A nontraditional account of the “irregular” verb paradigms in Seoul Korean

Renato Pamir Martini (13574795)

BA thesis Linguistics – University of Amsterdam

Supervisor: Cesko Voeten

14 July 2025 – Amsterdam, Netherlands

Abstract

This thesis investigates whether the Seoul Korean verb paradigms that do not fit the traditional phonological description of the language can be accounted for by an extension of such a description, with a focus on consonants. Four principles to find underlying forms starting from surface ones are first defined: *contrast*, i.e. the ability for different categories to form minimal pairs; *alternation*, i.e. the ability for the same underlying category to correspond to different surface forms in different environments; *neutralization*, i.e. the ability for different underlying categories to merge into the same surface category in the same environment; *behavioral difference*, i.e. the ability for different underlying categories in the same environment to affect their surroundings differently, which is then visible in the surface form. Subsequently, all the attested verbal conjugation patterns are discussed in order to find whether a suitable underlying form can be defined to explain the attested alternations. What emerges is that, yes, unique underlying forms can be assigned to each one of the patterns, and patterns traditionally deemed "irregular" also turn out to be far more prevalent in the lexicon relative to similar "regular" patterns, which seems to contradict their apparent idiosyncrasy. In conclusion, the uses of and meanings behind labels such as "(ir)regular" need be better justified, possibly with empirical data.

R. Pamir Martini

Table of contents

1. Introduction

Seoul Korean[1] is the Koreanic variety of the Korean capital Seoul (서울) and its surroundings, and it is the prestige variety at the basis of the modern spoken standard language. The traditional literature on this language sets aside some of its verbal declension patterns by labeling them as "irregular", i.e. synchronically unaccountable and phonologically unpredictable (Albright & Kang 2008; Lee, 1989; Lee & Ramsey, 2000; Martin, 1992; Sohn, 1999; i.a.). This label, however, overlooks the fact that the alternations found in the paradigms of such verbs are not uncommon across the lexicon and they seem to be conditioned by the same contextual factors that also trigger the alternations among the "regular" patterns.

Like all Koreanic varieties, Seoul Korean (henceforth just Korean) has strong constraints as to what segments can occur in coda positions. For example, of the three obstruent sets it features, i.e. plain, aspirated and glottalized[2], only the plain stops are found in codas. This leads to alternations in the surface forms of stems when their edges are found in different phonological environments. Before moving to verbs, this sort of contextual alternation is exemplified in (1) using nominal stems where the stem-final consonant surfaces as plain /k/[3] before consonant-initial /t'o/ ALSO, but before vowel-initial /e/ LOCATIVE as /k k^h k'/ depending on the lexical item.

(1) Alternation between prevocalic /k k^h k'/ and preconsonantal /k/ at the end of nominal stems. The stem-final consonant is in bold.

a. /pa**k**t'o/ /pa**k**e/ 'gourd'[4]
b. /puʌ**k**t'o/ /puʌ**k^h**e/ 'kitchen'
c. /pa**k**t'o/ /pa**k'**e/ 'outside'

In the literature, such alternations are analyzed as being the surface realizations of different underlying categories (Lee, 1989; Lee & Ramsey, 2000; Martin, 1992; Sohn, 1999); in the case of (1) above these are (a) |k|, (b) |k^h| and (c) |k'| respectively. These kinds of alternations, as mentioned

1 Glottocode: seou1239 (Hammarström, Forkel, Haspelmath & Bank, 2025).
2 In the literature, the three sets are variously named and notated; the following notation is used in this thesis: a bare "C" is used for an unspecified consonant, a dotted "C˙" for plain, "C^h" for aspirated and "C'" for glottalized obstruents. This choice has no particular reason and it reflects personal preferences, and it does not want to be a claim to the phonetic properties associated to each set or category, which is still a topic of discussion in the field.
3 The following notation, taken from the BiPhon notation (Boersma, 2009), is applied throughout this thesis: pipes "|...|" denote underlying forms (i.e. the phonological information stored in the lexicon), slashes "/.../" surface forms (i.e. the phonological information most proximate to the phonetics), and square brackets "[...]" phonetic forms. Furthermore, a plus "+" denotes morpheme boundaries, a hash "#" word boundaries.
4 For completeness and reference purposes, all the lexical items used as examples are also reported with their spellings in Appendix A.

already, are not restricted to nominal stems: verbal[5] stems exhibit analogous alternations as demonstrated in (2), where the stem-final consonants (in bold) all share the same surface form /t/ before the consonant-initial ending /k'o/ 'and', while they are distinguished before the vowel-initial ending /a/ or /ʌ/ (henceforth referred to as the A form of the verb).[6]

(2) Alternation between prevocalic /t tʰ s s' c/ and preconsonantal /t/ at the end of verbal stems. The stem-final consonant is in bold.

a.	/pa**t**k'o/	/pa**t**a/	'receive'
b.	/ka**t**k'o/	/ka**t**ʰa/	'same'
c.	/pʌ**t**k'o/	/pʌ**s**ʌ/	'take off'
d.	/i**t**k'o/	/i**s'**ʌ/	'existent'
e.	/i**t**k'o/	/i**c**ʌ/	'forget'

The verbal stem alternations in (2) are accounted for with different underlying categories, namely |t tʰ s s' c| respectively, in a similar fashion to the nominal examples in (1). However, the picture gets more complicated when (3) is taken into account: preconsonantal /t/ alternates with both prevocalic /t/ and prevocalic /l/, while prevocalic /l/ alternates with both preconsonantal /l/ and preconsonantal /t/. In (3), the one-to-many alternation pattern between preconsonantal and prevocalic environments of the verb-final stem consonants of (a) and (c) is, unlike in (1–2), accompanied by the many-to-one pattern of (b) and (c).

(3) Alternation patterns of preconsonantal and prevocalic /t l/. /k'o/ and /ko/ are equivalent and their difference is not relevant.

a.	/kʌ**t**k'o/	/kʌ**t**ʌ/	'roll up'
b.	/kʌ**l**ko/	/kʌ**l**ʌ/	'hang'
c.	/kʌ**t**k'o/	/kʌ**l**ʌ/	'walk'

The traditional literature analyzes (3a) and (3b) as having underlying stems ending in |t| and |l| respectively, which matches the analysis of the data in (1–2). As for the alternation seen in (3c), it is either left unanalyzed and mentioned as an oddity in the paradigm of certain verb stems (e.g. Martin, 1992) or, more commonly, called irregular and idiosyncratic in a certain way and left equally unanalyzed (e.g. Lee, 1989).

5 The difference between verbs (action-denoting) and adjectives (state-denoting and quality-denoting) is disregarded as it is not relevant to the present discussion: the term *verb(al)* is applied to all of them.

6 The distribution of /a/ and /ʌ/ is based on vowel harmony, which is beyond the scope of this work. See Kang (2012) for more details.

The present thesis investigates whether the Seoul Korean paradigms that do not fit the traditional phonological description of the language can be accounted for by an extension of such a description. The phonological categories assumed to exist by the traditional description have so far been determined *a priori* and were made to fit the grammar, rather than doing the reverse, that is looking at the available language data and then deriving the categories *a posteriori*. The focus of this thesis is on consonant-final verb stems because, throughout their paradigm, these verbs include alternations that have not yet been extensively researched and that clash with the traditional phonological description of Korean.

While previous analyses that delve into these "irregular" Korean verb alternations do exist (cf. Kim-Renaud, 1973; Lee, 2019), they focus on the sets of rules and their ordering that can describe each recorded alternation in the language, with the underlying forms being whatever best fits the posited rules. This thesis takes a different point of view and analyzes the data in a systematic way that can be applied to the entirety of Korean and further to other languages. This is done by first setting principles to follow, taken from existing literature and common practice, in order to define the underlying categories of a given language, which are then applied to the language data. Then, a list of systematically-derived underlying categories is formed based on the data, which is more extensive than that of the traditional account. A secondary discussion that arises is whether the label "irregular" can still have some use, which is also addressed. This discussion is lead by a token count of the frequency of each paradigm among Korean verbs.

It should be noted that Korean verbs are more of a case study rather than the only instance of seemingly incompatible alternation patterns in a language the synchronic status of which should be discussed. In fact, similar issues also exist in other languages, e.g. Turkish vowel length (Sezer, 1986) and Modern Hebrew "gutturals" (Enguehard & Faust, 2018), as well as other cases mentioned and presented further below.

The rest of the thesis is structured as follows: first the structure of the Korean surface-form data and its traditional account are presented in §2.1–.2, followed in §2.3 by the discussion of the principles to be applied. In §3 the token count of the size of each alternation paradigm is presented. Further, §4.1 explores the many alternations attested among Korean verbs and it applies the principles of §2.3 to output a list of detectable underlying categories and clusters, then §4.2 makes further remarks on the discussion and, finally, §5 presents a summary and a conclusion to the thesis.

2. Background

In this section, the surface structure of Seoul Korean is first described in §2.1, with the goal of providing the background for the following discussion. Next, §2.2 describes the traditional approach to the analysis of such surface structures. Finally, §2.3 defines the principles that will, further below, lead the discussion on the underlying forms of Korean verb stems.

2.1. The surface structure of Seoul Korean

Before describing the underlying forms that Seoul Korean is traditionally assumed to have, i.e. the phonological information stored in the lexicon of Seoul Korean speakers and listeners, which is the topic of discussion in this work, it is useful to first know what surface forms, i.e. the phonological information most proximate to the phonetics,[7] appear in Korean.

Nineteen surface consonants (C) /p pʰ p’ t tʰ t’ c cʰ c’ k kʰ k’ s s’ m n ŋ l h/ are widely reported in the literature as well as at least seven monophthongs (V), /a ʌ e o u ɯ i/, and about nine diphthongs /ja wa jʌ wʌ je ɥe jo ju ɥi/ (Kim-Renaud, 2022; Lee, 1989; Lee & Ramsey, 2000; Martin, 1992; Sohn, 1999).

There is no consensus on exactly how many vowels and diphthongs there are in the language, due to recent or still ongoing sound changes.[8] Table 1 lists the aforementioned surface vowel and diphthong inventory, of current young Seoulites, which is used here. While further analyses can be carried out on the vowel inventory, especially regarding the nature and consequences of the changes, this is beyond the scope of this thesis and shall not be further discussed.

On the other hand, there is more consensus on the amount of consonants. Among them, three sets of obstruents are distinguished: plain /p t c s k/[9] (C̥), aspirated /pʰ tʰ cʰ kʰ/ (Cʰ), and glottalized /p’ t’ c’ s’ k’/ (C’). In addition, a full nasal set /m n ŋ/ (N), one liquid /l/ (L) and the

7 Depending on the author and the framework, the surface form may be practically indistinguishable from a phonetic form. The two are here taken as distinct, but the present discussion would not be significantly influenced by merging them into a single surface/phonetic form.

8 To name a few: /e/–/ɛ/ have merged (Kang, Schertz & Han, 2015; Lee & Jongman, 2016), vowel length is no longer contrastive (Kang, Yoon & Han, 2015; Sohn, 1999: 157), [y ø] have broken into diphthongs [ɥi ɥe] (Martin, 1992: 24; Sohn, 1999: 156), the diphthong [ɯi~ɯj] has been reintroduced, possibly as a spelling pronunciation (Martin, 1992: 43), and [ɥiʌ] is being shortened, possibly to [ɥʌ] (Choe, 2021; Martin, 1992: 38).

9 The fricative /s/ has properties matching both the plain set (e.g. post-obstruent glottalization) and the aspirated set (e.g. high pitch on the subsequent vowel). However, it is different from the glottalized /s’/ (Cho, Jun & Ladefoged, 2002; Lee & Jongman, 2016).

aspirate /h/ are also found. Phonetically, the contrast between plain /p t c k/ and aspirated /pʰ tʰ cʰ kʰ/ is shifting from aspiration to pitch word-initially, which has gained some attention in the literature (cf. Cho, Jun & Ladefoged, 2002; Kang & Han, 2013; Silva, 2006; Watkins, 2022). Further, obstruents have no audible release /k t p/ [kˀ tˀ pˀ] in codas, and plain /p t c k/ are often voiced [b d d͡z~d͡ʑ g] word-internally. Nasals /m n/ are often denasalized word-initially (Kim, 2011; Yoo & Nolan, 2020), and /l/ is a tap [ɾ] intervocalically. The three /c cʰ cˀ/ are phonetically affricates [t͡s~d͡z t͡sʰ t͡sˀ] or [t͡ɕ~d͡ʑ t͡ɕʰ t͡ɕˀ], but there is no consensus on the exact quality.[10]

Table 1. The surface vowel and diphthong inventory of the current young Seoulites. The usual spelling is also given in angle brackets "⟨…⟩" for reference.

	monophthongs			diphthongs		
	front unrounded	**central/back unrounded**	**central/back rounded**	**front unrounded**	**central/back unrounded**	**central/back rounded**
high	/i/ ⟨ㅣ⟩	/ɯ/ ⟨ㅡ⟩	/u/ ⟨ㅜ⟩	/ɥi/ ⟨ㅟ⟩	/ɯi/ ⟨ㅢ⟩	/ju/ ⟨ㅠ⟩
mid	/e/ ⟨ㅐ ㅔ⟩	/ʌ/ ⟨ㅓ⟩	/o/ ⟨ㅗ⟩	/je, ɥe/ ⟨ㅒ ㅖ, ㅙ ㅚ ㅞ⟩	/jʌ, wʌ/, ([ɥʌ]) ⟨ㅕ, ㅝ⟩, ⟨ㅝㅓ⟩	/jo/ ⟨ㅛ⟩
low		/a/ ⟨ㅏ⟩			/ja, wa/ ⟨ㅑ, ㅘ⟩	

As for their distribution, Korean codas are notoriously strict: only the three plain obstruents /p t k/, the three nasals /m n ŋ/ and the liquid /l/ are found in codas. For most of the literature on Korean, including Albright & Kang (2008), Lee (1989), Lee & Ramsey (2000), Martin (1992), Sohn (1999), i.a., as well as prescriptively, surface clusters are not allowed or found as codas. Despite this, coda clusters with an initial liquid (/lC./) have been attested and described by Kim-Renaud (1973, 1974, 1977, 2022) and Cho & Kim (2009) in free variation with the shorter versions reported by other authors, but they have not yet received much attention. For the sake of simplicity, the more extensive description with /lC./ clusters is followed here, which will be further discussed in §4.1.4. Not all /CC/ combinations are attested: nasals can be followed by any consonant except /ŋ l/ and they can only be preceded by another nasal or /l/, while plain consonants never follow obstruents. Table 2a lists all the surface consonants attested in Korean, and the ones that can occur in coda position are in bold. Table 2b further summarizes the attested consonant cluster types.

10 The specific phonetic properties of the affricates are a matter of debate. Traditional impressionistic descriptions mention geographic variation, but other factors such as gender and age have also been found to play a role (Kim, 1999, 2001; Kong, Kang & Seo, 2014). Since this thesis discusses abstract categories, with arbitrary labels, the symbols /c cˀ cʰ/ will be used to refer to them regardless of their phonetic details.

Table 2a. The surface consonant categories of Korean. The consonants that occur in codas are in bold. The usual spelling is also given in angle brackets "⟨…⟩" for reference.

	labial	**c o r o n a l**			**dorsal**	**laryngeal**
glottalized **C'**	/p'/ ⟨ㅃ⟩	/t'/ ⟨ㄸ⟩	/c'/ ⟨ㅉ⟩	/s'/ ⟨ㅆ⟩	/k'/ ⟨ㄲ⟩	
aspirated **C^{h}**	/p^{h}/ ⟨ㅍ⟩	/t^{h}/ ⟨ㅌ⟩	/c^{h}/ ⟨ㅊ⟩	/s/ ⟨ㅅ⟩	/k^{h}/ ⟨ㅋ⟩	/h/ ⟨ㅎ⟩
plain **C˙**	**/p/** ⟨ㅂ⟩	**/t/** ⟨ㄷ⟩	/c/ ⟨ㅈ⟩		**/k/** ⟨ㄱ⟩	
nasal **N**	**/m/** ⟨ㅁ⟩		**/n/** ⟨ㄴ⟩		**/ŋ/** ⟨ㅇ⟩	
liquid **L**			**/l/** ⟨ㄹ⟩			

Table 2b. The attested surface consonant cluster types of Korean. A third consonant, /l/, can also be found at the very beginning of the cluster.

(l)C$_1$ **C$_2$**	**/C'/**	**/C^{h}/**	**/s/**	**/C˙/**	**/N/**	**/h/**	**/l/**
/(l)C˙/	✓	✓					
/(l)N/	✓	✓	✓	✓	✓	✓	
/l/	✓	✓	✓	✓	✓	✓	✓

Furthermore, /l/ has additional constraints to its occurrence, some stronger than others, which are relevant to the discussion. As shown by Lee & Ramsey (2011), /l/ has historically been disfavored before coronal consonants other than itself, as well as in certain other environments, namely word-initially and after other consonants. This has created certain morphophonological environments where the presence of a single surface liquid is avoided. These environments will be referred to as *hygrophobic*, while environments that tolerate the presence of the liquid *hygrotolerant*.[11] For example, the coronal nasal /n/ constitutes a hygrophobic environment, and in fact the cluster */ln/ is ill-formed despite /lN/ clusters being possible.

The possible synchronic phonological explanations, if they exist, for the two different environments are beyond the scope of the present discussion. The distinction is nonetheless relevant and it becomes useful when discussing liquid patterns (§4.1.3) and traditionally "irregular" patterns (§4.1.6).

11 These two terms literally mean "adverse to moisture or liquids" and "tolerant of moisture or liquids", from the prefix *hygro-* (from Ancient Greek ὑγρός *hugrós* 'wet, fluid, liquid') used in the Western scientific lexicon, and employed in reference to liquid consonants, for which the term υγρό *iγró* is also used in Modern Greek.

2.2. The traditional account of the underlying structure of Seoul Korean

The traditional account posits as many underlying categories as the surface ones, as glimpsed by the lists of phonemes provided by Lee (1989: 11), Lee & Ramsey (2000: 61), Martin (1992: 24), Sohn (1999: 152), i.a., which are reported in Table 3.

When, due to morphosyntactic reasons, disallowed surface clusters would be expected, surface alternations arise in the form of neutralizations (e.g. §4.1.2), deletions (e.g. §4.1.3) and coalescences (e.g. §4.1.5) of the nineteen underlying categories. Each category therefore has a number of conditional surface allophones that predictably occur in certain environments, also partially listed in Table 3. Anything that is not readily explainable by these nineteen categories and their interactions has been set aside as an oddity of the lexical items contradicting established rules.

Table 3. The underlying consonant categories of Korean according to the traditional description. The most basic preconsonantal /_C/, with the obstruent type they occur with, and prevocalic /_V/ surface allophones of each underlying category is also given.

	labial	**c o r o n a l**			**dorsal**	**laryngeal**
glottalized C'	|p'| /p'V/*	|t'| /t'V/*	|c'| /c'V/*	|s'| /tC' s'V/	|k'| /kC' k'V/	
aspirated C^h	|p^h| /pC' p^{h}V/	|t^h| /tC' t^{h}V/	|c^h| /tC' c^{h}V/	|s| /tC' sV/	|k^h| /kC' k^{h}V/	|h| /C^h (h)V/†
plain C˙	|p| /pC' pV/	|t| /tC' tV/	|c| /tC' cV/		|k| /kC' kV/	
nasal N	|m| /mC˙ mV/‡		|n| /nC˙ nV/‡		|ŋ| /ŋC˙ ŋV/	
liquid L			|l| /lC˙ lV/			

* These are never found morpheme-finally where they can variably end up before another consonant.

† At the end of verb stems, it only surfaces as aspiration on the following obstruent, elsewhere it is /hV/. This is further discussed in §4.1.5, however compare /n**aha**ko/ |n**a**+**ha**ko| 'with me' but /n**aa**/ |n**ah**+**A**| 'give birth'.

‡ At the very end of verb stems, the first obstruent of the ending is glottalized /mC' nC'/ rather than plain /mC˙ nC˙/. This is further discussed in §4.1.1, however compare /a**nc**a/ |a**nc**+A| 'sit down' but /a**nc**'a/ |a**n**+**c**a| 'let's hug'.

This can very well be exemplified by trying to assign underlying forms to the verbs in (2–3), as shown in (4). All the verb stems (a–f) end with /t/ in the 'and' form, which is not telling of what the A form may be, while the A form provides enough information to derive the 'and' form. As mentioned above, coda environments are highly constrained in Korean, so the coda neutralization into /t/ is posited for all coronal obstruents and then, based on the A form, an underlying

structure can be posited. As for (g), /l/ can occur both preconsonantally and prevocalically, so it does not alternate and it does not run into issues when reconstructing its underlying structure. Additionally, after another obstruent, and not after |l|, underlying plain obstruents correspond to surface glottalized obstruents, which also explains why 'and' is /k'o/ after /t/, but /ko/ after /l/; its underlying form must thus be |+ko|.

When it comes to (h), neither form is indicative of what the other one may be, and none of the available underlying categories in Table 3 is able to explain the alternation between /t/ and /l/. The verb 'walk' is thus labeled "irregular" and the alternation is justified as an allomorphy, a morphological phenomenon more than a phonological one, which nonetheless has a purely phonological conditioning. Needless to say, it is a bit awkward to give up on a phonological explanation for a number of alternations with phonological triggers.

(4) "Irregular" verbs require multiple underlying stems according to the traditional account of Korean.

a.	/pa**t**k'o/	/pa**t**a/	'receive'	|pa**t**+|
b.	/ka**t**k'o/	/ka**t^{h}**a/	'same'	|ka**t^{h}**+|
c.	/pʌ**t**k'o/	/pʌ**s**ʌ/	'take off'	|pʌ**s**+|
d.	/i**t**k'o/	/i**s'**ʌ/	'existent'	|i**s'**+|
e.	/i**t**k'o/	/i**c**ʌ/	'forget'	|i**c**+|
f.	/kʌ**t**k'o/	/kʌ**t**ʌ/	'roll up'	|kʌ**t**+|
g.	/kʌ**l**ko/	/kʌ**l**ʌ/	'hang'	|kʌ**l**+|
h.	/kʌ**t**k'o/	/kʌ**l**ʌ/	'walk'	|kʌ**t**+| and |kʌ**l**+|: "irregular"

This thesis proposes an alternative to the description just provided. In order to do so, a series of principles to build upon must be established in order to derive a more comprehensive model of the underlying structure of Korean based on the available surface data. These principles are discussed in the next section.

2.3. Defining the underlying forms

This section presents some principles that phonologists have been following to define the categories used in language descriptions and models, namely the following: (§2.3.1) *contrast*, i.e. the ability for different categories to form minimal pairs, (§2.3.2) *alternation*, i.e. the ability for the same underlying category to correspond to different surface forms in different environments, (§2.3.3) *neutralization*, i.e. the ability for different underlying categories to merge into the same surface category in the same environment, and (§2.3.4) *behavioral difference*, i.e. the ability for

different underlying categories in the same environment to affect their surroundings differently, which is then visible in the surface form. Crucially, these principles allow to reconstruct an underlying structure, segment by segment, given sufficient surface-form data.

2.3.1. First principle: contrast

The most basic principle is that of contrast: different surface categories that can form minimal pairs can be ascribed to different underlying categories. Considering the English examples in (5) with a /_ɪt/ structure, all of the surface categories /p b l n s f k h w/ are able to form minimal pairs, and even minimal sets in this case. Therefore, each one of these SF categories can be ascribed to a separate underlying category: |p b l n s f k h w|. Similarly, the Korean verbs in (6) form a minimal set with a /ca_a/ structure (as further discussed in §4.1.1).

(5) A minimal set of English words with a /_ɪt/ structure, which can be ascribed to unique underlying categories.

a.	/**p**ɪt/	|**p**ɪt|	'pit'	d.	/**n**ɪt/	|**n**ɪt|	'knit'	g.	/**k**ɪt/	|**k**ɪt|	'kit'
b.	/**b**ɪt/	|**b**ɪt|	'bit'	e.	/**s**ɪt/	|**s**ɪt|	'sit'	h.	/**h**ɪt/	|**h**ɪt|	'hit'
c.	/**l**ɪt/	|**l**ɪt|	'lit'	f.	/**f**ɪt/	|**f**ɪt|	'fit'	i.	/**w**ɪt/	|**w**ɪt|	'wit'

(6) A minimal set of Korean verbs, shown in their A form, with a /ca_a/ structure.

a.	/ca**p**a/	|ca**p**+A|	'hold'
b.	/ca**k**a/	|ca**k**+A|	'small'
c.	/ca**c**a/	|ca**c**+A|	'frequent (adjective)'

2.3.2. Second principle: alternation

In certain cases, the same morpheme appears slightly differently in different phonological environments, thus creating an alternation pattern with phonological conditions. The second principle deals with that: different surface categories that predictably alternate among each other in the same morpheme can be ascribed to the same underlying category.

Example (7) showcases the prefix /dis/ or /diz/ 'dis-' in Abruzzo Italian,[12] which alternates depending on the phonological environment it is found in: before a voiced consonant /z/ occurs, while /s/ is found before anything else (Krämer, 2009: 210). This means that /s/ and /z/ are allophones of the same underlying category and their alternation can be represented as follows:

12 The choice to use Abruzzo Italian is due to the absence of intervocalic /z/: compare Lombard Italian /pre**z**ente/ 'present' but /pre**s**ente/ 'hears/feels beforehand', which complicate the system. This issue is avoided is Abruzzo varieties, where both 'present' and 'hears/feels beforehand' would be /pre**s**ente/.

/sX/~/zÇ/, where /Ç/ is a voiced consonant and /X/ is any other segment, which can be visualized with the schema in Figure 1. The alternation /sX/~/zÇ/ can be ascribed to the underlying category |s|,[13] and the underlying form of the prefix /dis/~/diz/ can be said to be |dis+|.

(7) The /sX/~/zÇ/ alternation pattern of |s| in Abruzzo Italian.

a.	/armare/	|armare|	'to arm'	/di**s**armare/	|di**s**+armare|	'to disarm'
b.	/livɛllo/	|livɛllo|	'level'	/di**z**livɛllo/	|di**s**+livɛllo|	'difference in hight'
c.	/misura/	|misura|	'measure'	/di**z**misura/	|di**s**+misura|	'measurelessness'
d.	/d͡ʒɛlo/	|d͡ʒɛlo|	'frost'	/di**z**d͡ʒɛlo/	|di**s**+d͡ʒɛlo|	'defrost'
e.	/porre/	|porre|	'to put'	/di**s**porre/	|di**s**+porre|	'to arrange'

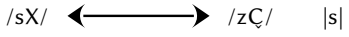


Figure 1. |s| alternates between /sX/ and /zÇ/ in Abruzzo Italian.

Examples from Korean verbs are provided in (8). If the TA and A forms are compared, many Korean verb stems display alternations near the morpheme boundary. Despite the different surface forms, each alternation can be linked to a single underlying form. Figure 2 schematically summarizes the alternations in (8).

(8) Different surface categories that predictably alternate can be ascribed to the same underlying category.

a.	/po**k**t'a/	/po**k'**a/	|po**k'**+|	'stir-fry'
b.	/ka**t**t'a/	/ka**t^{h}**a/	|ka**t^{h}**+|	'resemble'
c.	/si**p**t'a/	/si**p^{h}**ʌ/	|si**p^{h}**+|	'want'
d.	/nulɯta/	/nulɯ**l**ʌ/	|nułɯ**r**+|	'yellow'
e.	/a**n**t'a/	/a**nc**a/	|a**nc**+|	'sit'

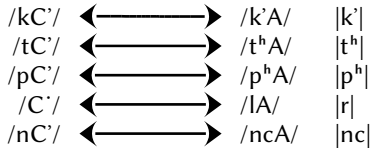


Figure 2. Surface differences do not necessarily correspond to underlying differences.

13 Where possible, the label for the underlying category will be chosen to be the same as the most common allophone or, when it sometimes neutralizes with other categories, an allophone unique to that alternation or pattern, like traditional phonology. This must not be taken to be a claim about the specific qualities or properties of the said category: in principle, any arbitrary yet distinct label is good, because the goal is defining categories, regardless of their internal complexity and composition. Further, since there is no need to transform one form into the other, but simply linking and matching them is sufficient, the similarity of the labels between surface and underlying forms is rather out of convenience more than anything else.

2.3.3. Third principle: neutralization

The third principle allows different underlying categories to merge under certain conditions in the surface form, i.e. to neutralize, and different neutralization patterns can be indicative of different underlying structures. Neutralizations create scenarios in which contrast between two alternation patterns is only partial. A Polish example of this is illustrated in (9), where both |p| and |b| correspond to /p/ before a voiceless consonant (/Ç/) like /k/ (Gussmann, 2007: 292), and they are thus neutralized in the /_Ç/ environment, but not in /_V/. This Polish voicing neutralization is visualized in Figure 3, where the /pÇ/ is shared by both alternation patterns.

(9) Polish voicing neutralization: |p| and |b| are neutralized as /p/ before /k/.

a.	/wa**p**a/	'paw'	/wa**p**ka/	'paw (diminutive)'	|wa**p**+|
b.	/ʒa**b**a/	'frog'	/ʒa**p**ka/	'frog (diminutive)'	|ʒa**b**+|

Figure 3. In Polish, both |p| and |b| share /pÇ/.

Similarly, Korean verb patterns often neutralize with each other, as demonstrated in (10) with /itt'a/, which is the TA form shared by three different verbs, the differences among which are evident in their A form. They form the three alternations /tC'/~/cʌ/, /tC'/~/s'ʌ/ and /tC'/~/ʌ/ that have /tC'/ in common.

(10) Partial neutralization in Korean verbs that share the same TA form /itt'a/.

a.	/i**tt**'a/	/i**c**ʌ/	|i**c**+|	'forget'
b.	/i**tt**'a/	/i**s'**ʌ/	|i**s'**+|	'existent'
c.	/i**tt**'a/	/iʌ/	|i**z**+|	'link'

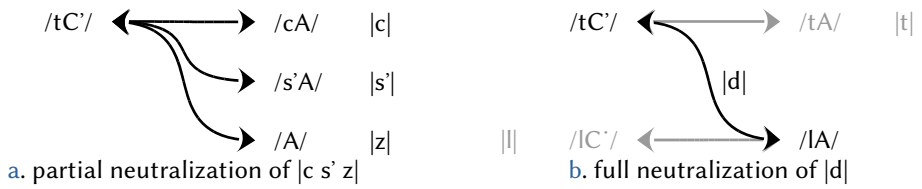
Nevertheless, this does not impede an alternation pattern to always neutralize with some other pattern and, thus, not have forms unique to itself. Despite this, the fact that the alternation pattern is different from the others may still point to a different underlying structure. This is exemplified in (11), which reports the examples in (3), adapted to the TA form, where neither /tC'/ nor /lʌ/ is unique to one pattern, yet the /tC'/~/lʌ/ alternation can still be linked to a single underlying |d| different from both |t| and |l| despite neutralizing with one or the other in both environments. Figure 4 compares the partially-neutralizing patterns in (10) with the fully-neutralizing one in (11); these are further discussed in §4.1.2 and §4.1.6 respectively.

(11) Neutralization pattern of |d|, which lacks a surface form of its own.

a. /kʌ**tt**'a/ /kʌ**t**ʌ/ |kʌ**t**+| 'roll up'

b. /kʌ**l**ta/ /kʌ**l**ʌ/ |kʌ**l**+| 'hang'

c. /kʌ**tt**'a/ /kʌ**l**ʌ/ |kʌ**d**+| 'walk'



a. partial neutralization of |c s' z| b. full neutralization of |d|

Figure 4. While |c s' z| neutralize into /tC'/, yet they contrast in /_V/, |d| always neutralizes with either |t| or |l|.

2.3.4. Fourth principle: behavioral difference

The last principle is about the surroundings of the segments that are being examined. Given the same underlying context, if the surface context is different around the same surface category, that surface category may correspond to different underlying categories, which behave differently despite merging. A well-known examples of this is the French word-initial "aspirated" *h*, presented in (12). Gabriel & Meisenburg (2009) show how an abstract underlying segment, which they label |ʔ̲|, contrasts with the absence of any segments, and how it affects the different surface forms of its surroundings. (c) looks like it is vowel-initial, but it behaves like consonant-initial (a), so it takes the articles /la/ and /le/ rather than /l/ and /lez/, which hints at the presence of an initial |ʔ̲|.

(12) French |ʔ̲|-initial words compared to consonant and vowel-initial words.

a.	/fij/	/**la**fij/	|fij|	'(the) girl'	/gaʁsɔ̃/	/**le**gaʁsɔ̃/	|gaʁsɔ̃|	'(the) boys'
b.	/abɛj/	/**la**bɛj/	|abɛj|	'(the) bee'	/ami/	/**lez**ami/	|ami|	'(the) friends'
c.	/os/	/**la**os/	|ʔ̲os|	'(the) rise'	/ɔ̃gʁwa/	/**le**ɔ̃gʁwa/	|ʔ̲ɔ̃gʁwa|	'(the) Hungarians'

Both in the French example in (12) and the Korean one in (13), the underlying categories that explain the surface differences do not surface themselves. As seen for French, the presence of |ʔ̲| leaves a recognizable trace, and (13) demonstrates how |ʔ| and |h| can analogously be distinguished by the way they influence their surroundings: the ending |+ta| in fact corresponds to /t'a/ after an otherwise-invisible |ʔ|, but to /tʰa/ after |h|. Compare (14), which presents another Korean example where the patterns of the three verbs are indistinguishable in the A form and all simply contain a liquid. However, when a hygrophobic environment is created through the

addition of the suffix |+(ɯ)nik'a| 'because', the stems that do not underlyingly end with a liquid retain it, while it is lost in the others.

(13) Korean |h|-final stem compared to a |ʔ|-final stem: only the following consonant hints at the difference.

a. /ana/ |an**ʔ**+A| /an**t'**a/ |an**ʔ**+ta| 'hug'

b. /ana/ |an**h**+A| /an**t^{h}**a/ |an**h**+ta| 'do not'

(14) Korean /l/ can be found even in hygrophobic environments if the verb stem does not actually end with an underlying liquid.

a. /mʌ**l**ʌ/ /mʌnik'a/ |mʌ**l**+| 'far'

b. /mu**l**ʌ/ /mu**l**ɯnik'a/ |mu**d**+| 'ask'

c. /si**l**ʌ/ /si**l**ɯnik'a/ |si**lh**+| 'unpleasant'

2.3.5. Combining the principles

In the messy reality of language, all principles must be combined to draw a full picture of a system. This is illustrated well in Underhill (1976), where Greenlandic is argued to have an underlying |ï| category distinct from |i|. This is done by noting how, despite it not forming any minimal pairs, |ï| sometimes corresponds to surface forms that |i| cannot correspond to in the same environment. In fact, |ï| leaves traces of its presence by letting an underlying |t| correspond to a surface /t/, which is atypical of a Greenlandic |i|, because it would require the |t| to surface as a /s/.

Arguments have also been made for Slavic jers, which surface as their unique vowel–zero alternations contrasting with full vowels, to exist as underlyingly different segments from the other vowels (cf. Rubach, 2016, for Polish; Hristovsky, 2021, for Bulgarian) on the basis of their different alternation patterns and behavior, among other things. Similarly, the discussion on Korean verb patterns will make use of all the principles combined, whenever they are relevant, and not individually.

3. Frequency of the patterns in the lexicon

This section is dedicated to the collection of the data about Seoul Korean verb patterns. While §3.1 explains how the list of patterns and their frequencies were obtained, §3.2 presents the results.

3.1. Methods

The online version of the Standard Korean Language Dictionary (표준국어대사전) (National Institute of the Korean Language, n.d.) was used as source material.[14] The alternation pattern of each verb was assessed and the token frequency of each detected pattern was counted automatically through the use of Python code, provided in Appendix B.

The `csv` files, converted from the `xls` filed provided by the National Institute of the Korean Language, were ingested using Python's `pandas` library. First, all entries that are not verbs were removed, as well as the negative copula (아니다 /anita/ 'not to be') due to its unique conjugation, and entries without sufficient conjugation information were also removed. Certain verbs can have two similar but different A forms, a longer one and a shorter one: these were separated and looked at independently. After the selection of only the relevant and usable entries, their Korean-script orthographic forms were transliterated with an *ad hoc* system using a modified version of the `transliter` library. The code then isolated the TA form (i.e. the citation form) and the A form (usually found as the first element in the conjugation section of the dictionary[15]), and it removed the information that was not needed.

Later, the code checked the final parts of the three isolated forms, and it recorded the partial pattern each form followed. The three partial patterns obtained were combined into a full pattern, one per dictionary entry, with the following form: `TA-form:A-form1/A-form2`. Afterwards, multimorphemic stems were removed in order to avoid recounting the same conjugated morpheme multiple times; this was only possible where the dictionary provides morphological information, so in a few cases this was not possible.[16]

Finally, the frequency of each full pattern was counted. The results of the counting are reported in the following section, both with and without removing the multimorphemic stems.

14 The full content of the dictionary is available on https://stdict.korean.go.kr/main/main.do (in Korean), downloadable as `xls` files.

15 The following entries were found to have typos in their conjugation section, which can prevent the code from running correctly: 끌어-모으다, 뒤-바르다, 신고-스럽다, 쌍-스럽다, 어르다(01), 증산-하다(01), 짜디-짜다, 윤-내다, 깨다(04), 답-쌔다, 드러-쌔다, 쌔다(02), 째다(03), 채다(02), 패다(05), 폐다, 헛방-놓다, 헛불-놓다, 휩쌔다, 깊어-지다, 나-굴다, 되잡-히다, 부탁-하다, 육보-하다, 증산-하다(01), 추출-되다, 토옥-하다. Their typos were manually corrected beforehand.

16 These are cases like the verb |is'+| 있+ 'existent' which is also found in |cem+is'+| 재ㅁ+ 있+ 'entertaining'. Due to the properties of the Korean writing system, the morpheme boundary between the abbreviated form of |cemi| 재미 'fun' and the target verb stem |is'+| cannot easily be represented on a dictionary, and the compilers have thus chosen to ignore it. The code cannot parse |cemis'+| 재밌+ and analogous entries, they are therefore counted as monomorphemic stems.

3.2. Results

This section presents the results of the counting described above. Table 4a reports the results without removing multimorphemic stems, while Table 4b reports the results excluding multimorphemic stems. Each pattern is given both as the tag produced by the code (i.e. `TA-form:A-form1/A-form2`) and by the surface alternations it stands for. The code detected and categorized a total of 67 295 verbs, including 1 961 monomorphemic stems, comprising both consonant and non-consonant patterns.

Table 4a. Results including multimorphemic stems. The output of the code is provided in the following format: `TA-form:A-form1/A-form2`, together with the alternation it represents and the section number for those that are further discussed.

consonant patterns

#	TA:A1/A2	pattern		#	TA:A1/A2	pattern	
1 061	b:w/w	/pC'/~/wA/	§4.1.6	45	nj:nj/nj	/nC'/~/ncA/	§4.1.4
622	l:l/l	/(l)C˙/~/lA/	§4.1.3	42	lg:lg/lg	/lkC'/~/lkA/	§4.1.4
234	g:g/g	/kC'/~/kA/	§4.1.1	32	s:s/s	/tC'/~/sA/	§4.1.2
162	j:j/j	/tC'/~/cA/	§4.1.2	29	k:k/k	/kC'/~/k'A/	§4.1.2
154	b:b/b	/pC'/~/pA/	§4.1.1	27	lh:lh/lh	/lC^{h}/~/lA/	§4.1.5
152	bs:bs/bs	/pC'/~/ps'A/	§4.1.4	24	lb:lb/lb	/lpC'/~/lpA/	§4.1.4
118	t:t/t	/tC'/~/t^{h}A/	§4.1.2	22	n:n/n	/nC'/~/nA/	§4.1.1
101	d:d/d	/tC'/~/tA/	§4.1.1	20	ɯ:l/l	/C˙/~/lA/	§4.1.3
71	s:q/q	/tC'/~/A/	§4.1.6	19	lm:lm/lm	/lmC'/~/lmA/	§4.1.4
70	m:m/m	/mC'/~/mA/	§4.1.1	14	ʂ:ʂ/ʂ	/tC'/~/s'A/	§4.1.2
61	d:l/l	/tC'/~/lA/	§4.1.6	11	c:c/c	/tC'/~/c^{h}A/	§4.1.2
56	nh:nh/nh	/nC^{h}/~/nA/	§4.1.5	7	lt:lt/lt	/ltC'/~/lt^{h}A/	§4.1.4
55	h:h/h	/C^{h}/~/A/	§4.1.5	1	lp:lp/lp	/lpC'/~/lp^{h}A/	§4.1.4
46	p:p/p	/pC'/~/p^{h}A/	§4.1.2	1	lb:w/w	/lpC'/~/lwA/	§4.1.6

mixed consonant and non-consonant patterns

#	TA:A1/A2	pattern	#	TA:A1/A2	pattern
120	h:e/e	/AC^{h}/~/e/	55	h:h/w	/oC^{h}/~/oA~wA/

non-consonant patterns

#	TA:A1/A2	pattern	#	TA:A1/A2	pattern
48 005	a:a/e	/aC˙/~/ajA~e/	184	ɯ:ɯ/ɯ & ɯ:q/q	/ɯC˙/~/A/
7 935	e:q/e	/eC˙/~/eA~e/	143	o:q/o	/oC˙/~/oA~wA/
6 172	i:q/i	/iC˙/~/iA~jA/	116	i:q/q	/iC˙/~/iA/
591	a:a/a & ʌ:ʌ/ʌ	/AC˙/~/A/	81	o:o/o	/oC˙/~/wA/
322	u:q/u	/uC˙/~/uA~wA/	17	ʌ:e/e	/ʌC˙/~/e/
296	ɯ:ll/ll	/ɯC˙/~/lA/	1	u:u/u	/uC˙/~/A/

Table 4b. Results excluding multimorphemic stems. The output of the code is provided in the following format: `TA-form:A-form1/A-form2`, together with the alternation it represents and the section number for those that are further discussed.

consonant patterns

#	TA:A1/A2	pattern		#	TA:A1/A2	pattern	
175	l:l/l	/(l)C˙/~/lA/	§4.1.3	12	k:k/k	/kC'/~/k'A/	§4.1.2
127	b:w/w	/pC'/~/wA/	§4.1.6	9	h:h/h	/C^{h}/~/A/	§4.1.5
47	nh:nh/nh	/nC^{h}/~/nA/	§4.1.5	9	p:p/p	/pC'/~/p^{h}A/	§4.1.2
31	g:g/g	/kC'/~/kA/	§4.1.1	8	lm:lm/lm	/lmC'/~/lmA/	§4.1.4
30	m:m/m	/mC'/~/mA/	§4.1.1	6	lb:lb/lb	/lpC'/~/lpA/	§4.1.4
29	j:j/j	/tC'/~/cA/	§4.1.2	4	ɯ:l/l	/C˙/~/lA/	§4.1.3
23	b:b/b	/pC'/~/pA/	§4.1.1	3	lt:lt/lt	/ltC'/~/lt^{h}A/	§4.1.4
18	d:d/d	/tC'/~/tA/	§4.1.1	2	nj:nj/nj	/nC'/~/ncA/	§4.1.4
18	s:q/q	/tC'/~/A/	§4.1.6	2	n:n/n	/nC'/~/nA/	§4.1.1
16	t:t/t	/tC'/~/t^{h}A/	§4.1.2	2	c:c/c	/tC'/~/c^{h}A/	§4.1.2
14	d:l/l	/tC'/~/lA/	§4.1.6	2	bs:bs/bs	/pC'/~/ps'A/	§4.1.4
14	lg:lg/lg	/lkC'/~/lkA/	§4.1.4	2	ʂ:ʂ/ʂ	/tC'/~/s'A/	§4.1.2
13	s:s/s	/tC'/~/sA/	§4.1.2	1	lp:lp/lp	/lpC'/~/lp^{h}A/	§4.1.4
13	lh:lh/lh	/lC^{h}/~/lA/	§4.1.5	1	lb:w/w	/lpC'/~/lwA/	§4.1.6

mixed consonant and non-consonant patterns

#	TA:A1/A2	pattern	#	TA:A1/A2	pattern
62	h:e/e	/AC^{h}/~/e/	1	h:h/w	/oC^{h}/~/oA~wA/

non-consonant patterns

#	TA:A1/A2	pattern	#	TA:A1/A2	pattern
664	i:q/i	/iC˙/~/iA~jA/	53	i:q/q	/iC˙/~/iA/
172	e:q/e	/eC˙/~/eA~e/	17	ʌ:e/e	/ʌC˙/~/e/
124	u:q/u	/uC˙/~/uA~wA/	10	o:q/o	/oC˙/~/oA~wA/
90	ɯ:ll/ll	/ɯC˙/~/lA/	3	o:o/o	/oC˙/~/wA/
74	ɯ:ɯ/ɯ & ɯ:q/q	/ɯC˙/~/A/	1	a:a/e	/aC˙/~/ajA~e/
58	a:a/a & ʌ:ʌ/ʌ	/AC˙/~/A/	1	u:u/u	/uC˙/~/A/

The results show that the two most common consonant patterns are /lC˙/~/lA/ and /pC'/~/wA/ by a large margin, both of which are peculiar in different ways: the former has certain vowel-like characteristics, while the latter is traditionally one of the "irregular" patterns, despite having a far higher frequency in the lexicon than similar patterns: it is about 5.5 times more common than /pC'/~/pA/ (about 6.9 times including multimorphemic stems), about 14.1 times more common than /pC'/~/p^{h}A/ (about 23.1 times including multimorphemic stems), and about 63.5 times more common than /pC'/~/ps'A/ (about 7.0 times including multimorphemic stems, probably due to |ʌps'+| 'not to have' being commonly used in compounds).

4. Discussion

This section discusses the various surface alternations patterns attested in Seoul Korean in light of the principles described in §2.3 and with the goal of expanding on the traditional account of the underlying form of Seoul Korean, introduced in §2.2. §4.1 specifically focuses on the alternation patterns, while §4.2 adds further remarks.

4.1. The declension patterns of Seoul Korean

Below, the criteria in §2.3 are applied to detect unique underlying forms that can describe each attested alternation pattern. Most of such patterns are not novel and they have been proposed or are generally robust in the literature, but here they are all reclassified and justified by the same measures. Table 5 shows the full list of underlying forms discussed.

Table 5. The underlying forms of Seoul Korean verb-stem-final consonants and consonant clusters as further discussed in the indicated sections.

		labial	**coronal**			**dorsal**	**laryngeal**
glottalized C'					|s'| (§4.1.2)	|k'| (§4.1.2)	
aspirated C^h		|p^h| (§4.1.2)	|t^h| (§4.1.2)	|c^h| (§4.1.2)	|s| (§4.1.2)		
plain C˙		|p| (§4.1.1)	|t| (§4.1.1)	|c| (§4.1.2)		|k| (§4.1.1)	
nasal N(+?)		|m(?)| (§4.1.1)		|n(?)| (§4.1.1)			
liquid L				|l| (or |ɬ| & |r|) (§4.1.3)			
other[17] **C^x**		|b| (§4.1.6)	|d| (§4.1.6)		|z| (§4.1.6)		|h| (§4.1.5)
clusters	**L+C^h**	|lp^h| (§4.1.4)	|lt^h| (§4.1.4)				
	L+C˙	|lp| (§4.1.4)				|lk| (§4.1.4)	
	L+N(+?)	|lm(?)| (§4.1.4)					
	L+C^x	|lb| (§4.1.6)					|lh| (§4.1.5)
	C+C	|ps'| (§4.1.4)		|nc| (§4.1.4)			|nh| (§4.1.5)

17 These labels are arbitrary and only loosely based on the historical sounds that ultimately created the modern patterns that are being discussed from an otherwise synchronic point of view.

Note, however, that a few of the underlying categories with a robust presence in the literature are absent from the following discussion, as no alternation that can be linked to them has been detected stem-finally in verbs. These are the nasal |ŋ|, the glottalized |p' t' c'|[18] and the aspirated |k^{h}|.

4.1.1. Contrasting but non-alternating patterns

The first group of patterns is that where verb-stem-final consonants do not feature an alternation between the prevocalic and preconsonantal environments. This group comprises the following surface patterns: /pC'/~/pA/, /mC'/~/mA/, /tC'/~/tA/, /nC'/~/nA/, /kC'/~/kA/, which can be ascribed to underlying |p m t n k| respectively due to the fact that they can form minimal sets, as shown in Table 6. Minimal pairs for |p|–|m|, |p|–|n| and |t|–|n| in this position have not been found.

Table 6. Minimal pairs of verb-stem-final |p m t n k|, which contrast throughout their paradigm but do not feature alternations. Minimal pairs for |p|–|m|, |p|–|n| and |t|–|n| in this position have not been found.

	k	**t**	**n**	**p**
m	|si_+| /si**m**t'a/ ~ /si**m**ʌ/ 'plant' /si**k**t'a/ ~ /si**k**ʌ/ 'cool down'	|ta_+| /ta**m**t'a/ ~ /ta**m**a/ 'put into' /ta**t**t'a/ ~ /ta**t**a/ 'shut'	|si_+| /si**m**t'a/ ~ /si**m**ʌ/ 'plant' /si**n**t'a/ ~ /si**n**ʌ/ 'wear (footwear)'	–
p	|ca_+| /ca**p**t'a/ ~ /ca**p**a/ 'hold' /ca**k**t'a/ ~ /ca**k**a/ 'little'	|ko_+| /ko**p**t'a/ ~ /ko**p**a/ 'numb' /ko**t**t'a/ ~ /ko**t**a/ 'straight'	–	
n	|si_+| /si**n**t'a/ ~ /si**n**ʌ/ 'wear (footwear)' /si**k**t'a/ ~ /si**k**ʌ/ 'cool down'	–		
t	|pa_+| /pa**t**t'a/ ~ /pa**t**a/ 'receive' /pa**k**t'a/ ~ /pa**k**a/ 'insert'			

When the first consonant after the stem is an obstruent, it is always glottalized. As mentioned in §2.2, this is expected for the three obstruents /p t k/. However, the nasals /m n/ can be followed by plain obstruents in Korean, though this is not the case across verb–ending boundaries. This might point to the fact that /m n/-final stems are not, in fact, |m n|-final, and that they are plausibly

18 To my knowledge, the three glottalized |p' t' c'| are not attested at the end of any stem, verbal or else.

clusters containing a glottalizing segment |ʔ|, *i.e.* |mʔ nʔ|, much like the aspirating segment |h| discussed in §4.1.5.

The additional glottalization is found in a few scenarios where plain and glottalized obstruents also alternate after sonorants, which is otherwise unexpected, such as in compounding (15a), after the modifier future ending (15b) or with /m n/-final verb stems (15c) (Martin, 1992: 12–15, Kim-Renaud, 2022). Different explanations are possible: further research is needed to confirm whether these indeed represent the same phenomenon or whether they are only superficially similar.

(15) Glottalized obstruents appear after sonorants across some, but not all, morpheme and word boundaries.

a. *compounding*: /pom/ 'spring' + /**p**am/ 'night' > /pom**p'**am/ 'spring night' |pom+**ʔ**+**p**am|
   cf. *compounding without glottalization*: /pjʌl/ 'star' + /**p**am/ 'night' > /pjʌl**p**am/ 'starry night' |pjʌl+**p**am|

b. *future*: /mʌkɯl/ 'that will eat' + /**k**ʌt/ 'thing' > /mʌkɯl**k'**ʌt/ 'something to eat' |mʌk+ɯl**ʔ#k**ʌs|
   cf. *past*: /mʌkɯn/ 'that has eaten' + /**k**ʌt/ 'thing' > /mʌkɯn**k**ʌt/ 'something eaten' |mʌk+ɯn**#k**ʌs|
   cf. *accusative*: /mʌkɯl/ 'inkstick (accusative)' + /**k**al/ 'rub' > /mʌkɯl**k**al/ 'rub an inkstick' |mʌk+ɯl**#k**al+|

c. *nasal-final verb stems*: /an/ 'hug' + /**t**a/ 'TA-form' > /an**t'**a/ 'to hug' |an**ʔ**+**t**a|
   cf. *liquid-final verb stems*: /al/ 'know' + /**t**a/ 'TA-form' > /al**t**a/ 'to know' (see §4.1.3) |al+**t**a|
   cf. *vowel-final verb stems*: /ka/ 'go' + /**t**a/ 'TA-form' > /ka**t**a/ 'to go' |ka+**t**a|

4.1.2. Partially neutralizing alternation patterns

The following patterns present an alternation between prevocalic and preconsonantal environments: /pC'/~/pʰA/, /tC'/~/tʰA/, /tC'/~/cA/, /tC'/~/cʰA/, /tC'/~/sA/, /tC'/~/s'A/, /kC'/~/k'A/; they all contrast prevocalically, but preconsonantally they neutralize with some other pattern, as shown in Figure 5. Since these alternations involve an allophone each unique to themselves that can contrast across different patterns, each alternation can be ascribed to its unique underlying category: |pʰ tʰ c cʰ s s' k'| respectively. However, due to the low number of verb stems following some of these patterns (§3.2), minimal pairs where the stem-final consonant is the only difference are difficult to find. Some minimal pairs are nonetheless shown in (16).

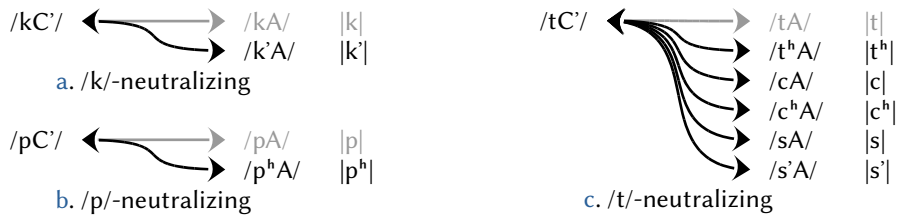


Figure 5. Schematic representation of partially neutralizing alternation patterns (black) together with the non alternating patterns they neutralize with (grey).

(16) Examples of minimal pairs for verb-stem-final |p pʰ t tʰ c cʰ s s’ k k’|. Among /t/-neutralizing patterns, minimal pairs are not all attested stem-finally in verbs, hence why not all combinations are shown.

a. |k|–|k’|:
/mu**k**t’a/ ~ /mu**k**ʌ/ ‘lodge’ |mu**k**+|
/mu**k**t’a/ ~ /mu**k’**ʌ/ ‘tie’ |mu**k’**+|

b. |p|–|pʰ|:
/ʌ**p**t’a/ ~ /ʌ**p**ʌ/ ‘carry’ |ʌ**p**+|
/ʌ**p**t’a/ ~ /ʌ**pʰ**ʌ/ ‘turn over’ |ʌ**pʰ**+|

c. |t|–|s|:
/pʌ**t**t’a/ ~ /pʌ**t**ʌ/ ‘stretch’ |pʌ**t**+|
/pʌ**t**t’a/ ~ /pʌ**s**ʌ/ ‘take off’ |pʌ**s**+|

d. |c|–|s|:
/pi**c**t’a/ ~ /pi**c**ʌ/ ‘brew’ |pi**c**+|
/pi**t**t’a/ ~ /pi**s**ʌ/ ‘comb’ |pi**s**+|

e. |c|–|cʰ|:
/c’o**c**t’a/ ~ /c’o**c**a/ ‘tie’ |c’o**c**+|
/c’o**c**t’a/ ~ /c’o**cʰ**a/ ‘go after’ |c’o**cʰ**+|

f. |c|–|tʰ|:
/ma**c**t’a/ ~ /ma**c**a/ ‘correct’ |ma**c**+|
/ma**c**t’a/ ~ /ma**tʰ**a/ ‘take on’ |ma**tʰ**+|

g. |c|–|t|:
/ku**c**t’a/ ~ /ku**c**ʌ/ ‘bad’ |ku**c**+|
/ku**t**t’a/ ~ /ku**t**ʌ/ ‘harden’ |ku**t**+|

h. |c|–|s’|:
/i**c**t’a/ ~ /i**c**ʌ/ ‘forget’ |i**c**+|
/i**t**t’a/ ~ /i**s’**ʌ/ ‘existent’ |i**s’**+|

4.1.3. Liquid patterns

The next two patterns both involve liquids: /(l)C˙/~/lA/ and /C˙/~/lA/. They both alternate between a surface liquid and nothing, they can nonetheless be ascribed to different underlying categories, |ɬ| and |r| respectively, due to their different neutralization patterns. Before a vowel-initial ending, they both neutralize and their stems ends with /l/. Meanwhile, before consonant-initial endings the scene is more complex: |r| never surfaces; |ɬ| only surfaces before hygrotolerant endings (e.g. |+ta|), and it does not surface before hygrophobic endings (e.g. |+(ɯ)nik’a| ‘because’). Table 7 provides examples of the two liquid patterns compared to a consonantless vowel pattern, and Figure 6 schematically illustrates their alternation and neutralization patterns.

Table 7. Liquid patterns compared to a consonantless pattern (first row). The stem-final consonant is in bold, if present.

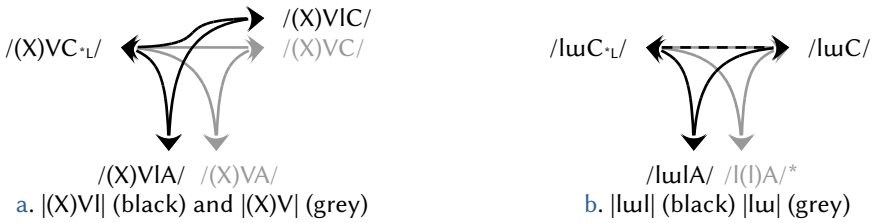
stem-final consonant	**underlying stem**	**hygrotolerant**		**hygrophobic**
		_A	**_C**	**_C˙ʟ**
	|...+|	**|+A|**	**|+ta|**	**|+(ɯ)nik’a|**
none	|cʰiɬɯ+| ‘pay’	/cʰilʌ/ (invisible |ɯ|)*	/cʰilɯta/	/cʰilɯnik’a/
|r|	|pʰuɬɯ**r**| ‘blue’	/pʰulɯ**l**ʌ/	/pʰulɯta/	/pʰulɯnik’a/
|ɬ|	|mantɯ**ɬ**+| ‘make’	/mantɯ**l**ʌ/	/mantɯ**l**ta/	/mantɯnik’a/

* |ɯ| does not surface in the A form of |ɯ|-final verbs.

a. |ɬ| (black) and no consonant (grey) b. |r| (black) and no consonant (grey)

Figure 6. Schematic representation of liquid patterns compared to a consonantless (vocalic) pattern (/A/~/C/~/C˙ʟ/). The neutralization patterns of (a) |ɬ| and (b) |r| differ in how they behave before hygrotolerant consonants (C).

It must also be noted that |ɬ| and |r| have not been found to form minimal pairs: |r| is restricted to stems that end with |ɬɯr|, while |ɬɯɬ| does not seem to occur.[19] They have mutually exclusive distributions and the only argument in favor of them being different underlying categories is their different behavior before hygrotolerant consonants. It is possible that |ɬ| and |r| are in fact one and the same liquid category |l|, with more complex and specific environmental conditions to its alternation pattern. Figure 6′ illustrates the same liquid patterns of Figure 6 above, but it shows them following this single-liquid analysis, with a more detailed context also provided in order to account for the different alternations attested: |lɯ_| occurs with a |r|-like pattern, while |(X)V_|, i.e. everything else, occurs with a |ɬ|-like pattern.



a. |(X)Vl| (black) and |(X)V| (grey)
b. |lɯl| (black) |lɯ| (grey)

Figure 6′. Schematic representation of the liquid patterns (black) compared to a consonantless vocalic pattern (grey). The context determining the surfacing or not of the liquid includes what comes before: |lɯ_| makes a difference.
* |lɯ|-final stems mostly have an A form ending with /llA/, but a few verbs have /lA/ like other |ɯ|-final stems that have a non-liquid consonant before. Their distinction is beyond the scope of this thesis and, regardless, there is no /lɯA/.

Furthermore, subsuming both |ɬ| and |r| under a single |l| simplifies the phonological system. A distinct |r| would otherwise only be detectable in five verbal stems (and their derivatives): |pʰałɯr+| and |pʰułɯr+| 'blue', |nołɯr+| and |nułɯr+| 'yellow', |iłɯr+| 'arrive'.[20] It is thus not very economical to maintain a distinct |r| with a very restricted distribution. Table 7′ thus reproposes the same data of Table 7 but in light of the single-liquid analysis.

Table 7′. Liquid patterns compared to a consonantless pattern (first row). The stem-final consonant is bold, if present.

end of the stem	**underlying stem**	**hygrotolerant**		**hygrophobic**
		_V	**_C**	**_C·L**
	|...+|	**|+A|**	**|+ta|**	**|+(ɯ)nikʾa|**
|(C)V|	|cʰilɯ+| 'pay'	/cʰilʌ/ (invisible |ɯ|)*	/cʰilɯta/	/cʰilɯnikʾa/
|lɯl|	|pʰulɯ**l**+| 'blue'	/pʰulɯ**l**ʌ/	/pʰulɯta/	/pʰulɯnikʾa/
|(C)Vl|	|mantɯ**l**+| 'make'	/mantɯ**l**ʌ/	/mantɯ**l**ta/	/mantɯnikʾa/

* |ɯ| does not surface in the A form of |ɯ|-final verbs.

19 |rɯr| and |rɯɬ| would surface in the same way as |ɬɯr| and |ɬɯɬ| respectively; the distinction is trivial.
20 See Kwon (2018) and Larsen & Heinz (2012) for a discussion on the iconic function of Korean vowel harmony.

Further research should investigate whether |ɬ| and |r| are indeed distinct. Unfortunately, the focus of this work on verb-stem-final consonants does not provide sufficient data to give a satisfactory answer and it only leaves room to speculation.

4.1.4. Patterns with surface clusters

In a number of patterns, the A form features a stem-final cluster. These are the following: /ps' nc lp lpʰ lm ltʰ lk/ and they are widely recorded in the literature (Kim-Renaud, 1974: 133–145; Lee, 1989: 34–36; Lee & Ramsey, 2000: 67–68; Martin, 1992: 101; Sohn, 1999: 170–172). On the other hand, the preconsonantal forms they alternate with do not all have a surface structure that the literature agrees upon. /pC'/~/ps'A/ and /nC'/~/ncA/, where the second element does not surface preconsonantally, are uncontroversial and can simply be posited to correspond to underlying |ps' nc| respectively,[21] as exemplified in (17). As for the other ones, Kim-Renaud (1973, 1974, 1977, 2022) and Cho & Kim (2009) attest to optional cluster simplification before consonants, while other authors only report the prescriptive simplified clusters. Such simplifications are reported in (18) with examples. What is notable is that they all begin with /l/, which is rather more vowel-like when it comes to verbal conjugation, as seen in the previous section.

(17) Alternation patterns of |ps' nc|. The second element of the cluster does not surface preconsonantally.

a.	/**p**C'/~/**ps'**A/	|ps'|	/ʌ**p**t'a/	/ʌ**ps'**ʌ/	'inexistent'
b.	/**n**C'/~/**n**cA/	|nc|	/a**n**t'a/	/a**nc**a/	'sit'

(18) Alternation patterns comprising /lC/ clusters: surface simplification. The consonant in parentheses "(…)" is the one being dropped when the cluster is optionally simplified. Example verbs are also given.

a.	/**(l)k**C'/~/**lk**A/*	/i**(l)k**t'a/	/i**lk**ʌ/	'read'
b.	/**l(t)**C'/~/**lt^{h}**A/†	/ha**l(t)**t'a/	/ha**lt^{h}**a/	'lick'
c.	/**(l)m**C'/~/**lm**A/	/cʌ**(l)m**t'a/	/cʌ**lm**ʌ/	'young'
d.	/**(l)p**C'/~/**lp^{h}**A/	/ɯ**(l)p**t'a/	/ɯ**lp^{h}**A/	'recite (poetry)'
e.	/**(l)p**C'/~/**lp**A/	/pa**(l)p**t'a/	/pa**lp**a/	'step'
f.	/**l(p)**C'/~/**lp**A/	/nʌ**l(p)**t'a/	/nʌ**lp**ʌ/	'wide'

* The simplification is /l(k)C'/ before /k'/: simplified |+ta| /i**k**t'a/, but |+ko| /i**l**k'o/; though this is a detail.

† Kim-Renaud justifies the absence of a phonetically unreleased /t/ [t˺] on articulatory grounds and she does not consider its simplification optional, which can be a good reason for [t˺] to never be pronounced but, given the rest of the variation, there is no reason to believe /t/ cannot (optionally) be there.

21 Due to the post-obtruent glottalization mentioned in §2.2 and preconsonantal neutralization (§4.1.2), |ps'| could possibly also be any of the following: |ps pʰs pʰs'| or, including the categories in §4.1.6, even |bs bs'|. Not much is gained by choosing one or the other, so the closest to the surface form is chosen: /ps'/ |ps'|.

What is striking, or knowing the rather weak nature of /l/ (§4.1.3) perhaps not, is that in almost all of the patterns in (18) it is the liquid being dropped. The only consistent exception is /l(t)C'/~/lt^{h}A/, though this can easily be explained articulatorily, as Kim-Renaud does, if the cluster simplification is a phonetic, rather than phonological, phenomenon (e.g. it is invariably /cʌlmt'a/ but variably [tɕʌlmt'a~tɕʌmt'a], alternating with an invariable /cʌlmʌ/ [tɕʌlmʌ]). If this is the case, then (18′) can be reasonably assumed. This, however, leaves the problem of the choice between deleting the [l] or the [p$^{\text{ˀ}}$], as the preference seems lexicalized.

(18′) Alternation patterns comprising /lC/ clusters: phonetic simplification.

a.	[(l)k$^{\text{ˀ}}$C']~[lgA]	/lkC'/~/lkA/	|lk|	[i**(l)**k$^{\text{ˀ}}$t'a]~[i**l**gʌ]	/i**lk**t'a/~/i**lk**ʌ/	|i**lk**+|	'read'
b.	[lC']~[lt^{h}A]	/ltC'/~/lt^{h}A/	|lt^{h}|	[ha**l**t'a]~[ha**lt^{h}**a]	/ha**l**tt'a/~/ha**lt^{h}**a/	|ha**lt^{h}**+|	'lick'
c.	[(l)mC']~[lmA]	/lmC'/~/lmA/	|lm(?)|*	[tɕʌ**(l)m**t'a]~[tɕʌ**lm**ʌ]	/cʌ**lm**t'a/~/cʌ**lm**ʌ/	|cʌ**lm**+|	'young'
d.	[(l)p$^{\text{ˀ}}$C']~[lp^{h}A]	/lpC'/~/lp^{h}A/	|lp^{h}|	[ɯ**(l)p**$^{\text{ˀ}}$t'a]~[ɯ**lp^{h}**ʌ]	/ɯ**lp**t'a/~/ɯ**lp^{h}**ʌ/	|ɯ**lp^{h}**+|	'recite'
e.	[(l)p$^{\text{ˀ}}$C']~[lbA]	/lpC'/~/lpA/	|lp|	[pa**(l)p**$^{\text{ˀ}}$t'a]~[pa**lb**a]	/pa**lp**t'a/~/pa**lp**a/	|pa**lp**+|	'step'
f.	[l(p$^{\text{ˀ}}$)C']~[lbA]	/lpC'/~/lpA/	|lp|	[nʌ**l(p**$^{\text{ˀ}}$**)**t'a]~[nʌ**lb**ʌ]	/nʌ**lp**t'a/~/nʌ**lp**ʌ/	|nʌ**lp**+|	'wide'

* The same "unexpected" glottalization after /m/ (§4.1.1) is also found after /lm/. If verb-stem-final /m/ is |m?|, then /lm/ must be |lm?|.

Alternatively, the cluster simplification is an optional phonological phenomenon, and positing two underlying clusters |lp$_1$| /(l)pC'/~/lpA/ [lp$^{\text{ˀ}}$C'~p$^{\text{ˀ}}$C']~[lbA] and |lp$_2$| /l(p)C'/~/lpA/ [lp$^{\text{ˀ}}$C'~lC']~[lbA] that differ minimally in some way would solve the issue, because different, though similar, underlying structures can correspond to different surface structures. Either way, empirical research on the topic is lacking and further research is needed.

4.1.5. Aspirating patterns

Aspirating patterns involve, as the label suggests, aspiration on the surface form that is otherwise hard to explain. More specifically, they involve the aspiration of the obstruent following the stem, while the stem itself does not seem to change on the surface: /C^{h}/~/A/, /lC^{h}/~/lA/, /nC^{h}/~/nA/. This is a clear case of behavioral difference pointing to different underling forms (§2.3.4). In fact, the stem itself of these verbs does not undergo any alternations and it never looks any different from a vowel, /l/ or /n/-final stem, but the difference is clear when an obstruent follows the stem. Examples comparing these patterns are in (19), and the pattern-neutralization patterns are visualized in Figure 7. This aspiration can be reconstructed as an underlying stem-final |h| that, like the French |ʔ| (Gabriel & Meisenburg, 2009), is itself invisible on the surface but leaves a trace of its presence, in this case in the form of aspiration.

(19) Aspirating alternation patterns compared to similar non-aspirating patterns.

a.	/co**t^{h}**a/	/co**a**/	|co**h**+|	'good'
a′.	/cu**ta**/	/cu**ʌ**/	|cu+|	'give'
b.	/i**lt^{h}**a/	/i**lʌ**/	|i**lh**+|	'lose'
b′.	/i**lt**a/	/i**lʌ**/	|i**l**+|	'happen'
c.	/a**nt^{h}**a/	/a**na**/	|a**nh**+|	'do not'
c′.	/a**nt'**a/	/a**na**/	|a**n(?)**+|	'hug'

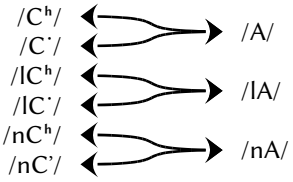


Figure 7. The aspirating patterns merge prevocalically with similar non-aspirating patterns.

It must be noticed that word-internally, other instances of |h| do seem to surface, cf. /nahant^{h}e/ 'to me', from |na| 'me' and |+hant^{h}e| 'to'. This might, however, be a product of the different environments |h| is found in: at the end of verb stems, it is before a morpheme boundary, while elsewhere it is never right before a morpheme boundary. Otherwise, two separate |h$_1$| and |h$_2$| must be posited, but this seems redundant since the contextual information is enough to resolve the issue.

4.1.6. Fully neutralizing patterns

Fully neutralizing patterns are patterns that the traditional literature deems "irregular". These patterns are the following: /pC'/~/wA/ (also in a /lC/ cluster: /l(p)C'/~/lwA/), /tC'/~/lA/, /tC'/~/A/. What is most characteristic about these patterns is that nowhere in their paradigm do they have forms unique to themselves, which is clearly illustrated in Figure 8 and further exemplified in (20). Despite this, the characteristics that enable these patterns to be distinctly linked to their unique underlying forms are that the alternation and neutralization patterns cannot be explained by any of the other underlying categories or clusters of categories that had been introduced already. This is, indeed, the issue mentioned in §2.3.3, about the lack of a unique surface form, and that can now be discussed with the right tools at hand. Each one of these patterns is underlyingly distinct: |b| /pC'/~/wA/ (and |lb| /l(p)C'/~/lwA/), |d| /tC'/~/lA/, |z| /tC'/~/A/.

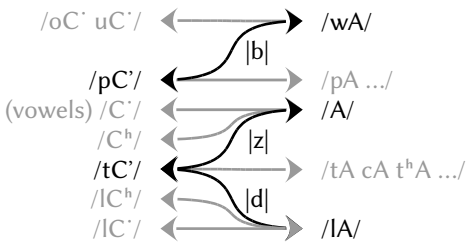


Figure 8. |b z d| (black) neutralize with other patterns (grey) in every possible context.

(20) Verbs following fully neutralizing patterns compared to similar verbs that follow other patterns. Additional forms in a hygrophobic environment are given for the examples involving a liquid.

a.	/me**pt'**a/	/me**wʌ**/		|me**b**+|	'spicy'
a′.	/me**ut**a/	/me**wʌ**/		|me**u**+|	'fill up'
a″.	/i**pt'**a/	/i**pʌ**/		|i**p**+|	'wear'
b.	/ta**tt'**a/	/ta**la**/	/ta**lɯn**ik'a/	|ta**d**+|	'run'
b′.	/ta**tt'**a/	/ta**ta**/	/ta**tɯn**ik'a/	|ta**t**+|	'shut'
b″.	/ta**lt^{h}**a/	/ta**la**/	/ta**lɯn**ik'a/	|ta**lh**+|	'wear down'
b‴.	/ta**lt**a/	/ta**la**/	/ta**n**ik'a/	|ta**l**+|	'hang; sweet'
c.	/ci**tt'**a/	/ci**a**/		|ci**z**+|	'make'
c′.	/ci**tt'**a/	/ci**ca**/		|ci**c**+|	'bark'
c″.	/ci**tt'**a/	/ci**t^{h}a**/		|ci**t^{h}**+|	'dense'
c‴.	/c'i**t^{h}**a/	/c'i**a**/		|c'i**h**+|	'hit'

It can further be noted, as additionally provided in (20b–b‴), that when /l/ corresponds to |d| (as well as |lh|) it can also be found in the hygrophobic environment created by |+(ɯ)nik'a|, however a true liquid like |l| does not appear.

4.2. Further remarks

The traditional account fails to capture the consistence that certain patterns have across the lexicon. The category that can now be labeled |b|, based on the alternation /pC'/~/wA/, and the liquid |l| are by far the most common stem-final consonants among Seoul Korean verbs. Since the former has mostly been considered "irregular", it leaves one to wonder what that actually means. If a pattern is irregular and no systematicity can be found, how can this pattern be so prevalent in the lexicon? The usage of such terms should perhaps be backed by empirical data, rather than being used out of convenience whenever a more complex phenomenon is encountered in a language. Perhaps criteria like productivity, child errors and reanalysis tendencies can be used to better define which of the patterns are better labeled as "regular" and which ones "irregular".

A factor that certainly influences the results of said criteria is the frequency of the alternations and patterns in a naturalistic environment, rather than a simple token count like the one carried out here. Note, for example, that both |is'+| 'existent' and |ɯlpʰ+| 'recite (poetry)' are unique for their alternation pattern among verb stems, but |is'+| is far more common than |ɯlpʰ+| in speech, since |ɯlpʰ+| is a verb with very specific semantics, while |is'+| is a function word with a wide range of uses. Furthermore, the /tC'/~/s'V/ alternation is also found in the past |+As'+| and future |+kes'+| endings, while the /lpC'/~/lpʰA/ alternation is not attested elsewhere; the frequency imbalance in the usage of the morphemes featuring such alternations affects the frequency of the alternations themselves in the language in use. This might have different strengthening or weakening effects on such patterns, which translates to a more or less "regular" pattern. It is also possible that each and every word is stored separately in the Korean user's lexicon, with no generalizable patterns, as the exemplar theory would suggest. This is however discredited by research on the generalizability of Korean verb patterns like that of Jun & Albright (2017).

Future research should investigate the gaps left or pointed at throughout this discussion. This may be done by introducing novel alternations in an experimental setting, which is possible thanks to some categories not being attested verb-stem-finally. Further, variation in the specific quality and structure has been observed among nominal stems (Jun, 2010; Kim, 2022) and verbal stems alike (Cho & Kim, 2009). As for clusters specifically, while Cho & Kim (2009) seem to show a spread of stem-final surface clusters from prevocalic environments to preconsonantal environments also, Kim (2022) explains that the opposite is true for nouns, where clusters tend to be simplified across all environments. This is a yet-unexplored difference and a comprehensive overview comparing the two is still missing.

5. Conclusion

This thesis has presented the traditional account of Seoul Korean phonology and it has showed how certain phonological alternations commonly found among verb stems cannot be explained by such an account. To close this gap, a renewed account has been discussed on the basis of four criteria, which were inspired by the phonological literature on similar issues in other languages and widely-used concepts. These are the principles of *contrast*, i.e. the ability for different categories to form minimal pairs, *alternation*, i.e. the ability for the same underlying category to correspond to different surface forms in different environments, *neutralization*, i.e. the ability for different underlying categories to merge into the same surface category in the same environment,

and *behavioral difference*, i.e. the ability for different underlying categories in the same environment to affect their surroundings differently, which is then visible in the surface form. Data has also been collected from a dictionary to give a quantitative dimension to the attested patterns and to make sure that no relevant alternation has been overlooked.

It has emerged that the labels "regular" and "irregular" traditionally used to refer to certain Korean conjugation patterns rather than others ought to be revisited, as phonological explanations for every alternation, at least those involving consonants, can easily be provided by following a few commonly used principles without having to appeal to morphological processes. It is not however clear yet whether all of the patterns described can indeed be called "regular", especially considering the very low number of verbs (as low as only one) following some of them. Further work is also required to fill certain gaps that have not been filled or have been brought to the light, such as whether certain phenomena are of phonetic or phonological nature and whether certain patterns can be grouped under a bigger one with the whole language as its domain, rather than only verb paradigms.

In conclusion, what tradition maintains may not always be the full picture, and different points of view must be taken, perhaps taking a step back, in order to better see the full scope of the matter.

6. Acknowledgements

I want to thank my supervisor, Cesko Voeten, for the patience and the help when guiding me through the process, as well as Silke Hamann for the useful comments and Marijn van 't Veer for helping me focus on this topic. I would also like to thank my family and my friends for the help and support offered while writing this thesis and during the years that brought me here.

7. References

Albright, Adam & Yoonjung Kang. 2008. Predicting innovative variants in Korean verb paradigms. *18th International Congress of Linguists (CIL 18), Seoul, Korea 2008*.

Boersma, Paul. 2009. Cue constraints and their interactions in phonological perception and production. in Paul Boersma & Silke Hamann (eds.), *Phonology in perception*. 55–110. Berlin & Boston (MA): Mouton de Gruyter. doi:10.1515/9783110219234.55

Cho, Taehong, Sun-Ah Jun & Peter Ladefoged. 2002. Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics* 30(2). 193–228. doi:10.1006/jpho.2001.0153

Cho, Taehong & Sahyang Kim. 2009. Statistical patterns in consonant cluster simplification in Seoul Korean: Within-dialect interspeaker and intraspeaker variation. *Phonetics and Speech Sciences* / 말소리와 음성과학 1(1). 33–40.

Choe, Ho-chol [최호철]. 2021. 사고와 표현 사이의 불일치 문제: "-귀어"의 준말, "같잖다"의 어근, "주말"의 의미, 수의 자릿점을 대상으로 [A problem of the discordance between thought and expression: Focused on the abbreviation of "-wieo", the root of "gatjanhda", the meaning of "weekend", the digit comma]. *Journal of Korean Culture* 53. 7–55. [in Korean] doi:10.35821/jkc.2021.05.53.7

Enguehard, Guillaume & Noam Faust. 2018. Guttural ghosts in Modern Hebrew. *Linguistic Inquiry* 49(4). 685–721. doi:10.1162/ling_a_00287

Gabriel, Christoph & Trudel Meisenburg. 2009. Silent onsets? An optimality-theoretic approach to French h aspiré words. in Frank Kügler, Caroline Féry & Ruben van de Vijver (eds.), *Variation and gradience in phonetics and phonology*. 163–184. Berlin & New York (NY): De Gruyter Mouton. doi:10.1515/9783110219326.163

Gussmann, Edmund. 2007. *The phonology of Polish* (The Phonology of the World's Languages). Oxford & New York (NY): Oxford University Press. doi:10.1093/oso/9780199267477.001.0001

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2025. *Glottolog* 5.2, https://glottolog.org. Leipzig: Max Planck Institute for Evolutionary Anthropology. (last accessed on 2025-06-27) doi:10.5281/zenodo.14006617

Hristovsky, Gueorgui. 2021. Vowel-zero alternation in root-final obstruent–sonorant clusters in Modern Bulgarian: Is [ə] an epenthetic vowel or a vocalised yer?. *Slavonica* 26(2). 164–173. doi:10.1080/13617427.2021.1996161

Jun, Jongho. 2010. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics* 19(2). 137–179. doi:10.1007/s10831-010-9058-8

Jun, Jongho & Adam Albright. 2017. Speakers' knowledge of alternations is asymmetrical: Evidence from Seoul Korean verb paradigms. *Journal of Linguistics* 53(3). 567–611. doi:10.1017/S0022226716000293

Kang, Hijo. 2012. *Diachrony in synchrony: Korean vowel harmony in verbal conjugation*. Stony Brook (NY): Stony Brook University (SBU) PhD dissertation.

Kang, Yoonjung & Sungwoo Han. 2013. Tonogenesis in early Contemporary Seoul Korean: A longitudinal case study. *Lingua* 134. 62–74. doi:10.1016/j.lingua.2013.06.002

Kang, Yoonjung, Jessamyn L. Schertz & Sungwoo Han. 2015. Vowels of Korean dialects. *Journal of the Acoustical Society of America* 137(4_Supplement): 2414. doi:10.1121/1.4920798

Kang, Yoonjung, Tae-Jin Yoon & Sungwoo Han. 2015. Frequency effects on the vowel length contrast merger in Seoul Korean. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 6(3–4). 469–503. doi:10.1515/lp-2015-0014

Kim, Hyunsoon. 1999. The place of articulation of Korean affricates revisited. *Journal of East Asian Linguistics* 8. 313–347. doi:10.1023/A:1008331010694

Kim, Hyunsoon. 2001. The place of articulation of the Korean plain affricate in intervocalic position: An articulatory and acoustic study. *Journal of the International Phonetic Association* 31(2). 229–257. doi:10.1017/S0025100301002055

Kim, Ji Yea. 2022. Variation in stem-final consonant clusters in Korean nominal inflection. *Glossa* 7(1). 1–45. doi:10.16995/glossa.5784

Kim, Young Shin. 2011. *An acoustic, aerodynamic and perceptual investigation of word-initial denasalization in Korean*. London: University College London (UCL) PhD dissertation. (available at https://discovery.ucl.ac.uk/id/eprint/1324541)

Kim-Renaud, Young-Key. 1973. "Irregular" verbs in Korean revisited. *Language Research* / 어학연구 9(2). 206–225. hdl:10371/85492

Kim-Renaud, Young-Key. 1974. *Korean consonantal phonology*. Honolulu (HI): University of Hawai'i PhD dissertation.

Kim-Renaud, Young-Key. 1977. Syllable-boundary phenomena in Korean. *Korean Studies* 1. 243–273. doi:10.1353/ks.1977.0000

Kim-Renaud, Young-Key. 2022. Phonology: An overview. in Sungdai Cho & John Whitman (eds.), *The Cambridge handbook of Korean linguistics*. 17–62. Cambridge & New York (NY): Cambridge University Press. doi:10.1017/9781108292351.003

Kong, Eun Jong, Soyoung Kang & Misun Seo. 2014. Gender difference in the affricate productions of young Seoul Korean speakers. *Journal of the Acoustical Society of America* 136(4). EL329–EL335. doi:10.1121/1.4896420

Krämer, Martin. 2009. *The phonology of Italian* (The Phonology of the World's Languages). Oxford & New York (NY): Oxford University Press. doi:10.1093/oso/9780199290796.001.0001

Kwon, Nahyun. 2018. Iconicity correlated with vowel harmony in Korean ideophones. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9(1). 1–18. doi:10.5334/labphon.53

Larsen, Darrell & Jeffrey Heinz. 2012. Neutral vowels in sound-symbolic vowel harmony in Korean. *Phonology* 29(3). 433–464. doi:10.1017/S095267571200022X

Lee, Hansol H. B. 1989. *Korean grammar*. Oxford & New York (NY): Oxford University Press.

Lee, Hyunjung & Allard Jongman. 2016. A diachronic investigation of the vowels and fricatives in Korean: An acoustic comparison of the Seoul and South Kyungsang dialects. *Journal of the International Phonetic Association* 46(2). 157–184. doi:10.1017/S0025100316000013

Lee, Iksop & S. Robert Ramsey. 2000. *The Korean language*. Albany (NY): State University of New York Press.

Lee, Ki-Moon & S. Robert Ramsey. 2011. *A history of the Korean language*. Cambridge: Cambridge University Press.

Lee, Sechang. 2019. A study of common behavior in /s/-irregular and /h/-final stems of Korean. *Linguistic Research* / 언어연구 36(2). 263–288. doi:10.17250/khisli.36.2.201906.005

Martin, Samuel E. 1992. *A reference grammar of Korean: A complete guide to the grammar and history of the Korean language*. Rutland (VT) & Tokyo: C. E. Tuttle.

National Institute of the Korean Language [국립국어원]. n.d. 표준국어대사전 [Standard Korean Language Dictionary], https://stdict.korean.go.kr [in Korean] (retrieved on 2025-05-11)

Rubach, Jerzy. 2016. Polish yers: Representation and analysis. *Journal of Linguistics* 52(2). 421–466. doi:10.1017/S0022226716000013

Sezer, Engin. 1986. An autosegmental analysis of compensatory lengthening in Turkish. in Leo Wetzels & Engin Sezer (eds.), *Studies in compensatory lengthening*. 227–250. Dordrecht & Riverton (NJ): Foris Publications. doi:10.1515/9783110821666-011

Sohn, Ho-min. 1999. *The Korean language*. Cambridge: Cambridge University Press.

Silva, David J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23(2). 287–308. doi:10.1017/S0952675706000911

Underhill, Robert. 1976. The case for an abstract segment in Greenlandic. *International Journal of American Linguistics* 42(4). 349–358. doi:10.1086/465439

Watkins, Michaela. 2022. *Production of Korean stops: A corpus analysis of the lenis–aspirated stop distinction in Standard Seoul Korean*. Amsterdam: University of Amsterdam (UvA) rMA thesis. (available on https://www.fon.hum.uva.nl/archive)

Yoo, Kayeon & Francis Nolan. 2020. Sampling the progression of domain-initial denasalization in Seoul Korean. *Laboratory Phonology* 11(1): 22. doi:10.5334/labphon.203

Appendix A. Lexical items in the examples by language and spelling

This appendix lists all the lexical items used throughout this thesis as examples. They are given in their common spelling accompanied by some phonological information and a brief gloss or translation.

English

bit	/bɪt/	
fit	/fɪt/	
hit	/hɪt/	
kit	/kɪt/	
knit	/nɪt/	
lit	/lɪt/	
pit	/pɪt/	
sit	/sɪt/	
wit	/wɪt/	

French

abeille	/abɛj/	bee
ami,-s	/ami/	friend
fille	/fij/	girl
garçon,-s	/gaʁsɔ̃/	boy
hausse	/os/	rise
Hongrois	/ɔ̃gʁwa/	Hungarian
l'	/l/	the (singular before a vowel)
la	/la/	the (feminine singular)
les	/le/~/lez/	the (plural)

Italian

armare	/armare/	arm
dis-	/dis/~/diz/	dis-
disarmare	/disarmare/	disarm
disgelo	/dizd͡ʒɛlo/	defrost
dislivello	/dizlivɛllo/	difference in hight
dismisura	/dizmisura/	measurelessness
disporre	/disporre/	arrange
gelo	/d͡ʒɛlo/	frost
livello	/livɛllo/	level
misura	/misura/	measure
porre	/porre/	put
presente	/presɛnte/	present
presentire	/presentire/	hear or feel beforehand

Korean

++	|+ʔ+|	*compounding morpheme* (also +ㅅ+)
가다,가	|ka+|	go
갈다·아	|kal+|	rub
같다·아	|kat^{h}+|	same, resemble, look like
걷다·어	|kʌt+|	roll up
걷다,걸어	|kʌd+|	walk
걸다·어	|kʌl+|	hang
것	|kʌs|	thing
+겠+	|+kes'+|	*future*, will ...
+고	|+ko|	... and
곧다·아	|kot+|	straight
곱다·아	|kop+|	numb
굳다·어	|kut+|	harden, solidify, stiffen
궂다·어	|kuc+|	bad (*weather*)
+ㄴ	|+n|	*past participle*, that ...ed (also +은)
나	|na|	I, me
낳다·아	|nah+|	give birth
넓다	|nʌlp+|	wide
노르다·러	|nolɯl+|	yellow (also 누르다)
누르다·러	|nulɯl+|	yellow (also 노르다)
+니까	|+nik'a|	because ... (also +으니까)
+다	|+ta|	*dictionary form*, TA *form*
닫다·아	|tat+|	shut, close
닫다,달아	|tad+|	run
달다·아(1)	|tal+|	hang
달다·아(2)	|tal+|	sweet

닳다·아	|talh+|	wear down
담다·아	|tam?+|	put into
+도	|+to|	also, ... too
+ㄹ	|+l?|	*future participle*, that will ... (also +을)
+를	|+lɯl|	*accusative* (also +을)
만들다·어	|mantɯl+|	make
맞다·아	|mac+|	correct
맡다·아	|mat^{h}+|	take on, undertake (*a task*)
맵다,매워	|meb+|	spicy
먹	|mʌk|	inkstick
먹다·어	|mʌk+|	eat
멀다·어	|mʌl+|	far
메우다·메워	|meu+|	fill up
묵다·어	|muk+|	lodge, stay
묶다·어	|muk'+|	tie (*a knot*), tie up, bind
묻다,물어	|mud+|	ask
박	|pak|	gourd
박다·아	|pak+|	insert
밖	|pak'|	outside
받다·아	|pat+|	receive, get
밟다·아	|palp+|	step
밤	|pam|	night
벋다·어	|pʌt+|	stretch, extend
벗다·어	|pʌs+|	take off (*clothes*)
별	|pjʌl|	star
별밤	|pjʌl+pam|	starry night
볶다·아	|pok'+|	stir-fry
봄	|pom|	spring
봄밤	|pom+?+pam|	spring night
부엌	|puʌk^{h}|	kitchen
빗다·어	|pis+|	comb
빚다·어	|pic+|	brew
+ㅅ+	|+?+|	*compounding morpheme* (also ++)
식다·어	|sik+|	cool down
신다·어	|sin?+|	wear (on the foot)
싫다·어	|silh+|	unpleasant
심다·어	|sim?+|	plant
싶다·어	|sip^{h}+|	want (*to do*)
+아	|+a| (|+A|)	A *form* (also +어, + ㅏ, + ㅓ)
+았+	|+as'+| (|+As'+|)	*past*, ...ed (also +었+, +ㅆ+, +ㅆ+)
안다·아	|an?+|	hug
앉다·아	|anc+|	sit
않다·아	|anh+|	do not
알다·아	|al+|	know
+어	|+ʌ| (|+A|)	A *form* (also +아, + ㅏ, + ㅓ)
+었+	|+ʌs'+| (|+As'+|)	*past*, ...ed (also +았+, +ㅆ+, +ㅆ+)
업다·어	|ʌp+|	carry
없다·어	|ʌps'+|	inexistent, not to have
엎다·어	|ʌp^{h}+|	turn over
+에	|+e|	*locative*, in/on/at/to ...
+으니까	|+ɯnik'a|	because ... (also +니까)
+은	|+ɯn|	*past participle*, that ...ed (also +ㄴ)
+을(1)	|+ɯl|	*accusative* (also +를)
+을(2)	|+ɯl?|	*future participle*, that will ... (also +ㄹ)
읊다	|ɯlp^{h}+|	recite (*poetry*)
이르다·러	|ilɯl+|	arrive, reach
일다·어	|il+|	happen
읽다·어	|ilk+|	read
잃다·어	|ilh+|	lose
입다·어	|ip+|	wear
잇다,이어	|iz+|	link
있다·어	|is'+|	existent, have
잊다·어	|ic+|	forget
+자	|+ca|	let's ...
작다·아	|cak+|	small, little
잡다·아	|cap+|	hold
잦다·아	|cac+|	frequent, occurring often
재미	|cemi|	fun
재미있다·어	|cemi+is'+|	entertaining, fun (also 재밌다)
재밌다·어	|cem+is'+|	entertaining, fun (also 재미있다)
젊다·어	|cʌlm?+|	young
좋다·아	|coh+|	good
주다;어/줘	|cu+|	give
짓다,지어	|ciz+|	make
짖다·어	|cic+|	bark
짙다·어	|cit^{h}+|	dense, thick, deep, dark
쫓다·아	|c'oc+|	tie (*a topknot*)
쫓다·아	|c'oc^{h}+|	go after, follow, pursue

찧다	|c'ih+|	pound, hit
치르다,치러	|c^{h}ilɯ+|	follow
파르다·러	|p^{h}alɯl+|	blue (also 푸르다) (only compounds)
푸르다·러	|p^{h}ulɯl+|	blue (also 파르다)
+하고	|+hako|	with ...
+한테	|+hant^{h}e|	to/for ...
핥다·아	|halt^{h}+|	lick
+ ㅏ	|+a| (|+A|)	A *form* (also +아, +어, + ㅓ)
+ㅆ+	|+as'+| (|+As'+|)	past, ...ed (also +았+, +었+, +ㅆ+)
+ ㅓ	|+ʌ| (|+A|)	A *form* (also +아, +어, + ㅏ)
+ㅆ+	|+ʌs'+| (|+As'+|)	*past*, ...ed (also +았, +었+, +ㅆ+)

Polish

łapa	/wapa/	paw
łapka	/wapka/	paw (diminutive)
żaba	/ʒaba/	frog
żabka	/ʒapka/	frog (diminutive)

Appendix B. Code

This appendix contains the code that was used to extract the data from the Standard Korean Language Dictionary (표준국어대사전).

```
############## NOTES #############
# ::: the following changes were made to the original content of the files before running anything :::
### - as of 2025-05-11, some verbs have typos in the "활용" (conjugation) column that must be manually
###    corrected beforehand or conjugation information that doesn't conform to that of the other verbs
###    and that must be normalized manually; this ensures that the code runs smoothly; these verbs are
###    the following (given in the form found under "어휘" (lexicon)):
###    - A-form not the first one: 끌어-모으다, 뒤-바르다, 신고-스럽다, 쌍-스럽다, 어르다(01), 증산-하다(01),
###       짜디-짜다
###    - expected long form missing: 윤-내다
###    - expected short form missing: 깨다(04), 답-쌔다, 드러-쌔다, 쌔다(02), 째다(03), 채다(02), 패다(05),
###       폐다, 헛방-놓다, 헛불-놓다, 휩쌔다
###    - other typos: 깊어-지다, 나-굴다, 되잡-히다, 부탁-하다, 육보-하다, 증산-하다(01), 추출-되다, 토옥-하다
# ::: further notes :::
### - the focus of this piece of code is consonant-final verbs, but it also works for vowel-final ones
### - vowel harmony is disregarded
### - this code relies on orthographic information, as there is more of it than the phonetic one
#################################

### import and select what's needed
import pandas as pd
import re

### import the files
data = <your_files>

### this chunk of code was adapted from the Transliter package files ko_jamo.py and transliter_ko.py
### (from https://github.com/elibooklover/Transliter/tree/main/transliter)
BASE_CODE, CHOSUNG, JUNGSUNG = 44032, 588, 28
CHOSUNG_LIST = ["ㄱ", "ㄲ", "ㄴ", "ㄷ", "ㄸ", "ㄹ", "ㅁ", "ㅂ", "ㅃ", "ㅅ", "ㅆ", "ㅇ", "ㅈ", "ㅉ", "ㅊ",
                "ㅋ", "ㅌ", "ㅍ", "ㅎ"]
JUNGSUNG_LIST = ["ㅏ", "ㅐ", "ㅑ", "ㅒ", "ㅓ", "ㅔ", "ㅕ", "ㅖ", "ㅗ", "ㅘ", "ㅙ", "ㅚ", "ㅛ", "ㅜ", "ㅝ",
                 "ㅞ", "ㅟ", "ㅠ", "ㅡ", "ㅢ", "ㅣ"]
JONGSUNG_LIST = [" ", "ㄱ", "ㄲ", "ㄳ", "ㄴ", "ㄵ", "ㄶ", "ㄷ", "ㄹ", "ㄺ", "ㄻ", "ㄼ", "ㄽ", "ㄾ", "ㄿ",
                 "ㅀ", "ㅁ", "ㅂ", "ㅄ", "ㅅ", "ㅆ", "ㅇ", "ㅈ", "ㅊ", "ㅋ", "ㅌ", "ㅍ", "ㅎ"]
CHOSUNG_LIST2 = [" ", "ㄱ", "ㄲ", "ㄴ", "ㄷ", "ㄸ", "ㄹ", "ㅁ", "ㅂ", "ㅃ", "ㅅ", "ㅆ", "ㅇ", "ㅈ", "ㅉ",
                 "ㅊ", "ㅋ", "ㅌ", "ㅍ", "ㅎ"]
JONGSUNG_LIST2 = ["ㄱ", "ㄲ", "ㄳ", "ㄴ", "ㄵ", "ㄶ", "ㄷ", "ㄹ", "ㄺ", "ㄻ", "ㄼ", "ㄽ", "ㄾ", "ㄿ", "ㅀ",
                  "ㅁ", "ㅂ", "ㅄ", "ㅅ", "ㅆ", "ㅇ", "ㅈ", "ㅊ", "ㅋ", "ㅌ", "ㅍ", "ㅎ"]
KOREAN_ALPHABET = {"ㄱ": "g", "ㄴ": "n", "ㄷ": "d", "ㄹ": "l", "ㅁ": "m", "ㅂ": "b", "ㅅ": "s",
                   "ㅇ": "q", "ㅈ": "j", "ㅊ": "c", "ㅋ": "k", "ㅌ": "t", "ㅍ": "p", "ㅎ": "h",
                   "ㄲ": "k", "ㄸ": "t", "ㅃ": "p", "ㅆ": "ş", "ㅉ": "ć", "ㅏ": "a", "ㅑ": "ya",
                   "ㅓ": "ʌ", "ㅕ": "yʌ", "ㅗ": "o", "ㅛ": "yo", "ㅜ": "u", "ㅠ": "yu", "ㅡ": "ɯ",
                   "ㅣ": "i", "ㅐ": "e", "ㅒ": "ye", "ㅔ": "e", "ㅖ": "ye", "ㅘ": "wa", "ㅙ": "we",
                   "ㅚ": "we", "ㅝ": "wʌ", "ㅞ": "we", "ㅟ": "wi", "ㅢ": "i"}
JONGSUNG_ALPHABET = {"ㄱ": "g", "ㄲ": "k", "ㄳ": "gs", "ㄴ": "n", "ㄵ": "nj", "ㄶ": "nh", "ㄷ": "d",
                     "ㄹ": "l", "ㄺ": "lg", "ㄻ": "lm", "ㄼ": "lb", "ㄽ": "ls", "ㄾ": "lt", "ㄿ": "lp",
                     "ㅀ": "lh", "ㅁ": "m", "ㅂ": "b", "ㅄ": "bs", "ㅅ": "s", "ㅆ": "ş", "ㅇ": "ŋ",
                     "ㅈ": "j", "ㅊ": "c", "ㅋ": "k", "ㅌ": "t", "ㅍ": "p", "ㅎ": "h"}
```

```
def convert(Ko_words):
    split_word_list = list(Ko_words)
    jamo_result = []
    for keyword in split_word_list:
        if re.match(".*[가-힣]+.*", keyword):
            char_code = ord(keyword) - BASE_CODE
            char1 = int(char_code / CHOSUNG)
            char2 = int((char_code - (CHOSUNG * char1)) / JUNGSUNG)
            jamo_result.append(CHOSUNG_LIST[char1])
            jamo_result.append(JUNGSUNG_LIST[char2])
            char3 = int((char_code - (CHOSUNG * char1) - (JUNGSUNG * char2)))
            if char3==0:
                jamo_result.append("")
            else:
                jamo_result.append(JONGSUNG_LIST[char3])
        else:
            jamo_result.append(keyword)
    return "".join(jamo_result)
def ko(text):
    text = convert(text)
    letter = list(text)
    transliterated_letters = []
    for i in range(len(letter)-1):
        transliterated_letter = ""
        try:
            if letter[i] in JONGSUNG_LIST2 and letter[i+1] in CHOSUNG_LIST2:
                transliterated_letter += JONGSUNG_ALPHABET[letter[i]]
            elif letter[i] in CHOSUNG_LIST or JUNGSUNG_LIST:
                transliterated_letter += KOREAN_ALPHABET[letter[i]]
            else:
                transliterated_letter += letter[i]
            transliterated_letters.append(transliterated_letter)
        except KeyError:
            transliterated_letter += letter[i]
            transliterated_letters.append(transliterated_letter)
    if len(letter) > 1:
        if letter[-1] in JONGSUNG_LIST2:
            transliterated_letters.append(JONGSUNG_ALPHABET[letter[-1]])
        elif letter[-1] in JUNGSUNG_LIST:
            transliterated_letters.append(KOREAN_ALPHABET[letter[-1]])
        else:
            transliterated_letters.append(letter[-1])
    transliterated_word = "".join(transliterated_letters)
    return transliterated_word

### isolate verbs/adjectives
data["class"] = data["품사"].str.contains("동사|형용사")
data = data[data["class"] == True]

### remove the negative copula 아니다
copula = data[(data.어휘 == "아니다")].index
data = data.drop(copula)

### remove the verbs for which the conjugation info is not given
data["활용"] = data["활용"].fillna("127")
data["conjugation"] = [0 if (x == "127") else 1 for x in data["활용"]]
data = data[data["conjugation"] == 1]
```

```
### transliterate and clean the transliteration
data["어휘0"] = data["어휘"].str.replace("-", "")
data["활용1"] = data["활용"].str.replace("-", "")
data["어휘0"] = data["어휘0"].str.replace(":", "")
data["활용1"] = data["활용1"].str.replace(":", "")
data["form0"] = data["어휘0"].apply(ko)
data["form1"] = data["활용1"].apply(ko)
data["form0"] = data["form0"].str.replace("(", "")
data["form0"] = data["form0"].str.replace(")", "")
data["form0"] = data["form0"].str.replace("0", "")
data["form0"] = data["form0"].str.replace("1", "")
data["form0"] = data["form0"].str.replace("2", "")
data["form0"] = data["form0"].str.replace("3", "")
data["form0"] = data["form0"].str.replace("4", "")
data["form0"] = data["form0"].str.replace("5", "")
data["form0"] = data["form0"].str.replace("6", "")
data["form0"] = data["form0"].str.replace("7", "")
data["form0"] = data["form0"].str.replace("8", "")
data["form0"] = data["form0"].str.replace("9", "")

### isolate the first form in the "활용" (conjugation) column, which should be the A-form
### other possible forms (mostly -(ɯ)ni) are ignored
### n.b.: correct the typos before running this piece of code
### an alternative to form1 is given in form1a, otherwise form1 is repeated
data["form1"] = data["form1"].str.split(",").str[0]
data["form1a"] = data["form1"].str.split("(").str[1]
data["form1"] = data["form1"].str.replace(")", "")
data["form1a"] = data["form1a"].str.replace(")", "")
data.form1a.fillna(data.form1, inplace=True)
data["form1"] = data["form1"].str.split("[").str[0]
data["form1a"] = data["form1a"].str.split("[").str[0]
data["form1"] = data["form1"].str.split("(").str[0]
data["form1a"] = data["form1a"].str.split("(").str[0]

### check whether form0 ends with "da" and remove those verbs where it doesn't: only 가라사대
### then extract stem0 by removing "da"
data["form0-da"] = [1 if (x[-2:] == "da") else 0 for x in data["form0"]]
data = data[data["form0-da"] == 1]
data["stem0"] = data["form0"].str[:-2]

### check whether form1·form1a ends with "ʌ/a/e" and remove those verbs where it doesn't as that means
###  there isn't enough usable data
data["form1-A"] = [1 if (x[-1:] == "a" or x[-1:] == "ʌ"
                         or x[-1:] == "e") else 0 for x in data["form1"]]
data["form1a-A"] = [1 if (x[-1:] == "a" or x[-1:] == "ʌ" or
                          x[-1:] == "e" or x[-1:] == "7") else 0 for x in data["form1a"]]
data = data[data["form1-A"] == 1]
data = data[data["form1a-A"] == 1]

### verbs now only have two (or three) forms:
### - form0: dictionary form, ends with "da" ---> stem0 (preconsonantal) is extracted by removing "da"
### - form1: full-vowel form, ends with "ʌ/a/e" ---> stem1 (prevocalic) is extracted by removing "ʌ/a"
###                                                    but sometimes it's fused
### - form1a: shorter variant of form1 that verbs can have
```

```
### some useful categories for later
consonant = ["q", "h", "ŋ", "g", "k", "ƙ", "s", "ʂ", "j", "c", "ƈ",
             "n", "d", "t", "ƭ", "l", "m", "b", "p", "ƥ"]
vowel = ["a", "ʌ", "e", "o", "u", "ɯ", "i"]
vwl = ["a", "ʌ", "e", "o", "u", "i"]
glide = ["y", "w"]
cons_n_glide = ["q", "h", "ŋ", "g", "k", "ƙ", "s", "ʂ", "j", "c", "ƈ",
                "n", "d", "t", "ƭ", "l", "m", "b", "p", "ƥ", "y", "w"]

### find the conjugation pattern of form0
data["stem0-q"] = data["stem0"].str.replace("q", "")
data["pat0"] = data["stem0-q"]
for i in range(len(data["stem0-q"])):
    x = data["stem0-q"].iloc[i]
    y = data["pat0"].iloc[i]
    if x[-1] in vowel:
        data["pat0"].iloc[i] = x[-1:]
    elif x[-1] in consonant:
        if x[-2] in vowel:
            data["pat0"].iloc[i] = x[-1:]
        elif x[-2] in consonant:
            if x[-3] in vowel:
                data["pat0"].iloc[i] = x[-2:]
            elif x[-3] in consonant:
                if x[-4] in vowel:
                    data["pat0"].iloc[i] = x[-3:]
                elif x[-4] in consonant:
                    if x[-5] in vowel:
                        data["pat0"].iloc[i] = x[-4:]
                    elif x[-5] in consonant:
                        data["pat0"] = data["pat0"].replace(y, "127")
    else:
        data["pat0"] = data["pat0"].replace(y, "127")
### there shouldn't be any instance of "127" under pat0

### find the conjugation pattern of form1
data["form1-q"] = data["form1"].str.replace("q", "")
data["form1a-q"] = data["form1a"].str.replace("q", "")
data["pat1"] = data["form1-q"]
data["pat1a"] = data["form1a-q"]
for i in range(len(data["form1-q"])):
    w = data["stem0-q"].iloc[i]
    x = data["form1-q"].iloc[i]
    y = data["pat1"].iloc[i]
    z = data["pat0"].iloc[i]
    if x[-1] == "e":
        data["pat1"] = data["pat1"].replace(y, "e")
    elif (x[-1] == "a" or x[-1] == "ʌ"):
        if x[-2] in vowel:
            data["pat1"] = data["pat1"].replace(y, "q")
        elif x[-2] in cons_n_glide:
            if z in vwl:
                data["pat1"].iloc[i] = data["pat0"].iloc[i]
            elif z == "ɯ":
                if x[-2] == "l":
                    if x[-3] in vowel:
                        if x[:-2] == w:
                            data["pat1"].iloc[i] = x[-2]
                        elif x[:-2] == w[:-2]:
```

```
                        data["pat1"].iloc[i] = data["pat0"].iloc[i]
                    else:
                        data["pat1"] = data["pat1"].replace(y, "127")
                elif x[-3] in consonant:
                    if x[-4] in vowel:
                        data["pat1"].iloc[i] = x[-3:-1]
                    elif x[-4] in consonant:
                        if x[-5] in vowel:
                            data["pat1"].iloc[i] = x[-4:-1]
                        else:
                            data["pat1"] = data["pat1"].replace(y, "127")
                    else:
                        data["pat1"] = data["pat1"].replace(y, "127")
                else:
                    data["pat1"] = data["pat1"].replace(y, "127")
            else:
                data["pat1"].iloc[i] = data["pat0"].iloc[i]
        elif z[-1] in consonant:
            if x[-2] in glide:
                data["pat1"].iloc[i] = x[-2]
            elif x[-2] in consonant:
                if x[-3] in vowel:
                    data["pat1"].iloc[i] = x[-2]
                elif x[-3] in consonant:
                    if x[-4] in vowel:
                        data["pat1"].iloc[i] = x[-3:-1]
                    elif x[-4] in consonant:
                        if x[-5] in vowel:
                            data["pat1"].iloc[i] = x[-4:-1]
                        else:
                            data["pat1"] = data["pat1"].replace(y, "127")
                    else:
                        data["pat1"] = data["pat1"].replace(y, "127")
                else:
                    data["pat1"] = data["pat1"].replace(y, "127")
            else:
                data["pat1"] = data["pat1"].replace(y, "127")
        else:
            data["pat1"] = data["pat1"].replace(y, "127")
    else:
        data["pat1"] = data["pat1"].replace(y, "127")
### there shouldn't be any instance of "127" under pat1
### find the conjugation pattern of form1a
for i in range(len(data["form1a-q"])):
    w = data["stem0-q"].iloc[i]
    x = data["form1a-q"].iloc[i]
    y = data["pat1a"].iloc[i]
    z = data["pat0"].iloc[i]
    if x[-1] == "e":
        data["pat1a"] = data["pat1a"].replace(y, "e")
    elif (x[-1] == "a" or x[-1] == "ʌ"):
        if x[-2] in vowel:
            data["pat1a"] = data["pat1a"].replace(y, "q")
        elif x[-2] in cons_n_glide:
            if z in vwl:
                data["pat1a"].iloc[i] = data["pat0"].iloc[i]
            elif z == "w":
```

```
                if x[-2] == "l":
                    if x[-3] in vowel:
                        if x[:-2] == w:
                            data["pat1a"].iloc[i] = x[-2]
                        elif x[:-2] == w[:-2]:
                            data["pat1a"].iloc[i] = data["pat0"].iloc[i]
                        else:
                            data["pat1a"] = data["pat1a"].replace(y, "127")
                    elif x[-3] in consonant:
                        if x[-4] in vowel:
                            data["pat1a"].iloc[i] = x[-3:-1]
                        elif x[-4] in consonant:
                            if x[-5] in vowel:
                                data["pat1a"].iloc[i] = x[-4:-1]
                            else:
                                data["pat1a"] = data["pat1a"].replace(y, "127")
                        else:
                            data["pat1a"] = data["pat1a"].replace(y, "127")
                    else:
                        data["pat1a"] = data["pat1a"].replace(y, "127")
                else:
                    data["pat1a"].iloc[i] = data["pat0"].iloc[i]
            elif z[-1] in consonant:
                if x[-2] in glide:
                    data["pat1a"].iloc[i] = x[-2]
                elif x[-2] in consonant:
                    if x[-3] in vowel:
                        data["pat1a"].iloc[i] = x[-2]
                    elif x[-3] in consonant:
                        if x[-4] in vowel:
                            data["pat1a"].iloc[i] = x[-3:-1]
                        elif x[-4] in consonant:
                            if x[-5] in vowel:
                                data["pat1a"].iloc[i] = x[-4:-1]
                            else:
                                data["pat1a"] = data["pat1a"].replace(y, "127")
                        else:
                            data["pat1a"] = data["pat1a"].replace(y, "127")
                    else:
                        data["pat1a"] = data["pat1a"].replace(y, "127")
                else:
                    data["pat1a"] = data["pat1a"].replace(y, "127")
            else:
                data["pat1a"] = data["pat1a"].replace(y, "127")
        else:
            data["pat1a"] = data["pat1a"].replace(y, "127")
    else:
        data["pat1a"] = data["pat1a"].replace(y, "127")
### there shouldn't be any instance of "127" under pat1a

### create the full pattern tags
data["pattern"] = data["pat0"] + ":" + data["pat1"] + "/" + data["pat1a"]
```

```
######### don't run this chunk of code if compounds/suffixes shouldn't be filtered #########
### count the same suffix and compounds only once
data["compound"] = data["어휘"].str.contains("-")
compounds = data[data["compound"] == True]
compounds["suffixoid"] = data["어휘"].str.split("-").str[1]
compounds["suffixoid"] = compounds["suffixoid"].str.replace("(", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace(")", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("0", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("1", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("2", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("3", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("4", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("5", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("6", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("7", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("8", "")
compounds["suffixoid"] = compounds["suffixoid"].str.replace("9", "")
compounds["sub_megapat"] = compounds["pattern"] + "-" + compounds["suffixoid"]
data["어휘_clean"] = data["어휘"]
data["어휘_clean"] = data["어휘_clean"].str.replace("(", "")
data["어휘_clean"] = data["어휘_clean"].str.replace(")", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("0", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("1", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("2", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("3", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("4", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("5", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("6", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("7", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("8", "")
data["어휘_clean"] = data["어휘_clean"].str.replace("9", "")
data["suffixoid"] = data["어휘_clean"].str.split("-").str[1]
data["어휘_clean"] = data["어휘_clean"].str.replace("-", "")
data["megapattern"] = data["pattern"] + "-" + data["어휘_clean"]
data["sub_megapat"] = data["pattern"] + "-" + data["suffixoid"]
suffixoids = list(set(compounds["sub_megapat"]))
amount = []
for s in set(compounds["sub_megapat"]):
    number = len(data[data["megapattern"] == s])
    amount.append(number)
dictionary_suffixes = dict(zip(suffixoids, amount))
suffixes = pd.DataFrame(dictionary_suffixes.items(), columns=["megapattern", "amount"])
suffixes = suffixes[suffixes["amount"] == 0]
suffixes["어휘_clean"] = suffixes["megapattern"].str.split("-").str[1]
suffixes["어휘"] = suffixes["어휘_clean"]
suffixes["pattern"] = suffixes["megapattern"].str.split("-").str[0]
data = pd.concat([data, suffixes], ignore_index=True)
data = data[data["compound"] == False]
###############################################################################################

### count the patterns and output results
patterns = list(set(data["pattern"]))
occurrence = []
for p in set(data["pattern"]):
    numbers = len(data[data["pattern"] == p])
    occurrence.append(numbers)
dictionary_patterns = dict(zip(patterns, occurrence))
results = pd.DataFrame(dictionary_patterns.items(), columns=["pattern", "occurrence"])
print(results)
```