SIGMORPHON–UniMorph 2023 Shared Task 0, Part 2: Cognitively Plausible Morphophonological Generalization in Korean

Canaan Breiss[1] **Jinyoung Jo**[2]
[1]Massachusetts Institute of Technology [2]University of California, Los Angeles
canaan@mit.edu jinyoungjo@ucla.edu

Abstract

This paper summarises data collection and curation for Part 2 of the 2023 SIGMORPHON-UniMorph Shared Task 0, which focused on modeling speaker knowledge and generalization of a pair of interacting phonological processes in Korean. We briefly describe how modeling the generalization task could be of interest to researchers in both Natural Language Processing and linguistics, and then summarise the traditional description of the phonological processes that are at the center of the modeling challenge. We then describe the criteria we used to select and code cases of process application in two Korean speech corpora, which served as the primary learning data. We also report the technical details of the experiment we carried out that served as the primary test data. [1]

1 Introduction

This paper summarises data collection and curation for Part 2 of the 2023 SIGMORPHON-UniMorph Shared Task 0, which focused on modeling speaker knowledge and generalization of a pair of interacting phonological processes in Korean. We briefly describe how modeling the generalization task could be of interest to researchers in both Natural Language Processing and linguistics, and then summarise the traditional description of the phonological processes that are at the center of the modeling challenge. We then describe the criteria we used to select and code cases of process application in two Korean speech corpora, which served as the primary learning data. We also report the technical details of the experiment we carried out that served as the primary test data.

1.1 Motivation

In this subtask, we sought to build on the success of the human-generalization subtasks (*wug*-tests) in the 2021 and 2022 SIGMORPHON Shared Tasks by creating a dataset that would be of interest to both researchers in NLP and those working in linguistic theory, with the goal of sparking further mutually beneficial collaboration and exchange of ideas between the two fields. The dataset that we gathered documented two phonological processes in Korean that sometimes overlap in their scope of application. Thus, the data bear on questions of linguistic interest about whether human language users generate language in a derivation-based (serial) or output-oriented (parallel) manner. This question of cognitive architecture has clear parallels in computational models of language, where there is a range of statistical, mathematical, and neural methods that embody both the extreme ends, and wide middle, of this architectural range.

Of relevance to both NLP researchers and linguistics is our finding that the disambiguating learning data is also quite sparse: in a child-directed speech corpus of 53,000 words, we found that the environment crucial to learn what happens when rule conditioning contexts overlap appears only 12 times (The Ko Corpus, Ko et al. (2020)). In a corpus of adult speech, the forms occur a total of about 1,000 times in 900,000 phrases (The NIKL Korean Dialogue Corpus; National Institute of Korean Language (2022)). This poses a challenge for models that need large amounts of data to reliably learn linguistic patterns. By pairing the generalization task with the curation of corpus data, we hope to shed light on what kind of generalizations human learners form in the face of such sparse data. These data can be then used to inform the further the development of cognitively plausible linguistic theories, and can also be used to benchmark the development of machine learning models that learn to generalize from sparse data to novel out-of-domain items in a human-like way.

[1]All data discussed here are available at: https://github.com/sigmorphon/2023InflectionST

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 126–131

2 Description of the phonological processes

We bring to bear data from the interaction of two phonological processes in Korean, Post-Obstruent Tensification (POT) and Cluster Simplification (CS). When their conditioning environments overlap, we can observe crucial evidence about how (or whether) the processes are ordered (Kim-Renaud, 1974; Sohn, 1999; Kim, 2003). Note that throughout we use the International Phonetic Alphabet for linguistic data, augmented with the symbol "*" to indicate the tense stop series in Korean; we use the symbol "C" to represent an obstruent consonant, and the symbol "V" to represent a vowel. We follow convention in the linguistic literature by using /slashes/ to represent underlying representations (URs) presumed to be represented in the speaker's mental lexicon, and [brackets] to represent surface representations (SRs) which are taken to be the intended phonetic targets of phonological computation.

2.1 Post-obstruent tensification (POT)

POT causes a lenis consonant to tensify after an obstruent; using SPE-style rewrite rules (Chomsky and Halle, 1968), the process can be expressed as: lax C → tense C / [p, t, k] _. For example, 잡다 /cap-ta/ is realized as [cap-t*a] 'to hold-DECL' ; 받고 /pat-ko/ → [pat-k*o] 'to receive-and' . POT is described as nearly categorical within the accentual phrase (Jun, 1998), a finding which we also observe in the data we report here, and applies in nearly all possible morphological and phrasal environments.

2.2 Cluster simplification (CS)

CS targets underlying consonant clusters in coda position, yielding simplification when followed by a C-initial suffix or a word boundary. The process can be expressed using SPE rules as: CC → C / _{#, C}. For example, in 앉는 /anc-nin/, the final /c/ is deleted in the surface form [an-nin] 'to sit-COMP' ; a similar outcome is seen in 굶나 /kulm-na/ → [kum-na] 'to starve-INTERROG' . The process also applies at word boundaries, such that underlying 닭 /talk/ surfaces as [tak] 'chicken'. CS is variable depending on verb identity and final consonant place (Kwon et al., 2023), and the conditioning context exists in verbs and nouns.

2.3 Overlapping contexts

When verbs that end in an /-lC/ consonant cluster are suffixed with a lax obstruent-initial affix (denoted /C-/), the conditioning contexts for both processes are met. In verbs, the majority outcome is that the /lC/ cluster is simplified to singleton [l], and the following stop is tensed. For example, 맑고 /malk-ko/ is realized as [mal-k*o] 'to be clear-and', with a tense [k*] in spite of the triggering context having been deleted; a similar example is 낡고 /nalk-ko/ → [nal-k*o] 'to be old-and'. These types of form suggest that the two processes apply "in sequence", with POT ordered before CS, as shown in table 1.

UR	/pat-ko/ to receive-and	/anc-nin/ to sit.COMP	/malk-ko/ to be clear-and
POT	pat-k*o	—	malk-k*o
CS	—	an-nin	mal-k*o
SR	[pat-k*o]	[an-nin]	[mal-k*o]

Table 1: Example of apparent ordering between POT and CS in Korean /-lC/-final verbs.

This type of process interaction is known in the phonological literature as *counter-bleeding opacity* (Kiparsky, 1968): CS would destroy the conditioning environment for POT (removing the obstruent in the cluster), but applies too late to do so, resulting in an apparent "overapplication" of CS – it seems to have applied outside its conditioning environment. Note that in general, post-liquid tensification is absent from the language (e.g. 줄다 /cul-ta/ →[cul-ta], not *[cul-t*a] 'to decrease-DECL'), so the observed outcome 맑고 /malk-ko/ → [mal-k*o] cannot be attributed to other phonological processes at work.

Although the opaque outcome, as in 맑고 /malk-ko/ → [mal-k*o], is the canonical and majority type, the literature contains reports of variability in how CS and POT apply when overlapping. For example, (Kim, 2003) reports that the target of CS may vary between coda /l/ and coda /C/; for example 밟고 /palp-ko/ → [pal-k*o]~[pap-k*o] 'to step on-and'; 낡지 /nalk-ci/ → [nal-c*i]~[nak-c*i] 'to be old-CONN'. Further, while in /-lC/-final verbs the opaque outcome obtains when an obstruent-initial suffix is attached, in nouns of the same shape the outcome is not opaque; CS always targets the /l/ rather than the /C/, yielding outcomes like 닭도 /talk-to/ → [tak-t*o] 'chicken-also' and 흙과 /hɨlk-kwa/ → [hɨk-k*wa] 'soil-and' (Tak, 2008). Thus, we suspect that further examination of more nat-

uralistic data in corpora and in the generalization task may surface a more complex pattern of variation.

3 Task description

The task was to predict human responses to a generalization task (a *wug*-test, cf. Berko (1958)), involving existing high-frequency verb stems, existing low-frequency verb stems, and novel verb stems. The stems were paired with affixes that created environments that were designed to condition POT alone (as in 막다 /mak-ta/ 'to block-DECL'), CS alone (as in 밟는 /palp-nin/ 'to step on-COMP'), the critical overlapping context (as in 밟고 /palp-ko/ 'to step on-and'), or designed to trigger neither process, so that the underlying consonant cluster is resyllabified across the syllable boundary and survives deletion (as in 넓어 /nʌlp-ʌ/ 'to be wide-DECL').

Training data came in two types: a list of the 53 /-lC/-final verbs in the frequency list of Korean from Kang and Kim (2004), and counts and hand-coding of the outcome of environments that could condition POT and/or CS in verbs from an adult-directed speech corpus and an infant-directed speech corpus.

The list and corpus counts were designed to be used as the primary training data, and results of the generalization task were divided up into *train*, *dev*, and *test* splits. The *train* and *dev* splits were intended to be used during model development, and model performance calculated on the *test* set.

4 Corpus data collection

To approximate the data that a learner of Korean might be exposed to while acquiring their phonology, we culled relevant data from two corpora of spoken Korean.

4.1 Adult-directed speech corpus

For adult-directed speech, we used the NIKL dialogue corpus (National Institute of Korean Language, 2022), which consists of approx. 900,000 phrases of semi-spontaneous speech, together with orthographic and phonemic transcription. We extracted each suffixed /-lC/ verb from the corpus (7,570 tokens, 1,395 types), and manually annotated them for pronunciation. We excluded words with /-lh/-final stems because they participate in additional processes, such as coalescence with the following stop that yields aspiration instead of tensification 잃다 /ilh-ta/ → [il-t^{h}a], not *[il-t*a] 'to lose-DECL') (Kim-Renaud, 1974; Sohn, 1999). We did not extract POT-only contexts (simple /-C/-final verbs with following /C-/-initial suffixes) because they were extremely frequent, and impressionistic judgements of the second author align with the literature (Jun, 1998) that POT applies nearly obligatorily within phrases. In the smaller infant corpus and the results of the generalization task, such environments were extracted and coded.

4.2 Infant-directed speech corpus

For infant-directed speech, we used the Ko corpus (Ko et al., 2020), collected from interactions of mother-child pairs in a free-play session. The corpus consists of approx. 53,000 words of spontaneous infant-directed speech, paired with orthographic and phonemic transcription. We extracted and hand-checked all affixed /-lC/ verbs in the corpus (289 tokens, 38 types), as well as all simple /-C/ verbs with a following /C-/-initial affix (1,083 tokens, 171 types). Exclusions were the same as for the adult-directed speech corpus.

5 Experimental data collection

To probe how adult speakers represent CS, POT, and their interaction in existing words, and how they generalize this knowledge to entirely novel contexts, we carried out a production task where speakers were asked to produce inflected forms of verbs, and record their productions.

5.1 Stimuli

Stimuli had two stem types (/-lC/ and /-C/), and three frequency levels (high-frequency, low-frequency, and nonce). Frequency levels were calculated using information from Kang and Kim (2004). We selected 10 stimuli in each of the six resulting categories, and paired each with three affix types (/-a, ʌ/ -아, 어 'DECLARATIVE, INTERROGATIVE, IMPERATIVE',[2] /-na/ -나 'INTERROGATIVE', and /-ta/ -다 'DECLATIVE'). This yielded 180 stimuli, selected to elicit the four types of contexts exemplified in section 3: contexts where POT and CS could apply non-overlappingly, contexts where we

[2]The distribution of these allomorphs is governed by vowel harmony which is unrelated to the consonantal phenomena under investigation here; see Ahn (1985) for a traditional description, and Jo (forthcoming) for a recent overview of the empirical landscape.

could observe the form of the stem with no phonological effects at all, and contexts where POT and CS overlap.

5.2 Participants

Our goal was to recruit 30 speakers of Korean who were born in Korea and grew up with Korean as their dominant language. We used a combination of recruitment on Prolific, word-of-mouth, and posting on online forums to recruit participants, and ended up with 23 by the deadline for data release. We released data from 12 speakers to teams at that time as *train* data and 4 for as *dev*, held back data from 7 speakers as *test* and released their demographic info and trial information without the right answers, and continued collecting data. By the time the due date for releasing test data came, we had collected data from 6 more speakers, and so the correct answers were released for the original 7 *test* subjects, plus 6 "surprise" speakers. 1 more participant's data was collected after test data were released, to reach the total target of 30. Participants recruited through Prolific were paid for their time.

5.3 Design and procedure

The design leveraged the fact that, although Korean has a number of phonological processes that cross both morpheme-boundaries within words and word-boundaries within prosodic phrases (Sohn, 1999), the standard practice in writing Korean using the Hangul orthography is to write each morpheme as though no phonological processes had applied to it (approximating phonological URs). In spite of this norm, however, the orthography is still capable of expressing and uniquely identifying the full range of phonetic realizations that these alternations give rise to (approximating the SRs). For example, the underlying form of 'to block-DECL' is /mak-ta/, and is written in Hangul as 막다 and when POT applies, it is produced as [mak-t*a]; this can be represented in the spelling as 막따, though the normal written form is 막다. These facts about Korean orthographic norms allowed us to rely on the "self-transcription" method of Moore-Cantwell (2020), where participants spoke their response out loud in response to the standard written form of the stimulus (indicating the UR), and then were asked to choose an orthographic form that most closely matched the form they produced where the different possible surface realizations (SRs) were disambiguated.

The experiment was carried out over the internet using the Labvanced experimental platform (Finger et al., 2017). Participants were instructed to find a quiet room to complete the experiment, and that it would take approximately an hour. They were told that they would be asked to read a series of inflected words out loud while being recorded, and then select one of several multiple-choice options that matched what they had said the most closely. After, they would be asked to indicate whether they knew the word or not.

The experiment began with four practice trials, after which each participant completed the 180 inflection trials in a random order. On each inflection trial, the target word would be shown to participants with a V-initial suffix not included in the experimental design (/-ajo, ʌjo/ -아요, -어요 'DECLARATIVE, INTERROGATIVE, IMPERATIVE-polite'), and they would be asked to say the word out loud with one of the three affixes (/-a, ʌ/, /-na/, /-ta/), depending on the trial. Then, after producing the form and the recording was complete, they were asked to choose which of a number of multiple-choice options they had said. The number of multiple-choice items differed from trial to trial based on the type of stem (/-lC/- or /-C/-final) and suffix (vowel-, sonorant-, or obstruent-initial). The options always included transcriptions where each expected phonological process (POT and/or CS, depending on the stem and affix) either applied or not independently, and also overlapped; in cases with sonorant-initial affixes, candidates also included outcomes for possible application of lateralization and nasalization (Sohn, 1999) – the latter two are not the focus of study here, but were included for the sake of completeness. On each trial, the prompt was shown while participants were being recorded, then when they stopped the recording, the display was changed to show only a button to allow a replay of their own production, and the range of possible outcomes. There was always an "other" case listed, where participants were allowed to write their pronunciation if none of the options provided matched their pronunciation. In practice this was extremely rare; see section 5.4 for details.

After the production task, the second phase of the experiment was a vocabulary test. On each screen, participants saw a stem with the same non-target vowel-initial affix as in the prompt on the production task, and then indicated using a 5-point

Likert scale how familiar they were with the word, ranging from 1 ("I don't know this word at all") to 5 ("I am extremely familiar with the word").

Finally, participants were asked to provide some background information about themselves, including whether they had begun speaking Korean in some context before the age of seven, and what other languages they spoke. No recruited participants were excluded on grounds of having a language background that did not meet our criteria for inclusion described in section 5.2.

5.4 Data coding

Spot-checks were carried out to make sure that the forms produced by the speakers were consistent with the forms that they indicated that they produced; in general, subjects were extremely accurate in reporting what they said; "other" responses were excluded, which comprised only an extremely small percentage of the data.

After data checking, existing stems that were rated as 1 (="I don't know this word at all") by a given subject were re-classified as novel for that subject. This was done to allow for accurate estimation of the knowledge of each subject, and to avoid making the assumption that all participants know all words in the study.

6 Discussion

As stated in section 1, the goal of this subtask was to spur collaboration and cross-talk between two communities; thus, we set aside here a discussion of the contents of the dataset, referring the reader to the paper by Jeong et al. (2023) to summarise the findings in of the one team that worked on this subtask. The model and discussion found in their paper notwithstanding, we hope the data we gathered in this subtask may also have broader utility in testing linguistic theories of learning and representation, and in benchmarking models that attempt to reach human-like levels generalization while maintaining human-like requirements in terms of data efficiency. It is our hope that it may continue to be of use outside the context of this subtask going forward.

Acknowledgements

Thanks to Ryan Cotterell, Omer Goldman, Roger Levy, Garrett Nicolai, Ekaterina Vylomova, and the audience at the 97th LSA Annual Meeting for helpful comments, feedback, and guidance. We are grateful to the MIT Computational Psycholinguistics Lab for funding the data collection; CB also acknowledges the MIT-IBM Watson AI Lab for individual funding.

References

Sang-Cheol Ahn. 1985. *The interplay of phonology and morphology in Korean*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Noam Chomsky and Morris Halle. 1968. The sound pattern of English.

Holger Finger, Caspar Goeke, Dorena Diekamp, Kai Standvoß, and Peter König. 2017. Labvanced: a unified javascript framework for online studies. In *International Conference on Computational Social Science* (*Cologne*).

Chongnam Jeong, Dominic Schmitz, Akhilesh Kakolu Ramarao, Anna Stein, and Kevin Tang. 2023. Linear discriminative learning: a competitive non-neural baseline for morphological inflection. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Jinyoung Jo. forthcoming. Korean vowel harmony has weak phonotactic support and has limited productivity. *Phonology*.

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

Beom-Mo Kang and Hung-Gyu Kim. 2004. *Hankwuke hyengtayso mich ehwi sayong pintouy pwunsek2* [*Frequency analysis of Korean morpheme and word usage2*]. Institute of Korean Culture, Korea University, Seoul.

Seoncheol Kim. 2003. *Phyocwun Palum Silthay Cosa II* [*A Survey of Standard Pronunciation II*]. National Institute of Korean Language, Seoul.

Young-Key Kim-Renaud. 1974. *Korean Consonantal Phonology*. Ph.D. thesis, University of Hawaii.

Paul Kiparsky. 1968. *Linguistic universals and linguistic change*.

Eon-Suk Ko, Jinyoung Jo, Kyung-Woon On, and Byoung-Tak Zhang. 2020. Introducing the ko corpus of korean mother–child interaction. *Frontiers in Psychology*, 11:3698.

Soohyun Kwon, Taejin Yoon, Sujin Oh, and Jeon-Im Han. 2023. Variable realization of consonant clusters in seoul and gyeongsang korean. Poster at HISPHONCOG 2023.

Claire Moore-Cantwell. 2020. Weight and final vowels in the English stress system. *Phonology*, 37(4):657–695.

National Institute of Korean Language. 2022. NIKL Korean Dialogue Corpus (audio) 2020(v.1.3).

Ho-Min Sohn. 1999. *The Korean Language*. Cambridge University Press, Cambridge, UK.

Jin-young Tak. 2008. A uniform analysis of tensification in korean: An optimality approach. *Korean Journal of Linguistics*, 33:545–564.