Phonological Trends in Seoul Korean Compound Tensification

Seoyoung Kim
Seoul National University

1 Introduction

In a compound composed of two nouns in Korean, if the initial onset of the second element is a lax obstruent, it often undergoes tensification, as in /san + pul/ [san.p*ul] 'wild fire'. Traditionally, this phenomenon has been analyzed through an epenthesis of a mono-segmental compound juncture marker, mostly /t/. This epenthetic /t/ first triggers automatic post-obstruent tensification, where a lax consonant gets tensified after an obstruent, and then deletes itself through cluster simplification (Lee 1972 and many others). In fact, this epenthetic /t/ has been orthographically represented as <s>. Because of this, this epenthetic /t/ itself has been called *saisiot* (*sai* = between; *siot* = the letter ㅅ(/s/)) and this tensification phenomenon has been termed *saisiot* phenomenon (Chung 1980 and many others).

This tensification process does not always occur. Some compounds undergo this tensification process while others do not. Then, what decides the occurrence of tensification?

The data typically analyzed by traditional generative phonology shows a categorical pattern. Phonological rules are considered to apply whenever the defined conditions are met. Previous studies tried to set up the tensification rule and the application conditions in order to offer a complete explanation on the distribution of compound tensification. It has been widely suggested and agreed upon that this *saisiot* used to act as a genitive marker in Middle Korean, which was affixed to the end of the first component of a phrase when the two components are in a genitive relation (Ramstedt 1939 and others). Considering this earlier function as a genitive marker, when predicting the occurrence of *saisiot* in Contemporary Korean compounds, most studies focused on the morphological, semantic, or syntactic properties of the two nouns involved (Chung 1980 and many others). However, none of the suggested factors clearly divide the compounds into those that can and cannot undergo tensification.

In recent phonological research, it has been widely accepted that the distribution of the exceptions themselves is phonologically patterned (patterned exceptions; Zuraw 2000, 2010). That is, the "exceptional" cases might not be a mere exception and might have phonological reasons. For example, under the tensification rule which states that tensification apply to sub-compounds where the two components are in a modification relation (Shim 1979), the sub-compound /tol + kye.tan/ [tol.**k**ye.tan] 'stone steps' would be marked an exception. However, this compound's not undergoing tensification might be driven by other factors from various modules, such as etymology, frequency, or number of syllables. It has been noted in the literature that compound tensification does not normally occur if the second element is a bisyllabic Sino-Korean word (Lee 1972 and many others). The second element *kye.tan* is a bisyllabic Sino-Korean word. Thus, although the morphological condition forces tensification, the etymology and the stem length of *kye.tan* block it.

The example of [tol.kye.tan] shows that a single (morphological) factor does not completely decide the occurrence of tensification. Rather, various factors interact in a way that one factor might override another. Among many factors, this study focuses on phonological factors. I will examine what phonological factors and how significantly these factors contribute to the overall probability of compound tensification.

There are two studies that inspired this study. Zuraw (2011) is the first systematic study on Seoul Korean compound tensification, suggesting both phonological and non-phonological factors. Here, the data

* I would like to thank anonymous reviewers and audiences at AMP 2016 for helpful comments and suggestions. I also want to thank Jongho Jun for advising and SNU Linguistics Department for the travel grant. This work has been earlier presented at the 24th Japanese/Korean Linguistics Conference.

Proceedings of *AMP 2016*

was collected from a dictionary (Kuklip kukə yənkuwən 1999) and it might not have reflected actual pronunciations. Ito (2014) carries out a similar study on Yanbian Korean, focusing on phonological factors, especially on the OCP (Obligatory Contour Principle; McCarthy 1986) effect. For the data collection, she conducted a survey on native Yanbian Korean speakers.

In the present study, I performed a survey on Seoul Korean speakers and collected their actual pronunciations. Based on the survey results, I will show a number of significant factors: frequency, W_A (the first element of a compound) final-segment type, the place of W_B (the second element of a compound) initial onset segment, and the presence of laryngeally marked consonant. And then, diagnosing the significance of each factor, the survey results are fitted in a mixed effect logistic regression model in R (R Development Core Team 2014).

Another goal of this study is to establish a formal analysis on the Seoul Korean compound tensification phenomenon, using Optimality-Theoretic constraints (OT; Prince & Smolensky 1993). In order to capture the variable pattern of compound tensification, the numerical weights are assigned to the adopted constraints in the model of Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003, Hayes & Wilson 2008).

This paper proceeds as follows. Section 2 introduces how the survey was performed. Section 3 reports the survey results and relevant factors. Section 4 presents an OT analysis. Section 5 provides a learning simulation using a MaxEnt learner. Section 6 concludes the paper.

2 Survey

2.1 *Materials* I selected 304 native Korean noun-noun compounds, whose W_B initial onset is a lax obstruent, from the two data sources: *Korean usage frequency* (Kang & Kim 2009) and a Korean dictionary (Kuklip kukə yənkuwən 1999).

2.2 *Procedure and participants* The 304 survey items were listed in a questionnaire. The two components of each compound were given in separate parentheses with the morpheme boundary symbol '+' in between. For each compound, two possible pronunciation forms written in standard Korean orthography, one with and the other without tensification, were suggested. Twenty-one native Seoul Korean speakers, whose ages ranged from 21 to 36, were asked to choose their pronunciations, when making a compound with the two given nouns.

3 Trends in existing words

3.1 *Variation* The survey generated 6384 datapoints in total. The overall rate of compound tensification in Seoul Korean was 57% (3644 / 6384).

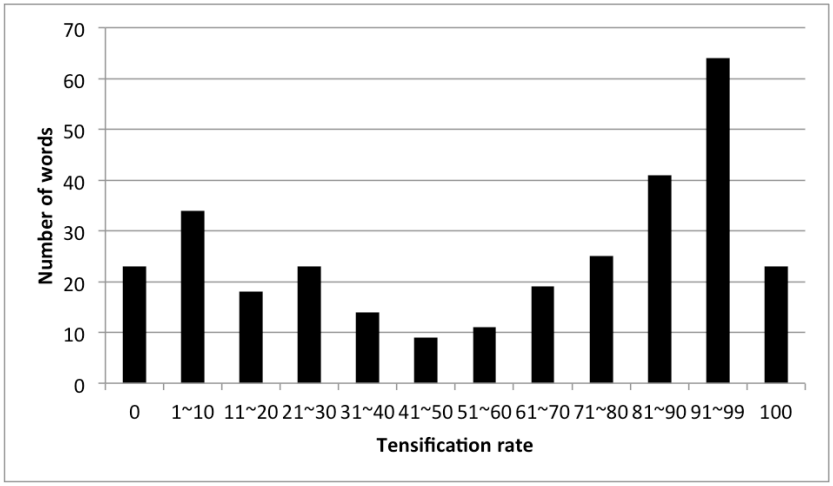

Figure 1 The number of compounds according to the tensification rates (%).

As shown in Figure 1, the tensification probability differs across the compounds.

3.2 *Frequency* It was often suggested in the literature that frequently used compounds are more likely to undergo tensification (Oh 1987, Lee 2009).

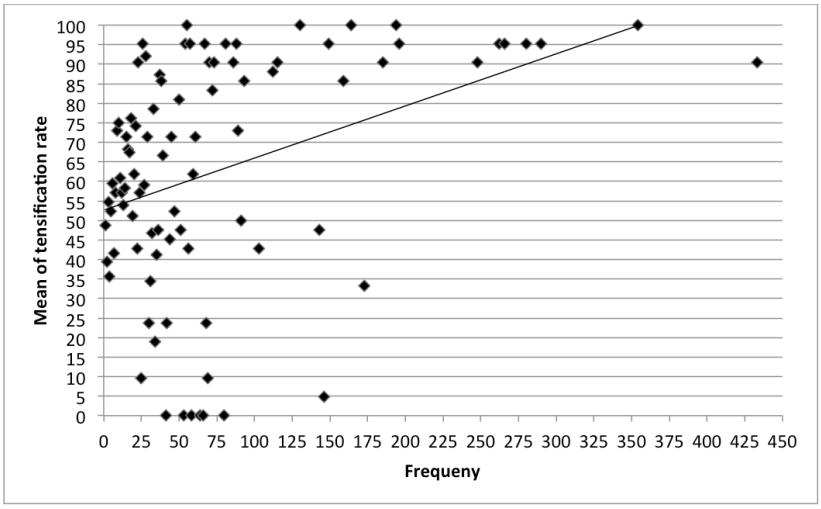

Figure 2 Averaged tensification rate (%) as a function of each usage frequency.

Figure 2 shows a positive correlation between the usage frequency and the averaged tensification rate. A linear regression analysis with the averaged tensification rate as the dependent variable and the frequency as the independent variable shows a significant result (est.=0.133, $p<0.001$). Thus, the usage frequency significantly affects the tensification probability.

3.3 *Syllable number* Some literature proposed that polysyllabic stems tend to resist compound tensification (Lee 1972 and many others).

W_A \ W_B	1	≥2
1	73.8	60.1
≥2	54.1	48.5

Table 1 Tensification rates (%) as a function of the syllable number of W_A and W_B.

As seen in Table 1, the tensification rate decreases as the syllable number increases in either W_A or W_B. A Pearson's χ^2 test with Yates' continuity correction shows that the differences between *monosyllabic* and *multisyllabic* were significant in all the contexts (when $W_A=1\sigma$: $\chi^2=28.6$, $p<0.001$; when $W_A\geq1\sigma$: $\chi^2=7.2$, $p<0.01$; when $W_B=1\sigma$: $\chi^2=44.9$, $p<0.00.1$; when $W_B\geq1\sigma$: $\chi^2=62.7$, $p<0.001$). Therefore, the syllable number of W_A and W_B is a significant factor in compound tensification.

3.4 *W_A final-segment type* It was defined in the literature that compound tensification is limited to the cases where W_A ends in a sonorant and W_B begins in a lax obstruent (Kim-Renaud 1974 and many others). In Table 2, the tensification rate according to each sonorous W_A final-segment type was calculated. Also, for the sake of comparison between sonorant and obstruent coda, the tensification rate for W_A obstruent final words was calculated.

Obstruent	86.1
Nasal	66.6
Liquid	47.9
Vowel	40.2

Table 2 Tensification rates (%) as a function of W_A final-segment type.

A Pearson's χ^2 test with Yates' continuity correction shows that obstruents and nasals, nasals and liquids, and liquids and vowels are all significantly different in their tensification effects (χ^2=120, p=2.2e-16; χ^2=135.3, p=2.2e-16; χ^2=21.1, p<0.001). Therefore, W_A final-segment type significantly influences the probability of tensification.

3.5 *Place of W_B onset consonant* Checking if W_A final coda and W_B initial onset interact, the tensification rates as a function of W_A final segments and W_B initial ones were figured.

W_B \ W_A	Obs	Nas	Liq	Vowel
Coronal	86.7	60.2	55.4	39.1
Non-coronal	85.1	73.0	39.7	41.3

Table 3 Tensification rates (%) according to the manner of articulation of W_A final segments and the place of articulation of W_B initial ones.

As seen in Table 3, if W_A ends in a liquid, the coronal consonants (*s*, *c* or *t*) are tensified much more than the non-coronal ones (*k* or *p*). A Pearson's χ^2 test with Yates' continuity correction shows that the tensification rates between the coronal and the non-coronal are significantly different after a liquid (χ^2=48, p<0.001). Thus, the place of W_B initial onset itself becomes a significant factor in tensification, when a liquid preceds it.

3.6 *Laryngeal co-occurrence restrictions* Japanese has a well-known phenomenon that occurs at the compound juncture, termed Rendaku (Ito & Mester 1986, Vance 1987), where the W_B initial obstruent undergoes voicing, as in /o.ri/ + /ka.mi/ [o.ri.**g**a.mi] 'paper folding'. It is also well known that this voicing is blocked when W_B has a voiced obstruent, as in /hi.to.ri/ + /ta.bi/ [hi.to.ri.**t**a.bi] 'traveling alone', which has been formally analyzed through co-occurrence restrictions or the OCP (Obligatory Contour Principle; McCarthy 1986) effect: voiced obstruents do not co-occur within a Japanese stem.

Given that Rendaku and compound tensification both occur at a compound juncture (Labrune 1999), the compounding process in Korean, which engenders another tense consonant, can be suppressed by the OCP constraint that bans a co-occurrence of tense consonants.

Zuraw (2011) showed that a tense consonant in either W_A or W_B blocks compound tensification to some extent. Ito (2014) proved that not only a tense but also an aspirated suppresses compound tensification. Given that tense and aspirated are both laryngeally marked, whereas lax is not (Hayes 2009b), that a tense or aspirated blocks compound tensification arises from the laryngeal co-occurrence restrictions (Ito 2014).

Also in my data, as seen in Table 4, the presence of a tense or aspirated consonant in W_B significantly lowers the tensification rate by half.

Lax	59.7
Aspirated	31.1
Tense	31.3

Table 4 Tensification rates (%) according to the type of consonant included in W_B.

A Pearson's χ^2 test with Yates' continuity correction shows that aspirated *vs.* lax, and tense *vs.* lax in W_B are significantly different in their tensification effects (χ^2=117.9, p<0.001; χ^2=135.5, p<0.001), whereas aspirated *vs.* tense is not (χ^2=0.3, p=0.5). Therefore, the presence of a laryngeally marked consonant in W_B shows a clear OCP effect.

Unlike in W_B, neither a tense nor an aspirated in W_A shows OCP effect.

Lax	58.1
Tense	62
Aspirated	56.4
Tense and aspirated (4)	4.7
Two tenses (4)	2.3

Table 5 Tensification rates (%) according to the type of consonants included in W_A.

As seen in Table 5, containing a tense rather promotes tensification (lax *vs.* tense; χ^2=4.2, p<0.05), and containing an aspirated makes no significant difference (lax *vs.* aspirated; χ^2=0.5, p=0.4). However, when there are two laryngeally marked consonants in W_A, although it is inappropriate to make a generalization in the present study due to the small sample size, a dramatic OCP effect seems to emerge. Four words containing both tense and aspirated in W_A and the other four words containing two tenses in W_A both have very low tensification rates. In the remainder of this paper, no additional analysis for this maximized OCP effect will be included, due to the low reliability. I only propose a conjecture here and leave its investigation for a future study.

3.7 *Mixed effect logistic regression model* So far, I have mentioned various factors affecting the probability of compound tensification. Investigating how significant each factor is and identifying these factors' relative strengths, I constructed a mixed effect logistic regression model using the *glmer* function from the lme4 package (Bates *et al.* 2014) in R (R Development Core Team 2014). The occurrence of tensification was the dependent variable (binary: *non-tensified* (ref), *tensified*) and the factors introduced in the previous sections were the independent variables: frequency (*log(frequency+1)* transformed), W_A final coda (backward difference coded, *vowel* > *liq* > *nas* > *obs*), W_B initial onset place (*non-coronal* (ref), *coronal*), W_A / W_B syllable number (*monosyllable* (ref), *multisyllable*), W_A consonant type (backward difference coded, *lax* > *tns* > *asp*), W_B consonant type (backward difference coded, *lax* > *tns* > *asp*). Random intercepts were set for item and subject. W_A final-segment type and W_B initial onset place were sum-coded. Table 6 shows the results.

	Est.	SE	Z	P(>|z|)
(Intercept)	- 1.72	0.63	- 2.70	0.006 **
Frequency	0.97	0.24	3.97	< 0.001 ***
W_A coda (liquid-vowel)	- 0.23	0.54	- 0.42	0.667
W_A coda (nasal-liquid)	1.99	0.47	4.18	< 0.001 ***
W_A coda (obs-nasal)	2.18	0.67	3.23	0.001 **
W_B onset place (cor)	- 0.04	0.28	- 0.16	0.867
W_A consonant (tense-lax)	0.00	0.39	0.00	0.998
W_A consonant (asp-tense)	0.16	0.57	0.29	0.768
W_B consonant (tense-lax)	- 2.57	0.58	- 4.40	< 0.001 ***
W_B consonant (asp-tense)	- 0.61	0.79	- 0.77	0.438
W_A syllable number (multi)	0.41	0.33	1.24	0.214
W_B syllable number (multi)	- 0.34	0.37	- 0.93	0.350
W_A coda (L-V): W_B onset (cor)	1.23	0.71	1.73	0.082
W_A coda (N-L): W_B onset (cor)	- 1.58	0.66	- 2.38	0.017 *
W_A coda (P-N): W_B onset (cor)	- 0.67	0.85	- 0.79	0.429

Table 6 Results of a mixed effect logistic regression model.

The result suggests that the more frequent a compound is, the more likely tensification is to apply. Regarding W_A final coda, an obstruent tensifies the following onset segment more readily than a nasal does, which in turn tensifies more than a liquid does. Regarding the OCP effect, the differences between aspirated, tense and lax consonants are insignificant in W_A, accurately mirroring the survey results where

one laryngeally marked consonant in W_A showed no significant OCP effect. On the contrary, the tensification probability becomes much lower when a tense consonant is contained in W_B than when only lax ones are in W_B. An aspirated is not different from a tense in its tensification effect in W_B, correctly reflecting the survey results where both a tense and an aspirated in W_B significantly lowered the tensification rates by a similar degree. The syllable number was not a significant factor. Lastly, there is an interaction between the W_A coda and the W_B onset. Coronals are more likely to be tensified after a liquid, compared to after a nasal.

4 Analysis

Establishing a formal analysis of Seoul Korean compound tensification, I will first propose the constraints that are responsible for the occurrence of tensification and the trends observed in the data.

4.1 *The occurrence of compound tensification* This study focuses on tensification but there is another phenomenon often observed in a compounding process: nasal lengthening. When W_A ends in a vowel and W_B begins in a nasal, the nasal lengthens as in /pæ + no.ræ/ [pæ**n**.no.ræ] <pæ**s**.no.ræ> 'sailor's song'. Since tensification and nasal lengthening both occur at the compound juncture and share the orthographic representation <s>, much of the previous literature attempted to propose a unified solution. First, most studies proposed that /t/ is inserted between two nouns (Ryu 1963 and many others) because this /t/ triggers either post-obstruent tensification or nasal assimilation depending on the following segment. Other studies proposed /s/, for <s> being the orthographic letter (Kim 1992, Moon 1997). This epenthetic /s/ can also explain both phenomena in a similar way, since /s/ is neutralized to [t] in coda anyway.

However, Chung (1980) does not posit a segment insertion. He claimed that if any segment is inserted, it creates otherwise impossible clusters (e.g. *mt*, *nt*, etc.). Therefore, instead of inserting a segment for the sake of a unified solution, he simply set up an independent rule for compound tensification: W_B initial onset is tensified when W_A final coda is sonorous.

Ito (2014) focuses on compound tensification and also adopts a way of simply tensifying the W_B initial onset, by assuming a floating [tense] feature as an underlying representation for the compound juncture marker. In the present study, I also assume that a [tense] feature is inserted between two nouns when they form a compound.

Regarding the realization of this [tense] feature, in accordance with Ito's (2014) analysis, I adopt REALIZEMORPHEME (RM; Kurisu 2001), which causes the [tense] feature to have phonological exponence, by being associated to the W_B initial onset. I also adopt IDENT(tense) which requires that the specification for the [tense] feature be identical in input and output. In this study, all the W_B initial consonants are underlyingly lax. Thus, IDENT(tense) basically blocks these onset segments from tensification. Variable occurrences of tensification in compounds come from the relative weights of these two constraints.

4.2 *Frequency of syllable number?* In §3.3, the tensification rate significantly dropped when the syllable number increased in either W_A or W_B. However, this syllable number factor was not significant in the regression model; see Table 6.

It is widely agreed upon that frequent words tend to be short. I assume that the syllable number effect was not confirmed in the regression model because this effect was already covered by the frequency factor. First, a linear regression analysis, with the syllable number as the dependent variable and the frequency as the independent variable, shows that these two factors are in a negative correlation in my data (est.=-0.002, $p<0.001$). Second, excluding either one of these two factors does not significantly drop the accuracy of the regression model, which implies that the syllable number and the frequency are not entirely separate factors. Using an *ANOVA* function for a likelihood-ratio test in R, I checked that the two mixed effect logistic regression models, one of which includes both W_A/W_B syllable number and frequency as the factors and the other which includes W_A/W_B syllable number but excludes frequency, are not significantly different ($\chi^2=11.8$, $p<0.001$, logLik=-2720.1; -2726). I also checked that the two models, one of which includes both W_A/W_B syllable number and frequency and the other which includes frequency but excludes W_A/W_B syllable number, are not significantly different ($\chi^2=6.3$, $p<0.05$, logLik=-2720.1; -2723.3).

Although these two factors account for one effect, what actually affects compound tensification is undoubtedly the frequency. Above all, the model including both factors showed that only the frequency is

significant. Also, the log likelihood, an important measure of the goodness of the model, was higher for the model including the frequency factor than that for the model including the syllable number factor (-2723.3 > -2726.0). Thus, I establish a constraint only for the frequency effect. Since my data showed a positive correlation between the frequency and the tensification rate, I adopt a constraint *TENSE/LOWFREQUENCY (*T/LF; slightly adapted from Ito 2014) that suppresses tensification for items with low frequency, which was defined to be below 33.7, an average value of all the frequencies from the survey items.

4.3 *W_A final-segment type* In §3.4, the tensification rates according to each W_A final-segment type were presented: obstruent > nasal > liquid > vowel.

Speaking of obstruents, post-obstruent tensification (POT) is obligatory in Korean, meaning that a lax obstruent after another obstruent always gets tensified. However, the tensification rate after an obstruent did not reach 100%, which might indicate that POT is not obligatory at the compound juncture. It was proven in Jun (1993, 1998) that POT is applied within an accentual phrase (AP), but not across an AP boundary. I speculate that the tensification rate for W_A obstruent-final words did not reach 100% in my survey because W_A and W_B were often produced as two independent APs and the application of POT failed. Nevertheless, the rate after an obstruent is still higher than the rates after the other sonorous codas, indicating that POT is still at work. Therefore, I adopt a *obs-lax constraint ('No lax obstruent after an obstruent in a single AP').

Regarding sonorous W_A final codas, Ito (2014) mentioned that the impetus for lenition might militate against the application of tensification because tensification is kind of fortition. And different preferences on lenition in various contexts would lead to different tensification rates. More specifically, if a given context is a preferred lenition target, a consonant would rather undergo lenition, instead of tensification. Since an intervocalic consonant is a widely attested lenition target, Ito (2014) assumes that tensification rate is the lowest when W_A ends in a vowel.

Meanwhile, Kirchner (1998) gives an insightful analysis on lenition, based on an effort-based approach. He argues that the motivation for lenition is basically to minimize the articulatory effort. The articulatory effort cost is measured by the displacement of the jaw. For example, when making vowel sounds, the jaw position is relatively lower than pronouncing consonants. Therefore, the displacement of the jaw is greater and hence, more effort is required when pronouncing a sequence of a vowel and a consonant, compared to a sequence of two consonants, because a change from a vowel to a consonant requires longer movement of the jaw. Since the purpose of lenition is the reduction of such articulatory efforts, depending on the effort required for making the gesture, lenition occurs with different probabilities; The more effort a gesture requires, the more likely lenition is to apply. For example, a consonant is more prone to undergo lenition when adjacent to a vowel, compared to when adjacent to another consonant, due to the higher degree of effort required for the former context. On this basis, he suggested the lenition-trigger hierarchy (vowels > liquids > … > nasals > stops > …), which reflects the degree of the lenition preference in each context. According to the hierarchy, the impetus for lenition is the highest when a target consonant is adjacent to a vowel, lower to a liquid and the lowest to a nasal. These different lenition preferences accord with the different tensification rates among post-vowel, post-liquid and post-nasal environments in my data. Thus, I establish different constraints that are responsible for blocking tensification in each context; *TENSE/VOWEL__ (*TNS/V_ 'After a vowel, no tensification'), *TENSE/LIQUID__ (*TNS/L_ 'After a liquid, no tensification'), and *TENSE/NASAL__ (*TNS/N_ 'After a nasal, no tensification'; all slightly adapted from Ito 2014).

According to the lenition-trigger hierarchy, the impetus for lenition is the greatest when W_A ends in a vowel, lesser in a nasal, and the least in a liquid. If the impetus for lenition is greater, tensification is more strongly blocked. Thus other things being equal, it is expected that the weights of these three constraints are as follows: *TNS/V_ > *TNS/L_ > *TNS/N_.

4.4 *Leakage* The survey result showed that the coronals (*s*, *c* and *t*) undergo tensification more readily than the non-coronals (*k* and *p*), after a liquid. This asymmetry between the coronal and the non-coronal might mirror the pattern of the existing lexicon. In the native Korean simplex word lexicon, the frequency of the tense coronal obstruents (*s**, *c** and *t**) is several times higher than that of the tense non-coronal ones (*p** and *k**) after *l* (Ko 1996). This is because in Middle Korean, after *l*, the immediately following coronal lax obstruents underwent tensification, while non-coronal ones remained intact.

Therefore, in Contemporary Korean, while sequences like *ls**, *lt**, and *lc** are easily found as in /kɨl.s*i/ 'handwriting', /p^{h}al.t*uk/ 'forearm', and /nal.c*a/ 'date', their lax counterparts *ls*, *lt*, and *lc* are not. Actually, this pattern has been formulated as a morpheme structure rule: /s, t, c/ -> [s*, t*, c*] / *l* __ (Ko 1996). Although this rule is active in the tautomorphemic domain, which is why it is called a morpheme structure rule, I hypothesize that its effect diffuses to the heteromorphemic domain as well.

Martin (2011) proved that a phonotactic constraint diffuses its weaker effect across morpheme boundaries. This process was named *leakage* in that a phonotactic generalization somewhat leaks from a tautomorphemic domain to heteromorphemic ones. Applying this concept, it can be stated that the weaker version of the phonotactic constraint that bans sequences of a liquid and a lax coronal, holds across morpheme boundaries, making coronals more prone to tensification than non-coronals. This can be captured by a general constraint *L(+)C ('A sequence of a liquid and a lax coronal is not allowed'; borrowed the constraint format from Martin 2011).

4.5 *OCP effect* In my data, as reported in §3.6, while a tense or an aspirated in W_B clearly showed the OCP effect, either one in W_A showed none. This is similar to Japanese Rendaku in that the presence of a voiced obstruent in W_B blocks voicing of the W_B initial consonant, while the presence of one in W_A does not show a substantial blockage effect (Ito & Mester 1998, Kawahara & Sano 2014). Thus, for these two phenomena, the blocking effect on marking the compound juncture comes from the OCP effect that holds within a stem. Capturing this OCP effect exerted by the presence of a laryngeally marked consonant in W_B, I adopt a constraint that blocks tensification if an aspirated or a tense consonant is contained in the same stem as the site of tensification, namely in W_B (OCP(stem) 'Do not allow a co-occurrence of laryngeally marked consonants in a stem'; adapted from Coetzee 2004).

5 Learning simulation

Formalizing the variable pattern of compound tensification, I adopt the model of MaxEnt Grammar, where constraints are assigned numerical weights. I found the specific weight of the adopted constraints, by conducting a learning simulation on the training data which are basically the survey result of mine, using Maxent Grammar Tool (Hayes 2009a).

OCP(stem)	1.676
*obs-lax	1.533
REALIZEMORPHEME	1.227
*TENSE/LIQUID__	1.102
*TENSE/VOWEL__	0.983
*L(+)C	0.785
*TENSE/LOWFREQUENCY	0.429
IDENT(tense)	0.142
*TENSE/NASAL__	$6*10^{-7}$

Table 7 Weights obtained using Maxent Grammar Tool.

The constraint OCP(stem) has the greatest weight, reflecting the survey results where the presence of a laryngeally marked consonant in W_B substantially suppressed tensification. Among the four constraints relevant to W_A coda, *obs-lax was weighted higher than the other three constraints *TNS/V_, *TNS/L_, and *TNS/N_, not only indicating that POT plays a significant role at the compound juncture, but also mirroring the survey results where the tensification rate was the highest with the W_A obstruent-final compounds. Meanwhile, considering that the tensification rate was lower with W_A vowel-final compounds than with W_A liquid-final ones, it was expected that *TNS/V_ would outrank *TNS/L_, but the result showed the opposite. This result is due to the other constraint *L(+)C.

(1) Analysis on W_A liquid-final and W_A vowel-final compound

a. /kæ.mi + cip/ 'an ants' nest'

/kæ.mi+[tense]+cip/	*TNS/L_	*TNS/V_	*L(+)C
i. kæ.mi.cip			
ii. kæ.mi.c*ip		*	

b. /tol + sot^h/ 'a stone pot'

/tol+[tense]+sot^h/	*TNS/L_	*TNS/V_	*L(+)C
i. tol.sot^h			*
ii. tol.s*ot^h	*		

In (1a), *TNS/V_ suppresses tensification by penalizing (1a.ii) and there is no opposing constraint that encourages tensification. By contrast, in (1b), *TNS/L_ suppresses tensification by penalizing (1b.ii) and there is the opposing constraint *L(+)C that promotes tensification by penalizing (1b.i). Thus, the learning simulation showed that *TNS/L_ outweighs *TNS/V_ because the tensification blocking effect of *TNS/L_ is largely offset by *L(+)C. Considering the intervening effect of *L(+)C, the greater weight of *TNS/L_ is understandable. Meanwhile, *TNS/N_ weighed almost zero. The effect of *TNS/N_ is to block tensification when W_A ends in a nasal. Considering that W_A nasal-final compounds had the higher tensification rate than W_A vowel-final and W_A liquid-final ones, it seems plausible that *TNS/N_ weighed about zero.

Lastly, I applied these weighted constraints to a test set of words which is identical to the training data. The result showed that this constraint set successfully reproduces the observed distribution of the candidates; A linear regression analysis with the observed distribution as the dependent variable and the predicted distribution as the independent variable showed a significant result (R^2=0.88, est.=1.09, p<2e-16). Thus, it is established that the grammar suggested for Seoul Korean compound tensification can be learned from the actual native speakers' pronunciation data and also can capture the various tendencies observed in this phenomenon.

6 Conclusions

In this study, Seoul Korean compound tensification is investigated. This tensification occurs in a compounding process with various probabilities. Based on the data collected through a survey, I showed that various factors play a significant role, and also that these factors contribute to the overall tensification probability. The tendencies caused by these factors are as follows: (i) tensification is more likely with high frequency items; (ii) tensification is more likely when W_A ends in an obstruent, followed by a nasal, liquid and a vowel in order; (iii) tensification is less likely when W_B has a laryngeally marked consonant; (iv) tensification is more likely when W_B ends in a liquid and W_A also begins in a coronal obstruent. In addition, I provided a formal analysis of this phenomenon. The tendencies mentioned above were formalized into the separate OT constraints. Capturing the variable pattern, MaxEnt OT was employed. The specific weights of each constraint are evaluated through Maxent Grammar Tool and this constraint set was highly successful in predicting the tensification probability.

References

Bates, Douglas M., Martin Maechler, Ben Bolker & Steven Walker. 2014. Package 'lme4': linear mixed-effects models using Eigen and S4. Version 1.0-6. http:// cran.r-project.org/web/packages/lme4/index.html.

Chung, Kook. 1980. *Neutralization in Korean: a functional view*. PhD dissertation, University of Texas at Austin.

Coetzee, Andries. 2004. *What it means to be a loser: Non-optimal candidates in Optimality Theory*. PhD dissertation, University of Massachusetts, Amherst.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenador, Andres Eriksson & Östen Dahl (eds.) *Proceedings of the Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379–440.

Hayes, Bruce. 2009a. Maxent Grammar Tool. Software package. [http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/]

Hayes, Bruce. 2009b. *Introductory Phonology*. Wiley-Blackwell. Chichester.

Ito, Chiyuki. 2014. Compound tensification and laryngeal co-occurrence restrictions in Yanbian Korean. *Phonology* 31(3): 349-398.

Ito, Junko & Armin Mester. 1986. The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Language Inquiry* 17: 47–73.

Ito, Junko & Armin Mester. 1998. Markedness and word structure: OCP effects in Japanese. Ms. University of California, Santa Cruz.

Jun, Sun-Ah. 1993. *The phonetics and phonology of Korean prosody.* PhD dissertation, Ohio State University.

Jun, Sun-Ah. 1998. The accentual phrases in the Korean prosodic hierarchy. *Phonology* 15(2): 189–226.

Kang, Beom-Mo & Hung-Gyu Kim. 2009. *Hankukə sayoŋ pinto* [Korean usage frequency]. Institute of Korean Culture.

Kawahara, Shigeto & Shin-ichiro Sano. 2014. Testing Rosen's Rule and Strong Lyman's Law. *NINJAL Research Papers* 7: 111–120.

Kim, Cha-Kyun. 1992. Phonology of Epenthetic /S^{h}/. *Journal of Korea Linguistics* 22: 191–236.

Kim-Renaud, Young-Key. 1974. *Korean consonantal phonology*. PhD dissertation, University of Hawaii.

Kirchner, Robert Martin. 1998. *An effort-based approach to consonant lenition*. PhD dissertation, UCLA.

Ko, Kwang-Mo. 1996. Two Sound Changes Related to r in Korean. *Eoneohag* 18: 31–50.

Kuklip kukə yənkuwən. 1999. *Pyocun kukə tæsacən* [Standard Korean dictionary]. Seoul: Tusantoŋa.

Kurisu, Kazutaka. 2001. *The phonology of morpheme realization*. PhD dissertation, University of California, Santa Cruz.

Labrune, Laurence. 1999. Variation intra et inter-langue: morpho-phonologie du rendaku en japonais et du sai-sios en coréen. *Cahiers de Grammaire* 24: 117–152.

Lee, Chung-Min. 1972. Boundary phenomena in Korean revisited. *Papers in Linguistics* 5: 454–474.

Lee, Ho-Young. 2009. Tensification preference of native Seoul speakers of Korean. *Journal of The Korean Society of Speech Sciences* 1(2): 151–162.

Martin, Andrew. 2011. Grammars leak: modeling how phonotactic generalizations interact within the grammar. *Language* 87(4): 751–770.

McCarthy, John J. 1986. OCP effects: gemination and antigemination. *Linguistic Inquiry* 17: 207–263.

Moon, Yang-Soo. 1997. The 'Sais-sori' Phenomena of Korean. *Journal of Humanities* 37: 63–80.

Oh, Jung-Ran. 1987. Fortition within Korean compounds. *Korean Journal of Linguistics* 12(1): 35–53.

Prince, Alan & Paul Smolensky. 1993. *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [http://www.R-project.org.]

Ramstedt, Gustav J. 1939. *A Korean Grammar*. Helsinki: Société Finno-Ougrienne.

Ryu, Chang-Don. 1963. On the additional /t/ phenomenon. *The Journal of Korean Studies* 7: 1–19.

Shim, Jae-Kee. 1979. Semantic functions of modifications in Korean. *Language Research* 15(2): 109–121.

Vance, Timothy. 1987. *An Introduction to Japanese Phonology*. Albany: State University of New York Press.

Zuraw, Kie. 2000. *Patterned exceptions in phonology*. PhD dissertation, UCLA.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory* 28(2): 417–472.

Zuraw, Kie. 2011. Predicting Korean sai-siot: phonological and non-phonological factors. Handout from paper presented at the 21st Japanese/Korean Linguistics Conference. Seoul National University.