Extension of phonotactic constraints across morphological subdomains: Evidence from Korean

It has been argued and shown by Martin (2011) that a phonotactic generalization that holds for monomorphemic words can also become effective for heteromorphemic words over time. For instance, in English, geminate consonants such as /pp/, which are not attested in the monomorphemic lexicon, are also significantly underrepresented among compounds. Thus, a constraint prohibiting geminate consonants (*GEM), which holds categorically within a morpheme, is also effective, though gradiently, across morphemes. Martin attributed such an extension of a phonotactic generalization ("leakage" in his terminology) to a Maximum Entropy (MaxEnt) learning algorithm that includes a smoothing term.

Given that there are multiple heteromorphemic domains, it is not known whether the amount of leakage differs across these domains. As suggested in the theory of Lexical Phonology (Kiparsky 1982 and many others), phonology may differ depending on the morphological domain, and the strength of the morpheme boundary, which is reflected in the order of affixes, generally decreases from derivation through compounds to regular inflection. In this study, we explore whether and how the amount of grammatical leakage correlates with the morpheme boundary strength by conducting a corpus study on a phonotactic constraint prohibiting lenis coronal obstruents after a liquid (*LT) in Korean.

To investigate the distribution of LT sequences in the lexicon of Korean, we conducted a corpus study using the lexicon created by Jun et al (in prep.) which is based on Sejong Corpus (National Institute of Korean Language 2007). Only nouns are collected for the analysis since the optional compound tensification rule in Korean, which tensifies the lax obstruent at the initial position of the second morpheme, is only applied to noun-noun compounds. Also, only native or Sino-Korean nouns are included in the analysis. We observed the tensification rate in the heteromorphemic domain to observe the effect of *LT. Both compounding and derivation have a variable process of tensification for LT so tensification can be used as a repair strategy to measure the avoidance of the LT.

In the tautomorphemic domain, among 35,179 native and Sino-Korean nouns, only five exceptions were in the lexicon. Considering that the frequency of those words is less than four times, Korean native speakers might not consider words with LT sequences to be well-formed. To confirm whether LT sequences are significantly underattested in the monomorphemic lexicon, we conducted a computational learning simulation using UCLA Phontactic Learner (Hayes and Wilson 2008). The monomorphemic words were given as the input, and the learner was asked to find bigram constraints up to 300. The O/E ratio threshold was set to 0.3. The Learner learned constraints avoiding LT sequences and the weights for those constraints are high. According to the weights learned by the Learner, the penalty score for /lt/ is 3.96, 2.94 for /ls/, 2.89 for /ltɕ/, and 3.26 on average (≈20th out of 187 constraints in total). This implies that the knowledge of *LT is likely to be present in the Korean phonological grammar. This result is also predicted by Ko (1996), who reported the unattestedness of LT in Korean monomorphemic words.

Both in derivational and compound domain, LT tensification rule, which tensifies lax coronal obstruent after a liquid, is used as a repair strategy for LT avoidance. To compare the tensification rate of L-T and non-L-T in derivational domain, we measured the tensification rates of the L-T and non-L-T context: nasal+lenis coronal (abbreviated to N-T), vowel+lenis coronal (abbreviated to V-T), liquid+lenis labial (abbreviated to L-P), and liquid+lenis dorsal (abbreviated to L-K). The overall tensification rate for derived words is lower than for compounds (29.9% for compounds and 12.6% for derived words). However, the tensification rate for LT in derivation (71.6%) is much higher than for non-LT sequences (6.2% for N-T, 6.3% for V-T; 16.9% for L-P, 9.5% for L-K).

In compounding, there is a variable process of compound tensification, which occurs regardless of LT avoidance. We compared tensification rates of LT context and non-LT contexts. L+T shows a higher rate of tensification (53.9%) than other context (33.7% for V+T, 16.6% for N+T; 43.9% for L+K, 40.0% for L+P). This accords with Kim (2016)'s observation, which found the higher rate of tensification at LT sequences.

In inflection, LT sequences are categorically allowed. Inflectional suffixes beginning with lenis coronal obstruents can be freely combined with the stems ending with liquids as in *tol+to* 'stone-postposition' and *al-ta* 'know-declarative.' This suggests that *LT does not play any role in inflection. Through the observation that inflection is closer to syntax than to the lexicon, the Split Morphology Hypothesis (Anderson 1982, 1988) proposes that inflection and derivation are separate grammar components. Considering Lexical phonology presumes separate domains between word formation and syntax, Korean speakers might not apply *LT in the syntactic domain. The avoidance of LT is correlated to the morpheme boundary strength. Comparing the tensification rates of LT in derivation, compounding, and inflection, we can say that words with morphemes tightly bound are more tensified when having LT sequences at the morpheme boundary. It means that words with strong morpheme boundary strength are more likely to be subject to the *LT constraint.

We performed a mixed-effect Bayesian regression analysis to ascertain the effects of *LT in derivation and compounding. The results show that the lax obstruent is more tensified when the first morpheme ends with a liquid and the second morpheme begins with a coronal obstruent and even more tensified when such sequences are contained in derived words.

As Martin shows the extension of the effect of the tautomorphemic constraint by the MaxEnt model, we formalize the relevant constraints and perform a computational simulation. We propose markedness constraints to avoid LT sequences by how blind each constraint is to the morphological boundary structure as Martin did. Unlike Martin, who divided lexicon into the tautomorphemic domain and the heteromorphemic domain, we set three markedness constraints discriminating the tautomorphemic, derivational, and compounding domain. Those three constraints are defined to have a stringency relation. Constraints driving compound tensification and requiring the identity of [tense] between the input and the output are also included to the model. MaxEnt Grammar Tool (Hayes 2007) was used for the simulation and all constraints had the same mean of 0 and standard deviation of 500. The results show that compounds with LT are overpredicted to be tensified (0.34%p), and derived words with LT are more overpredicted to be tensified than the compounds (2.84%p). This results from the presence of the general constraint in the MaxEnt model. This might result from the presence of the constraint penalizing monomorphemic and derived words. Since the MaxEnt model prefers to distribute constraint weights to all relevant constraints, the constraint in the intermediate level might penalize the LT candidate more.

We have seen that the effect of *LT shows gradience with respect to the morphological domain. LT sequences are avoided categorically morpheme-internally, but such avoidance is not categorical in the heteromorphemic domain. In inflection, LT sequences are always allowed. In the heteromorphemic words, *LT has more effect on the derivation than compounding. We propose that grammatical leakage from the tautomorphemic domain is strongly correlated with the morpheme boundary strength. The effect of the constraint in the tautomorphemic domain declines as the morpheme boundary strength goes weaker from the derivational domain to compounding and no effect was observed in inflection. Computational simulation supports this generalization by proving the rate of the grammatical leakage differs by the morphological domain.