /ɨ/-initial suffixes give the appearance of epenthesis after /C/-final stems. This is

mirrored by the fact that in loanword adaptation, epenthesis of [ɨ] is actually the preferred repair for illegal /CC/ combinations—e.g., English *picnic* adapted as [pʰikʰɨnik]. It is possible that the preference for preserving [ɨ] in this context reflects a form of epenthesis, at least in child Korean. In this connection, it is relevant to note that children do epenthesize in contexts where adult Korean would have clusters, and that they do so less often in /lC/ clusters than in other /CC/ clusters (Lee and Im 2004). This difference may be due the fact that in their learning data, /l/-final stems pattern with vowel-final stems and take [ɨ]-less form of some /ɨ/-initial suffixes such as (ɨ)myən, (ɨ)lə etc. Alternatively, this asymmetry may be due to some articulatory difference, or perhaps it is an effort to avoid inserting vowels in contexts where they are relatively more perceptible (Fleischhacker 2001). Either way, the fact that our simulations assume perfect knowledge of cluster reduction may give it an unfair advantage in producing /CC/ reanalyses.

A second possible explanation of the discrepancy is that /ɨ/ may be inserted specifically in order to maintain laryngeal properties of C_1, by avoiding neutralization in pre-consonantal position. Consistent with this idea, Oh (2004) hypothesizes that the use of [ɨ] is favored Output-Output faithfulness constraints which ban alternations in aspiration and tenseness. We might conjecture that [l] ~ [ɾ] alternations are considered less serious than C ~ C^h alternations (a difference that is also seen, to a limited extent, in loanword adaptation). Alternatively, it could be that preservation of the [ɨ] is favored by Paradigm Contrast constraints (Kenstowicz and Sohn, to appear), since it helps to maintain lexical contrasts between verbs that end in lax, tense, and aspirated obstruents. Under this account, we might expect to find fewer /lɨ/ ~ /l/ contrasts than other /Cɨ/ ~ /C/ contrasts in the lexicon, creating less pressure to maintain /ɨ/ after /l/.

The third hypothesis is that this [ɨ] insertion is indeed a reflection of the lexical pattern. As it turns out, in present day Korean a high proportion of /Cɨ/-final stems actually involve laryngeally marked C's, making /Cɨ/ a very strong pattern when C is aspirated or tense. Although this trend is weak enough that the model did not pick up on and extend it, perhaps speakers notice it more reliably for some reason, and upon hearing [...C^hə] forms they infer /...C^hɨ/. Unlike the previous two hypotheses, this account does not seek to explain why the existing lexicon has this pattern, but merely connects the current lexical statistics with the observed behavior of speakers. We currently have no basis for deciding among these competing hypotheses.

There is one final type of phonological error that the model predicts, but which is not reflected in human errors. As we saw above the model, like humans, occasionally reinterprets elided vowels as coming from a different source (⋆[p^hɨda] instead of [p^huda] ‘scoop’). In some cases, however, the result of elision is unambiguous due to vowel harmony of the suffix vowel (-*a* after [a], [o]). For instance, the form [p^ha] ‘dig’ is unambiguously /p^ha+ə/ with harmony of the suffix vowel, since underlying /p^hɨ+ə/ would yield surface [p^hə] (no harmony). The model is extremely limited in its ability to encode vowel harmony, since it encodes rules that refer to the immediately adjacent phonological context and cannot encode long-distance conditioning environments. Therefore, it occasionally inflects forms like [p^ha] as ⋆[p^hɨda], rather than as correct [p^hada]. It appears that speakers do not produce similar innovations. We anticipate that a better ability to encode and learn the relation between the stem vowel and the suffix

vowel would elimiate such errors.

3.4. Predictions for innovative regularizations

The reanalysis based on A-suffix forms also lead to regularization of many irregular verbs. A good example can be found in p-irregular verbs such as [ki**w-**ə] ~ [ki**p-**t'a] 'sew' ((15c) above), for the model predicts regularized C-initial forms, mirrored also in human innovations: ⋆[kiu-da]. Similarly, for verbs like [naː] ~ [na**t'**a] 'get better', which trigger tensification of a suffix obstruent, the model predicts reanalysis to the attested innovation ⋆[na**t^{h}**a], or secondarily to [nada], which is also attested in American Korean (Choi 2003).

The model also predicts some changes to irregular verbs which imperfectly resemble attested innovations. For so-called t-irregular verbs ((15h) above) such as [mu**ɾ**ə] ~ [mu(t)**t'**a] 'ask', the model predicts reanalysis to a regular liquid-final verb: ⋆[mulda]. Such verbs are indeed partially rebuilt in their C-initial forms, but the innovative form typically preserves the tense stop in the suffix (⋆[mu**lt'**a]) rather than regularizing all the way to ⋆[mulda]. The innovative form ⋆[mu**lt'**a] innovation is particularly interesting because it creates a verb type that is not found in the pre-existing lexicon. One possibility is that retention of the tense [t'] reflects partial preservation of the older form [mu(t)t'a], perhaps through hypercorrection as suggested by Kang (2006). It is also worth noting, however, that tensification after sonorants is also seen quite regularly with nasal-final stems, where all verb stems cause a suffix-initial consonant to become either aspirated or tense (that is, no verbs like [an-a] ~ [an-da], only [an-t^{h}a] 'do not' or [an-t'a] 'hug'). Therefore, it seems possible that the tense stop in ⋆[mult'a] is part of a broader trend towards tense stops in post-sonorant position—reflecting either a lexical trend that the model is not picking up on correctly, or motivated by a phonotactic constraint against sonorant+voiced stop sequences (Pater 1999, Hyman 2001).

Another example of a minor discrepancy concerns "lə-irregular" verbs like [iɾɨ**ɾ**ə] ~ [iɾɨda] 'reach for' ((15i)). For these, the model predicts regularization to a liquid-final stem (⋆[iɾɨlda]), while human learners prefer to reanalyze them as regular l**ɨ**-final verbs (⋆[iɾɨɾɨda]). In this case, the number of existing regular /l**ɨ**/ verbs is quite small, leading the model to prefer to treat ambiguous verbs as /l/-final. It is possible that human learners are motivated to retain the [ɨ] in order to avoid the [ld] sequence in hypothetical ⋆[iɾɨlda]. It may also be significant that the preferred pattern involves a perceptually minimal [ɨ] ~ ∅ alternation ([iɾɨ**ɾ**-ə] ~ [iɾɨ**ɾɨ**-da]), rather than a [l] ~ [ɾ] alternation ([iɾɨ**ɾ**-ə] ~ [iɾɨ**l**-da]).

Another discrepancy between the attested innovation and the prediction of the model is found in l**ɨ**-irregular verbs. For example, the irregular verb [hɨllə] ~ [hɨɾɨda] 'flow' show the extension of geminate [ll] from the A-suffix form throughout the paradigm: ⋆[hɨllɨda], ⋆[hɨllɨmyən]. As Jun (2007) points out, this innovation is particularly puzzling given the fact that the l**ɨ**-irregular verbs outnumber /ll**ɨ**/-final verbs by 160 to 21, (according to Jun (2007) based on Kang and Kim (2004)), and by 49 to 1 in our learning data. One possibility is that the form ⋆[hɨllɨda] may be encouraged by a desire for elimination of irregular allomorphy or for non-alternating paradigms (Kim 1972, Huh 1985, Choi 1993, Park 2002, Oh 2006, Kenstowicz and Sohn in press, among others), which overrides the analogical pull to robust existing alternations. Another

possibility, suggested by Jun (2007), is that the ⋆[hɨllɨda] is derived from [hɨllə] by a general mapping rule *ə → da* and the resulting illegal cluster [lld] is repaired by [ɨ] insertion (See section 3.3.).

4. Base selection in the broader context of Korean inflection

The results of the preceding section support the idea that A-forms are, in fact, the most informative forms in predicting properties of other inflected forms. However, this result is based on a schematic comparison of just three affixes (*-ə/-a*, *-ta*, *-(ɨ)myən*), standing in for broader classes (A-suffixes, C-suffixes, ɨ-suffixes). This is clearly an idealization, since in actuality, each class of affixes has many members, with its own segments and frequencies. The size and frequency of these classes is potentially an issue, since if some affix shapes are much more common and more widely used than others, the ubiquitous need to predict their form could make it preferable to choose a base form accordingly. To see how this could have an effect, consider the schematic example in (17), in which there is just a single A-suffix alongside three different C-suffixes. In this example (like in actual Korean), the *-a* suffix is better at predicting C-suffixes than vice versa. However, since the C-suffixes are perfectly mutually predictable, there is an overall advantage to selecting a C-suffix as base, since the larger number of C-forms makes them better on average.

(17) The role of affix class size, 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%***

In fact, this situation is not unlike actual Korean. In (18) we provide frequency counts from the National Institute of the Korean Language[ix], showing that there are many frequent C-initial and ɨ-initial suffixes. If comparisons are weighted to take into account the number of relevant inflected forms as well as their relative frequency, there is the danger that this could tip the balance (incorrectly) in favor of choosing a form other than an A-suffix as base.

(18) Inventory of most frequent verbal affixes

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

Why does the large number of C-initial suffixes not influence base selection? One possibility is that learners abstract over broad classes of affixes, much as in the idealized simulation. That is, instead of seeking the most informative affix, perhaps learners seek the best affixal context, grouping sets of affixes that behave alike with respect to phonological and morphological context. We assume that 'behaving alike' involves a combination of taking the same stem allomorph in case of irregularity, and also

inducing the same set of phonologically predictable alternations and neutralizations. If bases are selected in this more abstract fashion, then the learner may indeed conclude that A-suffixes are the most predictive, even though there happen to be many individual C-suffixes that are mutually predictable.

This idea of "affix grouping" has some intuitive appeal, but it also raises a mystery, since in other known cases the frequency of individual inflected forms does appear to matter.[x] We therefore consider a second possibility, which is that the corpus counts in (18) are simply not representative of spoken child-directed Korean. In fact, this seems quite likely, since the intimate or informal *-ə/-a* form is highly underrepresented in written texts, while the declarative *-ta* form is strongly overrepresented. Kim and Phillips (1998) show that in child-directed speech, informal *-ə/-a* is actually 6.6 times more frequent than declarative *-ta*.[xi] We conclude that in colloquial speech (and especially in child-directed speech), *-ə/-a* forms are by far the most frequent. Therefore, it is not necessarily advantageous to select a C-initial form in order to be able to predict the large number of other C-initial forms correctly.

This conclusion now turns the question about the role of frequency on its head: could it be the case that the high frequency of *-ə/-a* alone that creates the observed asymmetry, and that there is no role for informativeness at all? We believe that this conclusion is not warranted, for several reasons. First, it is important to bear in mind that although the *-ə/-a* form is very frequent, other, more neutralizing forms are also relatively frequent in spoken speech: *-ta* 'declarative', *-(nɨ)n* 'progressive', *-ko* 'and', and so on. It is not at all unlikely that a child might hear a particular verb for the first time used with one of these more neutralizing affixes. However, childrens' own productions overwhelmingly (80%–100%) involve *-ə/-a* forms, especially in the earliest stages (Kim and Phillips 1998; Lee, Lee and Im 2003). Logically, this means that there should be words that have been heard only in the context of a C-initial suffix, for which the child wants to produce a *-ə/-a* form. This predicts the possibility of innovative reanalyses based on C-initial forms. However, as Kang (2006) shows, these tend not to occur. By imposing a paradigm structure in accordance with the single surface base restriction and by giving the model no means for 'back-formation' to infer unknown base forms, we correctly prevent the model from making such reanalyses.

A more subtle hypothesis is that frequency is not the sole explanation of the asymmetry in innovations, but that it is the reason why *-ə/-a* forms are selected as the base, without any need to compare the relative informativeness of different forms. This hypothesis is compatible with the Korean facts, if we take Kim and Phillips' counts to be representative of the learning data. It also coincides with the more general observation that analogical change tends to favor more frequent base forms (Manczak 1958). This account is unlikely to be sufficient in the long run, however, since there are numerous other cases in which the direction of reanalysis is not predicted straightforwardly by frequency (Hock 1991; Albright 2002). Tellingly, both of the acquisition studies on child morphophonology cited above (Clahsen, Aveledo, and Roca 2002; Clahsen, Prüfert, and Eisenbeiß 2002) involve errors based on less frequent forms, and innovations on more frequent 3SG forms. It appears that frequency alone is not sufficient, and that the relative informativeness of inflected forms is also a crucial factor in motivating the direction of reanalysis.

5. Conclusion

The model we have presented here attempts to explain a striking asymmetry in the reanalyses seen in Korean verbal inflection, both in child errors and in historical change. In particular, attested reanalyses are overwhelmingly based on ambiguities in A-forms, rather than in other affixal contexts (Kang 2006). This asymmetry is attributed to the structure of the morphological grammar that Korean speakers use to project inflected forms, which uses A-forms to project the remaining inflected forms. We hypothesize that this directionality is learned based on the fact that A-forms provide a better basis for predicting other forms than vice versa. Computational modeling confirms that this predictability relation is in fact true, making the analysis of Korean compatible with other cases investigated so far.

This study leaves a number of open questions. First, the role of frequency has been seen to cooperate with phonological predictiveness in guiding base selection, but the mechanism by which frequency is taken into account requires further clarification. One obstacle to investigating this issue is that the frequency counts of child-directed spoken language available to us are at best rough estimates, and more comprehensive data is needed. Furthermore, we observed several cases in which the model predicted errors that are not mirrored by attested innovations. In some cases, these discrepancies may be explained by phonological considerations that are not incorporated into the model; others require further empirical investigation.

Notes

[*] We would like to thank the following people for helpful comments and discussion: Bruce Hayes, Jongho Jun, Michael Kenstowicz, Hyang-sook Sohn, and the CIL18 audience. All remaining errors are, of course, our own.

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

[ii] Such reanalyses are frequently seen in nouns: unaffixed [kap] ‘price’ corresponds to [kaps’-i] ~ ⋆[kab-i] ‘price-NOM’ (Kenstowicz 1996; Ko 2006).

[iii] The Spanish and German examples cited above are a good example of this: highly frequent 3SG forms are reanalyzed on the basis of less frequent plural or non-third person forms; see Albright (2002) for additional examples.

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

[v] An optional process of glide formation that may apply in some A-suffixed forms ([kiə] ~ [kjəː] ‘crawl’) was omitted. In actual learning data, the availability of such variants make for A-suffix forms likely make the A-suffix form less ambiguous and more informative than in the current simulation.

[vi] That is, we abstract away from the fact that Korean verb forms may involve long sequences of suffixes, since the only thing that is relevant for determining stem allomorphy is the immediately following suffix, and the only thing that determines affix allomorphy is the immediately preceding context.

[vii] Sources: The AKS (1990-1995), 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).

[viii] Jun (2007), in an acceptability judgement experiment, similarly finds that

Korean speakers find the innovative tensification and aspiration to be quite acceptable (e.g. [i-ə] ~ ⋆[i-k^{h}o] (norm: [i-k'o])'to connect', [s'a-a] ~ ⋆[s'a-k'o] (norm: [s'a-k^{h}o]). Interestingly, such innovation is found even for regular s-final verbs such innovation is unexpected based on an A-form based reanalysis.

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

[x] For example, see Albright (2008) for discussion of how frequency influences the direction of leveling in Korean noun paradigms.

[xi] Unfortunately, Kim and Phillips (1998) provide data only for mood markers, so it is not possible to determine the relative frequency of other common C-suffixes, or of other common A-suffixes such as the past tense marker *-əs'/-as'*.

References

Albright, Adam. (2002). *The Identification of Bases in Morphological Paradigms*. Ph. D. thesis, UCLA.

Albright, Adam. (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, Adam and Bruce 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, Adam and Bruce Hayes (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition 90*, 119–161.

Aronoff, Mark. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Bak, Suk-Hui. 2004. Ekan caykwucohwauy twu yoin ('Two factors of the restructuring of verbal stems'). *Hangul* 265. 135-169.

Cho, Taehong. (1999) Intra-dialectal variation in Korean consonant cluster simplification: A stochastic approach. Chicago Linguistic Society 35. 43-57.

Choi, Hye-Won. (2003) Paradigm leveling in American Korean. *Language Research* 39, 183–204.

Choi, Hye-Won. 2004. Explaining variation in Optimality Theory: the case of L-irregular verbs. *Studies in Modern Grammar* 38. 169-194.

Choi, Myeng-Ok. 1988. li, lə-, ɛ(jə)-, h- pyenchiktongsauy umwunhyensangey tayhaye: pyenchiktongsalul cungsimulo ('On the phonology of irregular verbs: with special attention to li, lə-, ɛ(jə)-, and h-irregular verbs'). *Language Research* 24, 1. 41-68.

Choi, Myeng-Ok. 1993. Ekanuy caykwucohwawa kyochayhyenguy tanilhwa panghyang ('Restructuring of stem and direction of unification of alternants'). *Sengkoknonchong* 24. 1599-642.

Clahsen, Harald, Fraibet Aveledo, and Iggy Roca (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language* 29, 591–622.

Clahsen, Harald, Peter Prüfert, Sonja Eisenbeiß, and Joana 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: Akadamie Verlag.
Fleischhacker, Heidi. (2005). *Similarity in Phonology: Evidence from Reduplication and Loan Adaptation*. Ph. D. thesis, UCLA.
Hayes, Bruce. (1999). Phonological restructuring in Yidiɲ 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, Hans Henrich. (1991). *Principles of Historical Linguistics* (2nd edition). Mouton de Gruyter.
Huh, Wung. (1985). *Kwukeumwunhak: Wulimal soliuy onul-ecey* (*Korean Phonology: Present and Past of our language*). Seoul:Saymmunhwasa.
Hyman, Larry. (2001). The limits of phonetic determinism in phonology: *NC revisited. In E. Hume, K. Johnson, E. Hume, and K. Johnson (Eds.), *The Role of Speech Perception in Phonology*. San Diego: Academic Press.
Jun, Jongho. 2007. Stem-final variation in Korean verbal paradigm. *Korean Journal of Linguistics* 32(2). 265-292.
Kang, Beom-Mo and Hung-Gyu Kim. (2004). *Frequency analysis of Korean morpheme and word usage 1.*[In Korean]. Institute of Korean Culture, Korea University, Seoul.
Kang, Yoonjung. (2006). Neutralization and variations in Korean verbal paradigms. In *Harvard Studies in Korean Linguistics XI*, pp. 183–196. Hanshin Publishing Company.
Kenstowicz, Michael. (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, Michael and Charles Kisseberth (1977). *Topics in Phonological Theory*. New York: Academic Press.
Kenstowicz, Michael and Hyangsook Sohn (in press) Paradigmatic Uniformity and Contrast: Korean Liquid Stems. To appear in *Phonological Studies* 2008 vol. 11, Phonological Society of Japan.
Kim, Boong-Gook. (2003). Pokswukicehyenguy yuhyeng (1): hyengseng yoinuy kwancemeyse ('The pattern of plural underlying forms (1): from a generative point of view'). *Cintanhakpo* 95. 165-199.
Kim, Hung-Gyu. and Beom-Mo Kang (2000). Frequency analysis of Korean morpheme and word usage. Technical report, Institute of Korean Culture, Korea University, Seoul.
Kim, Hyun. (2001). Hwalyonghyenguy caypunsekey uyhan yongen ekan caykwucohwa: 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") *Kwukehak* 37.85–113.
Kim, Hyun. (2002). Hwalyonghyenguy caypunsekey uyhan caykwucohwawa pulmyenghwaklon ('Abduction and restructuring by reanalysis of conjugated forms'). *Language Research* 38:779-799.
Kim, Meesook and Colin Phillips (1998). Complex verb constructions in child Korean: Overt markers of covert functional structure. In A. Greenhill et al. (Eds.), *BUCLD22*. Somerville, MA: Cascadilla Press.

Kim, Wanjin. (1972). Hyengthayloncek hyenanuy umwunloncek kukpokul wihaye. (‘A phonological solution to a morphological phenomenon) *Tongamunwha 11*.

Ko, Heejeong. (2006). Base-output correspondence in Korean nominal inflection. *Journal of East Asian Linguistics 15*(3), 195–243.

Lee, June-Yub (1994). Hcode: Hangul code conversion program, version 2.1. ftp://ftp.kreonet.re.kr/pub/hangul/cair-archive/code/hcode/.

Lee, Phil-Young and Yoo-Jong Im. (2004). Emi hwalyong olyulu thonghay pon yuauy ene suptuk (‘a study on the acquistion of language through the mistake of inflection in childhood’). *Kwukekyoyuk 115*, 65–85.

Lee, Sam-Hyung, Phil-Young Lee, and Yoo-Jong 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]’). *Kwukekyoyukhakyenkwu 18*, 320–346.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, and Fei Xu (1992). *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development.

Mikheev, Andrei. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics 23*(3), 405–423.

Oh, Mira. (2004). Output-to-Output Correspondence in variants. In *Harvard Studies in Korean Linguistics X*, Susumo Kuno et al (Eds.) Seoul: Hanshin Publishing, 269–289.

Oh, Mira. (2006). Output-based variants in Korean inflectional morphology. *Korean Journal of Linguistics* 31.1, 77-110.

Park, Sun-Woo. 2002. Hyentaykwuke 'lu' pulkyuchik hwalyongey tayhan kochal (‘A study of Korean '-lɨ' irregular predicate’). *Hanmalyenkwu*10. 23-41.

Park, Sun-Woo. 2004. Pulkyuchikhwalyonguy pulkyuchiksengey tayhan kemtho (‘On the irregularity of Korean unsystematic conjugation’). *Chengnamemunkyoyuk* 30. 223-249.

Pater, Joe. 1999. Austronesian nasal substitution and other NC effects In *The Prosody-Morphology Interface*, René Kager, Harry van der Hulst and Wim Zonneveld eds., 310-343. Cambridge: Cambridge University Press.

Paul, Hermann. (1920). *Prinzipien der Sprachgeschichte* (5th ed.). Halle: Niemeyer.

Tesar, Bruce and Alan Prince (2007). Using phonotactics to learn phonological alternations. In J. E. Cihlar, A. L. Franklin, D. W. Kaiser, and I. Kimbara (Eds.), *CLS 39–2: The Panels: Papers from the 39th Annual Meeting of the Chicago Linguistic Society*, pp. 209–237. Chicago Linguistic Society.

The Academy of Korean Studies (AKS). 1990-1995. hankwukpangencalyocip ('Korean dialectal data'), vols. 1, 2, 4, 6, 8, 9. Sengnam, Korea: The Academy of Korean Studies.

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.