Probability Matching versus Probability Maximization in Morphophonology: The Case of Korean Noun Inflection *

Brent de Chene
Waseda University and Shoin Institute for Linguistic Sciences
dechene[at]waseda.jp

Abstract

The assumption that speakers take isolation forms as basic in Korean noun inflection is confronted with an alternative according to which historically conservative forms are basic. That alternative is shown to be problematic, validating the original assumption. With that conclusion in place, the question of whether speakers analyze cases in which basic X alternates with multiple Y_i by postulating multiple stochastic rules or by taking only the most frequent alternation as regular is raised. Predictions regarding the presence or absence of variation in innovative and established stems are generated for each mode of analysis, and searches of a new Korean subtitle corpus and of the internet are used to test those predictions. The conclusion is reached that only the most frequent alternation is phonologized; minor alternations are represented in the lexicon rather than as probabilistic rules.

韓国語名詞の屈折において接語のない形が基底形であるという分析に対し、歴史的に保守的な形が基底形である分析が紹介され、２つの分析が比較される。後者に問題があることが示され、前者の妥当性が確認される。その結論を確立した上で、基底形 X が複数の Y_iと交替する場合に、話者が複数の確率規則を立てるか最も頻度の高い交替のみを規則とするかという問題が検討される。既存の語幹、新しい語幹に対し、変異の有無に関する予測がそれぞれの分析方法に応じて生成され、インターネットおよび新しい韓国語字幕コーパスにおける検索をもとに、その予測が検証される。最も頻度の高い交替のみが音韻化され、少数交替は確率規則としてではなく、語彙目録の情報として表示されることが結論になる。

Key Words: categorical, stochastic, isolation form, reanalysis, subtitle corpus

1. Introduction

Ongoing changes in Korean noun inflection have received considerable attention in the phonological literature (see Albright 2008, Jun 2010, and references cited there). In the present paper, I take up an issue raised by the treatment of those changes in Jun 2010, namely whether

*A version of sections 3 and 4 of this paper was presented at the 21st Manchester Phonology Meeting, 23-25 May 2013. I thank participants for their feedback. For invaluable assistance of various kinds, I thank Sang-Cheol Ahn, Jieun Bark, Jae-Woong Choe, Sukhoon Choo, Yoonjung Kang, and Yunjin Nam. I am particularly indebted to Ricardo Bermúdez-Otero for illuminating correspondence concerning the material of section 2 and to Kevin Tang for construction of the subtitle corpus introduced in section 3. Remaining mistakes are my responsibility.

speakers postulate multiple probabilistic rules to account for cases in which basic X alternates with multiple Y_i. Relying on the results of searches in a newly created corpus of Korean movie and television subtitles, supplemented by searches of the internet, I conclude that they do not. Rather, I claim, speakers react to the set of alternations $X \sim \{Y_i\}$ by taking the most frequent of those alternations as regular, with the remaining alternations analyzed as irregular and therefore subject to elimination over time. In decision theoretic terms, then, the claim is that morphophonological analysis relies on probability maximization rather than on probability matching in cases of the relevant type.

Asking how speakers analyze basic X alternating with multiple Y_i presupposes, of course, that we have grounds for the judgment that X is in fact basic—that is, computationally prior—with respect to $\{Y_i\}$. In the case of Korean noun inflection, it is often assumed that speakers take as basic isolation forms rather than the conservative prevocalic alternants that preserve a wider range of contrasts. To my knowledge, however, this assumption has not been systematically examined in the literature. Before taking up the issue of probability matching versus probability maximization, then, I will argue in some detail that isolation forms are indeed basic, as against an alternative analysis according to which basic forms coincide with conservative prevocalic alternants. Section 2, after introducing the relevant data, deals with this question of underlying or lexical representations, while section 3 introduces the subtitle corpus and treats the issue of probability matching. Section 4 suggests directions for future research.

2. The Underlying Representations of Noun Stems

2.1 Basic Data

Where [^{h}] represents aspiration (generally analyzed as [+Spread Glottis]), [’] represents “tenseness” (generally analyzed either as [+Constricted Glottis] or as gemination), and [c] is an alveopalatal affricate, Korean has the following 15 obstruents: /p p^h p’ t t^h t’ s s’ c c^h c’ k k^h k’ h/. Its consonant system is completed by the six sonorants /m n ŋ l y w/. Historically, Korean noun and verb stems end in a variety of consonants and clusters. In contemporary Korean, the historically expected forms appear before vowel-initial clitics and suffixes, in which environment stem-final consonants can be resyllabified into syllable onsets. Pre-consonantally and (in nouns) word-finally, however, they are subject to (a) reduction of clusters to single consonants and (b) neutralization of laryngeal and manner contrasts in obstruents. As a result, /p t k m n ŋ l/ are the only permissible syllable codas. More specifically, Cluster Reduction (CR) in Seoul Korean deletes the first of two adjacent consonants if the second is a noncoronal stop (in practice, one of /p p^h k m/) and the second consonant otherwise, while Coda Neutralization (CN) replaces [+Spread Glottis], [+Constricted Glottis], [+Continuant] and [+Delayed Release] (below, [+SG], [+CG], [+Cnt], [+DR]) with the corresponding negative feature specifications, reducing the obstruent inventory to the three unmarked obstruents /p t k/. CR feeds CN in the case of the underlying cluster /lp^h/ at the end of verb stems, which reduces to /p^h/ by CR and to /p/ by CN.

CR and CN result in alternations in noun and verb stem finals depending on the following environment. Consider as an example the conjugation of the verb stem /coc^h-/ “follow”. Before an ending beginning with a consonant, CN will reduce stem-final /c^h/ to [t]. Taking account of another process that tensifies an obstruent immediately following a stop, /coc^h-ko/, a conjunctive form of the verb sometimes called the “gerund” in English, thus surfaces as [cot.k’o], where the period represents syllable boundary. Before an ending beginning with a vowel, however, the /c^h/ of /coc^h-/ is resyllabified into the onset of the following syllable, bleeding CN. Thus /coc^h-a/, another conjunctive form of the verb sometimes called the “infinitive” in

English, surfaces as [co.c^{h}a]. The stem-final consonant thus alternates between [c^{h}] before a vowel and [t] before a consonant. Verb stems ending in other marked obstruents show parallel alternations. Alternations due to CN are stable in verb inflection, showing no variation and no tendency to change.

One might expect that noun stems ending in marked obstruents would alternate in the way we have just seen for the verb stem /coc^{h}-/ depending on whether the following clitic, typically a case particle or the copula, begins with a consonant or with a vowel. This expectation is satisfied before consonan*t*-initial clitics: CN guarantees that in that context, an obstruent at the end of a noun stem must be one of the set /p t k/. For example, the noun stems "knee" and "kitchen", which historically have the forms **muri̯p^{h}* and **puək^{h}*, respectively, show the expected alternants [muri̯p] and [puək] before a consonan*t*-initial clitic: the results of combining these stems with /to/ "also" surface as [mu.ri̯p.t'o] and [pu.ək.t'o]. Since noun stems, unlike verb stems, can appear in isolation—that is, with zero inflection—they also show that alternant word-finally. What is unexpected is that the neutralized alternants [muri̯p] and [puək] may also occur before a vowel-initial clitic. Thus the results of combining the stems in question with *-i̯l* "accusative" may be either [mu.ri̯.p^{h}i̯l] and [pu.ə.k^{h}i̯l], as the model of verb inflection would lead us to expect, or [mu.ri̯.bi̯l] and [pu.ə.gi̯l], where the [p] of [muri̯p] and the [k] of [puək] have undergone a further automatic process voicing plain stops when both the preceding and following sounds are voiced.

There is thus synchronic variation between [mu.ri̯.p^{h}i̯l] and [mu.ri̯.bi̯l] as the accusative of "knee" and between [pu.ə.k^{h}i̯l] and [pu.ə.gi̯l] as the accusative of "kitchen", and the same variation is observed before other vowel-initial clitics. Furthermore, there is a diachronic directionality to this variation: the conservative forms [mu.ri̯.p^{h}i̯l] and [pu.ə.k^{h}i̯l], which are historically expected and sanctioned by Korean orthography, are gradually losing out to the innovative forms [mu.ri̯.bi̯l] and [pu.ə.gi̯l]. Abstracting away from the assimilatory voicing of plain stops, there is a tendency, that is, for the alternations [p^{h}] ∼ [p] and [k^{h}] ∼ [k] at the end of a noun stem to be replaced over time by the null alternations [p] ∼ [p] and [k] ∼ [k]. Alternatively, the alternations [p^{h}] ∼ [p] and [k^{h}] ∼ [k] display a tendency to be leveled in favor of [p] and [k], respectively. With qualifications that we will see below, parallel remarks apply to noun stems that end historically in other marked obstruents or in consonant clusters.

It is widely assumed that these ongoing changes are consequent to a reanalysis of lexical representations as coinciding with isolation allomorphs. On this account, the lexical representations of "knee" and "kitchen" have been restructured so that their default shapes coincide with the forms that occurs in isolation—that is, without any following clitic—and are thus now /muri̯p/ and /puək/. The alternants [muri̯p^{h}] and [puək^{h}] will be included in the nouns' lexical entries as well, but they will be qualified by the environmental specification /__V, indicating that they occur only before a vowel. If we assume that, within a lexical entry, information in excess of a simple pairing of sound and meaning (here, /muri̯p/ "knee" or /puək/ "kitchen") tends to be difficult to retrieve from memory and is thus subject to loss over time, we predict correctly that [muri̯p^{h}] and [puək^{h}] will tend to be lost, leading to the use of /muri̯p/ and /puək/ in all contexts.

On another understanding of the changes in question, however, lexical representations have not been restructured, so that the underlying forms of "knee" and "kitchen" are still the historically conservative /muri̯p^{h}/ and /puək^{h}/, and the alternants [muri̯p] and [puək] result from the application of CN. On this account, the ongoing change in the realization of /muri̯p^{h}/ and /puək^{h}/ before vowel-initial clitics is the result of lexical diffusion of CN in the phonological domain defined by a noun stem. In the remainder of section 2, we take up the question of the

choice between these two interpretations, discussing evidence involving CN in section 2.2 and evidence involving CR in section 2.3.

2. 2 URs as Isolation Forms I: Coda Neutralization

According to Kiparsky's (2003:320) model of lexical diffusion, the lexical diffusion of rule R proceeds through loss from lexical entries of marked feature specifications and their replacement by unmarked specifications supplied by R. In the case of Korean CN, the marked specifications involved, as noted above, will be [+SG], [+CG], [+Cnt], [+DR]. As these are lost from lexical representations, they will be replaced by the corresponding negative values supplied by CN, the latter operating as a structure-building rule or in "lexical redundancy mode" (Bermúdez-Otero 2012:28). Alternatively, the values [–SG], [–CG], [–Cnt], and [–DR] can be thought of as supplied by universal default rules that correspond one-to-one with features (Kiparsky 2003:319), an observation which can be taken as showing that Korean CN has in fact no language-specific content.

Note that this interpretation of lexical diffusion requires us to believe that there is a general tendency for marked feature specifications to be lost from lexical entries, an assumption whose justification is unclear. But given that assumption, it appears to capture quite accurately the variation between [muɾip^{h}] and [muɾip] as the prevocalic allomorph of "knee" and between [puək^{h}] and [puək] as the prevocalic allomorph of "kitchen" as well as that variation's diachronic directionality. What shows most clearly that it is nevertheless unlikely to be the correct account of that variation is the behavior of stems that end historically not in marked labial or velar obstruents (**muɾip^{h}*, **puək^{h}*), but in marked coronal obstruents.

For example, the name of a popular traditional game played with four sticks (below, "yut") is **juc^{h}*, so that on the lexical diffusion account of variation in prevocalic stem allomorphs we would expect the corresponding accusative form to vary between [ju.c^{h}ɨl] and [ju.dɨl], with the marked specifications [+SG +DR] of *c^{h}* being replaced by their unmarked counterparts in the latter variant (cf. the variation [mu.ɾɨ.p^{h}ɨl]/[mu.ɾɨ.bɨl], [pu.ə.k^{h}ɨl]/[pu.ə.gɨl]). In fact, however, the conservative accusative [ju.c^{h}ɨl] varies with, and is being replaced by, not [ju.dɨl], but [ju.sɨl]. More generally, while there are complex subtendencies that we will touch on below, the alternations [t^{h}] ∼ [t], [c^{h}] ∼ [t], and [c] ∼ [t] at the end of noun stems, where in each case the first alternant appears prevocalically and the second elsewhere, are all in the process of being replaced by the alternation [s] ∼ [t]; for stems that originally showed the null alternation [t] ∼ [t], this process is already complete (see Ito 2010:363).[1]

There is no way, then, to claim that it is the rule CN whose lexical diffusion is responsible for the variation displayed by stems that originally ended in **t^{h}*, **c^{h}*, and **c*. If we retain Kiparsky's proposal that lexical diffusion begins with the loss of marked feature specifications, the variation between [ju.c^{h}ɨl] and [ju.sɨl] as the accusative of "yut" can be captured only by postulating a rule that, for stem-final coronals, inserts the marked feature value [+Cnt] (and, redundantly, [+DR]), turning an unmarked coronal into *s*. This rule will have the consequence that at the point in time where the conservative prevocalic variant [ju.c^{h}] has been totally eliminated, the former /juc^{h}/ will become /jus/, in the same way that /mu.ɾɨ.p^{h}/ and /pu.ə.k^{h}/ will become /mu.ɾɨ.p/ and /pu.ə.k/ when the conservative prevocalic variants [mu.ɾɨ.p^{h}] and [pu.ə.k^{h}] have been completely eliminated.

There is a problem with a rule that turns unmarked coronals resulting from the deletion of marked feature specifications into *s*, however, namely that it is stipulatory. There are no

[1]The one historical **t*-stem claimed by Ito for Modern Korean, *nat* "grain", is no longer a free noun (Martin 1992:108).

candidate principles of any generality and plausibility, that is, that would explain why such a rule should exist. More generally, this objection applies to any account of the ongoing replacement of the alternations $t^h/c^h/c \sim t$ by the alternation $s \sim t$ which relies on the reanalysis of $*t^h/*c^h/*c$-stems as *s*-stems at any level of representation. Further, all such accounts confront a second problem, parallel to the first, occasioned by the treatment of *t*-final loanwords such as [int^{h}ənet] "internet". Before vowel-initial clitics, the final *t* of all such loanwords, with complete regularity, alternates with *s*, so that the accusative of "internet" is [in.t^{h}ə.ne.sɨl]. To say that $t^h/c^h/c$-stems are being reanalyzed as *s*-stems is to claim that, in an abstract sense relative to the contemporary state of affairs, the obstruents permissible at the end of a noun stem are /p s k/. Applying this claim to *t*-final loanwords will entail that the lexical form of [int^{h}ənet] must be /int^{h}ənes/. But again there would appear to be no principled way to explain why the final consonant should be altered in this way in the borrowing process.

Above, we have introduced the lexical diffusion account of variation and change in the prevocalic forms of Korean nouns ending historically in marked obstruents and claimed that it is subject to an objection that applies more generally to any account of that variation that postulates reanalysis of $*t^h/*c^h/*c$-stems as *s*-stems as well to any account of apparent *t*-final loanwords that treats them as *s*-stems. This objection is that such accounts are stipulatory in the sense that they cannot be motivated by general principles. Let us now see what the reanalysis account of the variation in question has to say about these two topics and whether it fares better with regard to the question of motivation.

Since the reanalysis account claims that the default lexical form of a noun coincides with its isolation form, it will postulate underlying /int^{h}ənet/ for "internet". The final consonant of that form, as we have already noted, is subject to a productive mapping taking *t* to *s* before a vowel-initial clitic. That mapping can be stated as (1) below, which assumes the results of resyllabification of a stem-final consonant, when prevocalic, into the onset of the following syllable.[2]

(1) $\mathrm{t} \rightarrow \mathrm{s} \; / \; . \_\_ \; _{\mathrm{N}}]$

In the existing vocabulary, the effect of (1) is that the regular alternation for noun stems ending in coronal obstruents, that which over time tends to replace other alternations, is $s \sim t$. For example, under the reanalysis account, the noun "yut", discussed above, will have the default lexical representation /jut/, accompanied by an environmentally restricted allomorph [juc^h] /__V that tends to be lost over time. Insofar as that occurs, the innovative prevocalic allomorph [jus] will be the automatic result of (1).

Turning now to the question of whether the reanalysis account can be motivated in terms of general principles, let us divide the speaker's analytic task with respect to the alternations in question into two parts, choice of basic or underlying forms and choice of regular alternation. With regard to choice of underlying forms, it seems clear that there is a tendency, given an alternation, for speakers to take as basic alternants that occur unsuffixed—that is, isolation forms. While the exact scope of this principle is subject to further research, it provides a plausible motivation for the choice of underlying forms postulated by the reanalysis account.

Having taken as basic the unsuffixed forms that reflect the result of CN, however, speakers are faced with a situation where stem-final [p], depending on the stem, alternates either with [p] or with [p^h] before a vowel; where stem-final [k], depending on the stem, alternates either

[2]The proposal that there is a rule like (1), in conjunction with the postulation of isolation forms as basic, goes back at least to Ko 1989.

with [k], with [k^h], or with [k']; and where stem-final [t] alternates with [s], [t^h], [c^h], or [c]. In each case, as we will see in more detail in section 2, the alternation that has been chosen as regular—the null alternation for labials and velars, and the [t] ∼ [s] alternation for coronals—is the most frequent of the candidate alternations, the one whose stem count in the lexicon is the highest. A simple quantitative criterion, then, provides the answer to why we observe leveling in favor of [p] and [k] for labials and velars, but extension of the [t] ∼ [s] alternation in the case of coronals.

We saw above that the account on which ongoing changes in Korean noun inflection are due to the lexical diffusion of CN in the domain of the noun stem captures the fact that marked labial and velar stem finals revert to their unmarked counterparts [p] and [k], but, in the absence of a rule that turns an unmarked coronal into [s] by inserting the marked feature value [+Cnt], predicts incorrectly that marked coronal stem finals should revert to [t]. In explaining the special status of *s*, the lexical diffusion account is of course free to refer to the lexical imbalance just noted—specifically, the fact that *s*-stems are more frequent in the lexicon than stems ending in any other coronal obstruent. If lexical frequency is the crucial factor determining the choice of a regular alternation in the case of coronals, however, it suggests that the lexical diffusion account of leveling in favor of [p] and [k] is in the end mistaken, since the lexical frequency criterion gives the correct result for all three points of articulation. Further, this mode of explanation would have to confront the fact that lexical imbalances in the frequency of stem-types are entirely typical, perhaps universal, in the languages of the world, but are in the vast majority of cases phonologically inert, failing to occasion the large-scale reassignment of stem-types that Korean shows.

The preponderance of *s*-stems among coronal-final stems in the lexicon is thus insufficient, in and of itself, to motivate the assimilation of all coronal alternations to *t* ∼ *s* that is in progress in the existing vocabulary. In the same way, it is insufficient to motivate the treatment of innovative items like [int^{h}ɔnet] as *s*-stems. There is no general principle of loanword adaptation, that is, that would justify concluding that words that to all appearances end in *t* are to be lexicalized as *s*-final simply on the basis of the fact that *s*-stems are common in the lexicon. The lexical preponderance of *s*-stems acquires explanatory force only in the context of the postulation of underlying representations that coincide with isolation forms, because that analytic decision places the alternations of preconsonantal *t* with *s, t^h, c^h*, and *c* in direct competition with each other for the role of regular alternation.

In section 2.2, focusing on evidence related to the rule of CN, we have compared two interpretations of ongoing change in Korean noun inflection, crucially differentiated by their claims about underlying forms. We have seen that while either interpretation can handle the changes displayed by stems that end historically in marked labial and velar obstruents, the changes shown by stems ending in marked coronal obstruents argue for the interpretation under which default lexical representations of noun stems coincide with their isolation forms. If that interpretation is correct, the results of CN have now been incorporated into lexical representations, and CN no longer plays a role in the computation of nominal inflected forms. The rule that does play an important role, we have argued, is (1), which is the result of inversion of the basic/derived relationship in one case of CN. In section 2.3, we look at further evidence, this time involving Cluster Reduction, for the conclusion that the default lexical representations of noun stems coincide with isolation forms.

2.3 URs as Isolation Forms II: Cluster Reduction

Let us first note that if there were a stem with two alternants A and B distributed in the same way as are [murɨp^h] and [murɨp] for "knee" or [puək^h] and [puək] for "kitchen" but which bore an irregular phonological relationship to each other or, in the limiting case, no phonological relationship at all, that situation would be describable unproblematically in terms of the reanalysis account: the gradual eclipse of the environmentally restricted alternant A in favor of the default alternant B would be the result of precisely the same mechanism as in the case of [murɨp^h] or [puək^h], namely elimination from lexical entries of environmentally restricted allomorphs. No account of this change in terms of lexical diffusion would be available, however, because there would be no phonological rule involved.

Korean noun inflection has undergone leveling of irregular alternations in favor of the isolation form in the past; thus in Middle Korean the stem "tree", contemporary *namu*, was *namk-* before a vowel-initial clitic and *namo* otherwise (Martin 1992:101, 238-239, Lee 1975:174). In the contemporary language, the best example of an irregular alternation is probably the numeral "eight", which is [jətəlp] ∼ [jətəl] prevocalically and [jətəl] otherwise. Since the regular reduction of *lp* is *p* rather than *l*, there is no rule such that the spread of [jətəl] at the expense of [jətəlp] in the prevocalic environment can be expressed as the lexical diffusion of that rule in the domain of the noun stem. There is no reason to believe, however, that the ongoing spread of [jətəl] differs in nature from the ongoing spread of [murɨp] and [puək] prevocalically at the expense of [murɨp^h] and [puək^h]. It thus follows that the lexical diffusion account of the latter changes is of insufficient generality.

The stem "eight" also provides evidence concerning whether the conservative alternant of a noun stem undergoing the changes described above is lexically limited to the environment /__V, as claimed by the reanalysis account of those changes, but not the by lexical diffusion account or any account according to which the conservative alternant is the unique lexical form. This is because the automatic post-stop tensification process described above is counterbled by cluster reduction, so that a verb form like /halt^h-ta/ "lick (Declarative)" surfaces as [hal.t'a] (Yun 2008:450), with tensification induced by the subsequently deleted stem-final stop. If the conservative alternant [jətəlp] of "eight" were available for insertion in any environment, we would expect that tensification of a clitic-initial obstruent would be possible after that stem. Yun (loc.cit.), however, points out that the result of suffixing "eight" with /kwa/ "and" surfaces without tensification (and abstracting away from voicing) as [jə.təl.kwa], and consultation with several native speakers confirms the judgement that tensified clitic-initial obstruents are systematically impossible in that combination as well as with /to/ "also" and /coc^{h}a/ "up to". It would thus seem that the (default) lexical representation of the stem must be /jətəl/, with [jətəlp] limited to the prevocalic environment, just as claimed by the reanalysis account of the ongoing changes.[3]

Returning to the lexical diffusion account of those changes, it is unclear how a rule that effects a deletion, such as Korean CR, is to be dealt with in the model of that phenomenon introduced above: it would seem necessary to postulate a feature [Extant] such that generalization of a deletion rule can be mirrored as insertion of the unmarked value of that feature after the marked value has been lost from lexical representations. More importantly, however, we have seen in this section that lexical diffusion in the domain of the noun stem is suspect as an account of the full range of simplifications that are in progress in Korean nominal inflection both because it cannot deal with the loss of marked allomorphs that bear an irregular phonological relationship to their unmarked counterparts and because there is evidence that marked

[3] Yun pursues a computational rather than a representational account of these facts.

allomorphs are environmentally restricted.

Below, then, we will assume that default lexical representations of Korean noun stems coincide with their isolation forms. With that conclusion that in place, we are in a position to move to the question of how speakers respond to a situation in which basic X alternates with multiple Y_i.

3. Categorical versus Stochastic Generalizations in Korean Noun Inflection

3.1 Background: two theories of morphophonology

Recent research has made it clear that speakers internalize many statistical patterns in the lexicon, as shown, for example, by their ability to extend those patterns to nonce forms in an experimental setting ("wug-testing"). Extrapolating from this fact, a number of researchers have suggested that the process of phonological analysis for nonautomatic or "irregular" patterns consists of assessing lexical statistics and transferring them as rules or constraints, perhaps filtered by universal restrictions, directly into the phonology. This view of phonological analysis predicts that when confronted with novel forms, speakers will display, for each of a competing set of patterns or generalizations, a propensity to extend that pattern whose strength is directly proportional (modulo statistical adjustments) to the pattern's lexical frequency. Thus Albright (2009:185), speaking of formal models of analogy, claims that the phonology must be able "to adjudicate between multiple conflicting patterns by assessing the relative strength of each, and to generalize them to novel forms based on their relative strength." Similarly, Zuraw (2000:xiv) claims that "speakers' behavior on novel words probabilistically reflect the lexical regularities.", and Becker (2009:viii) notes that the same "grammar that applies deterministically to known items ... applies stochastically to novel items." Let us call this concept of phonological analysis the "proportional representation" (below, "PR") theory, since it claims that lexical trends are phonologized and extended in proportion to their frequency. In the terms of decision theory, as we have noted, the PR theory is an example of probability matching.

The recent literature, however, also displays an alternative to the idea that the non-automatic phonology of a language, for a given input configuration, incorporates a multiplicity of competing generalizations whose strength is proportional to their lexical frequency. This alternative is the proposal that, after the relative lexical frequency of such a set of generalizations is determined, only the most frequent or reliable generalization is phonologized, the others being treated as irregular and as a result becoming subject to elimination over time. With regard at least to the question of how novel forms are treated, this viewpoint is clear in Albright 2005: "When the grammar is used to derive new forms, ... the rule with the highest reliability value is the one that applies." (loc.cit.:25); and again, "The model of paradigm learning advocated here always extends the strongest pattern, regardless of whether it is alternating or uniform." (loc.cit.:41). Let us call this theory of how speakers treat non-automatic alternations the "Winner-Take-All" (below, "WTA") theory, since, in contrast to the PR theory, it claims that only one of a set of competing alternations becomes part of the phonology. On the WTA theory, then, speakers' treatment of non-automatic alternations is based on categorical rather than probabilistic generalizations; in decision-theoretic terms, it is an example of probability maximization.

PR and WTA make sharply distinct predictions about variation and its absence in innovative forms and the existing vocabulary. In PR theories (e.g. Zuraw 2000), stochastic rules or constraints that reflect lexical statistics dictate the treatment of innovative forms. Established forms are held to have fixed, lexicalized pronunciations, so that the stochastic grammar must

be prevented from applying to them. In sum, PR makes the predictions in (2).

(2) a. Innovative forms: variation according to lexical statistics
b. Established forms: no variation

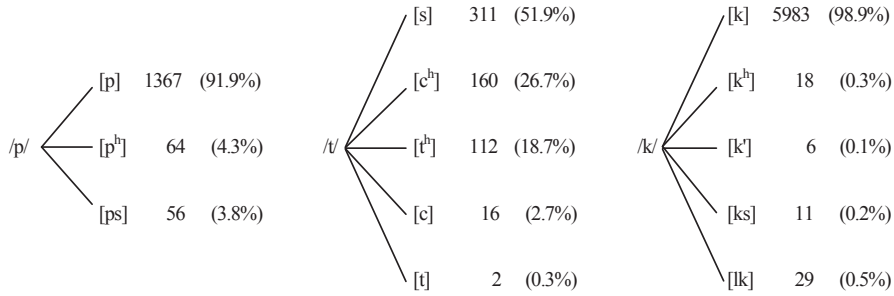
Under WTA, in contrast, patterns of alternation are either regular or irregular. Regular patterns other than the null alternation are represented as rules, irregular patterns as excess information (e.g. listed allomorphs) in lexical entries, as discussed in section 2.1. Assuming that such excess information tends to be lost over time, WTA predicts that items displaying irregular alternations will tend to vary with regularly derived substitutes. Innovative forms will normally lack excess information in their lexical entries and will therefore fail to show variation. WTA thus makes the predictions in (3).

(3) a. Innovative forms: no variation
b. Established forms: variation if and only if irregular

3.2 Korean Noun Inflection as a Test Case

As we saw in section 2, Korean noun stems end historically in a variety of consonants and clusters that appear before V-initial suffixes but are subject in codas to cluster reduction and neutralization of laryngeal and manner contrasts, so that the only permissible coda obstruents are /p t k/. With the reanalysis of isolation forms as basic, /p t k/ have come to alternate unpredictably, when prevocalic, with the original stem-final consonants and clusters in the proportions shown in (4).[4]

(4) Distribution of obstruent-final stem-types



A PR analysis of these alternations will postulate stochastic rules or constraints with reliability/confidence levels that to a first approximation reflect the lexical statistics above. On the WTA analysis, in contrast, the regular alternations will be $t \sim s$ for coronals and the null alternation for labials and velars; the only rule postulated will be (1) above. As proposed in section 2, minority alternations will be represented in lexical entries, with an irregular alternant constituting a special case and the basic form a default.

[4]The stem counts of (4) were obtained using the concordance program Geuljabi II as applied to the .txt files of Kim and Kang 2000, which is based on the Sejong Corpus (`www.sejong.kr`). They correspond closely to stem counts reported by Albright (2008:171) except with regard to stems showing the $t \sim s$ alternation, for which Albright's count is higher; similar counts are reported by Jun (2010:149) (neither Albright nor Jun include alternations involving clusters).

The essentials of the WTA analysis, as mentioned in note 2 above, go back to Ko 1989; a PR analysis based on the minimal generalization model of Albright and Hayes 2003 and relying for the most part on speaker well-formedness judgements is proposed by Jun 2010. Jun (2010:156), following Albright (2005:41), is well aware that the Korean situation appears to call for a WTA condition, in that only the strongest of a set of competing alternations tends to be extended. He is nevertheless at pains to argue for the productivity of minority alternations and in particular (Jun and Lee 2007) for their activity in the phonology of loanwords, and we will take his analysis as representative of a PR or probability matching approach to the alternations of Korean noun inflection.

3. 3 Data and Results

To obtain a full picture of ongoing change in Korean noun inflection, we would ideally like to have a corpus of transcribed spontaneous speech along with information on the sociolinguistic characteristics (age, gender, etc.) of the individual speakers represented therein. We would then be able to compute the rate of regularization of individual irregular stems and classes of stems and determine how rate of regularization varies with sociolinguistic parameters. In particular, it would be possible to check for the existence of apparent-time effects that could indicate the progression of regularization over time. Such a corpus, needless to say, would be extremely expensive to construct in terms of time and effort.

In the absence of a corpus of spontaneous speech, one might think to make use of standard corpora, the best-known of which for Korean is the Sejong Corpus, mentioned in note 4 above. The problem with standard corpora for our purposes, however, is that they are typically based for the most part on published texts. Since the regularized forms we are interested in are, from a normative point of view, mistakes, they are virtually guaranteed to be absent from texts that have undergone the editing process associated with publication. Searches in the online Sejong Corpus confirm that regularized forms for nouns like **murɨp^h* "knee", **puəkʰ* "kitchen", and **yucʰ* "yut" are systematically absent.

There is no question that in informal, unedited writing, Korean speakers sometimes employ regularized forms; this can quickly be confirmed with internet searches for such forms, which typically turn up in texts like blog entries and photo captions. For our purposes, then, the ideal corpus of written Korean would be one based on such informal writing. In answer to this need, Kevin Tang has recently compiled a corpus SUBTLEX-KR of about 90 million words from files of Korean language movie and television subtitles, building on his experience in creating a similar corpus for Brazilian Portuguese (Tang 2012). Subtitles are an ideal text type for our purposes because the writers who create them will in general be aiming to reproduce colloquial speech. The idea of a subtitle corpus goes back at least to New et al. 2007, and such corpora have now been created for a number of languages.

In verifying whether it is the predictions of the PR theory or those of the WTA theory that accord with the facts, the crucial question to be answered is what classes of stem are stable and what classes are variable. Searches of SUBTLEX-KR and analysis of search results are still in progress, so it is not possible to make a final report at this stage, but two major conclusions are clear. First, the corpus does include regularized forms for a number of stems that are reported in the literature to show variation. It thus overcomes the problem of normalization that makes standard corpora like the Sejong Corpus uninformative for our purposes. Second, the corpus fails to show variation for large classes of stems, including historical **s*-stems and all loanwords. It thus suggests strongly that those stems are in fact invariant, in accordance with the predictions of the WTA theory.

This pair of conclusions is reinforced by searches of the internet. While internet search results clearly cannot be taken as reliable indications of frequency of occurrence, search engines do function as sensitive detectors of variation. In the Korean case, they reveal even more variation associated with ongoing regularization than is apparent in the subtitle corpus. This sensitivity to variation, again, means that when search results show zero variation for the inflection of a given stem, we can conclude with some confidence that that stem is indeed of invariant inflection. In the present context, the crucial conclusions from online searches are (5a) and (5b), which accord, respectively, with predictions (3a) and (3b) above.

(5) a. Innovative stems, primarily loanwords from English, show essentially no variation: automatic low-level processes apart, stem-final *p* and *k* are nonalternating, and stem-final *t* alternates predictably with *s* before a vowel-initial clitic as a result of rule (1).

b. Historical **p-/*s-/*k*-stems, likewise, show essentially no variation, while stems ending historically in other consonants and in clusters typically show some degree of variation between the historically expected alternation and the regular one.

There would seem, then, to be no evidence from either innovative stems or regular stems of the established vocabulary that the minority alternations of (4) above are phonologically active; the productive generalizations speakers make about the system of alternations are categorical rather than stochastic. If minority alternations are not part of the phonological computational system, they will be part of the lexicon. We have already claimed above that conservative prevocalic alternants like [muriɸ^{h}], [puɘk^{h}], and [yuc^{h}] are accompanied in lexical entries by the environmental condition /__V; as a result of the principle of proper inclusion precedence (elsewhere condition), they will take priority over default forms at the time of lexical insertion even as they show a tendency to be eliminated from lexical entries. More detailed patterns of allomorphy, such as the fact that some **c^{h}*-stems tend to assimilate to **t^{h}*-stem conjugation rather than to the regular $t \sim s$ alternation, can also be handled in terms of (multi-layered) proper inclusion precedence formalism, although there is not space to demonstrate that in detail here. No stochastic rules, then, are required by such allomorphic patterns, which crucially are not extended to innovative forms.

In this section, we have taken care to appeal for evidence concerning the phonological computational system only to forms that speakers have unselfconsciously produced. A methodological lesson that emerges from the discussion, then, is that reliance instead on speaker well-formedness judgements, as in Jun 2010 or Jun and Lee 2007, may create the appearance of more variation than is actually attested and thus obscure categorical generalizations that are in force.

4. Conclusion

We have seen that the course of ongoing change in Korean noun inflection appears to support the WTA theory of morphophonology—equivalently, the idea that speakers employ probability maximization rather than probability matching in analyzing morphophonological alternations, making categorical rather than stochastic generalizations about those alternations. To conclude on a further methodological point, the WTA theory, in assigning minority alternations to the lexicon rather than the grammar, has the merit of bringing into focus the distinction between cases in which speakers do generalize a nonautomatic alternation and those in which they gradually eliminate it. Spanish Diphthongization (e/o → ie/ue), which has been leveled far

more frequently than extended, and Portuguese Lowering (e/o → ɛ/ɔ), which has been fully generalized, constitute a suggestive minimal pair in this regard: given that the Portuguese alternants differ by only a single degree of vowel height, the difference between the two cases could be explained if there is an input-output phonological distance condition on the induction of morphophonological rules. The WTA theory, then, has the potential to open up interesting issues for future research.

References

Albright, Adam (2005). The morphological basis of paradigm leveling. In *Paradigms in phonological theory*, ed. Laura Downing et al., 17-43. Oxford: Oxford University Press.

Albright, Adam (2008). Explaining universal tendencies and language particulars in analogical change. In *Linguistic universals and language change*, ed. Jeff Good, 144-181. Oxford: Oxford University Press.

Albright, Adam (2009). Modeling analogy as probabilistic grammar. In *Analogy in Grammar*, ed. James P. Blevins and Juliette Blevins, 185-213. Oxford: Oxford University Press.

Albright, Adam, and Bruce Hayes (2003). Rules vs. analogy in English past tenses. *Cognition* 90:119-161.

Becker, Michael (2009). *Phonological Trends in the Lexicon: The Role of Constraints*. Doctoral dissertation, University of Massachusetts, Amherst.

Bermudez-Otero, Ricardo (2012). The architecture of grammar and the division of labor in exponence. In *The Morphology and Phonology of Exponence*, ed. Jochen Trommer, 8-83. Oxford: Oxford University Press.

Ito, Chiyuki (2010). Analogy and lexical restructuring in the development of nominal stem inflection from Middle to Contemporary Korean. *Journal of East Asian Linguistics* 19:357-383.

Jun, Jongho (2010). Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics* 19:137-179.

Jun, Jongho, and Jeehyun Lee (2007). Multiple stem-final variants in Korean native nouns and loanwords. *Eoneohag* 47:159-187.

Kim, Heung-gyu and Beom-mo Kang (2000). *Frequency Analysis of Korean Morpheme and Word Usage* 1. Seoul: Institute of Korean Culture, Korea University.

Kiparsky, Paul (2003). The phonological basis of sound change. In *The Handbook of Historical Linguistics*, ed. Brian D. Joseph and Richard D. Janda, 313-342. Oxford: Blackwell.

Ko, Kwang-Mo (1989). Explaining the noun-final change $t > s$ in Korean. *Eoneohag* 11:3-22.

Lee, Ki-Moon (1975). *Kankokugo no Rekisi* [The History of the Korean Language], translated from the Korean by Yukio Fujimoto. Tokyo: Taishukan.

Martin, Samuel (1992). *A Reference Grammar of Korean*. Tokyo: Tuttle.

New, Boris, Marc Brysbaert, Jean Veronis, and Christophe Pallier (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28:661-677.

Tang, Kevin (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *University College London Working Papers in Linguistics* 24:208-214.

Yun, Jiwon (2008). Noun-verb asymmetries in Korean phonology. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 449-457. Somerville, MA: Cascadilla Proceedings Project.

Zuraw, Kie (2000). *Patterned Exceptions in Phonology*. Doctoral dissertation, University of California, Los Angeles.

(Received: 2014.1.10)