Title

Can a Transformer Model Learn Sublexical Groups in Korean Like Us?

Journal

Japanese/Korean Linguistics, 32(1)

Author

Nam, Stanley

Publication Date

2026-07-31

DOI

10.5070/J7.66638

Can a Transformer Model Learn Sublexical Groups in Korean Like Us?*

STANLEY NAM
University of British Columbia

1 Introduction

L-Tensification (LT) is a phonological process in Korean in which coronal obstruents /t, s, tɕ/ are tensified after a lateral /l/. Tense sounds are represented with an asterisk, e.g., [t*, s*, tɕ*]. A rule for LT is provided in (1), adapted from Kim-Renaud (1974: 174).

(1) L-Tensification (adapted from Kim-Renaud (1974))
[−son, +cor] → [+tense] / l ___
'A coronal obstruent becomes tense following a /l/.'
(Note: Kim-Renaud explicitly states that this rule only applies to Sino-Korean words.)

LT applies selectively to a subset of the lexicon. Even when all phonological conditions are met, LT only applies to a certain group of lexical items. Reference grammars have traditionally attributed this selectivity to etymology, claiming that LT is limited to Sino-Korean items (Kim-Renaud 1974: 171-174, Shin et al. 2012: 203, Bae 2013: 315-319, Kim 2022: 211-212). This study challenges this view by examining how speakers and a transformer model apply LT to nonce words which have no etymology. It reports that LT applies differently depending on phonotactics.

* I thank Kathleen Currie Hall and Gunnar Ólafur Hansson for discussions and feedback throughout this project, and Miikka Silfverberg for guidance on the transformer model. I also thank Sunhoi Kim for supporting the experiment and the audience of the 32nd Japanese/Korean Linguistics Conference for comments. Any remaining errors are mine.

Japanese/Korean Linguistics 32.
Edited by Youngdong Cho, Young-Hoon Kim, and John Whitman.

Although previous studies have framed LT as an etymology-based process, drawing largely on diachronic facts about morphemes, such accounts face two major problems. First, etymological information is difficult to learn, especially in Korean, since it lacks easy ways to identify a word's etymological membership, unlike Japanese or Michif. Korean does not use separate writing systems to hint etymology as Japanese does (Martin 1972: 90, Itô and Mester 1999), nor does it have distinct phoneme inventories for different etymological groups as Michif does (Bakker and Papen 1997).

Second, there are also cases in which LT and etymology do not align. As shown in (2), not all Sino words undergo LT and not all non-Sino words do not undergo LT. Words in (2a–b) are Sino by origin but LT does not apply, contrary to what etymological accounts would predict. Words in (2c–d) are not Sino but LT applies, contrary to what etymology predicts.[1]

(2) Misalignment between etymology and the application of L-Tensification

a. /hol**t**ɛ/	[hol**t͡ɕ**ɛ]	'neglect, Sino' (cf. [hol**t***ɛ] expected)
b. /mol**tɕ**ikɑk/	[mol**t͡ɕ**igɑk]	'thoughtless, Sino' (cf. [mol**tɕ***ikɑk] expected)
c. /tɕɑŋpɑl**tɕ**ɑŋ/	[tɕɑŋbɑl**tɕ***ɑŋ]	'Jean Valjean, loan' (cf. [tɕɑŋbɑl**t͡ɕ**ɑŋ] expected)
d. /nopɛl + sɑŋ/	[nobɛl**s***ɑŋ]	'Nobel, loan + prize, Sino' (cf. [nobɛl**s**ɑŋ] expected)

In light of these learnability and empirical issues, it is implausible that speakers rely on etymology for LT. Rather, I hypothesize that phonotactics is a better predictor of LT. Previous research on etymological subclassing has shown that phonotactic generalizations correlate to some degree with etymology in Korean (Kang 1998, Park et al. 2013, Park 2020, Park 2023), and it is possible to learn etymological subgroups from phonotactics (Morita 2018, Morita and O'Donnell 2020, Nam 2023).

Based on that, this study investigates whether nonce words that reflect phonotactic generalizations of lexical subgroups show differences in LT. Specifically, this study focuses on the word-initial and word-medial cues drawing upon Park (2020). It turns out that both speakers and a transformer model selectively applied LT, but the aspects were different.

Section 2 describes the nonce word generation process. Sections 3 and 4 report LT judgments from speakers and from the phonotactic transformer model. Section 5 concludes with discussion.

2 Nonce Words

The nonce words utilized two of the phonotactic patterns that separate Sino and non-Sino words as reported in Park (2020): word-initial /h/ vs. /l/, and word-medial /w/ vs. /ɯ/ after a bilabial sound. Words with initial /h/ are overrepresented in the Sino class, whereas those with initial /l/ are more likely to be non-Sino. Likewise, the presence of medial /pwa/ or /pwʌ/ is associated with Sino, while /pɯ/ is non-Sino.

The set of nonce words consisted of 144 trisyllabic words in three subgroups that reflect the Sino, non-Sino and neutral phonotactics. As an overview of the process, a custom Python program first generated 24 etymologically vague base words with an /lt/ sequence and another 24 containing

[1] The tensification in /nopɛl + sɑŋ/ 'Nobel Prize' is not a realization of the morphological boundary '+' as the same boundary does not induce tensification elsewhere: /tɛtɕoŋ + sɑŋ/ [tɛtɕoŋsɑŋ] 'Grand Bell Awards' and /tʰɛmpʰɯltʰʌn + sɑŋ/ [tʰɛmpʰɯltʰʌnsɑŋ] 'the Templeton Prize.'

/ltɕ/.[2] Then I applied two types of phonotactic manipulations to each base word so that each has two derivatives, one toward Sino and the other toward non-Sino.

Specifically, the program first generated base words using a second-order (trigram) Markov model trained on 6,826 monomorphemic nouns in phonemic transcription.[3] Then, for each sequence of symbols, the program calculated the likelihood of being Sino or non-Sino. I selected only those sequences whose Sino-likelihood values fell between 0.499 and 0.501. These base words were designed to have two different structures: One designed to start with a vowel to allow insertion of /h/ or /l/ word-initially, and the other to contain a bilabial-vowel sequence to allow the word-medial manipulations. Additionally, they varied in LT location as a secondary factor. Once the base words were created, two types of phonotactic manipulations were applied. I added /h/ and /l/ word-initially to vowel-starting base words, and for the second type of base words, I changed the vowel to either a /w/-diphthong (i.e., /wɑ/ or /wʌ/) or /ɯ/, after a bilabial.

Table 1 illustrates the organization of the target words. Two types of base words were manipulated differently to reflect either Sino or non-Sino phonotactics: Type 1 underwent word-initial manipulation and Type 2 word-medial manipulation. Also note the secondary LT location factor: *al.tun.kuk* and its derivatives have the LT-applicable environment between the first and the second syllables, whereas *ta.mil.taŋ* and its derivatives have it between the second and the third. However, LT location was not tied to manipulation type. Type 1 also included forms like *ʌn.kal.tun* where the /lt/ sequence occurs between the second and the third syllables.

Type 1: **Word-initial manipulation**	Neutral base word	ɑl.tun.kɯk
	Toward Sino (add /h/)	**h**ɑl.tun.kɯk
	Toward non-Sino (add /l/)	**l**ɑl.tun.kɯk
Type 2: **Word-medial manipulation**	Neutral base word	tɑ.mil.tɑŋ
	Toward Sino (change /i/ → /wʌ/)	tɑ.m**wʌ**l.tɑŋ
	Toward non-Sino (change /i/ → /ɯ/)	tɑ.m**ɯ**l.tɑŋ

Table 1: Two types of nonce words

3 Production Experiment

3.1 Method

In the production experiment, speakers read nonce words that contain the /…lt…/ or /…ltɕ…/ sequence to which LT may selectively apply. I measured relative unvoiced duration for the plosive /t/ and relative closure duration for the affricate /tɕ/. Since tense productions have higher values for the two acoustic measures (Shin et al. 2012: 63-65, 79), higher values after /l/ were regarded as a result of LT.

The stimuli consisted of 288 trisyllabic words: 144 target items and the same number of distractors. Each participant was presented with only a third of the total stimuli, namely 48 target words and 48 distractors, in a Latin square design to avoid influence across stimuli. To illustrate, a participant would read the target item 다믈당 *ta.mul.taŋ* in Table 1, which might be produced

[2] https://github.com/stannam/PyScripts-dissertation/blob/main/existing_words_ngram_calc.py

[3] The raw set included 342,498 words derived from Park (2020). I removed items of other parts of speech, polymorphemic words and rare words, i.e., words with a frequency less than 100 in Kang and Kim (2009).

as [tɑmɯlt̚ɑŋ] (no LT) or [tɑmɯlt*ɑŋ] (LT applied). This participant would not see its derivatives, i.e., 다밀당 *ta.mil.taŋ* and 다뭘당 *ta.mwʌl.taŋ*, but would see the distractor 개집만 *kɛ.tɕip.man*.

Forty-three native speakers of Seoul Korean participated (gender: 27 female, 15 male, 1 preferred not to say; age: range [19–39], average 24.95, median 25). Each participant received KRW 15,000 in compensation. The experiment was conducted in a quiet room at Chung-Ang University between May 31 and June 13, 2023. To mitigate the "college sophomore" bias (Sears 1986), recruitment included local community boards, and sessions were offered until 9 p.m.

The whole procedure was designed and conducted using PsychoPy v2022.2.5 (Peirce et al. 2019). Productions were recorded through a Shure WH20 microphone at a sampling rate of 44.1 kHz and 16-bit depth. The microphone was mounted on the participant's head while they read aloud the stimuli written in standard Korean orthography on a 13-inch computer screen.

A total of 6,192 tokens (144 items × 3 repetitions × 43 speakers ÷ 3 groups) were collected. Of these, 339 (5.47%) tokens were excluded from analysis for various reasons, and 5,853 were analyzed. All 144 tokens from one participant (ID: 10917) were excluded due to poor recording quality; 190 from others were excluded due to production errors (e.g., mispronunciation, stutter or cough); and five due to other technical errors (e.g., cropped recording). Of the 5,853 tokens analyzed, /t/ productions comprised 2,935 (50.15%), and /tɕ/ productions made up 2,918 (49.85%).

For acoustic analysis, relative unvoiced duration was measured for /t/ after /l/ and relative closure duration for /tɕ/. Both are known correlates of tensification. Higher values indicate tense production, that is, LT application. Tokens were first segmented using the Montreal Forced Aligner (McAuliffe et al. 2017), and then focus windows were set as the target and the two flanking segments. After measuring the absolute durational values, each value was divided by the duration of the entire window, so that the resulting relative values were comparable across items.

Variable	Type	Levels	Reference level	Description
Fixed effects				
`phntcs`	Categorical	neutral, sino, non-sino	neutral	Phonotactic bias
`man_loc`	Categorical	initial, medial	initial	Location of phonotactic manipulation
`LT_loc`	Categorical	12, 23	12	Location of LT environments: 12 for syllables 1–2, 23 for syllables 2–3.
Random effects				
`participant_id`	Grouping factor	N/A		Variation by participants
`base_id`	Grouping factor	N/A		Variation by base words

Table 2: Fixed and random effects in the mixed-effects models.

For statistical analysis, linear mixed-effects models were fit using `lme4` (Bates et al. 2015) in R (R Core Team 2024) with two response variables, i.e., relative unvoiced duration for plosives and relative closure duration for affricates. Due to technical issues, `afex` (Singmann et al. 2024) was used to include random intercepts and slopes. Table 2 provides an overview of the variables.

For the affricate data, the most complex model was selected with all possible three-way interactions and random intercepts and slopes. However, a full model was singular for the plosive data, so a simpler model was fit without the three-way interaction involving `LT_loc`, because it was not the primary focus of the current study.

3.2 Results

The effects of `LT_loc` and `phntcs2` were significant in the plosive data, as shown in Table 3. The estimate for `phntcs2` means that nonce words manipulated to reflect non-Sino phonotactics showed a reduction of relative unvoiced duration by 0.0053, indicating less tensification after /l/. Additionally, nonce words with an LT environment between the second and third syllables exhibited longer unvoiced durations regardless of phonotactics. These results suggest that non-Sino phonotactics *suppress* LT application. See Table 3 and the top left panel of Figure 1 for details.

In the affricate data, in contrast, neither `phntcs2` nor `phntcs3` showed a significant main effect. Instead, the three-way interaction among `phntcs3`, `man_loc` and `LT_loc` was significant, indicating conditional effects emerging in specific combinations. Sino phonotactics increased LT likelihood under certain structural conditions. Particularly, words with medial manipulations toward Sino phonotactics (i.e., /pwa/ or /pwʌ/) showed longer closure durations when the LT environment was between the second and third syllables. Additionally, `man_loc` and `LT_loc` showed significant main effects. Medial manipulations and LT environments between syllables 2 and 3 showed longer relative closure durations. None of the phonotactic manipulation types showed a significant main effect. These results show that Sino phonotactics conditionally promote LT. Table 4 summarizes fixed-effects estimates, and the top right panel of Figure 1 shows the general pattern in affricates.

In summary, the production experiment shows a nuanced picture, in which non-Sino phonotactics suppresses LT in plosives, while Sino phonotactics promotes it in affricates, but only under specific conditions. To exclude other factors speakers might have brought to the task, the next section examines a purely phonotactic transformer model's predictions for the same data.

	Estimate	Std. Error	df	t value	Pr(>|t|)	Sig.
(Intercept)	0.0760	0.0065	53.32	11.78	< .001	
`phntcs2`	−0.0053	0.0020	23.12	−2.66	0.014	*
`phntcs3`	0.0019	0.0032	38.82	0.58	0.562	
`man_loc`	0.0055	0.0050	30.78	1.10	0.281	
`LT_loc`	0.0481	0.0046	22.08	10.43	< .001	***
`phntcs2:man_loc`	0.0069	0.0038	31.12	1.81	0.080	
`phntcs3:man_loc`	0.0006	0.0056	28.74	0.11	0.910	
Note:	`phntcs2` = Neutral → non-Sino, `phntcs3` = Neutral → Sino					
Significance:	* p < .05, ** p < .01, *** p < .001					

Table 3: LMM fixed-effects estimates for relative unvoiced duration in the plosive data

	Estimate	Std. Error	df	t value	Pr(>|t|)	Sig.
(Intercept)	0.0639	0.0055	51.20	11.57	< .001	
`phntcs2`	−0.0045	0.0028	35.22	−1.65	0.108	
`phntcs3`	−0.0007	0.0023	33.16	−0.29	0.773	
`man_loc`	0.0092	0.0040	29.08	2.27	0.031	*
`LT_loc`	0.0230	0.0040	30.03	5.71	< .001	***
`phntcs2:man_loc`	0.0010	0.0050	27.86	0.20	0.843	
`phntcs3:man_loc`	−0.0064	0.0042	29.67	−1.52	0.138	
`phntcs2:LT_loc`	0.0002	0.0050	29.12	0.04	0.972	
`phntcs3:LT_loc`	0.0007	0.0043	30.92	0.16	0.874	
`man_loc:LT_loc`	0.0033	0.0081	28.83	0.41	0.687	
`phntcs2:man_loc:LT_loc`	−0.0157	0.0100	27.59	−1.58	0.126	
`phntcs3:man_loc:LT_loc`	0.0212	0.0084	29.30	2.53	0.017	*

Note: `phntcs2` = Neutral → non-Sino, `phntcs3` = Neutral → Sino
Significance: * p < .05, ** p < .01, *** p < .001

Table 4: LMM fixed-effects estimates for relative closure duration in the affricate data

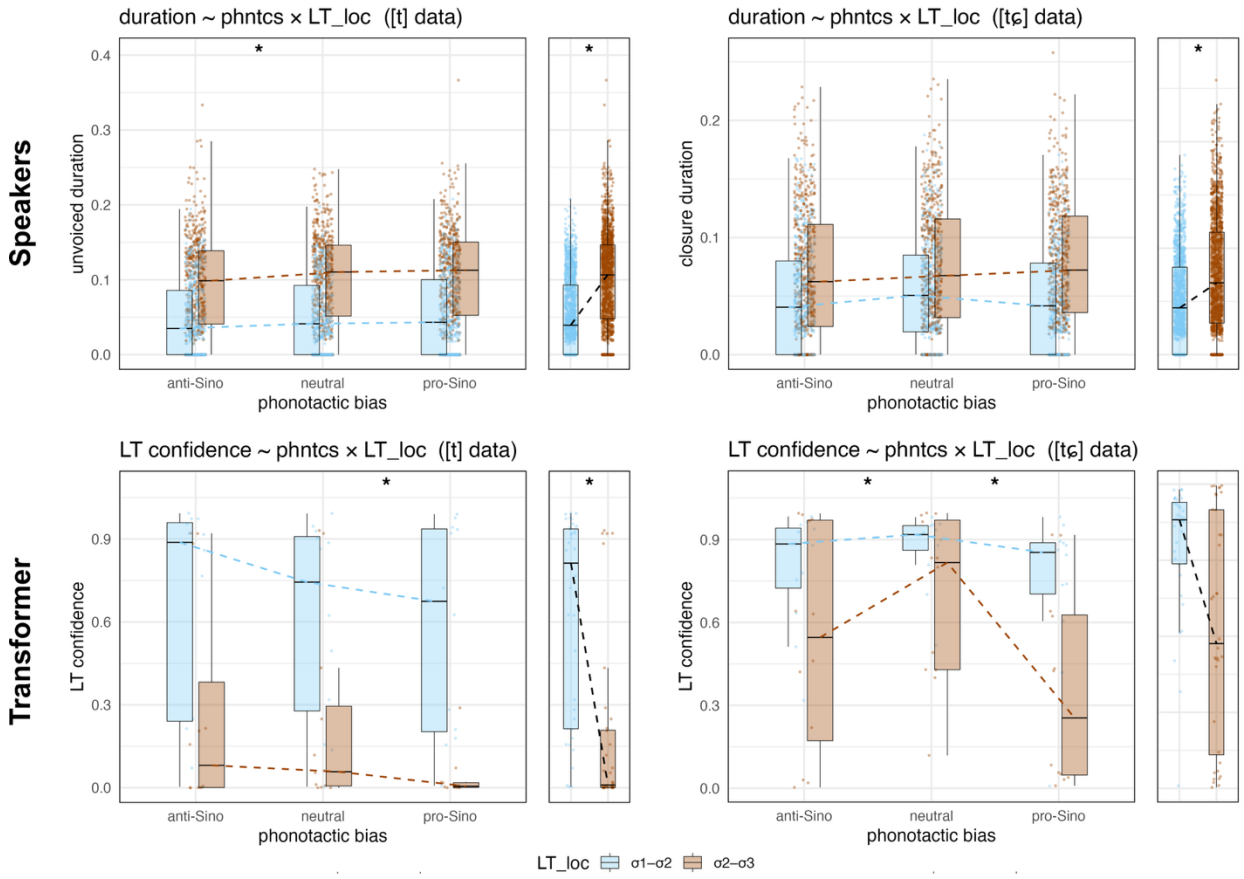

Figure 1: Interaction of `phntcs` and `LT_loc` on relative durations (top) and LT confidence (bottom) (the asterisk on the main panel for sig. phonotactics; the one on the side panel for sig. LT location)

4 Transformer model

4.1 Model Training and Evaluation

A neural model with the transformer architecture (Vaswani et al. 2017) was trained on Korean nouns using Fairseq (version 0.12.2; Ott et al. 2019). I adopted Klimaszewski's (2023) settings except for several parameters, whose values I adjusted through manual tuning informed by the model's loss trajectory.[4] The final model was configured with four encoder layers, four decoder layers and two attention heads for both the encoder and decoder. See Table 5 for major changes.

Parameter	Fairseq flag	Klimaszewski (2023)	This study
Architecture	`--arch`	`transformer_wmt_en_de`	`transformer`
Half precision	`--fp16`	(included)	(removed)
# of encoder layers	`--encoder-layers`	N/A	4
# of decoder layers	`--decoder-layers`	N/A	4
Encoder attention heads	`--encoder-attention-heads`	N/A	2
Decoder attention heads	`--decoder-attention-heads`	N/A	2
Batch size	`--batch-size`	N/A	32
Evaluation metric	`--best-checkpoint-metric`	`bleu`	`loss`
Loss function	`--criterion`	`label_smoothed_cross_entropy`	`cross_entropy`

Table 5: Fairseq parameter settings for training a transformer model

The training dataset consisted of 31,422 input-output pairs in phonemic transcriptions, derived from Park (2020). Note that nonce words described in Section 2 were not included in training. The output forms were results of all phonological processes in Korean, including but not exclusively L-Tensification. Importantly, the dataset did not include etymological information. The data were divided into an 80-20 training-development split.

As the loss curves in Figure 2 show, both training loss (in blue) and validation loss (in orange) decreased rapidly during the early phases of training. The validation loss plateaued after Epoch 100, and the gap widened. Checkpoint #110 was chosen as the final model since it had the highest validation accuracy of 96.56% among the five checkpoints with the lowest validation loss.

Epoch	Validation loss	Validation accuracy (%)
Checkpoint #103	0.034	96.18
Checkpoint #105	0.030	96.34
Checkpoint #110	0.032	96.56
Checkpoint #130	0.032	96.34
Checkpoint #132	0.032	96.32

Table 6: Validation loss and accuracy of candidate checkpoints

[4] As listed in Table 5, cross-entropy was used as the loss function. The loss is higher if the model's prediction diverged more from the gold standard. The loss trajectory refers to the change of this value over the course of training, which can be illustrated like Figure 2.

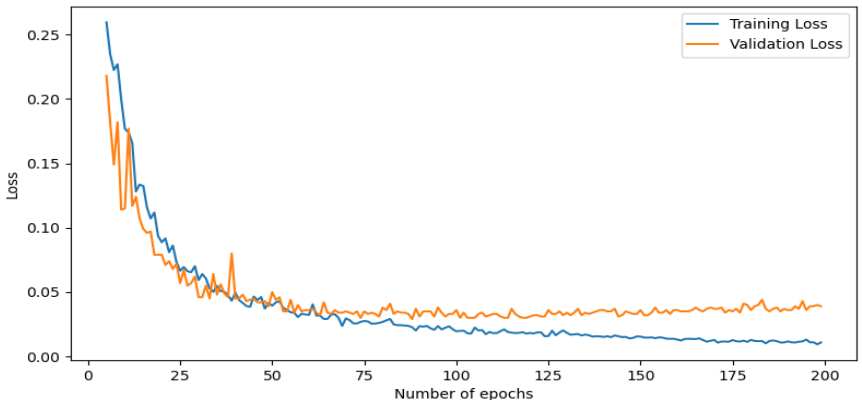

Figure 2: Training and validation loss curves

4.2 Performance on Existing Words

To understand how well Checkpoint #110 performed on real Korean words that it did not see during the training session, its performance on 6,284 items in the validation set was compared to the baseline of a rule-based model.[5] It sequentially applied a series of phonological rules from Chapter 8 of Shin et al. (2012), which do not require extra-phonological information. Table 7 lists the rules with an example and the corresponding section in Shin et al. (2012).

The non-neural baseline model showed a validation accuracy of 93.73%, or 394 errors out of 6,284 words. In contrast, all in Table 6 exceeded 96%, including the selected model. The 2.83%-point difference between Checkpoint #110 and the baseline suggests that the transformer model captures generalization beyond simple sequential patterns, likely including the one related to LT.

Rule	Example		Shin et al. (2012)
Aspiration	/to**kh**ɑk/	→ [to**kʰ**ɑk] ‘self-study’	§ 8.2.6
Obstruent Nasalization	/kɑ**km**ɑk/	→ [kɑ**ŋm**ɑk] ‘cornea’	§ 8.2.2
Liquid Nasalization	/tɑ**ml**on/	→ [tɑ**mn**on] ‘discussion’	§ 8.2.3
Lateralization	/si**ln**ɛ/	→ [si**ll**ɛ] ‘interior’	§ 8.2.4
Post-Obstruent Tensification	/mʌ**kt**ɑ/	→ [mʌ**kt***ɑ] ‘to eat’	§ 8.2.1
Complex Coda Simplification	/sɑ**lm**/	→ [sɑ**m**] ‘life’	§ 8.1.2
Coda Neutralization	/pi**tɕʰ**, pi**tɕ**, pi**s**/	→ [pi**t**] ‘light, debt, comb’	§ 8.1.1

Table 7: Phonological rules considered in the baseline model from Shin et al. (2012)

Now focusing on LT, Checkpoint #110 made correct predictions for 113 of the 121 LT-applicable items in the validation set, and it made eight incorrect predictions (accuracy: 93.39%). Table 8 lists the eight mistakes. In the ‘Error’ column, the upward arrow (▲) indicates the case where the model overapplied LT when it should not, and conversely, the downward arrow (▽) indicates underapplication. Overapplication happened near word-initial positions or in bisyllabic words, while underapplication happened near word-final positions.

[5] https://github.com/stannam/hangul_to_ipa/tree/dissertation.

Input	Predicted	Gold (Park 2020)	Error
tɕul.toŋ	tɕul.**t*** oŋ	tɕul.**ʈ**oŋ	▲
no.sɯ.tʰɛl.tɕi.ʌ	no.sɯ.tʰɛl.**tɕ*** i.ʌ	no.sɯ.tʰɛl.**ʈɕ**i.ʌ	▲
kwɑl.sɛ	kwɑl.**s*** ɛ	kwɑl.**s**ɛ	▲
pʰɑl.tɕaŋ	pʰɑl.**tɕ*** aŋ	pʰɑl.**ʈɕ**aŋ	▲
nɑm.njʌ.sɑŋ.jʌl.tɕi.sɑ	nɑm.njʌ.sɑŋ.jʌl.**ʈɕ**i.sɑ	nɑm.njʌ.sɑŋ.jʌl.**tɕ*** i.sɑ	▽
pɑ.lɑ.mil.tɑ	pɑ.lɑ.mil.**ʈ**ɑ	pɑ.lɑ.mil.t*ɑ	▽
il.ku.wʌl.sim	il.ku.wʌl.**s**im	il.ku.wʌl.**s***im	▽
il.tɕaŋ.il.tɑn	il.tɕ*ɑŋ.il.**ʈ**ɑn	il.tɕ*ɑŋ.il.**t***ɑn	▽

Note: ▲ indicates overapplication; ▽ indicates underapplication.
The periods mark syllable boundaries, which were not included in the training.

Table 8: Transformer model's prediction errors on existing Korean words

To summarize, the transformer model performed well on doing phonology with existing words that it did not see before. It showed a better performance than a rule-based baseline model, and errors in terms of LT application were limited.

4.3 Nonce Word Predictions

The analyses so far show that Checkpoint #110 generalizes beyond its training data and satisfactorily handles phonology in real words. It also predicts LT with considerable accuracy, suggesting that phonotactics, in the absence of etymological cues, provided enough information for the selective process. Now I turn to evaluate how well the model performs with respect to predicting LT in the same 144 nonce words that were used in the production experiment.

For the 144 nonce words, LT confidence scores were calculated as a softmax-normalized probability of the LT-applied output, so that the model's predictions can be compared with the results from the production experiment. As with the durational measures in Section 3, these confidence values ranged from 0 to 1, and the higher value indicates higher confidence that the words should undergo LT.

Two linear mixed-effects models were fit to predict these confidence scores. The model structure is similar to Section 3, with the same fixed effects, but the random effects were simplified: the by-participant term was excluded as there was no variation by participants, and the random slopes were removed for the by-item term as there is only one observation per phonotactics × base word. There is no statistical necessity *a priori* to split the models for plosives and affricates, but I followed the structure of Section 3 for comparability.

In the plosive data, the main effects of `phntcs3`, `LT_loc` and `man_loc` were significant. Manipulations toward Sino phonotactics decreased LT confidence scores by 0.1115. Regardless of phonotactics, nonce words with LT environment between the second and third syllables exhibited lower LT confidence, and the ones with a medial phonotactic manipulation exhibited higher LT confidence. Additionally, a marginal interaction between `phntcs2` and `man_loc` indicates that non-Sino phonotactics, only when applied word-medially, enhanced LT confidence by 0.1794. See Table 9 for fixed-effects estimates for plosives.

	Estimate	Std. Error	df	t value	Pr(>|t|)	Sig.
(Intercept)	0.4185	0.0603	34.64	6.94	< .001	
phntcs2	0.0483	0.0440	48.00	1.10	0.277	
phntcs3	−0.1115	0.0440	48.00	−2.54	0.015	*
man_loc	0.2588	0.1224	34.64	2.12	0.042	*
LT_loc	−0.3721	0.1207	34.64	−3.08	0.004	**
phntcs2:man_loc	0.1794	0.0892	48.00	2.01	0.050	*
phntcs3:man_loc	−0.1490	0.0892	48.00	−1.67	0.101	
phntcs2:LT_loc	−0.0100	0.0880	48.00	−0.11	0.910	
phntcs3:LT_loc	−0.1717	0.0880	48.00	−1.95	0.057	
man_loc:LT_loc	−0.0699	0.2447	34.64	−0.29	0.777	
phntcs2:man_loc:LT_loc	0.1333	0.1784	48.00	0.75	0.458	
phntcs3:man_loc:LT_loc	−0.0508	0.1784	48.00	−0.29	0.777	
Note:	phntcs2 = Neutral → non-Sino, phntcs3 = Neutral → Sino					
Significance:	* p < .05, ** p < .01, *** p < .001					

Table 9: LMM fixed-effects estimates for LT confidence in the plosive data

In the affricate data, both phntcs2 and phntcs3 showed significant main effects, decreasing LT confidence scores by 0.1170 and 0.2358, respectively. The two-way interaction between phntcs2 and man_loc increased LT confidence by 0.3287, suggesting non-Sino phonotactics, when applied medially, increased the confidence despite the overall negative main effect. In addition, the two-way interaction between phntcs3 and LT_loc decreased the response variable by 0.2036. Finally, the three-way interaction among phntcs2, man_loc and LT_loc was significant, indicating that non-Sino phonotactics increased LT confidence under certain structural conditions. Particularly, words with a word-medial non-Sino manipulation showed higher LT confidence scores, only when the LT environment was between the second and third syllables. See Table 10 for fixed-effects estimates for affricates.

	Estimate	Std. Error	df	t value	Pr(>|t|)	Sig.
(Intercept)	0.7788	0.0483	37.15	16.13	< .001	
phntcs2	−0.1170	0.0383	48.00	−3.05	0.004	**
phntcs3	−0.2358	0.0383	48.00	−6.16	< .001	***
man_loc	0.0958	0.0979	37.15	0.98	0.334	
LT_loc	−0.1782	0.0965	37.15	−1.85	0.073	
phntcs2:man_loc	0.3287	0.0777	48.00	4.23	< .001	***
phntcs3:man_loc	0.0347	0.0777	48.00	0.45	0.657	
phntcs2:LT_loc	−0.0866	0.0766	48.00	−1.13	0.264	
phntcs3:LT_loc	−0.2036	0.0766	48.00	−2.66	0.011	*
man_loc:LT_loc	0.1259	0.1958	37.15	0.64	0.524	
phntcs2:man_loc:LT_loc	0.3650	0.1554	48.00	2.35	0.023	*
phntcs3:man_loc:LT_loc	0.0059	0.1554	48.00	0.04	0.970	
Note:	phntcs2 = Neutral → non-Sino, phntcs3 = Neutral → Sino					
Significance:	* p < .05, ** p < .01, *** p < .001					

Table 10: LMM fixed-effects estimates for LT confidence in the affricate data

These results contradict the direction predicted by phonotactics and the production results. If LT follows phonotactic patterns, and if the phonotactic asymmetry between Sino and non-Sino extends beyond etymology, nonce words with Sino phonotactics should show higher LT likelihood, and those with non-Sino phonotactics should show lower LT likelihood. The results of the production experiment support that this generally holds: non-Sino phonotactics suppressed LT application in plosives, while Sino phonotactics promoted LT under particular structural conditions in affricates. However, the transformer model behaved quite differently. Both Sino and non-Sino manipulations suppressed LT in affricates, and Sino phonotactics suppressed LT in plosives as well. Figure 1 makes this comparison visually clear. The top panels show the speakers' results, and the bottom panels show the transformer model's behaviour.

4.4 Attention Analysis

To answer 'how' the model reached the predictions described in 4.3, cross-attention weights were analyzed. At each decoder layer, the model pays different degrees of attention to each input segment (Vaswani et al. 2017). Cross-attention weights indicate how much attention each input position gets. In other words, for a given output segment, the cross-attention distribution provides an approximation of how strongly individual input segments contribute to that output segment, although it cannot be understood as transparent evidence of the model's 'internal reasoning' (Serrano and Smith 2019).

Due to practical limitations, only CVC-CVC-CVC nonce words (n=25) were used in the attention analysis. These items were divided into two structural types: those with the LT-applicable environment between the first and second syllables, and the others with the environment between the second and third syllables. For each type, cross-attention weights were averaged across items and visualized as heatmaps.

Figure 3 visualizes the cross-attention weights in Layer 0, Head 0, representing early decoder layers. The left and right panels in Figure 3 correspond to the two structural types by LT-applicable location. Each panel includes red rectangles marking the LT-applicable consonant sequence. In both types, the model exhibited block-like attention concentrated to particular positions in the source sequence. Specifically, both panels show that the model attends to the initial consonant (C1) and the second vowel (V2), suggesting that these input positions treated as important at the early stage.

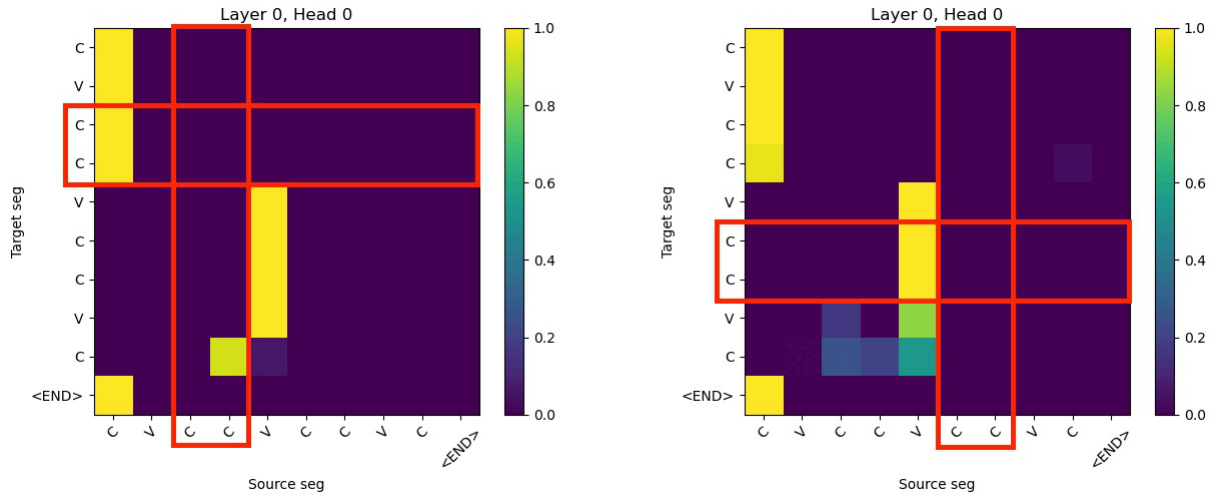

Figure 3: Cross-attention heatmaps from decoder Layer 0 Head 0, averaged over CVC-CVC-CVC items

By contrast, attention patterns in Layer 3 (both Heads 0 and 1) are more monotonic and diagonal. Figure 4 shows the cross-attention distribution in Layer 3, Head 0, using the same panel layout and red rectangles to mark the LT-applicable environments. At this higher layer, the model exhibits a largely position-to-position alignment, with each target segment attending primarily to the corresponding segment in the input sequence.

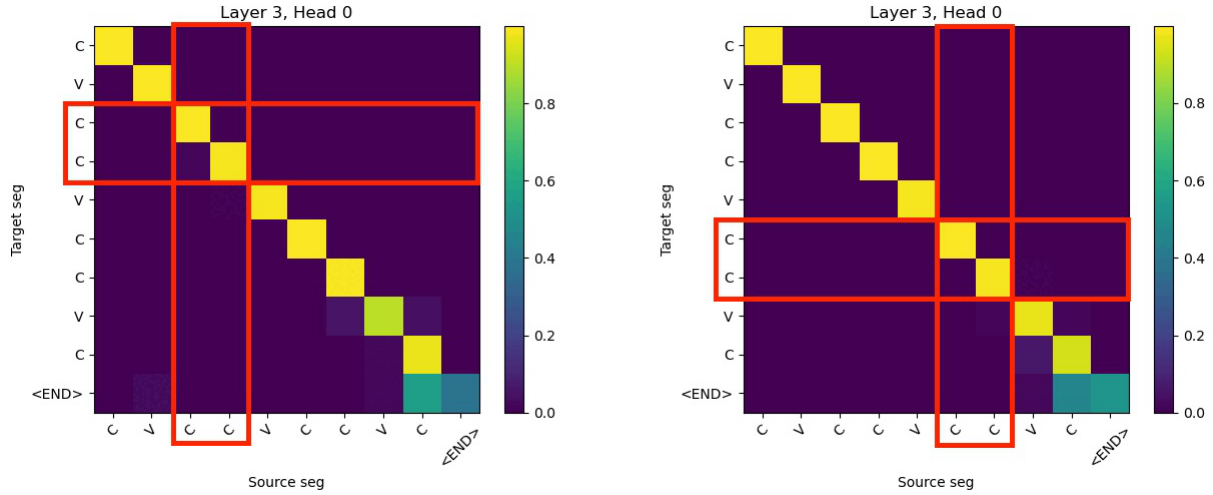

Figure 4: Cross-attention heatmaps from decoder Layer 3 Head 0, averaged over CVC-CVC-CVC items

In summary, the early decoder layer shows higher attention to particular positions, namely the initial consonant and the second vowel, while the higher layer appears to handle sequential alignment. Importantly, no layer-head pair appeared to be dedicated to LT, since no particular heatmap exhibited attention distribution localized horizontally inside the red rectangles that mark

the LT-applicable positions.[6] Given that LT affects only a featural specification of a single consonant, namely the laryngeal feature, the lack of a dedicated layer-head is not surprising. It is plausible that the model relies on the general phonotactic representation it forms during the early phase (e.g., Layer 0, Head 0) when it actualizes the target consonant later.

5 Discussion and Conclusion

This study examined how speakers and a pure phonotactic transformer model assess the applicability of L-Tensification (LT), a selective phonological process traditionally attributed to etymological conditioning. In Korean, Sino words generally undergo LT, while non-Sino do so less, and I hypothesized that this non-uniformity does not emerge from etymology per se, but from phonotactic patterns that correlate with etymology. The production experiment and machine learning model prediction involved 144 nonce words that were carefully controlled to measure influences of word-initial and word-medial phonotactic cues identified by Park (2020) as correlating with etymological classes. A wug-test design was well suited for determining whether etymology or phonotactics better predicts LT, because nonce words lack etymology. Etymological accounts would predict no LT application or random variation in the nonce words. In contrast, if phonotactics plays a role, nonce words with distinct phonotactic profiles should show different LT patterns.

The results showed that both speakers and the transformer model applied LT selectively. However, the aspect of this selectivity differed between speakers and the model. Speakers were sensitive to non-Sino phonotactic cues in plosives and, under specific structural conditions, to Sino phonotactic cues in affricates. The direction of the effects was consistent with the hypothesis: Sino phonotactics increased LT likelihood, whereas non-Sino phonotactics decreased it. The transformer model also showed selective application, but its pattern diverged from speakers' behaviour in a manner that ran counter to the hypothesized directions. Specifically, both Sino and non-Sino phonotactics decreased LT confidence in the transformer results. Taken together, these findings suggest that phonotactics influence LT patterns and that conventional etymology-based accounts cannot adequately explain the LT application.

As for the discrepancy between speakers and the transformer model, there are at least two possible explanations: (i) the task may have been too difficult for the model, or (ii) speakers may not rely solely on phonotactics. First, the nonce words presented to the transformer model may have deviated too far from the data it was trained on. The common assumption in machine learning is that the training and test data are from the same underlying distribution. When they are not, performance can deteriorate in what is commonly known as out-of-distribution (OOD) data degradation (Daumé III 2017: 104-115). If the nonce words in this study were indeed OOD, this might explain why the model performed poorly. Recall from Section 2 that the phonotactically biased nonce words were synthetic, rather than being purely sampled from a trigram Markov model. Because of how these items were generated, they might deviate substantially from the training data, potentially explaining why the model performed differently from speakers. On top of that, the OOD issue is likely exacerbated by the model's intentionally simple configuration. Recall from Section 4.1 that it had only four encoder and four decoder layers and two attention heads each.

[6] Suppose there were a specific layer-head dedicated to LT that uses the word-initial consonant in addition to /l/. In that case, the heatmap for that layer-head would show increased attention at two source positions within the red horizontal rectangle: the left-most source segment corresponding to the word-initial consonant, and the segment corresponding to /l/, the direct trigger of LT.

Although such a simple architecture was intended to facilitate interpretability, this choice leaves little redundancy that might otherwise have made the model more resilient to distribution shifts.

A second possible explanation for the speaker-model discrepancy, raised during the discussion at Japanese/Korean Linguistics 31, is that speakers may rely on orthographic form in their decisions. Due to the setting of the experiment, speakers were exposed to the Hangul letters, and the visual shapes might have shown influence. Confirming this speculation would require a dedicated psycholinguistic study on the role of visual cues in LT judgments.

Future studies can proceed in two main directions: (i) refining the diachronic and synchronic understanding of LT within Korean linguistics, and (ii) extending phonotactic approaches to other non-uniform processes cross-linguistically.

Regarding the first direction, LT itself requires more attention in Korean linguistics. One might easily connect LT to Post-Obstruent Tensification, especially because of the fact that the coda /l/ in the Sino morphemes is historically an obstruent. However, word-medial tensification had not been phonologized in Korean until the 17th century (Jung 2003, Kim 2013), and a survey in *Hunmongcahoe* (1527) reveals that all contemporary /l/-final morphemes had already shifted to the coda /l/ by that time. Thus, why coda /l/ triggers tensification remains unexplained. Besides, if LT is the norm and non-Sino phonotactics prevent its application as the speakers' results suggest, /l/ would be the only sonorant that appears to induce tensification independent of a morpheme boundary. This is something that calls for further attention. A final thread of possible expansion with regards to LT would be a follow-up nonce-word experiment. This study did not include /s/, so future work should investigate how LT applies selectively to nonce words with /s/. Moreover, a follow-up experiment should control materials with respect to manipulation location and LT location, given the `man_loc` and `LT_loc` effects observed here.

Second, a broader cross-linguistic investigation of selective processes may help clarify how phonotactic and etymological factors interact. Take Maltese, for example. It is similar to Korean (or Japanese) in terms of historical stratification of the lexicon and misalignment between etymology and (etymologically-expected) selective morphological processes. It exhibits historical strata of Semitic, Romance and English (Borg and Azzopardi-Alexander 1997), comparable to the strata of native, Sino and loanwords in Korean or Japanese. Although etymology predicts that non-concatenative morphology should occur only in Semitic words, counterexamples reveal a mismatch between etymology and selective rule application (Mifsud 1995). Michif is another language that can provide interesting insight (Papen 2003, 2014). As a mixed language between Cree and French, Michif contains selective phonological rules that apply differently to French- and Cree-origin lexical items. However, it lacks diachronic layering characteristic of stratified languages like Korean, Japanese or Maltese. Since its sublexical classes are not the result of diachronic development, Michif can be an ideal testing ground for investigating whether purely phonotactically-defined sublexical classes alone, without historical factors, can give rise to distinct phonologies within a single language (Becker and Gouskova 2016).

References

Bae, J. 2013. *Hankwukeuy Palum.* [The Pronunciation of Korean]. Seoul: Samkyengmunhwasa.

Bakker, P., and R. A. Papen. 1997. Michif: A Mixed Language Based on Cree and French. *Contact Languages: A Wider Perspective,* ed. S. G. Thomason, 295–363. Amsterdam: John Benjamins Publishing.

Bates, D. M., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1): 1–48.

Becker, M., and M. Gouskova. 2016. Source-Oriented Generalizations as Grammar Inference in Russian Vowel Deletion. *Linguistic Inquiry* 47(3): 391–425.

Borg, A., and M. Azzopardi-Alexander. 1997. *Maltese.* Milton: Routledge.

Daumé III, H. 2017. *A Course in Machine Learning.* Self-published. http://ciml.info/

Itô, J., and A. Mester. 1999. The Phonological Lexicon. *The Handbook of Japanese Linguistics*, ed. N. Tsujimura, 62–100. Malden: Wiley-Blackwell.

Jung, S.-C. 2003. Etwucaumkwunuy Kyengumhwawa Kyekumhwa. [On the Phonological Change of the Word-Initial Consonantal Clusters in Korean]. *Hankwukmunhwa* 32: 31–48.

Kang, B.-M., and H. Kim. 2009. *Hankwuke Sayongpinto.* [The Usage Frequency of the Korean Language]. Seoul: Hankwukmunhwasa.

Kang, Y. 1998. Hankwuke Ehwipu Kwuco. [The Organization of the Korean Lexicon]. *Studies in Phonetics, Phonology and Morphology* 4: 55–67.

Kim, G.-A. 2022. *Hankwuke Umunlon.* [Korean Phonology]. Seoul: Hankwukmunhwasa.

Kim, H. 2013. Cangayum Twi Kyengumhwaey Tayhan Thongsiloncek Kochal. [Diachronic Research on Post-Obstruent Tensification]. *Kwukekyoyuk* 141: 1–29.

Kim-Renaud, Y.-K. 1974. Korean Consonantal Phonology. Doctoral dissertation, University of Hawai‘i.

Klimaszewski, M. 2023. Fairseq 101 - train a model: Train your first Fairseq model - tutorial for NLP@WUT class. https://mklimasz.github.io/blog/2023/fariseq-101-train-a-model/

Martin, S. 1972. Nonalphabetic Writing Systems: Some Observations. *Language by Ear and by Eye: The Relationships Between Speech and Reading*, eds. J. F. Kavanagh, and I. G. Mattingly, 81–102. Cambridge: The MIT Press.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Paper presented at the *18th Conference of the International Speech Communication Association*, Stockholm, Sweden.

Mifsud, M. 1995. *Loan Verbs in Maltese: A Descriptive and Comparative Study.* Leiden: Brill.

Morita, T. 2018. Unsupervised Learning of Lexical Subclasses from Phonotactics. Doctoral dissertation, Massachusetts Institute of Technology.

Morita, T., and T. J. O’Donnell. 2020. Statistical Evidence for Learnable Lexical Subclasses in Japanese. *Linguistic Inquiry* 53(1): 87–120.

Nam, S. 2023. Unsupervised Learning of Sublexica in Korean. Paper presented at the *23rd Biennial Meeting of the International Circle of Korean Linguistics*, Olomouc, Czech Republic.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Paper presented at *2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota.

Papen, R. A. 2003. Michif: One Phonology or Two? *Proceedings of the Eighth Workshop on Structure and Constituency in Languages of the Americas*, eds. Y. Chung, C. Gillon, and R. Wojdak, 47–58. Vancouver: University of British Columbia Working Papers in Linguistics.

Papen, R. A. 2014. Hybrid Languages in Canada Involving French: The Case of Michif and Chiac. *Journal of Language Contact* 7: 154–183.

Park, N. 2020. Hankwuke Umsopayyelceyyakuy Thongkyeycek Haksupkwa Cekhyengseng Phantan. [Stochastic Learning and Well-Formedness Judgement in Korean Phonotactics]. Doctoral dissertation, Seoul National University.

Park, S., S.-H. Hong, and K. Byun. 2013. Hankwukeuy Ehwikyeychungkwa Umunloncek Pokcapseng. [Lexical Strata in Korean and Phonological Complexity]. *Studies in Phonetics, Phonology and Morphology* 19(2): 255–274.

Park, S. 2023. Machine Learning Classification of Native Korean, Sino-Korean and Loanwords Based on Phonotactics. *Studies in Phonetics, Phonology and Morphology* 29(3): 329–349.

Peirce, J., J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. 2019. PsychoPy2: Experiments in Behavior Made Easy. *Behavior Research Methods* 51(1): 195–203.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. (Version 4.4.1) [Computer software]. https://www.R-project.org/

Sears, D. O. 1986. College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology* 51(3): 515–530.

Serrano, S., and N. A. Smith. 2019. Is Attention Interpretable? Paper presented at the *57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Shin, J., J. Kiaer, and J. Cha. 2012. *The Sounds of Korean.* New York: Cambridge University Press.

Singmann, H., B. Bolker, J. Westfall, F. Aust, and M. S. Ben-Shachar. 2024. *afex: Analysis of Factorial Experiments*. (Version 1.4-1) [R package]. https://CRAN.R-project.org/package=afex

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems 30*, eds. I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008.