

Evidence for base-driven alternation in Tgdaya Seediq

Abstract.

Standard morphophonological analysis allows URs to “cobble” together information from multiple slots of a paradigm (Kenstowicz and Kisseberth, 1977). In contrast, under the **single surface base hypothesis** (Albright 2002 *et seq.*), the input to morphophonology must be a single slot in a paradigm. In this paper, I compare the two approaches by examining verb paradigms in Tgdaya Seediq. In a corpus study of the Seediq lexicon, I find that isolation stems are much more informative than suffixed forms in predicting other slots of the paradigm. This asymmetry is argued to support the surface-base approach. Corpus results are backed up a production experiment, where speakers productively extended alternations from the isolation stem. Interestingly, speakers over-generalised certain patterns instead of matching lexical statistics. Based on these results, I propose a surface-base model for Seediq alternations.

1 Introduction: Two approaches to morphophonology

The classical approach to morphophonological analysis, laid out by Kenstowicz and Kisseberth (1977), involves setting up underlying forms (URs) which preserve as many contrastive phonological properties as possible. When all forms of a paradigm are affected by neutralization, the resulting UR must be ‘cobbled’, in the sense that it combines information from multiple slots of a paradigm. For example, in Tonkawa verbal paradigms, verb roots display extensive morphophonemic alternations as illustrated in (1). Different vowels of the verb stem surface depending on the phonological properties of its affixes. Crucially, for trisyllabic stems like the ones in (1), no surface form has all three vowels. Instead, URs must cobble together information about the first vowel from slots ‘A’ or ‘C’ of the paradigm, and information about other vowels from other slots (e.g. slot ‘D’) (Kenstowicz and Kisseberth, 1977: 33). Under this approach, the UR can only be found by looking at multiple forms of a paradigm, and will often not correspond directly to any single existing surface form.

- (1) *Verbal alternations in Tonkawa* (Hoijer, 1946: cited by Noske, 2011)

A	B	C	D		
notx -o?	we- ntox -o?	notxo -n-o?	we- ntoxo -n-o?	‘hoe’	/notoxo/
netl -o?	we- ntal -o?	netle -n-o?	we- ntale -n-o?	‘lick’	/netale/
picn -o?	we- pcen -o?	picna -n-o?	we- pcena -n-o?	‘cut’	/picena/

Albright (2002: *et seq.*) proposes an alternative approach, called the *single surface base hypothesis*, where the UR must be based on a single surface form in the paradigm. A slot in the paradigm is selected as a ‘privileged base’. This base form is constrained to be the same slot of paradigm for all lexical items of a given category, and serves as the input for morphophonology.

In the Tonkawa example, the input to morphophonology would therefore have to be one of slots A-D of the paradigm. Under this approach, the grammar will have fewer informational resources

available and be more prone to exceptions, as no allomorph can perfectly predict all three vowels of a verb stem. For example, if slot D were chosen to be the base, Tonkawa speakers would have to memorize the fact that in slots A and C of the paradigm, ‘hoe’ surfaces with the initial vowel [o], while ‘lick’ surfaces with the vowel [e]. Despite this limitation, the process of UR building is less complex and more restrictive, as there is no need to reference multiple slots of a paradigm.

In Tgdaya Seediq (henceforth Seediq), processes of vowel and word-final consonant neutralization cause all forms of a paradigm to suffer loss of contrasts, making it a good test case for comparing the two theories of morphophonology. The current study presents evidence from Seediq in support of Albright’s single surface base hypothesis. In particular, a survey of the Seediq lexicon reveals asymmetries which are predicted under a surface-base model.

This paper is organized as follows. §2 describes Seediq verb paradigms alternations, which result in a loss of contrasts in all slots of the paradigm. In §3, I conduct a survey of a Seediq corpus, and find asymmetries which are consistent with the surface-base approach. In particular, stem forms are found to predict suffixed forms with much higher accuracy than suffixed forms can predict the stem forms. This provides indirect evidence for a surface-base approach, and suggests that Seediq speakers could have designated the stem to be the base. §4 confirms the stem-suffix asymmetry using a model of morphological mappings, which quantifies the accuracy of mappings from different base forms. In §6, I find more evidence for a stem base from a production experiment; speakers were found to productively extend generalizations from the stem base, but also learned generalizations in a way that did not veridically represent lexical statistics. In §7, I propose a surface-base model for Seediq alternations, and briefly discuss how speaker’s non-veridical learning could be accounted for using a generality learning bias. §8 concludes the paper.

2 Phonological alternations in Seediq verbal paradigms

Seediq is an Austronesian (Atayalic) language spoken in Central and Eastern Taiwan. There are around 6500 Seediq people living in Nantou, where the Tgdaya dialect, the focus of the current study, is primarily spoken (Council of Indigenous People, 2020). However, the number of fluent speakers is thought to be much fewer than this, due to high rates of language attrition.

The Seediq phoneme inventory is given in (2) and (3); where the orthography that I adopt differs from standard IPA, phonetic transcription is given in brackets. Seediq verbs are almost always inflected for voice, mood, and aspect; verbal inflection can take the form of prefixes, infixes or suffixes (Holmer, 1996). These affixes are summarised in Table 1. Crucially, distributional restrictions cause there to be extensive vowel and consonant alternations between the non-suffixed and suffixed forms of a verb paradigm.

During elicitation of verb paradigms (described in §2.1), all verbs were elicited with the /su-/, /-an/, /-un/, and /-i/ affixes. Because the patterns reported in the paper were found to be consistent across affixes, all examples (unless otherwise specified) will only compare the bare stem forms (which are representative of all non-suffixed slots of the paradigm) to forms suffixed with

/-an/ ‘LOCATIVE FOCUS.PRES.’ (which are representative of all suffixed slots).

Although there is one disyllabic suffix /-ani/ ‘LOCATIVE FOCUS.IMPERATIVE’, my discussion will focus on suffixed forms that take monosyllabic suffixes. As will be discussed in the following section, forms that take disyllabic suffixes are assumed to not be relevant in the discussion of possible base slots, as they always experience more neutralization than forms that take monosyllabic suffixes.

	AGENT FOCUS	LOCATIVE FOCUS	PATIENT FOCUS	INSTRU. FOCUS
PRES	-m-/mu-	-an	-un	su-
PRET	-mun-	-n-, -an	-un-	
FUT	mu(pu)-	RED-an	RED-un	
IMP		-ani	-i	

Table 1: Inflectional morphology of Seediq

(2) *Seediq consonant inventory*

Stops	<i>p b</i>	<i>t d</i>	<i>k g</i>	<i>q</i>	<i>ʔ</i>
Fricatives		<i>s</i>	<i>x</i>		<i>h</i>
Affricates		<i>c</i>	<i>[ts]</i>		
Nasals	<i>m</i>	<i>n</i>		<i>ŋ</i>	
Approximants		<i>r</i>	<i>[ɾ]</i>	<i>y</i>	<i>[j]</i>
Laterals		<i>l</i>		<i>w</i>	

(3) *Seediq vowel inventory*

<i>i</i>	<i>u</i>
<i>e</i>	<i>o</i>
<i>a</i>	

2.1 Data collection

The alternations to be described in the rest of this section are based both on existing descriptive work by Yang (1976), as well as a corpus of 340 verbal paradigms. These 340 paradigms were drawn from (1) the Taiwan Aboriginal e-Dictionary (Council of Indigenous Peoples, 2020), and (2) fieldwork with three Seediq speakers (ages 69-78), carried out by the author in Puli Township, Nantou, Taiwan. Data was collected over the course of three weeks in July 2019. There is a high rate of language attrition in Seediq communities, such that fluent speakers are mostly above age 40, and only speakers around age 60 and above consistently use Seediq in daily conversation. As such, the speakers consulted in this study likely represent a more conservative variant of Seediq. All three consultants reported speaking Mandarin and Seediq regularly at roughly equal rates.

184 paradigms were collected from the online dictionary, and the remaining 156 paradigms were collected from native speaker consultants. Verb paradigms taken from the dictionary were confirmed with consultants, and omitted if my consultant(s) did not recognise the word, or provided conflicting inflected forms. Three forms were omitted under these criteria.

(4) *Discrepancies in dictionary and consultant responses*

	STEM	SUFFIXED	
		<i>dict.</i>	<i>consultant</i>
(a) ‘to hook’	'daquc	du'qut-an	NA
(b) ‘to increase’	'uman	'mal-an	'man-an
(c) ‘to seal/close’	'sepuy	su'puy-an	su'puw-an

2.2 Stress-driven vowel alternations

Seediq stress is always penultimate; suffixation shifts stress rightwards (Yang, 1976), giving rise to alternations such as ['bunuh~bu'**nu**han] ‘wear hat’. Crucially, stress interacts with vowel neutralization, resulting in vowel alternations between the stem and suffixed forms of the paradigm.

Pretonically, all vowel contrasts are neutralised. Onsetless pretonic vowels are deleted. This pattern was found for all 35 vowel-initial words in the data; as illustrated in (5), the stem’s initial vowel is deleted when stress shifts to the second syllable in the /an/-suffixed form. The pretonic vowel will assimilate to an adjacent stressed vowel if the two are separated by [ʔ] or [h] (see (6)); 25 verbs were found to match this description. Otherwise, vowels are reduced to [u] pretonically, as in (7). This last process of reduction to [u] is by far the most common, occurring in 276 stems. All three pretonic vowel neutralization processes are exceptionless.

(5) *Onsetless vowels delete* (35/35)

	STEM	SUFFIX	COBBLED UR	GLOSS
(a)	'awak	'wak-an	/awak/	‘lead (by a leash)’
(b)	'eyah	'yah-an	/eyah/	‘come’
(c)	'uyas	'yas-an	/uyas/	‘sing’

(6) *Vowel assimilation across [ʔ] or [h]* (25/25)

(a)	'leʔiŋ	li'ʔiŋ-an	/leʔiŋ/	‘hide (an object)’
(b)	'saʔis	si'ʔis-an	/saʔis/	‘sew’

(7) *Vowel reduction to [u]* (276/276)

(a)	'gedaŋ	gu'daŋ-an	/gedaŋ/	‘die’
(b)	'biciq	bu'ciq-an	/biciq/	‘decrease’
(c)	'barah	bu'rah-an	/barah/	‘rare’

Pretonically, vowels are also optionally deleted between nasals and stops, as in (8). Although Yang (1976) describes this process as obligatory, my consultants accepted forms with or without vowel deletion.

(8) *Optional vowel deletion between nasals and stops* (2/2)

	STEM	SUFFIXED	
(a)	qu'nedis	qun'dis-an (~qun u dis-an)	‘lengthen’
(b)	gu'natuk	gun'tuk-an (~gun u 'tuk-an)	‘peck’

Pretonic vowel neutralization always results in a loss of contrasts in the *suffixed* forms. For example, consider examples (7c-d). They are distinctive in the isolation stem, but homophonous in the suffixed form due to reduction of the stem’s initial vowel.

Post-tonically, similar but more restricted processes of vowel reduction are observed. Specifically, /e, o, u/ reduce to [u] in post-tonic closed syllables. This results in alternations where a post-tonic [u] in the stem form may surface as [e], [o] or [u] when stressed in the suffixed form. Examples of such alternations are given in (9).

(9) *Post-tonic reduction of /e,o/ to [u]*

- | | | | | |
|-----|------------------------------------|---------|------------------|-------------|
| (a) | 'rem <u>ux</u> ~ ru'm <u>ux</u> an | /remex/ | 'enter' | (u~u, n=60) |
| (b) | 'pem <u>ux</u> ~ pu'm <u>ex</u> an | /pemex/ | 'hold' | (u~e, n=36) |
| (c) | 'doʔ <u>us</u> ~ doʔ <u>os</u> -an | /doʔos/ | 'refine' (metal) | (u~o, n=3) |

In addition, with the exception of /uy/, diphthongs are prohibited in post-tonic (i.e. word-final) position. /ay/ and /aw/ are respectively monophthongized to [e] and [o] as in (10a-b), while /ey/ is monophthongized to [u] as in (10c).

(10) *Word-final monophthongization*

- | | | | | |
|-----|------------------------------------|---------|---------------------|--------------|
| (a) | 'ra <u>pe</u> ~ ru'ɲa <u>y</u> -an | /ranay/ | 'play' | (e~ay, n=7) |
| (b) | 'sino ~ su'naw-an | /sinaw/ | 'to drink (alcohol) | (o~aw, n=1) |
| (c) | 'de <u>ju</u> ~ du'ɲe <u>y</u> -an | /deɲey/ | 'to dry (food)' | (u~ey, n=12) |
| (d) | 'se <u>ku</u> ~ su'ku <u>w</u> -an | /seku/ | 'to store' | (u~u, n=13) |

Stem-final [e] only results from monophthongization of /ay/. In other words, final [e] predictably alternates with [ay] in the suffixed form, barring a small subset of irregular alternations. On the other hand, final [u] sometimes surfaces as non-alternating, as in (10d); note that in (10d), [w] is inserted to resolve vowel hiatus; hiatus resolution by either glide or glottal stop insertion is a regular and predictable process in Seediq. Additionally, as will be discussed in §2.3, stem-final consonants sometimes neutralize to [u] and [o] in the non-suffixed form. Consequently, stem-final [o] and [u] have multiple possible alternants in the suffixed form.¹

Post-tonic neutralizations result in a loss of contrasts in the isolation stem, so that it is not possible to predict how a final vowel will alternate from just the stem. For example, the stem-final vowels of (11a) and (11b) are contrastive in the suffixed form, but both reduce to [u] in the isolation stem.

(11) *Contrast neutralization in stem form due to post-tonic reduction*

- | | STEM | SUFFIXED | |
|-----|----------------|-------------------|---------|
| (a) | 'pem <u>ux</u> | pu'm <u>ex</u> an | 'hold' |
| (b) | 'rem <u>ux</u> | ru'm <u>ux</u> an | 'enter' |

¹Note that [o] has a limited distribution in Seediq; it surfaces post-tonically as the result of post-tonic neutralization, but there are very few stems that surface with phonemic stressed [o] as in (9c). In the current data, only four were found.

2.3 Final consonant alternations

In addition to the vowel neutralization processes described so far, Seediq has phonotactic constraints against word-final [p b m t d l g], motivating various processes of word-final neutralization. /p, b, m, t, d, l/ are neutralised with other consonants as outlined in (12).

(12) *Processes of final consonant alternations*

- (a) /p/, /b/, /k/ → [k]
- (b) /d/, /t/, /c/ → [c]
- (c) /m/, /ŋ/ → [ŋ]
- (d) /l/, /n/ → [n]

As a result of (12a), the final [k] of a stem could surface as [k] in the suffixed form, or alternate with either [p] or [b]. Examples of each possibility are provided in (13a-c). In (13d-j), similar examples are provided for the other final consonant alternations.

As will be discussed further in §3, rates of alternation differ depending on the identity of the final consonant. For example, stem-final [ŋ] tends not to alternate; the [ŋ]~[m] alternation is only observed in three out of 35 ŋ-final forms (13h). In contrast, stem-final [c] almost always alternates with [t] (13e).

(13) *Alternation of final /p, b, m, t, d, l/*

	ALTERNATION		STEM	SUFFIXED	
(a)	[k~k]	(n=19)	'tatak	tu'tak-an	'chop'
(b)	[k~p]	(n=6)	'patak	pu'tap-an	'cut'
(c)	[k~b]	(n=1)	'eluk	'leb-an	'close'
(d)	[c~c]	(n=1)	bu'cebac	bucu'bac-an	'slice'
(e)	[c~t]	(n=16)	'damac	du'mat-an	'for eating'
(f)	[c~d]	(n=4)	'harac	hu'rad-an	'build (a wall)'
(g)	[ŋ~ŋ]	(n=32)	'gilan	gu'lan-an	'mill (rice)'
(h)	[ŋ]~[m]	(n=3)	'talan	tu'lam-an	'run'
(i)	[n~n]	(n=3)	'durun	du'run-an	'entrust'
(j)	[n~l]	(n=19)	'dudun	du'dul-an	'lead'

Stem-final /g/ shows more complicated patterns of alternation than those discussed above. As summarised in (14), /ag/ neutralizes with [o] word-finally (14a), /eg, ug/ both neutralize to [u] (14b), and /ig/ becomes [uy] (14c). These alternations are historically a result of /g/ weakening to [w] word-finally, followed by monophthongization of the resulting diphthong (Li, 1981).

(14) *Alternation of final /g/*

	ALTERNATION		STEM	SUFFIXED	
(a)	/ag/→[o]	(n=9)	'hilo	hu'lag-a	'cover with blanket'
(b)	/eg, ug/→[u]	(n=9)	'lihu	lu'hug-an	'string together'
(c)	/ig/→[uy]	(n=3)	'baru	bu'rig-an	'buy/sell'

2.4 Alternations involving disyllabic suffixes

In forms with the disyllabic suffix /-ani/, stress shifts to the first vowel of the suffix, and all of the stem's vowels are neutralized. This is demonstrated in (15). In other words, longer suffixed forms also experience more neutralization than other suffixed forms, and are in this sense less informative. For this reason, they are assumed to not be a possible base form in Seediq morphophonology.

	STEM	STEM-an	STEM-ani	COBBLED UR	
(15) (a)	'gedaŋ	gu'daŋ-an	guduŋ-ani	/gedaŋ/	'die'
(b)	'pemux	pu'mexan		/pemex/	'hold'

2.5 Irregular alternations

In addition to the alternations discussed so far, Seediq has a subset of irregular alternations; these are irregular in the sense that (i) are not motivated by compliance to phonotactic constraints, (ii) lack generality and apply to very few lexical items. Under the cobbled UR approach, these alternations must be treated as exceptions. In the surface-base approach, they must be explained using narrowly defined rules or constraints that are not generalisable to many forms.

In the data used for this study, 29 irregularly alternating stem-suffix pairs were found; a subset of these are listed below, along with the expected suffixed form (given the observed isolation stem). The majority of these irregularities (n=11) involve unexpected alternations of the stem's final vowel, as in (16a). In some forms (n=5), stem-final open vowels are deleted in the suffixed form (16b). Examples of other less common irregular alternations are provided in (16c).

(16) *Verbs showing irregular alternations*

(a) *Irregular vowel alternations* (n=11)

STEM	SUFFIXED	GLOSS	EXPECTED SUFFIXED
'huruc	hu'ridan	'come to a stop'	(hu'rudan, hu'redan, hu'rodan)
'raguh	ru'wahan	'open'	(ru'guhan, ru'gehan, ru'gohan)
'raqic	ru'qutan	'hook'	(ru'qitan)
'patis	pu'tasan	'write'	(pu'tisan)
'pehiŋ	pe'hejan	'destroy'	(pi'hijan)

(b) *Irregular final vowel deletion* (n=5)

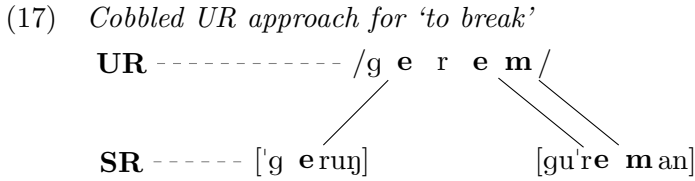
'hado	'hadan	'deliver'	(hu'dawan, hu'dagan)
'taje	'tajan	'rely on'	(tu'jayan)

	'kesa	'kesan	'tell (someone)'	(kusaʔan ²)
(c)	<i>Other irregular alternations</i>			
	'qera c	qu'ra pan	'grab'	(qu'racan, qu'ratan, qu'radan)
	'lawa	la'an-an	'invite'	(lu'waʔan)
	'bege	'biqan	'give'	(bu'gayan)

2.6 Two approaches to morphophonology in Seediq

Overall, as a result of vowel reduction and word-final neutralization, Seediq verbs undergo extensive alternations, and all forms of a Seediq verbal paradigm suffer from some form of neutralization. This complicates the task of analyzing Seediq verbal paradigms, and poses a potential challenge for Seediq learner. This is because, when given just one form of a paradigm (either a non-suffixed or suffixed form), there is no way to perfectly predict the other slots of the paradigm.

Earlier work on Seediq by Yang (1976) adopted the standard approach of the time, and resolved this issue using cobbled URs. Specifically, URs are set up by cobbling information from the non-suffixed forms (which are not affected by pretonic vowel neutralization) and the suffixed forms (which are not affected by post-tonic neutralizations). For example, consider the verb ['geruŋ]~[gu'reman] 'to break'. Given this paradigm, the learner can construct a UR /gerem/ which takes its initial vowel from the non-suffixed form, and its final vowel and consonant from the suffixed form; this is illustrated in (17).




Assuming, then, a cobbled UR approach, the majority of forms in Seediq (excluding irregularly alternating ones) can be derived using phonotactically motivated markedness constraints. Example (18) below demonstrates how this can be done in Optimality Theory (Prince and Smolensky, 1993).

A highly ranked markedness constraint *m]_w, suitably ranked against FAITHFULNESS, rules out candidates (a)-(b). Post-tonic vowel reduction to [u] is enforced by a positional licensing constraint, LICENSE(nonperipheral/stress), which limits non-peripheral vowel qualities to stressed syllables (Crosswhite, 2004). This constraint rules out candidate (c), where the vowel surfaces faithfully as [e]. Although relevant candidates are not shown here, the specific patterns of alternation observed (e.g. /m/→[ŋ]) result from the interaction of faithfulness constraints.

(18) *Derivation of ['geruŋ] under a cobbled UR approach*

²[ʔ] is inserted between identical vowels to resolve hiatus.

/gerem/	*m] _w	LIC-NONPER	ID-LAB	ID[back]
a. 'gerem	*!	*		
b. 'gerum	*!			*
c. 'gereŋ		*!	*	
 d. 'geruŋ			*	*

For now, the key point to note is that these constraints hold true across the entire Seediq lexicon, with few to no exceptions. As such, although UR discovery is more complex under the cobbled UR approach, the resulting grammar is elegant and relatively simple. Moreover, this approach makes empirically testable predictions about the range of possible alternations in Seediq.

In contrast, under the surface-base approach to UR construction, the Seediq learner would designate a surface allomorph to be the base (which is constrained to a single slot of the paradigm). In the case of Seediq, this means that the base would have to be either the non-suffixed form for all verbs, or the suffixed form for all verbs.

The resulting grammar is more complicated because the base, whether it is the suffixed or non-suffixed form, suffers from neutralization. For example, if the stem (non-suffixed) form were the base, the grammar would need to somehow ‘undo’ final consonant neutralization, which is impossible to achieve with perfect accuracy. As a result, any constraints (or rules) in the grammar will have exceptions which must be dealt with through methods like diacritics or lexical listing.

On the other hand, as noted in §1, UR discovery under the single surface base hypothesis is relatively easier. In addition, there is increasing evidence in support of the single surface-base hypothesis from various sources, including historical change in languages like Korean (Kang, 2006) and Yiddish (Albright, 2010). This historical evidence is further supported by results of wug tests (Jun, 2010) and surveys of child errors (Kang, 2006) for Korean.

Both the cobbled UR and surface-base approaches are able to account for the Seediq data (with relative strengths and weaknesses). However, we can compare the two approaches by examining their predictions about the type of mislearning (and analogical change) language learners will make. This approach, of using language learning errors to understand the language-specific grammatical structure imposed by learners, has been employed since Kiparsky (1978).

The cobbled UR approach predicts that when the learner has incomplete data (resulting in reanalysis of paradigms), the UR will be determined on the basis of whatever surface forms are available. Assuming a relatively straightforward mechanism where the language learner simply takes whatever surface form they heard to be the UR, they might posit the UR /geruŋ/ from the non-suffixed ['geruŋ], and project the suffixed form [gu'ruŋan]. On the other hand, if the learner hears the suffixed form [gu'reman], they might posit the UR /gurem/, and infer that the isolation stem is ['guruŋ]. Language learners may also utilize a more sophisticated approach, such as by forming the UR on the basis of relevant lexical frequencies (Jun, 2010). Regardless of the learner's strategy, reanalyses in both directions are plausible, and the resulting Seediq lexicon should reflect this.

The surface base approach makes markedly different predictions compared to the cobbled UR

approach with respect to how a learner behaves given incomplete data. Namely, reanalyses will always be projected from the designated base. This predicts that the resulting Seediq lexicon will have asymmetries in paradigm structure, reflecting asymmetries in reanalysis.

3 Stem-suffix asymmetries in Seediq

In this section, I show through a survey of the Seediq lexicon that although neither the stem nor suffixed forms can perfectly predict the rest of the verb paradigm, statistical tendencies in the data make it so that suffixed forms are highly predictable from stems, but the stems are not as predictable from the suffixed forms. This asymmetry supports the single surface-base approach.

This is because, under a system where speakers have selected one cell in the paradigm to be a base, verb paradigms whose other cells are poorly predicted by the base will be gradually leveled. This process acts as a feedback loop, in that reanalyses will continue to increase the informativeness of the base forms. If one cell in a paradigm is much more informative than the other, and this asymmetry cannot be attributed just to phonological neutralization processes (e.g. vowel reduction), restructuring from a single base form has likely happened.

For Seediq, the stem-suffix asymmetry suggests that speakers have designated the stem form to be the base, and that restructuring over time has exaggerated statistical tendencies which cause the stem base to be much more informative than the suffixed forms of the paradigm.

Note that, as described in §2, Seediq verbal paradigms have prefixed forms. Because the isolation stems and most prefixed forms show the same patterns of neutralization, there is no way to differentiate between stem and prefixed forms in terms of their suitability as bases. The data presented will use the ISOLATION STEM form to represent all non-suffixed slots of a paradigm, but in principle, any non-suffixed slot could be the base.

The following analysis is based on TYPE frequency. Due to lack of corpus data for Seediq, it was not possible to obtain data on token frequency. In any event, the literature suggests that when studying speakers' productive knowledge of morphophonological patterns, type frequency is the more relevant measure, and a better predictor of speakers' linguistic intuitions (Albright, 2002; Bybee, 2003; Pierrehumbert et al., 2003; Edwards et al., 2004, etc.).

3.1 Predictability from the isolation stem

The Seediq isolation stem has three potentially alternating environments: (i) post-tonic [u] in closed syllables (due to post-tonic reduction of mid vowels), (ii) [c, n, k, ŋ] in stem-final position (due to final consonant neutralization), and (iii) [o, u, e] in stem-final position (due to monophthongization and final-[g] neutralization). In this section, I will discuss how statistical regularities allow these alternations to be relatively predictable from just the isolation stem. In the interest of space, only the first two environments are discussed; for a discussion of final vowel alternations, refer to Kuo (2020).

3.1.1 Vowel matching

As described in §2.2, /e, o/ are reduced to [u] in post-tonic closed syllables. As a result, the final [u] of a CVCuC stem could surface as [e], [o], or [u] in the suffixed form. Although this alternation is not completely predictable, it turns out that the vowel which surfaces in the suffixed form is strongly correlated with the identity of the isolation stem’s stressed vowel.

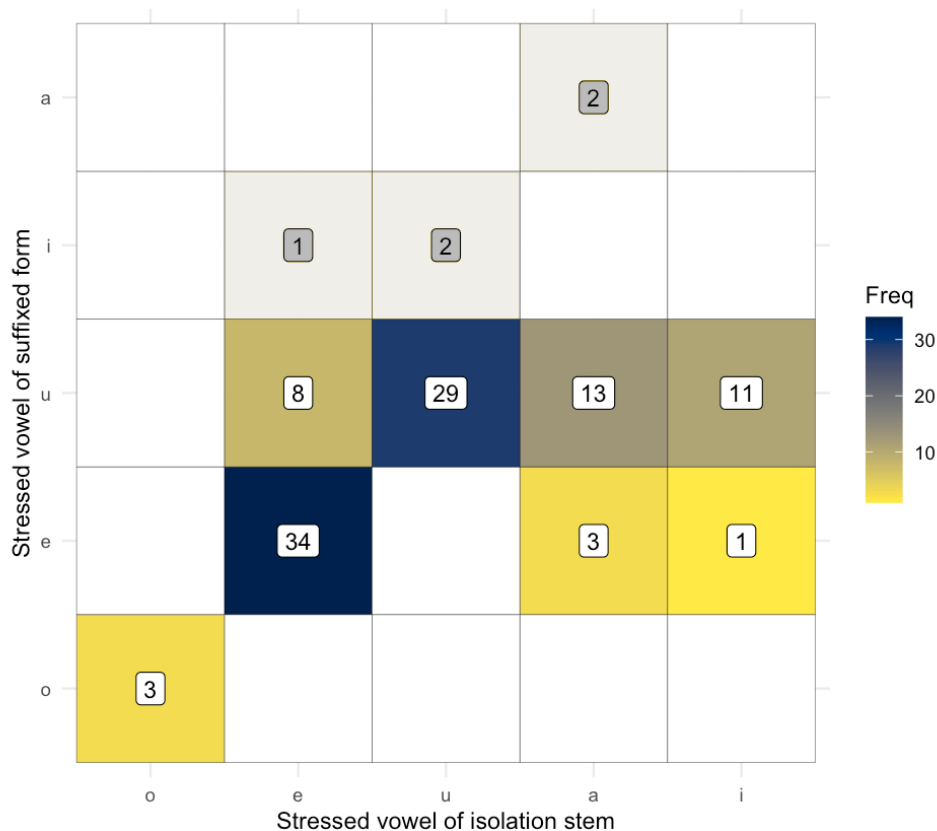


Figure 1: How the reduced [u] of non-suffixed CVCuC is realised when stressed under suffixation

Specifically, there is a tendency for VOWEL MATCHING, where the stressed vowel of the suffixed form ‘matches’ the stressed vowel of the isolation stem. This pattern is illustrated in Fig. 1, which shows the distribution of stressed vowels in CVCuC stems. Note that the data contains a few irregular alternations, where post-tonic [u] does not alternate with [u], [e] or [o]; these are shown in the top two rows of Fig. 1 (where the stressed vowel of the suffixed form is [i] or [a]).

Exceptions aside, if the stem stressed vowel is [o], the reduced [u] surfaces as [o] in the suffixed form (3/3, 100%). Similarly, if the stem stressed vowel is /u/, the reduced vowel will surface as [u] in the suffixed forms (29/31, 93%). For [e], there is similarly a strong tendency for vowel matching for around 79% of the relevant forms (34/43). Otherwise, if the stem stressed vowel is /a/ or /i/, the reduced vowel is usually non-alternating, and surfaces as [u].

This pattern was confirmed using Fisher’s exact tests; the suffixed form’s stressed vowel was significantly more likely to be [e] if the stem’s stressed vowel was [e], $p < 0.001$ ($= 9.0 \times 10^{-14}$,

odds ratio = 48). Similarly, the suffixed form’s stressed vowel was significantly more likely to be [u] if the stem’s stressed vowel was [u], $p < 0.001$ ($= 2.8 \times 10^{-8}$).

In general, for each CVCuC form, there is a clear majority pattern (depending on the identity of the stem’s stressed vowel). This makes it so that a speaker can predict, with relatively high accuracy, what a post-tonic [u] will surface as in the suffixed form. In other words, given a novel stem like ['putus], a speaker can predict that the suffixed form will be [pu'tusan]. Given a word like ['petus], the speaker could in principle predict that the suffixed form will most likely be [pu'tesan]. As a whole, the picking the majority variant in each vowel condition correctly predicts the stressed vowel of the suffixed form for 84% of CVCuC forms.

3.1.2 Final consonant alternations

Recall that word-finally, consonants /p, b, t, d, m, l/ are prohibited, resulting in the patterns of final consonant neutralization described in §2.3 and summarised in (19). As a result of these alternations, when given just the isolation stems, it is not possible to perfectly predict whether final [c, k, n, ŋ] will alternate in the suffixed form.

- (19) STEM SUFFIXED
- | | | |
|-----|---|-----------|
| [c] | ~ | [t, d, c] |
| [k] | ~ | [p, b, k] |
| [n] | ~ | [l, n] |
| [ŋ] | ~ | [m, ŋ] |

	Cons.	Alternates?	Alternant	Example	Frequency
(a)	c	Yes	t	patic ~ putitan	16 (76%)
		Yes	d	patic ~ putidan	4 (19%)
		No		patic ~ putican	1 (5%)
(b)	n	Yes	l	patin ~ putilan	18 (75%)
		No		patin ~ putinan	6 (25%)
(c)	k	Yes	p	patik ~ putipan	6 (23%)
		Yes	b	patik ~ putiban	1 (4%)
		No		patik ~ putikan	19 (73%)
(d)	ŋ	Yes	ŋ	patiŋ ~ putiman	2 (6%)
		No		patiŋ ~ putiŋan	33 (94%)

Table 4: Rates of final consonant alternation (irregularly alternating forms are excluded)

However, final consonants tend to either almost always or almost never alternate, as summarised in Table 4 and Fig. 2. Final [ŋ] almost never alternates with [m]; the hypothetical stem ['patiŋ] will surface as [pu'tiŋ-an] (i.e. with a non-alternating [ŋ]) about 94% of the time. For final [k], rates of alternation are more intermediate, but there is still a tendency towards non-alternation, with 72% of stem-final [k] surfacing faithfully as [k] in the suffixed form.

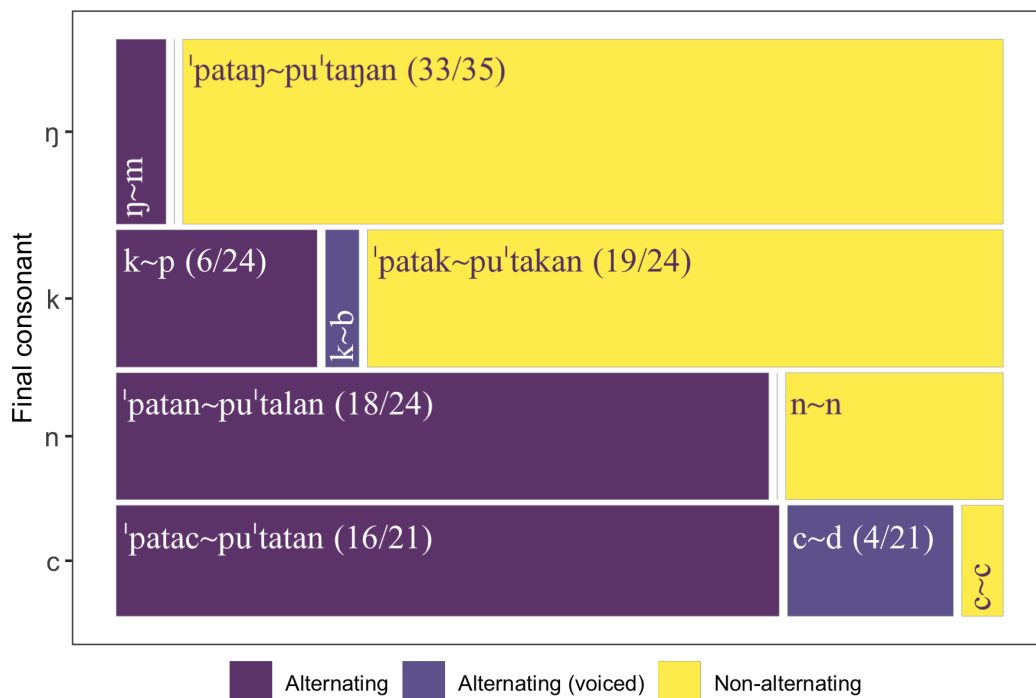


Figure 2: Rates of final consonant alternation

In contrast, final [c] and [n] show a strong preference for alternation; final [c], in particular, alternates with either [t] or [d] 95% of the time. Only a single [c]-final stem in the data was found to be non-alternating. In addition, for stem-final [c] and [k], which each have two possible alternants, there is a strong tendency to alternate with the voiceless variant (e.g. [c] alternates with [t] more than with [d]).

Because of these asymmetries in the alternation rate of final consonants, the final consonant of a suffixed form is actually highly predictable from the isolation stem. For example, most [ŋ]-final stems (around 94%) will **not** show the [ŋ]~[m] alternation. The speaker, if given a novel [ŋ]-final stem like [patiŋ], can rationally guess that the suffixed form will be [putiŋan], with a non-alternating final [ŋ]. This generalization predicts the wrong result for the small percentage of alternating forms, but still has a very high chance of being correct.

3.2 Predictability from the suffixed form

Having shown that stems can predict suffixed forms with fairly high accuracy, I now consider whether suffixed forms can also be used to predict stem forms with comparable levels of success.

Suffixed forms are not affected by the processes which resulted in contrast neutralization in the stem forms. Given a suffixed form, the identities of the final consonant and vowel are completely predictable. For example, final [m] will always be neutralised to [ŋ] in the non-suffixed stem. As

such, given a novel form [pu'tim-an], a Seediq speaker should in principle know with certainty that the isolation stem form will surface with a final [ŋ].

However, *pretonic* vowel reduction causes neutralization of contrasts in the suffixed forms; in the suffixed form, the penultimate vowel of *all* stems either reduce to [u], assimilate to the stressed vowel, or get deleted. As such, given a suffixed form like [pu'tis-an], it is impossible to perfectly predict what vowel the [u] will surface as when stressed in the stem form.

Vowel deletion further complicates the situation. Deletion of pretonic onsetless vowels affects onsetless disyllabic stems, resulting in alternations such as ['uyas~'yas-an], where the suffixed form surfaces with a monosyllabic stem. Phonemically monosyllabic stems, though rare, do exist (e.g. ['req ~ 'reqan] 'to swallow'). As a result, for a novel suffixed form like ['tis-an], where the stem appears as monosyllabic, speakers must also predict whether the non-suffixed stem is monosyllabic, or if there is an initial (onsetless) vowel.

In the case of post-tonic vowel alternations (§3.1.1), a correlation between the stressed vowels of the stem and suffixed forms made it possible to 'undo' post-tonic vowel reduction alternation with relatively high accuracy. For pre-tonic vowel reduction, however, there is less of a clear pattern of predictability in vowel distribution. This is demonstrated in Fig. 3 and 4.

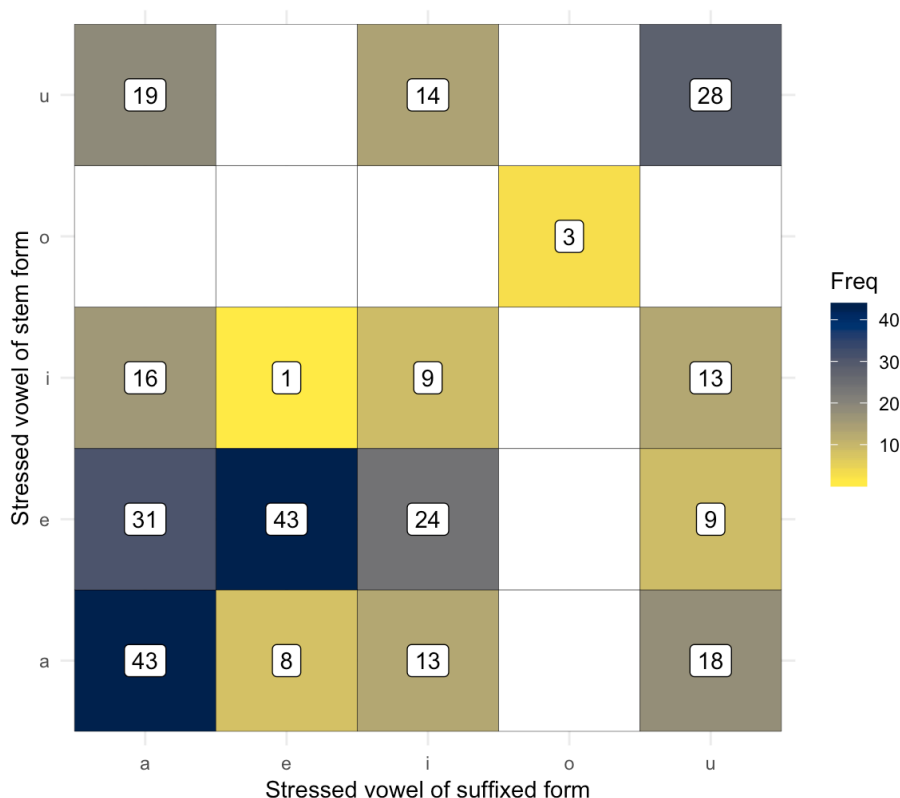


Figure 3: Distribution of stressed vowels in non-monosyllabic suffixed forms

Fig. 3 compares the stressed vowel of the suffixed form against the stressed vowel of the isolation stem in forms which surface as *disyllabic* in the suffixed form. As seen in the figure, for each stressed

vowel in the stem, there is some predictability between the stressed vowel of the stems and the stressed vowel of the suffixed forms. However, these trends appear to be relatively weak compared to the patterns observed in §3.1.1, which looked only at contexts in which post-tonic /e,o/ are reduced (i.e. stems of the form CVCuC).

To get a general measure of the ‘predictability’ of the stem stressed vowel from the suffixed form’s stressed vowel, we can look each column of Fig. 3, and select the majority variant, or cell with the largest number of forms. For example, looking at the first column of Fig. 3, the stressed vowel of the stem is most likely to be [a] if the stressed vowel of the suffixed form is also [a]. This is true for 43 out of 109 (43+31+16+19) verbs where the suffixed form’s stressed vowel is [a]. In other words, given a suffixed form [pu'tas-an], a speaker could pick the majority variant, and infer that the isolation stem form is ['patas]. This choice predicts the correct output 39% (43/109) of the time. In the third column, we see that if the suffixed form’s stressed vowel is [i], the stem form’s stressed vowel is most likely to be [e]. In other words, for suffixed forms like [pu'tisan], the stem form is most likely to be ['petis]. Applying this principle, we can correctly predict the stem stressed vowel for 39% (24/62) of relevant forms.

Table 5 summarizes the proportion of forms predicted by picking the majority variant for each column in Fig. 3 (i.e. predictability based on the suffixed form’s stressed vowel). The last column of Table 5 indicates the overall proportion of forms correctly predicted. Based on these figures, predictability from some vowels (/i/, /a/, and /u/) is fairly low. As a whole, picking the ‘best’ option based on statistical tendencies in the data only predicts the correct vowel 48% of the time.

<i>Suff vow.</i>	<i>Predicted</i>	<i>Total</i>	<i>% correct</i>
/a/	43	109	39%
/e/	43	52	83%
/i/	24	60	39%
/o/	3	3	100%
/u/	28	68	41%
Total	141	292	48%

Table 5: Predictability of stem stressed vowels from suffixed forms in disyllabic verbs

Fig. 4 shows the distribution of stems which surface as *monosyllabic* in the suffixed form. For these forms, the stem could either be monosyllabic in the isolation stem form (labeled \emptyset in Fig. 4), or surface with a stressed onsetless vowel. There is a general tendency for the stem to be disyllabic (monosyllabic stems are rare, with only four found in the current data), but once again, patterns of predictability are relatively weak.

The results so far suggest that the alternations which result from pretonic vowel neutralization are less predictable than the processes which result in loss of information in the stem form. Moreover, pretonic vowel neutralization affects more forms than other neutralization processes; 336 verbs (the entire corpus, excluding 4 monosyllabic stems) are affected. In contrast, post-tonic vowel reduction affects 107 verbs, final consonant neutralization affects 101 verbs, and final monophthongization affects 97 verbs.

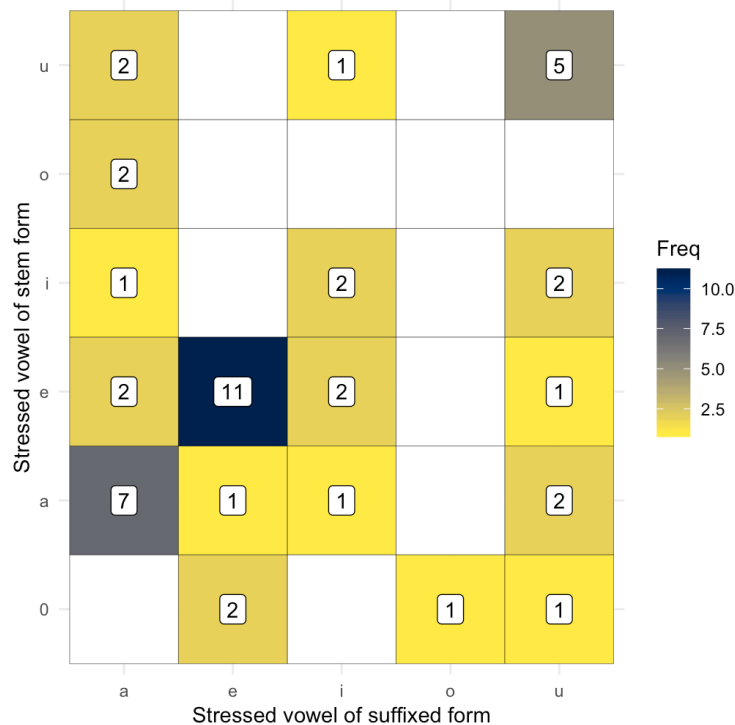


Figure 4: Realization of stressed vowels in monosyllabic suffixed forms

4 Quantifying the Seediq stem-suffix asymmetry

In the previous section, I provided a qualitative account of the stem-suffix asymmetry in Seediq verb paradigms. This section explicitly demonstrates the asymmetry using a surface-base model which learns surface mappings between inflected forms. This model I employ is from Albright (2002); Albright and Hayes (2003). It takes as its training data a set of pairs of morphologically related surface forms (in this case stem and suffixed forms), and attempts to learn the set of stochastic morphological mappings that project one from the other. The model learns grammars in both directions (stem→suffixed and vice versa), using the Minimal Generalization Learner algorithm (MGL; Albright, 2002; Albright and Hayes, 2003).

Note that the core argument of the single surface base hypothesis is that surface forms serve as the input to morphophonology. This theory of surface bases has primarily been implemented using the MGL-based model, but is in fact compatible with other theories of morphophonology. The model employed in this section is not meant to be a theoretical model, but rather a model of quantitative assessment, for testing whether the stem-suffix asymmetry exists in Seediq. In fact, as will be discussed in §6, in a wug test, Seediq speakers extended vowel copying beyond the environments predicted by a model of only morphological correspondences. This suggests that a theoretical model of Seediq alternations needs additional mechanisms.

The MGL parses each stem-suffix pair, and attempts to learn a grammar that predicts which change each form will take. It does so by comparing forms that share the same change, discovering

what phonological features they have in common, and generalizing rules based on shared features. The model is minimal in that it will retain specific rules, and only generalize more broadly defined rules when segments can be grouped using shared features. For details on the MGL, refer to Albright and Hayes (2003, p. 123-128).

For example, in learning stem→suffixed mappings, the model could compare pairs the two inputs in (20a) to learn a rule of suffixation after non-continuant dorsals. Including an additional input beras~burasan would result in the model learning a more general rule (20b). Eventually, consideration of a broader range of forms would result in a general suffixation rule $\emptyset \rightarrow \text{an}$. To learn a grammar of the reverse mapping, from suffixed to stem forms, the model repeats this same algorithm with the suffixed allomorph as input.

(20) *Examples of rules in the stem-to-suffix mapping*

	input	rule
a.	beli~bulijan, betaq~butaqan	$\emptyset \rightarrow \text{an} / [+DORSAL, -continuant]_-$
b.	beli~bulijan, betaq~butaqan, beras~burasan	$\emptyset \rightarrow \text{an} / [-continuant]_-$

4.1 Competing generalizations

In stem-suffix pairs where neutralization results in ambiguity, the most general rule learned will not correctly predict all outputs. For example, neutralization of final [p] to [k] results in stem-suffix pairs like [kayak]~[ku'yapan] ‘to cut’. The general stem-to-suffixed mapping of $\emptyset \rightarrow \text{an}$ would predict the wrong output *[kuyakan]. For such cases, the model learns minority patterns (e.g. k→pan), which exist alongside and compete with more general mapping.

Each rule is assessed for its *accuracy* (proportion of relevant forms correctly predicted by the rule). Accuracy values are then adjusted downwards using lower confidence limit statistics, such that rules with fewer data points will be penalized (Mikheev, 1997). This adjusted value, called *confidence*, better captures the fact that rules with very little evidence tend to be less reliable (Albright, 2002). Confidence determines the probability of a rule applying to each input.

The resulting grammar learned by the MGL is a system of competing rules that vary in generality, and are each assigned a confidence value. When the grammar is invoked to produce an inflected form, all applicable rules are tried, resulting in a set of output candidates, each given a confidence score. For example, as shown in (21), three rules apply to the input *birac*, resulting in three output candidates of varying confidence.

	input	output	rule	confidence
(21)	birac	buracan	$\emptyset \rightarrow \text{an} / _$	0.65
		buratan	$c \rightarrow \text{tan} / _$	0.67
		buradan	$c \rightarrow \text{dan} / _$	0.13

4.2 Model evaluation and implementation

We can assess the relative informativeness of different mappings learned by the model, to see which slot of the paradigm is on average better at predicting the other slots. The **base**, under this approach, is the slot that yields grammars which are best able to predict the lexicon.

To compare the informativeness of Seediq stem and suffixed forms, I trained the model on the 340-word corpus. The informativeness of a paradigm slot was taken to be its accuracy in predicting the other slot. The accuracy for each mapping was taken to be the proportion of forms correctly predicted by the grammar. Specifically, for each word in a mapping, the generalization with the highest reliability was taken to be the model’s prediction.

The stem-to-suffix mapping in Seediq has two largely independent processes of allomorphy: post-tonic vowel alternation and final consonant alternations. The MGL cannot straightforwardly deal with this because it assumes that there is exactly one morphological mapping for each stem-suffix pair, and cannot concurrently treat vowel and consonant alternations as independent processes.

The algorithm also assumes strictly local contexts, and is not able to learn non-local conditioning, such as the vowel matching pattern observed in mapping from stem to suffixed forms. This issue could potentially be resolved by modifying the MGL to search for non-local environments, as in Albright and Hayes (2006).

I opt for a simpler approach, and resolve the above issues by training two separate grammars for each mapping; one “segmental tier” grammar and one “vowel tier” grammar (Hayes and Wilson, 2008). A stem-suffix mapping is judged as being correctly predicted by the model only if both the main grammar and vowel tier grammar predict the correct output.

4.3 Results: comparing accuracy of different mappings

Table 22 compares the relative accuracy of the stem-to-suffix mapping and suffix-to-stem mapping; full details of model implementation and results are given in the Supplementary Materials.

It is much more accurate to project from the stem form to the suffixed form (72.7%) than vice versa (41.3%). The discrepancy between the two mappings is largely because pretonic neutralization renders the penultimate vowel of the stem unpredictable in the suffix-to-stem mapping. In fact, the vowel tier model is only 41.3% accurate in the suffix-to-stem mapping. These results back up the findings of §3, and show that the stem form is much more informative than the suffixed allomorph.

(22) *Model results: accuracy of mappings between Seediq stem and suffixed forms*

	grammar		
	segmental tier	vowel tier	overall
stem→suffix	78.2%	86.3%	72.7%
suffix→stem	90.0%	48.3%	41.3%

4.4 Predictions for reanalyses and novel forms

Table 6 shows a subset of the rules learned in the stem-base segmental tier grammar, starting with a general suffixation rule (R1). Assuming that a rule is productive if it has a higher confidence than R1, the only final consonant alternations predicted to be productive are [c]~[t] (R4) and [n]~[l] (R3). The [n]~[l] alternation is also predicted to be productive only following [i]. By design, the grammar also does not the [k]-[b] alternation, because it is only observed in one form ([^heluk]~[^hleban])

	rule	scope	hits	reliability	confidence
R1.	$\emptyset \rightarrow \text{an}$	344	230	0.67	0.65
R2.	$\text{o} \rightarrow \text{agan}$	13	11	0.85	0.75
R3.	$\text{n} \rightarrow \text{l/i_}$	8	7	0.88	0.74
R4.	$\text{c} \rightarrow \text{tan}$	23	17	0.74	0.67
R5.	$\eta \rightarrow \text{m}$	35	2	0.06	0.04
R6.	$\text{c} \rightarrow \text{dan}$	23	4	0.17	0.13
R7.	$\text{k} \rightarrow \text{pan}$	28	7	0.25	0.20

Table 6: Examples of rules learned in the segmental stem-base grammar

Overall, the model also learns the same vowel matching generalizations described in §3. Table 6 shows a subset of rules learned in the vowel tier grammar. Here, each rule predicts what the second vowel (V2) of a stem will surface as in the suffixed form. For ease of reading, rules have been schematized to more closely reflect Seediq surface forms, with the relevant vowel in bold. When V2 of the stem is /a/ or /i/, the grammar learns rules R1 and R2, where the post-tonic vowel of the stem is non-alternating, regardless of the identity of the preceding vowel.

However, when V2 of the stem is /u/, the rule predicting non-alternation (R3) has low confidence (0.44). Instead, R4 and R5, which predict vowel matching in CeCuC and CuCoC forms, are assigned higher confidence. In other words, the stem-base vowel model predicts that speakers will extend [u]~[e] and [u]~[o] alternations in CeCuC and CoCuC stems. The grammar also learns that in CVCuC stems where V1 is not [e] or [o], the post-tonic vowel surfaces as [u]. This is seen, for example, by comparing the competing rules R6 and R7, where R6 has a much higher confidence.

	rule	scope	hits	reliability	confidence
R1.	$\text{CVCaC} \rightarrow \text{CuCaCan}$	110	101	0.92	0.90
R2.	$\text{CVCiC} \rightarrow \text{CuCiCan}$	64	60	0.94	0.91
R3.	$\text{CVCuC} \rightarrow \text{CuCuCan}$	141	76	0.54	0.44
R4.	$\text{CeCuC} \rightarrow \text{CuCeCan}$	60	46	0.77	0.72
R5.	$\text{CeCuC} \rightarrow \text{CuCoCan}$	3	3	1.00	0.72
R6.	$\text{CaCuC} \rightarrow \text{CuCuCan}$	28	19	0.68	0.61
R7.	$\text{CaCuC} \rightarrow \text{CuCeCan}$	28	7	0.25	0.20

Table 7: Examples of rules learned in the vowel stem-base grammar

4.5 Interim summary

Comparison of the stem and suffixed forms revealed an asymmetry in the relative informativeness of the two paradigm slots. Specifically, the stem forms can be used to predict the suffixed forms with much higher accuracy than the other way around. This asymmetry does not have a purely phonological explanation; it isn't the case that the stem form is more informative than the suffixed form only because it undergoes fewer phonological neutralization processes. Instead, the informativeness of the stem form is in part due to the very skewed rates of alternation in neutralised segments.

This asymmetry is not predicted by a cobbled UR analysis; as discussed in §2.6, under the cobbled UR approach, reanalyses of verb paradigms can be based on all cells of the paradigm. As such, this approach makes no predictions about asymmetries between the stem and suffixed forms of Seediq verb paradigms. In contrast, under the single surface base approach, such an asymmetry is expected.

Specifically, it is possible that an older system of Seediq had a more symmetrical distribution of segments. However, as discussed in the beginning of this section, there could have been a gradual restructuring of paradigms, whereby generations of Seediq-learning children have replaced suffixed forms with new forms that obey the pattern of predictability given under the single-surface base hypothesis. The result is the new system observed in the current study, where distributions of segments are strikingly asymmetrical.

For example, statistical patterns in the modern Seediq lexicon reflect a strong dispreference for [ŋ]-[m] alternation. Historically, this dispreference may have been present as a weak statistical tendency. If Seediq speakers have designated the stem to be the base, paradigms which showed the dispreferred [ŋ]-[m] alternation would gradually have been restructured, resulting in the skewed rates of alternation that we see today. Although there is limited historical comparative data available for Seediq, I have elicited one example which suggests this type of reanalysis. As seen in (23), the verb 'to burn' is historically [m]-final (Li, 1981; Greenhill et al., 2008), and is therefore expected to show the [ŋ]-[m] alternation. Instead, the suffixed form surfaces with a non-alternating [ŋ].

- (23) 'lauŋ~lu'uŋan (<*l-um-aum) 'to burn'
(Li, 1981; Greenhill et al., 2008)

Restructuring of this variety is not unlikely; there are various documented cases where a regular pattern rendered unpredictable by historical change is either partially or completely restructured. One example involves the Oceanic 'thematic consonants'; for many Oceanic languages, the loss of word-final consonants in intransitive verbs and nouns resulted in C/∅ alternations, where a consonant of unpredictable quality surfaces in a subset of suffixed forms (Hale, 1973; Blevins, 2008). Examples from Maori are shown in (24).³

- (24) C/∅ alternations in Maori (Hale, 1973: 414)

³Note that only a subset of the possible thematic consonant are shown here, and that the shape of the passive suffix is conditioned by minimal word requirements.

STEM	PASSIVE	GLOSS
maka	maka-ia	‘throw’
awhi	awhi- t -ia	‘chase’
kimi	kimi- h -ia	‘seek’
tohu	tohu- ŋ -ia	‘point out’

For Maori, the identity of the consonant for many stems cannot be traced back to a historic stem-final consonant. Instead, there is an asymmetrically large number of stems which take a ‘default consonant’. The default consonant can be [t, h, ŋ] depending on the dialect of Maori (Blevins, 2008), and is the one used for derived verbs and loanwords (Hale, 1973: 417). Based on this, Hale argues that the consonant has been reanalyzed as belonging to the suffix /-Cia/. Moreover, although the resulting suffix has many allomorphs ([tia, hia, ŋia], etc.), and there is still a degree of lexical idiosyncrasy in which allomorph surfaces, forms are leveling in the direction of a single default consonant.

Blevins (2008) suggests that the default consonant tends to be more frequent one, as /-tia/ has the highest type frequency while /ŋia, hia/ are associated with common words. For Seediq, it is reasonable to conjecture that a parallel process happened, in which rates of alternation (or non-alternation) were exaggerated over time based on statistical tendencies already present in the lexicon. For example, it is possible that historically, [m] was already less frequent than [ŋ]; over time, reanalyses based on the stem form would have exaggerated this tendency, resulting in the current lexicon’s strong dispreference for [ŋ]-[m] alternation.

Unfortunately, there is almost no direct evidence for this type of historical re-analysis, and we cannot rule out the possible that current asymmetries in the lexicon are an artifact of historical sound distributions. In particular, the post-tonic u-e alternation in Tgdaya Seediq results from a sound change of Proto-Austronesian (PAN) *ə to [u] in the final syllable, and to [e] in other environments (Li, 1981). Modern Seediq’s tendency towards vowel matching alternation could be an artifact of historical distributions, if historically Seediq had much more CəCəC forms (than CəCuC/CəCiC/CəCaC forms).

PAN schwa has reduced to *u in the final position of *all* languages in proto-Atayalic, which encompasses both Seediq and Atayalic (Li, 1981). In addition, there are very few Seediq stems with established PAN protoforms. This makes it virtually impossible to systematically examine the degree to which reanalysis of post-tonic vowels has been driven by vowel matching. However, the argument for a stem base does not depend on there being direct historical evidence. In §6, I present the results of a productivity test, which suggest that regardless of its exact historical origins, in current Seediq, speakers productively extend vowel matching to novel suffixed forms.

5 An alternative cobbled UR analysis

So far, I have argued that the asymmetry in informativeness of stem and suffix bases supports the single surface-base approach to morphophonology. In this section, I briefly discuss the alternative

cobbled UR approach, and argue that it does not adequately account for the Seediq data.

My discussion will focus on vowel alternations, because there is vowel neutralization in both stem and suffixed forms, making it possible to compare expected reanalyses from both directions. The case for final consonant alternations is less clear; final consonants are perfectly predictable given a suffixed form, and therefore all reanalyses will be from the stem form.

The cobbled UR approach differs primarily from the surface-base approach in that reanalyses from both the stem and suffixed forms are possible. There are many possibilities for how learners construct URs when faced with uncertainty in allomorph selection, and exploration of all the possibilities is beyond the scope of the current paper.

Instead, I will discuss one approach, which is that when a language learner is faced with unpredictability, they guess the UR on the basis of relevant lexical frequencies (Jun and Albright, 2017). For instance, when a Seediq speaker hears [ˈgeruŋ], they would first posit its UR to be /gereŋ/ because Seediq stems with underlying /e/ in V1 position are most likely to have underlying /e/ in V2 position. From the UR /gereŋ/, they could then infer the suffixed form [guˈreŋan].

Reanalyses should only arise in alternating environments. This means that reanalyses from stems, on the basis of lexical frequencies, should affect URs that surface with a post-tonic [u] in the stem (/CVCuC/, /CVCeC/, /CVCoC/). These reanalyses will specifically result in a preference for (i) /CeCeC/ relative to /CeCoC/ and /CeCuC/, (ii) /CoCoC/ relative to /CoCuC/ and /CoCeC/, and (iii) /CuCuC/ relative to /CuCeC/ and /CuCoC/. Crucially, URs such as /CeCaC/, which go against the vowel matching principle, should nevertheless be observed because the corresponding stem form [CeCaC] has a non-neutralized post-tonic vowel that is not vulnerable to reanalysis.

On the other hand, reanalyses from the suffixed form affects all stem-suffix pairs, as V1 is always neutralized in the suffixed form. Reanalysis from the suffixed form on the basis of lexical frequencies is expected to increase vowel matching across more vowel contexts. Table 8 shows the most frequent V1 given a specific V2; in nearly all vowel contexts, reanalyses the suffixed forms are expected to increase vowel matching. The only exception is that when V2 is /i/, V1 is more likely to be /e/. In general, a frequency-based reanalysis of URs from suffixed forms should result in general tendencies towards vowel matching for /a/, /e/, /o/, and /u/. Overall, reanalyses from both stem and suffixed forms (as predicted by the cobbled UR approach) should collectively result in vowel matching across most vowel contexts.

V2	V1	%	UR
a	a	0.380	(CaCaC)
e	e	0.833	(CeCeC)
i	e	0.368	(CeCiC)
o	o	0.75	(CoCoC)
u	u	0.418	(CuCuC)

Table 8: Most frequent V1 given target V2; the third column gives the proportion of forms with the target V2 that have the predicted V1; column 4 gives the corresponding UR

To test whether UR vowel distributions reflect the predictions of the cobbled UR approach, I

implemented a Maximum Entropy (MaxEnt) model of vowel phonotactics in Seediq URs Hayes and Wilson (2008). MaxEnt is a stochastic implementation of Optimality Theory (Prince and Smolensky, 1993). Although rule-based approaches to morphophonology are also compatible with the cobbled UR analysis, MaxEnt is used here as a way to capture vowel distributions, because it can straightforwardly capture gradient distributional restrictions, and be used to test these restrictions for statistical significance.

In this model, GEN was based on **cobbled URs** (constructed from the corpus of 340 stem-suffix pairs). The GEN adopted was a simplified list of all possible two-vowel combinations (e.g. /ea/ represents the UR /CeCaC/). The constraint set was chosen to be relatively theory-neutral, and is meant to represent distributional facts of vowels in Seediq cobbled URs. Constraints included (i) general phonotactic constraints on each vowel in each position (e.g. the candidate /ea/ incurs violations of *e/V1 and *a/V2); (ii) MATCH-V, which penalizes all sequences of vowels where V1 and V2 are different; (iii) vowel-specific vowel matching constraints like MATCH-e, (iv) constraints on vowel-vowel sequences such as *eu and *ee, which penalize specific sequences of vowels. Following the method adopted in Hayes et al. (2012), all constraint weights were tested individually for significance with the Likelihood Ratio Test, by comparing a maximal model (with all constraints included) against one with the target constraint excluded.

As summarized in (25), all vowel matching constraints were non-significant except for MATCH-o, which is marginally significant. In contrast, (26) shows the weights learned for a subset of segment-specific constraints on vowel sequences, where V1 is either /e/ or /u/. Notably, the two which tested significant are the ones that penalize exactly the vowel sequences which would be dispreferred under a base-driven reanalysis account. The constraint *eu penalizes URs of the form /CeCuC/, which in turn correspond to stem-suffix pairs ['CeCuC]~[Cu'CuC-an], where a post-tonic [u] in the stem doesn't show the alternation predicted by vowel matching.

Similarly, *ue penalizes URs of the form /CuCeC/, corresponding to SRs ['CuCuC]~[Cu'CeC-an]; once again, this is a case where post-tonic [u] shows a dispreferred alternation. In contrast, a constraint like *ea, which penalizes URs of the form /CeCaC/, is non-significant even though the UR disobey vowel matching. /CeCaC/ corresponds to SRs ['CeCaC]~[Cu'CaC-an], where the post-tonic [a] is non-alternating, and therefore not prone to reanalysis from the stem form.

(25) *Constraints on vowel agreement in a model of Seediq URs*

Constraint	w	p
MATCH-V	0.61	0.93
MATCH-a	0	1.0
MATCH-e	0.39	1.0
MATCH-i	0.08	1.0
MATCH-o	4.85	0.046*
MATCH-u	0.13	1.0

(26) *Constraints on vowel sequences where V1 is /e/ or /u/ in a model of Seediq URs*

Constr.	w	p		Constr.	w	p	
*eu	2.75	8.2×10^{-6}	(p<0.001)	*ue	4.82	0.04	(p<0.05)
*eo	3.59	0.70	ns.	*uo	3.95	0.61	ns.
*ei	0.01	0.07	ns.	*ua	0.00	0.18	ns.
*ea	0.71	0.92	ns.	*ui	0.21	1.0	ns.
*ee	0.00	1.0	ns.	*uu	3.51	0.11	ns.

In summary, vowel-vowel sequences are dispreferred only in environments where they would be reanalyzed from the **surface stem**, but not from the suffixed form. This provides indirect evidence that re-analyses are overwhelmingly from the stem, rather than from the suffixed form. This asymmetry cannot be directly explained by a cobbled UR approach, where reanalyses from both stem and suffixed forms are assumed to be possible.

6 Productivity of base-driven alternations

The surface-base hypothesis predicts that when given novel stems, speakers should be able to apply alternations in a way that makes suffixed forms more predictable from stems. The stem-base grammar learned in §4 makes more specific predictions; in particular, as outlined in §4.4, [u]~[e/o] alternations are predicted to be extended in stems with a stressed [e/o].

This section discusses the results of a production experiment testing predictions for post-tonic vowel alternation. Results suggest that speakers productively apply vowel matching alternations to post-tonic [u] but not other vowels, in line with the predictions of the surface-base approach. However, speakers have also learned vowel matching non-veridically, extending it to environments not predicted by the model developed in §4.

6.1 Methodology

The experimental methodology adopted was a modified version of a nonce-word task (i.e. wug test; Berko, 1958). Speakers participated in a production task, where they were given stems and asked to produce the inflected suffixed form. Production experiments following this paradigm have been

shown to elicit responses that, when averaged over several speakers, replicate distributional facts about the lexicon (e.g. Zuraw, 2000; Ernestus and Baayen, 2003: and many others).

6.1.1 Participants

Participants were adult native speakers of Tgdaya Seediq ($N = 10$; 7 female; ages 42-76). All speakers were paid 500NTD (around \$17) for their time. Of the 10 participants, 7 notably had some slight experience training to be Seediq language teachers. They had metalinguistic awareness of suffixes and their functions, but were not taught explicitly about the consonant and vowel alternation processes.

6.2 Procedure

Since the experiment took place during the COVID-19 pandemic (August 2020), it was conducted remotely by the author, through video conferencing software.

Stimuli were presented in Seediq orthography using Microsoft PowerPoint; each word was given in a separate slide with its gloss (in Chinese orthography).

The experimenter prompted speakers to give suffixed forms for stimuli by providing an existing paradigm (e.g. [hediq]~[hudiqan]), and asking them to fill out the paradigm of the test item. This method worked well for my participants, as most have had some (limited) experience training to be Seediq language teachers, and therefore had metalinguistic knowledge about the suffixes.

If a participant failed to produce suffixed forms, the experimenter prompted them by providing more real stem-suffix examples. To minimize priming effects, example stem-suffix pairs always had /i/ as V2 (since the experiment included no stimuli where V2 was /i/). Subjects occasionally fluctuated between -an and -i suffixes, but this never resulted in variation of the stem allomorph. When needed, the experimenter would also provide a meaning for the stimulus verb. For example, the inflected form for *daruk* ‘oil, fat’ could have the meaning ‘to render the fat (out of food)’.

Prior to the experiment, speakers were asked to read a list of Seediq nouns to confirm their fluency with the orthography. Starting with two real-word practice items, speakers were asked to provide the /-an/ suffixed form for each word. After each item, the experimenter checked whether the speaker already knew the inflected forms of the stem (items known to speaker were excluded).

6.2.1 Stimuli

In a pilot experiment, speakers raised concerns that the use of nonce words in experiments would interfere with ongoing language revitalization efforts. In response to these concerns, the current study used ‘gapped forms’, or stems with no known suffixed forms, in place of nonce words. A full list of stimuli is given in the Appendix.

Gapped forms were selected using the following methods. First, most stimuli were formed by affixing noun stems with a verbalizer prefix ‘pu-’⁴. For example, the stimulus [pu'gakac] ‘VERB-chair’

⁴The prefix [pu-] can act as a “verbalizer” that derives a verb from a non-verb class Holmer (1996) describes this

can be interpreted as meaning ‘to build a chair.’ To sufficiently cover all experimental conditions, I also included some rare (or relatively unknown) verbs.

I worked with a primary consultant (age 76, female) to confirm that all test stimuli had no known suffixed forms. A stem was determined to have no known suffixed form if she had never heard it before and had never heard her elders using it before. According to my consultant, Seediq speakers rarely use innovative suffixed forms, and therefore have a clear intuition of whether a stem is gapped. My primary consultant also has certification as a Seediq language teacher, so it was relatively straightforward to ask her whether specific suffixed forms existed.

Stems were determined to be plausibly suffixable by running a pilot experiment with two consultants; words they judged to be impossible to suffix were omitted from the final stimuli.

Stimuli consisted of disyllabic stems ending in closed syllables (i.e. CVCVC), where the first vowel (V1) was one of /a, e, u/ and the second vowel (V2) was one of /a, u/. This results in six possible vowel combinations, summarized in Table 9. These vowel combinations were selected to elicit a range of environments in which post-tonic [u] is expected to either alternate with [e] or not alternate. Stems with a post-tonic /a/ are expected to never show V2 alternations.

There were 8 test items for each vowel combination, as well as 24 filler items (4 per vowel combination), which were Seediq words with known suffixed forms. This resulted in a total of 72 stimuli ($8 \times 6 + 24$).

6.3 Predictions

Predicted responses, on the basis of the rule-based model in §4, are summarized in Table 9; the rightmost column gives an example stimulus for each condition, with the expected preferred outcome given in parentheses. If speakers generalize the vowel matching pattern, they should apply the [u]~[e] alternation to most CeCuC stimuli. In CaCuC stems, [u]~[e] alternation should be very infrequent. In CuCuC stems, vowel alternation should never be observed, since the faithful non-alternating outcome already satisfies vowel matching.

V1	V2	Prediction	Example
a	u	disprefer alternation	'daruk (du'ruk-an) ‘oil’
e	u	[u]~[e] alternation	'keruŋ (ku'reŋ-an) ‘wrinkles’
u	u	never alternate	'cuguk (cu'guk-an) ‘Bidens plant’
a	a	V2 never alternates	'sabak (su'bak-an) ‘dregs, pulp’
e	a		'rehak (ru'hak-an) ‘seed’
u	a		ku'suwak (kusu'wak-an) ‘yawn’

Table 9: Experimental conditions: vowel alternations

as a causative prefix, but I found that it could be used somewhat productively to form denominal verbs; the same verbalizer prefix is found in closely related languages like Squliq Atayal (Huang and Hayung, 2008).

6.4 Results

The following section summarizes the experimental results. For 56 responses (12%), speakers did not provide any responses; these are labeled ‘no response’ in the figures below. In a small subset of tokens (n=8, 2% of total responses), instead of inflecting the provided stem, speakers would provide the inflected form of an existing, segmentally similar verb; these are labeled ‘other’.

6.4.1 Results

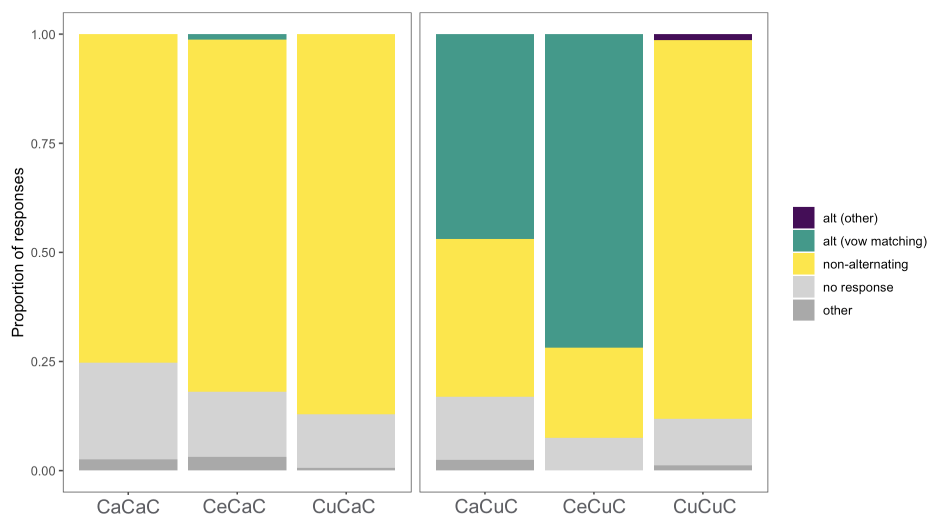


Figure 5: Results: vowel alternation rates

Results for final vowel alternations are given in Fig. 5, which shows the proportion of response types by vowel condition. Cases where vowel alternation obeyed the vowel matching pattern (i.e. resulted in the stem and suffixed forms having the same stressed vowel) are indicated in green.

First, looking at the left-hand column, which shows results for stems with a post-tonic [a], we see that as expected, final /a/ almost never alternates. There was one exception ([hu'renaŋ~huruneŋi]); in this case, an [a~e] alternation consistent with the vowel matching pattern was observed. On the right-hand column, consistent with predictions, CeCuC stems prefer the [u]~[e] alternation, and CuCuC stems never alternate (with one exception, ['cuguk~cu'gak-an]).

However, speakers deviated from the lexicon in stems of the form CaCuC. [u]~[e] alternation was not observed at all. Instead, for around half of the stems in this category, speakers applied a [u~a] alternation (e.g. ['daruk~du'rak-an]. This alternation is irregular and novel, in the sense that it is predicted by neither lexical statistics nor the stem-base grammar. Instead, it appears that speakers have extended the vowel matching pattern to CaCuC stems.

6.5 Discussion and interim summary

Overall, results suggest that speakers have productively learned to apply vowel alternations in a way that renders suffixed forms more predictable from stem forms. Specifically, speakers learned a vowel matching pattern, but appear to have generalized it beyond CeCuC and CoCuC stems, resulting in [u~a] alternations for CaCuC stems. The fact that speakers did not just match lexical statistics, but instead overgeneralized, further supports the conclusion that speakers are not just generating suffixed forms from analogy, but have learned a productive process for predicting vowel alternations using a stem base.

Speakers responses also suggest that a grammar of morphological mappings formed purely on the basis of lexical distributions (as was done in my stem-base model) is insufficient; speakers' grammars may have in part been shaped by other phonological principles (Hayes et al., 2009; Becker et al., 2011; Moore-Cantwell, 2013). With post-tonic vowel alternations, speakers seem to have learned a general vowel matching principle, instead of more specific principles about alternation of [u] with [e] or [o]. This could reflect a learning bias towards less complex, more general patterns. This possibly is explored and discussed in more detail in §7.5.

Note that this analysis implicitly assumes that speakers have learned a vowel matching pattern across all vowels. The stimuli doesn't contain stems with stressed [i], so potential follow-up work could confirm whether vowel matching holds for CiCuC stems.

7 Modeling stem-suffix alternations with phonological constraints

§4 assessed the stem-suffix asymmetry in Seediq using a model which learns morphological correspondences from lexical distributions. Experimental results suggest that this model does not provide a sufficient *theoretical* account of Seediq verbal alternations. In particular, speakers applied a novel [u]~[a] alternation to CaCuC forms. I argue that this is because, instead of directly learning morphological mappings of [u]→[e,o], speakers have learned a more general vowel matching principle.

In this section, I aim to unify lexical and experimental results, by proposing a stem-base model of Seediq vowel alternations, which uses a phonological constraint on vowel matching to motivate alternation. The analysis will be set in the framework of Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater and Johnson, 2003), a stochastic variant of Optimality Theory (Smolensky, 1986; Prince and Smolensky, 1993). I will also briefly discuss how the non-viridical extension of vowel matching could be modeled as the result of a GENERALITY BIAS.

Due to space constraints, this section and subsequent modeling will discuss only vowel alternations, where effects of stem-based reanalysis are backed up both lexical statistics and experimental evidence. An in-depth MaxEnt analysis of consonant alternation can be found in Kuo (2020).

7.1 MaxEnt

MaxEnt is a probabilistic variant of Harmonic Grammar (Legendre et al., 1990; Pater, 2009), which are themselves variants of OT that use weighted (instead of ranked) constraints. MaxEnt generates a probability distribution over the set of candidate outputs based on their violations of a set of weighted constraints.

Unlike classic OT, where strict ranking ensures that losing candidates never surface, all candidates in MaxEnt grammars receive some probability. However, if constraint weights are sufficiently different, MaxEnt produces results that are functionally very similar to classic OT, where the winning candidate gets near-perfect probability, while losing candidates get near-zero probability.

MaxEnt models are associated with learning algorithms that have been proved to converge on one optimal solution (Berger et al., 1996), which has the maximum (least negative) log-likelihood. To learn model constraint weights, I use Excel’s Solver add-in (Generalized Reduced-Gradient Algorithm; Fylstra et al., 1998).

7.2 A base-driven analysis of Seediq vowel alternations

This section introduces the constraints needed under a base-driven approach to Seediq post-tonic vowel and final consonant alternations. For simplicity, all tableau will use hand-ranked constraints. Unless otherwise noted, models will be fit to the lexical corpus (rather than the experimental results). §7.3 will present weights learned algorithmically in models trained on the lexicon.

7.2.1 Faithfulness using *Map (Zuraw, 2013)

Faithfulness is enforced using *MAP constraints instead of classic feature-based faithfulness (McCarthy and Prince, 1995). The *MAP family of constraints, proposed by Zuraw (2007, 2013), is formalised in (27). Unlike classic faithfulness constraints, *MAP constraints can describe the correspondence between any two natural classes of sounds, even when the two classes differ in more than one feature. It is possible to have both constraints such as *MAP([+high],[−high]) (which is functionally equivalent to IDENT[high]), and segment-specific constraints like *MAP(k,p).

- (27) *MAP(a, b): assess a violation to a candidate if output **a** is mapped to a corresponding output **b**, where **a** and **b** can be non-minimal (Zuraw, 2013).⁵

Although *MAP constraints are much more powerful than feature-based faithfulness, Zuraw (2007, 2013) constrains them using the P-MAP (Steriade, 2001). The P-MAP, which represents the perceptual distance between two sounds in a phonological context, can be used to establish *a priori* rankings of correspondence constraints in a way which biases smaller perceptual changes. In other words, *MAP constraints have a default ranking (or weighting), where correspondences between perceptually more dissimilar sounds are penalized more. These default preferences can

⁵Although it is not necessary here, *MAP constraints can be specified for the context in which the pair of sounds are in correspondence (e.g. natural class x in context A_B should not correspond to natural class y in context C_D).

be subverted given enough language-specific evidence, giving rise to marked alternations (Hayes and White, 2015). Because of its connection to the P-map, *MAP constraints should be limited to evaluating correspondences between surface forms. This is consistent with my stem-base analysis, where the input to phonology is surface allomorphs.

*MAP constraints are adopted because a subset of alternations in Seediq, including post-tonic vowel alternation, are saltatory. Saltation is where one sound A alternates with a sound C, but “leaps” over a phonetically and featurally ‘intermediate’ sound B (which remains invariant) (White, 2013; Hayes and White, 2015). Saltatory alternations are problematic for classic OT (Łubowicz, 2002; Ito and Mester, 2003), but can be dealt with using *MAP (Hayes and White, 2015).

Post-tonic vowel alternation is saltatory in the sense that when post-tonic [u] alternates, it always prefers to alternate with [e] rather than with [o], even in CaCuC and CiCuC forms, where alternation is not modulated by vowel matching. Of the [u]~[e] and [u]~[o] alternations, [u]~[e] is arguably more phonetically distinct, and involves a superset of the featural changes needed for [u]~[o] alternation. Trying to capture the preference for [u]~[e] alternation using classic feature-based faithfulness would result in constraint conflicts. Using *MAP constraints, we can avoid this by assuming constraints *MAP(u,e) and *MAP(u,o), where *MAP(u,o) is more highly weighted.

Compared to classic faithfulness constraints, *MAP constraints also more straightforwardly allow for different types of bias to be modeled (Wilson, 2006). Adopting *MAP constraints will allow for the model to be more easily tested for different kinds of bias in future modeling work.

7.2.2 Constraints for pre-tonic vowel neutralizations

Pretonically, vowels either delete, assimilate to an stressed vowel, or reduce to [u]. All three patterns can be motivated by fairly standard markedness constraints. In this section, I discuss only pretonic vowel reduction; analyses for vowel assimilation and deletion are found in Kuo (2020).

Pretonic vowel reduction to [u] is enforced by a positional licensing constraint (Crosswhite, 2004). LICENSE[u]/pretonic (LIC[u]/pret) is defined in (28), and essentially penalises non-[u] syllables in pretonic position. This is demonstrated in tableau (29) for ['barah~bu'rahan].

The faithful candidate (a) fatally violates the highly weighted LIC[u]/pret, and is assigned near-zero probability. Candidate (b), which repairs the LIC[u]/pret violation by incurring a violation of *MAP(u,a), has the highest probability. Candidate (c), which repairs the markedness violation by deleting the stem’s initial syllable, is ruled out due to violation of higher-weighted faithfulness constraints.

- (28) LICENSE[u]/*pretonic*: non-[u] vowels cannot appear in pretonic syllables.

(29) *Pretonic vowel reduction.* Very small probabilities (on the order of 10^{-5}) are listed as zero.

			$Lic([u], pret.)$	$*M_{AP}(C, \emptyset)$	$*M_{AP}(V, \emptyset)$	$*M_{AP}(u, a)$
['barah]	P	\mathcal{H}	20	10	10	1
a. ba'rah-an	0	20	1			
b. bu'rah-an	1	1				1
c. 'rah-an	0	20		1	1	

7.2.3 Constraints for post-tonic vowel alternations

Post-tonically, [u] alternates with [e/o] to satisfy VOWEL MATCHING; alternation is preferred only when it results in the stressed vowel of the suffixed form matching the stressed vowel of the stem.

This process is reminiscent of copy epenthesis (Stanton and Zukoff, 2018), which can be analyzed as feature spreading in autosegmental accounts (e.g. Clements, 1986; Kawahara, 2007), or using correspondence theory (McCarthy and Prince, 1995), by enforcing faithfulness between a host segment and its copy. This correspondence approach is particularly attractive to the current analysis, because it predicts prosodic identity effects (McCarthy and Prince, 1988).

Prosodic correspondence, which compares prosodic elements of input and output forms rather than segmental elements, is fleshed out in Crosswhite's (1998) analysis of Chamorro gemination. I adopt Crosswhite's approach, and use the constraint NUC-IDENT-OO(V) to enforce vowel matching. This constraint, defined in (30), sets up a correspondence relation between the stressed syllable nuclei of related output forms, and requires that these positions have the same vowel. In other words, if two outputs are related, they should share the same stressed syllable head.

(30) NUC-IDENT-OO(V): For α , a stressed nucleus of the base, and β , a stressed nucleus of an output, where α corresponds to β , α and β must be the same.

Unlike the MGL implementation in §4, the current analysis uses a general phonological constraint instead of morphological correspondence constraints. This choice is motivated by the experimental results. Recall that speakers extended the vowel matching pattern, applying [u]~[a] alternations to CaCuC stems that was not predicted by the MGL (§6.4.1). This suggests that they learned the more general phonological constraint.

Tableau (31) demonstrates how NUC-IDENT(V) is used to derive outputs that match lexical frequencies.⁶ For the input ['putus], the faithful candidate satisfies NUC-IDENT(V), and therefore receives high probability (≈ 1). For ['petus], the high weight of NUC-IDENT(V) relative to $*M_{AP}(u, e)$ causes candidate (e), which undergoes [u]~[e] alternation, to be preferred over the faithful candidate. Candidate (f) is ruled out because vowel alternation violates both segmental and prosodic

⁶The constraint SWP will not be introduced until the following section, but is not crucial to this tableau.

faithfulness. For ['patus], candidate (j), which undergoes post-tonic [u]~[a] alternation to resolve NUC-IDENT(V) violations, is ruled out by the high weight of *MAP(u,a). As will be addressed in §7.5, this differs from the experimental results, where speakers did in fact apply the [u]~[a] alternation to around 50% of CaCuC forms.

- (31) *Tableau: post-tonic vowel alternations for CVCuC inputs.* The probability of each candidate in the lexicon (Obs.) is shown alongside model predictions (P).

				$NUC-IDENT(V)$				
	Obs.	P	\mathcal{H}		SWP	$*MAP(u,a)$	$*MAP(u,e)$	$*MAP(u,o)$
	3	0.1	5.6	1.7	6.7			
['putus]								
a. pu'tusan	1	0.99	0		2			
b. pu'tesan	0	0.01	4.6	1	1		1	
c. pu'tosan	0	0.00	9.7	1	1			1
['petus]								
d. pu'tusan	0.2	0.20	4.6	1	2		1	
e. pu'tesan	0.8	0.80	3.2		1		2	
f. pu'tosan	0	0.00	11.3	1	1		1	1
['patus]								
g. pu'tusan	0.76	0.77	8.6	1	2	1		
h. pu'tesan	0.18	0.16	10.2	1	1	1	1	
i. pu'tosan	0	0.00	15.3	1	1	1		1
j. pu'tasan	0.06	0.07	11.2			2		

7.2.4 Introducing Stress-to-weight

In the experimental results for post-tonic vowel alternation, speakers applied a novel alternation to post-tonic [u], resulting in [u]~[a] alternations. However, they never applied [a]~[u] alternations to post-tonic [a], even when doing so would resolve violations of NUC-IDENT(V). There are multiple possible explanations for this; speakers could, for example, have learned a source-oriented generalization about the type of vowels allowed to undergo alternation Becker and Gouskova (2016).

I chose to explain this directionality as a preference for more sonorous vowels in stressed positions; [u]~[a] alternations are preferred over [a]~[u] alternations because the former increases the sonority of the stressed syllable, while the latter does the opposite. A similar preference for sonorous vowels in prosodic heads has been observed in various languages, including Zabiče Slovene (Crosswhite and Jun, 2001) and Chamorro (Chung, 1983).

To capture sonority effects, I use a single constraint STRESS-TO-WEIGHT (SWP), defined in (32). This constraint, taken from (Crosswhite, 1998), has been modified to be gradient; mid vowels in stressed syllables incur 1 violation, while high vowels incur 2 violations.⁷

- (32) STRESS-TO-WEIGHT: The head of the prosodic word should be a heavy syllable (Crosswhite, 1998). Incur a penalty of 1 if the nucleus of the stressed syllable is a mid vowel, and of 2 if it is a high vowel.

Although SWP is needed to explain experimental results, it has negligible effect in a lexicon-trained model, as excluding some irregularities, only [u] alternates post-tonically. Consequently, there is no evidence to disambiguate between alternations which improve or reduce the stressed syllable’s sonority.

To better illustrate the effects of SWP, tableau (33) is fit to the *experimental results*. Given the input ['patus], candidate (b), which exhibits [u]~[a] alternation, receives high weight because it satisfies both NUC-IDENT(V) and SWP. In contrast, for the input ['putas], the [a]~[u] alternating candidate (e) receives low weight despite satisfying NUC-IDENT(V). This is because alternation increases violations of SWP.

(33) *Tableau: effect of SWP in experimental results*

				$LIC[u]/pret$	$NUC-IDENT(V)$	SWP	$*MAP(u,e)$	$*MAP(u,a)$	$*MAP(a,e)$
	Obs.	P	\mathcal{H}	36.4	12.0	6.4	17.2	24.6	9.7
['patus]									
a.	pu'tasan	0.56	0.56	49.1				2	
b.	pa'tasan	0	0	61.0	1			1	
c.	pu'tusan	0.44	0.44	49.4		1	2	1	
d.	pu'tesan	0	0	60.2		1	1	1	
['putas]									
e.	pu'tusan	0	0	37.5			2		1
f.	pu'tasan	1.00	1.00	12.0		1			
g.	pu'tesan	0	0	28.1		1	1		1

This analysis makes predictions that should be tested in future work. Specifically, speakers are expected to disprefer extension of vowel matching when doing so reduces the stressed vowel’s sonority. For example, the post-tonic vowel in CiCaC stem could undergo [a]~[i] alternation to satisfy NUC-IDENT(V). However, because alternation of [a] to [i] increases violations of SWP, [a]~[i] alternation should occur at a lower rate than [u]~[a] alternation did for CaCuC stems.

⁷More sophisticated models of gradient constraints on syllable weight are explored in work like Flemming (2001).

7.3 Model fit to lexicon

The constraints presented in §7.2 were used to train a MaxEnt model. All relevant faithfulness constraints were included even if they ended up receiving zero weight.

The input was schematised to be CVCVC forms with all possible surface vowel combinations (i.e. the first vowel V1 was one of [a, e, i, o, u], while V2 was one of [i, a, u]). Vowel-final forms (e.g. ['qene]) were omitted. Since all pre-tonic vowel neutralization processes are exceptionless, I ignored the difference between i) stems with an initial onsetless syllable (e.g. ['atak]), ii) CVHVC stems (where H is [h] or [ʔ]), and iii) CVCVC stems. Idealised training data was used for ease of interpretation, and so that results could be compared with experimental data.

The vowel alternation model has 10 constraints and a log-likelihood of -78.05 (null=-301.43). The optimal weights assigned to the training data are shown in (34). Fig. 6 compares model predictions against the lexicon; each point represents a candidate's mean predicted probability against its observed probability. In this figure, the model closely fits the lexicon ($r^2 = 0.94$); the only clear outliers, which involve 'CoCuC inputs (labeled “potus”), will be discussed below.

(34)	LIC[u]/pret	31.22	*MAP(a,e)	5.04	*MAP(i,u)	5.71
	NUC-IDENT(V)	2.93	*MAP(u,o)	3.25	*MAP(i,e)	5.33
	SWP	0.29	*MAP(u,e)	1.89	*MAP(u,a)	6.42
	*MAP(i,a)	4.89				

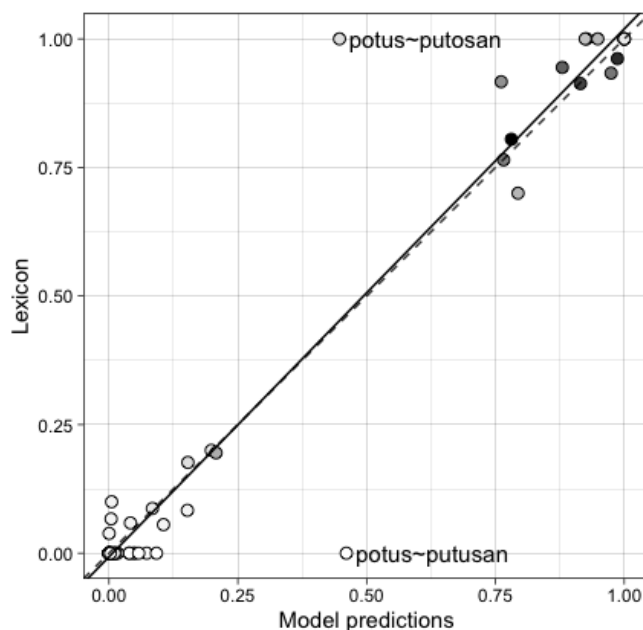


Figure 6: Vowel alternation model: predicted vs. lexical probabilities. Darkness of each point indicates frequency (black=highest freq). Solid lines are fitted regression lines; dotted lines plot the line of best fit.

The results of the vowel alternation model are exemplified by the tableaux in (35-37). First,

consider (35), which has the input 'CuCuC. As discussed in §7.2.3, the faithful candidate (a) has near-certain probability since it already satisfies NUC-IDENT(V). Note that the lexicon has one irregular [u]~[i] alternation, giving candidate (b), [Cu'CiC-an], an observed probability of 0.04. The model, on the other hand, predicts a probability of 0 for this candidate.

(35)

				LIC[u]/pret	NUC-IDENT(V)	SWP	*MAP(u,e)	*MAP(u,o)	*MAP(i,u)
['CuCuC] (n=26)	Obs.	P	\mathcal{H}	31.22	2.93	0.29	1.89	3.25	5.71
a. Cu'CuC-an	0.96	0.99	0.57			2			
b. Cu'CiC-an	0.04	0	9.2		1	2			1
c. Cu'CeC-an	0	0.01	5.1		1	1	1		
d. Cu'CoC-an	0	0	6.5		1	1		1	

The tableaux in (36) show results for the inputs ['CeCuC] and ['CoCuC]. For ['CeCuC], candidate (b) is assigned the highest probability because it satisfies NUC-IDENT(V), and because *MAP(u,e) has a relatively low weight. Candidates like (d) are once again ruled out because of the relatively low weight of SWP relative to relevant faithfulness constraints.

For ['CoCuC] inputs, the model undermatches the distribution of candidate (h), where [u]~[o] alternation is motivated by NUC-IDENT(V). Although (h) is exceptionless in the lexicon, the model assigns it only 0.45 probability. This is because [u]~[o] alternation is never observed outside of CoCuC stems, and consequently *MAP(u,o) has a high weight (3.25). Due to the small quantity of CoCuC stems in the lexicon (n=3), it is unclear if this is a large error on the model's part. Model fit for CoCuC stems could be improved by introducing segment-specific constraints (e.g. NUC-IDENT-OO[o]), but as I argued in §7.2.2, experimental results support the adoption of a more general constraint.

(36)

					LiC[u]/pret	NUC-IDENT(V)	SWP	*MAP(u,e)	*MAP(u,a)	*MAP(u,o)
	Obs.	P	\mathcal{H}		31.22	2.93	0.29	1.89	6.42	3.25
['CeCuC] (n=41)										
a.	Ce'CeC-an	0	0	40.4	1		1			
b.	Cu'CeC-an	0.80	0.78	4.1			1	2		
c.	Cu'CoC-an	0	0.01	8.3		1	1	1		1
d.	Cu'CaC-an	0.00		11.2		1		1	1	
e.	Cu'CuC-an	0.20	0.21	5.4		1	2	1		
['CoCuC] (n=3)										
f.	Co'CoC-an	0	0	26.1	1		1			1
g.	Cu'CeC-an	0	0.09	8.4		1	1	1		1
h.	Cu'CoC-an	1.00	0.45	6.79			1			2
i.	Cu'CuC-an	0	0.46	6.76		1	2			1

Finally, for the input ['CaCuC], shown in (37), candidate (e) is ruled out by the high weight of *MAP(u,a), even though it is the only candidate which satisfies both NUC-IDENT(V) and SWP.

(37)

					LiC[u]/pret	NUC-IDENT(V)	SWP	*MAP(u,e)	*MAP(u,a)	*MAP(u,o)
	Obs.	P	\mathcal{H}		31.22	2.93	0.29	1.89	6.42	3.25
['CaCuC] (n=17)										
a.	Ca'CuC-an	0	0	26.05	1	1	2			
b.	Cu'CeC-an	0.18	0.15	11.54		1	1	1	1	
c.	Cu'CoC-an	0	0.04	12.90		1	1		1	1
d.	Cu'CuC-an	0.76	0.77	9.93		1	2		1	
e.	Cu'CaC-an	0.06	0.04	12.85					2	

7.4 Fit of vowel alternation model to experimental results

Experimental results suggest that speakers learned base-driven alternations non-viridically. In this section, I discuss the implications of these findings on my model of Seediq alternations, by comparing weights from a model trained on the lexicon, and one trained on experimental data.

Fig. 7a plots the overall fit of a model trained on and assessed against the experimental results. This model in itself is not informative, but is included to show that the constraint set is sufficient to account for the experimental results; a model trained on the experimental results has an extremely good fit ($r^2 = 1$). Fig. 7b plots a model trained on the *lexicon*, but fit to the experiment

results. This model generally does well, but it severely under-predicts rates of [u]~[a] alternation (i.e. ['patus]~[pu'tasan] response). Recall that post-tonic [u]-[a] alternation was irregular in the lexicon, but in the experiment, speakers applied it at a high rate to CaCuC forms. In other words, a model trained on the lexicon predicts much lower rates of [u]~[a] alternation for CaCuC stems.

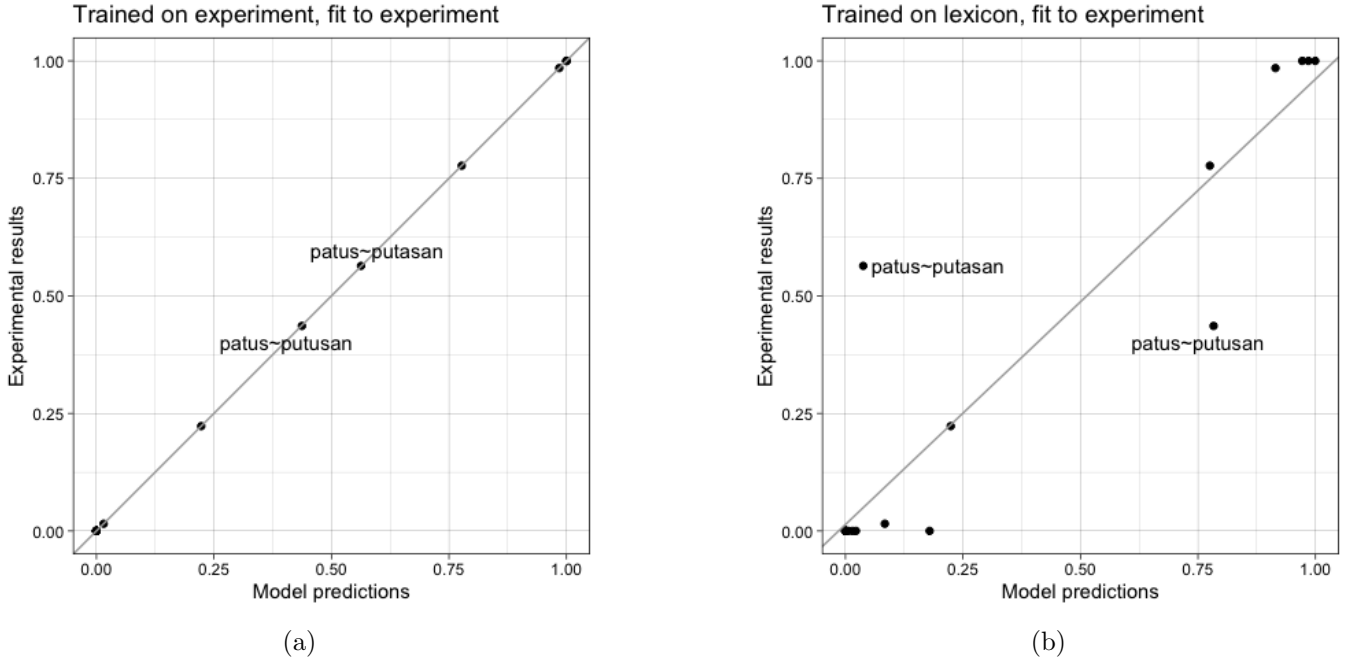


Figure 7: Model predictions plotted against experimental results. Fitted regression lines are included.

Table 10 compares weights learned from the lexicon and from experimental results; only the subset of constraint weights relevant to the experimental conditions are shown. The two models differ crucially in the weights learned for NUC-IDENT(V), SWP, and *MAP(u,a). In the model trained on the lexicon, NUC-IDENT(V) and SWP have much lower weights than *MAP(u,a); this causes [u]~[a] alternating candidates to be assigned very low probability.

In the model trained on experimental results, NUC-IDENT(V) and SWP still have lower weights than *MAP(u,a). However, both constraints have relatively higher weights than before. NUC-IDENT(V) in particular has a fairly high weight, closer to the weight of *MAP(u,a). As a result, given an input of the form CaCuC (e.g. ['patus]), NUC-IDENT(V) and SWP will gang up to give non-alternating candidate [pu'tusan] a higher harmony score, and cause it to be assigned slightly less probability than the [u]~[a] alternating candidate.

7.5 Explaining experimental results with a generality bias

In the experimental results, speakers successfully learned generalizations which made alternations more predictable from a stem base. At the same time, speakers appear to have preferentially learned the NUC-IDENT(V) constraint, assigning it a higher weight than expected given the lexical

Constraint	Lexicon	Experiment
LIC[u]/pret	17.37	36.39
NUC-IDENT(V)	2.93	12.66
SWP	0.28	1.63
*MAP(u,e)	1.89	13.04
*MAP(u,a)	6.41	15.66
*MAP(a,e)	5.06	15.20

Table 10: Vowel alternation model trained on lexicon vs. experiment

statistics. I propose that the results follow from a learning bias.

In principle, many types of bias could affect morphophonological learning, including substantive bias Wilson (e.g. 2006); White (e.g. 2017) and complexity bias (Moreton and Pater, 2012). Generality bias is closely related to complexity bias; in classic rule-based phonology, featural complexity can be thought of as one way of measuring a rule’s generality (Chomsky and Halle, 1968). Later on, Albright and Hayes (2003) similarly used a rule-based grammar, but quantified generality in turns of rule scope and accuracy.

For Seediq, I speculate that speakers’ preferential learning of NUC-IDENT(V) is rooted in a generality bias. NUC-IDENT(V) potentially affects all non-monosyllabic stems. It therefore has a large scope, and is in this sense more general than competing faithfulness constraints like *MAP(u,e), which affects only stems containing stressed [e]. Speakers should have more ‘evidence’ for NUC-IDENT(V) (relative to less general constraints), and be able to better learn it. Although a full model of generality bias is beyond the scope of the current paper, bias can be directly implemented to my MaxEnt model as a soft prior, in the form of a Gaussian distribution over each constraint weight. This approach was developed by Wilson (2006), and has since been explored in various work on bias learning, including White (2013) and Kimper (2016).

8 Conclusion

Based on a survey of 340 Seediq verb paradigms, the current study finds that Seediq paradigms show a striking asymmetry, whereby the non-suffixed slots of the paradigm can be used to predict the suffixed forms with much higher accuracy than the other way around. This asymmetry was demonstrated with a morphological mapping model which uses the Minimal Generalization Learner (Albright, 2002).

This asymmetry is expected if there has been a gradual restructuring of Seediq verb paradigms based on the non-suffixed forms. Asymmetric restructuring is unexpected in a cobbled UR approach, where reanalyses from all slots of a paradigm are possible. In contrast, it is the natural outcome under the single surface-based hypothesis, if Seediq speakers have selected a non-suffixed paradigm slot to be the base.

These results are supported by a production experiment, where speakers were found to extend generalizations about vowel alternations from the isolation stem, and productively apply a vowel

matching pattern when provided novel suffixed forms. Interestingly, speakers also extended the vowel matching pattern to more forms than was predicted by a grammar that only learns morphological mappings based on statistical generalizations.

Based on these findings, I offer a stem-base grammar of Seediq morphophonology, where alternation is driven by phonological constraints rather than morphological correspondence. This model is implemented in MaxEnt, allowing it to account for gradient rates of alternation in Seediq. NUC-IDENT(V) is used to enforce post-tonic vowel alternation.

Additionally, I tentatively propose that speakers' non-viridical extension of vowel matching is rooted in a generality bias, which could be implemented in my MaxEnt model as a Gaussian prior. A full implementation of this biased model is beyond the scope of this paper. Nevertheless, the possibility that a generality bias is present in Seediq, and more broadly in morphophonological learning, is an interesting topic for further study.

References

- Adam Albright. Base-driven leveling in Yiddish verb paradigms. *NLLT*, 28(3):475–537, 2010.
- Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161, 2003.
- Adam C Albright. *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles, 2002.
- Michael Becker and Maria Gouskova. Source-Oriented Generalizations as Grammar Inference in Russian Vowel Deletion. *Linguistic Inquiry*, 47(3):391–425, 07 2016. ISSN 0024-3892. doi: 10.1162/LING_a_00217. URL https://doi.org/10.1162/LING_a_00217.
- Michael Becker, Nihan Ketrez, and Andrew Nevins. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, pages 84–125, 2011.
- Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.
- Jean Berko. The child's learning of English morphology. *Word*, 14(2-3):150–177, 1958.
- Juliette Blevins. Consonant epenthesis: natural and unnatural histories. *Language universals and language change*, pages 79–109, 2008.
- Joan Bybee. *Phonology and language use*, volume 94. Cambridge University Press, 2003.
- Noam Chomsky and Morris Halle. *The sound pattern of English*. Harper & Row New York, 1968.
- Sandra Chung. Transderivational relationships in Chamorro phonology. *Language*, pages 35–66, 1983.

- G.N. Clements. Syllabification and Epenthesis in the Barra Dialect of Gaelic. In Koen Bogers, Harry van der Hulst, and Maarten Mous, editors, *The Phonological Representation of Suprasegmentals*, pages 317–336. Foris, Dordrecht, 1986. doi: 10.1515/9783110866292-017.
- Council of Indigenous People. Aboriginal household registration statistics data analysis in oct 2020. <https://www.cip.gov.tw/portal/docDetail.html?CID=940F9579765AC6A0>, 2020.
- Council of Indigenous Peoples. Online dictionary of aboriginal languages. <http://e-dictionary.apc.gov.tw/Index.htm>, 2020. Accessed: 2020-09-30.
- Katherine Crosswhite. Segmental vs. prosodic correspondence in Chamorro. *Phonology*, pages 281–316, 1998.
- Katherine Crosswhite. Vowel reduction. *Phonetically based phonology*, pages 191–231, 2004.
- Katherine Crosswhite and Alexander Jun. *Vowel reduction in optimality theory*. Psychology Press, 2001.
- Jan Edwards, Mary E Beckman, and Benjamin Munson. The interaction between vocabulary size and phonotactic probability effects on children’s production accuracy and fluency in nonword repetition. 2004.
- Mirjam Ernestus and R Harald Baayen. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, pages 5–38, 2003.
- Edward Flemming. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, pages 7–44, 2001.
- Daniel Fylstra, Leon Lasdon, John Watson, and Allan Waren. Design and use of the microsoft excel solver. *Interfaces*, 28(5):29–55, 1998.
- Sharon Goldwater and Mark Johnson. Learning of constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111120, 2003.
- Simon J Greenhill, Robert Blust, and Russell D Gray. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283, 2008.
- Kenneth Hale. Deep-surface canonical disparities in relation to analysis and change: An Australian example. *Current trends in linguistics*, 11(19731):401–458, 1973.
- Bruce Hayes and James White. Saltation and the p-map. *Phonology*, 32(2):267–302, 2015.
- Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *LI*, 39(3):379–440, 2008.

- Bruce Hayes, Péter Siptár, Kie Zuraw, and Zsuzsa Londe. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, pages 822–863, 2009.
- Bruce Hayes, Colin Wilson, and Anne Shisko. Maxent grammars for the metrics of Shakespeare and Milton. *Language*, pages 691–731, 2012.
- Harry. Hoijer. Tonkawa. *Linguistic structures of Native America*, 249, 1946.
- Arthur Holmer. *A parametric grammar of Seediq*. PhD thesis, Lund University, 1996.
- Lillian M Huang and Tali’ Hayung. Syntax and semantics of p- in Squliq Atayal. *Language and Linguistics*, 9(3):491–521, 2008.
- Junko Ito and Armin Mester. On the sources of opacity in OT: Coda processes in German. *The syllable in optimality theory*, pages 271–303, 2003.
- Jongho Jun. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics*, 19(2): 137–179, 2010.
- Jongho Jun and Adam Albright. Speakers’ knowledge of alternations is asymmetrical: Evidence from seoul korean verb paradigms 1. *Journal of Linguistics*, 53(3):567–611, 2017.
- Yoonjung Kang. Neutralizations and variations in Korean verbal paradigms. *Harvard Studies in Korean Linguistics*, 11:183–196, 2006.
- Shigeto Kawahara. Copying and spreading in phonological theory: Evidence from echo epenthesis. In Ehren Reilly Leah Bateman, Michael O’Keefe and Adam Werle, editors, *University of Massachusetts Occasional Papers 32: Papers in Optimality Theory III*, pages 111–143. Amherst: GLSA, 2007.
- Michael Kenstowicz and Larry M Kisseberth. *Topics in phonological theory*. New York: Academic Press, 1977.
- Wendell Kimper. Asymmetric generalisation of harmony triggers. In *Proceedings of the Annual Meetings on Phonology*, volume 3, 2016.
- Paul Kiparsky. Analogical change as a problem for linguistic theory. *Studies in the Linguistic Sciences Urbana, Ill*, 8(2):77–96, 1978.
- Jennifer Kuo. Evidence for base-driven alternation in Tgdaya Seediq. Master’s thesis, UCLA, 2020.
- Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. Harmonic grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the twelfth annual conference of the Cognitive Science Society*, pages 884–891. Citeseer, 1990.
- Paul Jen-kui Li. Reconstruction of Proto-Atayalic phonology. *Bulletin of the Institute of History and Philology*, 52, 1981.

- Anna Lubowicz. Derived environment effects in optimality theory. *Lingua*, 112(4):243–280, 2002.
- John J McCarthy and Alan Prince. Quantitative transfer in reduplicative and templatic morphology. *Linguistics in the Morning Calm 2*, 1988.
- John J. McCarthy and Alan S. Prince. Faithfulness and Reduplicative Identity. In Laura Walsh Dickey Jill N. Beckman and Suzanne Urbanczyk, editors, *Papers in Optimality Theory*, pages 249–384. Amherst: GLSA, 1995.
- Andrei Mikheev. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423, 1997.
- Claire Moore-Cantwell. Over-and under-generalization in learning derivational morphology. In *Proceedings of NELS*, volume 42, 2013.
- Elliott Moreton and Joe Pater. Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass*, 6(11):686–701, 2012.
- Roland Noske. *A theory of syllabification and segmental alternation: with studies on the phonology of French, German, Tonkawa, and Yawelmani*, volume 296. Walter de Gruyter, 2011.
- Joe Pater. Weighted constraints in generative linguistics. *Cognitive science*, 33(6):999–1035, 2009.
- Janet Pierrehumbert et al. Probabilistic phonology: Discrimination and robustness. *Probabilistic linguistics*, pages 177–228, 2003.
- Alan Prince and Paul Smolensky. Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in Phonology*, page 3, 1993.
- Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.
- Juliet Stanton and Sam Zukoff. Prosodic identity in copy epenthesis. *NLLT*, 36(2):637–684, 2018.
- Donca Steriade. The phonology of perceptibility effects: the p-map and its consequences for constraint organization. *Ms.*, *UCLA*, 2001.
- James White. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a p-map bias. *Language*, 93(1):1–36, 2017.
- James C. White. *Bias in phonological learning: Evidence from saltation*. PhD thesis, *UCLA*, 2013.
- Colin Wilson. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982, 2006.
- Hsiu-fang Yang. The phonological structure of the paran dialect of Sediq. *Bulletin of the Institute of History and Philology Academia Sinica*, 47(4):611–706, 1976.

Kie Zuraw. *Patterned Exceptions in Phonology*. PhD thesis, UCLA, 2000.

Kie Zuraw. The role of phonetic knowledge in phonological patterning: corpus and survey evidence from tagalog infixation. *Language*, pages 277–316, 2007.

Kie Zuraw. *map constraints. Master’s thesis, UCLA, 2013.

Appendix: stimuli

Test items			
V1	V2	word	gloss
a	a	papak	foot
		sabak	pulp (of fruit)
		kkarang	walk on all fours
		dayaN	make firebreak
		tanah	red
		tapaq	pat, slap
		slmadac/hlmadac	hunting knife
		gakac	chair
a	u	tatuk	xylophone
		daruk	fat, oil
		rapung	mold
		halung	gun
		damux	rooftop
		aguh	call over
		ahuc (paahuc)	hoe
		lapuc	lint
e	a	rehak	seed
		tpetak	conflict
		hrenang	sound
		gelang	string of, bundle of
		qseyaq	cough
		phepah	flower
		ngerac	outside
		sepac	four
e	u	etuk	contain
		thbehuk	stuffy
		kerung	wrinkles
		bngabung	grill

		gebuh knedux deluc peeruc	granules thick attached to pillar
u	a	ksuwak ptkurak kurang rubang hrulas srmusaq hunac murac	yawn month-old celebration for infant gums hunting tool spit turbid point compress
u	u	cuguk kduruk bukung ubung btunux pnunuh gukuc hukuc	Bidens plant forehead hunch-backed loom rock breastfeed wheel cane
Filler items			
V1	V2	word	gloss
a	a	awak qbahaN qamas qaras	lead (by leash) listen pickle happy
a	u	saruk balung rahuq haNuc	burn (fur, hair) egg leak away cook, boil
e	a	pkepak gedaN lepax bcebac	touch, feel lose grind cut
e	u	eluk geruN remux keruc	close split, break enter (cut with) saw

u	a	sdurak	chase
		duraN	rope hunting trap
		luqah	injure
		squwaq	noisy
u	u	suyuk	twist into thread
		putuN	light (fire)
		plukus	wear (clothes)
		lutuc	ancestry, descendants