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 & 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 & 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 & 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-oʔ	we-**ntox**-oʔ	**notxo**-n-oʔ	we-**ntoxo**-n-oʔ	'hoe'	/notoxo/
netl-oʔ	we-**ntal**-oʔ	**netle**-n-oʔ	we-**ntale**-n-oʔ	'lick'	/netale/
picn-oʔ	we-**pcen**-oʔ	**picna**-n-oʔ	we-**pcena**-n-oʔ	'cut'	/picena/

Albright (2002b: 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. Distributional restrictions cause there to be extensive vowel and consonant alternations between the unsuffixed and suffixed forms of a verb paradigm.

During elicitation of verb paradigms (described in Section 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 unsuffixed slots of the paradigm) to forms suffixed with /-an/ 'LOCATIVE FOCUS.PRES.' (which are representative of all suffixed slots).

Seediq suffixes are monosyllabic, and the majority of forms can take only one suffix. There is only exception, which arises when a stem is suffixed with /-an/ 'LOCATIVE FOCUS' followed by /-i/ 'IMPERATIVE'. As will be discussed in the following section, these doubly suffixed forms are assumed to not be relevant in the discussion of possible base slots, as they always experience more neutralization than forms that take just one suffix.[1]

---

[1]   These doubly suffixed forms are traditionally analyzed as a single suffix /-ani/ 'LOCATIVE.IMPERATIVE' (Holmer, 1996), but I choose

	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		-an-i	-i	

Table 1: Inflectional morphology of Seediq

(2) *Seediq consonant inventory*

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

(3) *Seediq vowel inventory*

i		u
e		o
	a	

2.1 Data collection

The descriptive generalizations to be outlined in the rest of this section are all taken from Yang (1976), supplemented with counts from the current study. In particular, I confirm these generalizations using a corpus of 341 verbal paradigms. 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.

188 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, leaving a total of 341 paradigms to be analyzed. The omitted words are given in (4); (4a) was not recognised by my consultants, and for the other words, consultants disagreed with the dictionary on the suffixed form. For the newly collected forms, there was a high degree of inter-speaker agreement; unless otherwise specified, all three consultants agreed on the forms collected.

(4) *Discrepancies in dictionary and consultant responses*

		STEM	SUFFIXED	
			dict.	*consultant*
(a)	'to hook'	ˈdaquc	duˈqut-an	NA
(b)	'to increase'	ˈuman	ˈmal-an	ˈman-an
(c)	'to seal/close'	ˈsepuy	suˈpuy-an	suˈpuw-an

to analyze /-ani/ as composed of /-an/ 'LOCATIVE FOCUS' and /-i/ 'IMPERATIVE'

2.2 Stress-driven vowel alternations

Seediq stress is always penultimate; suffixation shifts stress rightwards (Yang, 1976), giving rise to alterna-
tions such as ['bunuh∼bu'nuhan] '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; in the following examples, the neutralized segment in
each paradigm is highlighted in grey. First, onsetless pretonic vowels are deleted; 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]*

	STEM	SUFFIXED	UR	GLOSS	
(a)	ˈremux	ruˈmuxan	/remex/	'enter'	(u~u, n=60)
(b)	ˈpemux	puˈmexan	/pemex/	'hold'	(u~e , n=36)
(c)	ˈdoʔus	doˈʔos-an	/doʔos/	'refine' (metal)'	(u~o, n=4)

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*

	STEM	SUFFIXED	UR	GLOSS	
(a)	ˈraŋe	ruˈŋay-an	/ranay/	'play'	(e~ay, n=7)
(b)	ˈsino	suˈnaw-an	/sinaw/	'to drink (alcohol)	(o~aw,n=1)
(c)	ˈdeŋu	duˈŋey-an	/deŋey/	'to dry (food)'	(u~ey, n=12)
(d)	ˈseku	suˈkuw-an	/seku/	'to store'	(u~u, n=13)

Stem-final /Vg/ also neutralizes to vowels, resulting in final vowel alternations. As summarised in (11), /ag/ neutralizes with [o] word-finally (11a), /eg, ug/ both neutralize to [u] (11b), and /ig/ becomes [uy] (11c). These alternations are historically a result of /g/ weakening to [w] word-finally, followed by monophthongization of the resulting diphthong (Li, 1981).

(11) *Alternation of final /g/*

	STEM	SUFFIXED	UR	GLOSS	
(a)	ˈhilo	huˈlag-an	/hilag/	'cover'	(o~ag, n=9)
(b)	ˈlihu	luˈhug-an	/hilug/	'string together'	(u~ug/eg, n=9)
(c)	ˈbaruy	buˈrig-an	/burig/	'buy/sell'	(uy~ig, n=3)

Stem-final [e] only results from monophthongization of /ay/, so final [e] predictably alternates with [ay] in the suffixed form, barring a small subset of irregular alternations. On the other hand, the final [u] of an unsuffixed form could alternate with [ey], [ug], or [eg]. There are also stems where final [u] is non-alternating, as in (10d). Note that in this example, [w] is also inserted to resolve vowel hiatus; hiatus resolution by either glide or glottal stop insertion is a regular and predictable process in Seediq. In other words, the final [u] of an unsuffixed form has multiple possible alternants in the unsuffixed form. Similarly, stem-final [o] has multiple possible alternants in the suffixed form, and can alternate either either [aw] or [ag].[2]

In general, 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 (9a) and (9b) are contrastive in the suffixed form, but both reduce to [u] in the isolation stem.

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).

---

[2]   Note 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 su.rface with phonemic stressed [o] as in (9c). In the current data, only four were found.

(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/*

	STEM	SUFFIXED	UR	GLOSS	
(a)	ˈtatak	tuˈtak-an	/tatak/	'chop'	(k~k, n=19)
(b)	ˈpatak	puˈtap-an	/patap/	'cut'	(k~p, n=6)
(c)	ˈeluk	∅ˈleb-an	/eleb/	'close'	(k~b, n=1)
(d)	buˈcebac	bucuˈbac-an	/bucebac/	'slice'	(c~c, n=1)
(e)	ˈdamac	duˈmat-an	/damat/	'for eating'	(c~t, n=16)
(f)	ˈharac	huˈrad-an	/harad/	'build (a wall)'	(c~d, n=4)
(g)	ˈgilaŋ	guˈlaŋ-an	/gilaŋ/	'mill (rice)'	(ŋ~ŋ,n=32)
(h)	ˈtalaŋ	tuˈlam-an	/talam/	'run'	(ŋ~m, n=3)
(i)	ˈdurun	duˈrun-an	/durun/	'entrust'	(n~n, n=3)
(j)	ˈdudun	duˈdul-an	/dudul/	'lead'	(n~l, n=19)

2.4 Alternations involving doubly suffixed forms

In the LOCATIVE IMPERATIVE, stems are suffixed with /-an-i/. As a result, stress shifts to the first vowel of the suffix, and all of the stem's vowels are neutralized. This is demonstrated in (14). Consequently, doubly suffixed forms always 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	GLOSS
(14)	(a)	ˈgedaŋ	guˈdaŋan	guduˈŋani	/gedaŋ/	'die'
	(b)	ˈpemux	puˈmexan	pumuˈxani	/pemex/	'hold'

2.5 Irregular alternations

In addition to the alternations discussed so far, Seediq has a subset of alternations that are irregular in the sense that they (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 generalizable 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 (15a). In some forms (n=5), stem-final open vowels are deleted in the suffixed form (15b). Examples of other less common irregular alternations are provided in (15c).

(15)   *Verbs showing irregular alternations*

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

STEM	SUFFIXED	GLOSS	EXPECTED SUFFIXED
ˈhuruc	huˈridan	'come to a stop'	(huˈrudan, huˈredan, huˈrodan)
ˈraguh	ruˈwahan	'open'	(ruˈguhan, ruˈgehan, ruˈgohan)
ˈraqic	ruˈqutan	'hook'	(ruˈqitan)
ˈpatis	puˈtasan	'write'	(puˈtisan)
ˈpehiŋ	peˈheŋan	'destroy'	(piˈhiŋan)

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

ˈhado	ˈhadan	'deliver'	(huˈdawan, huˈdagan)
ˈtaŋe	ˈtaŋan	'rely on'	(tuˈŋayan)
ˈkesa	ˈkesan	'tell (someone)'	(kusaʔan[3])

    (c)   *Other irregular alternations*

ˈqerac	quˈrapan	'grab'	(quˈracan, quˈratan, quˈradan)
ˈlawa	laˈan-an	'invite'	(luˈwaʔan)
ˈbege	ˈbiqan	'give'	(buˈgayan)

## 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 learners. This is because, when given just one form of a paradigm (either an unsuffixed 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 unsuffixed forms (which are not affected by pretonic vowel neutralization) and the suffixed forms (which are not affected by post-tonic

---

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

neutralizations). For example, consider the verb [ˈgeruŋ]~[guˈreman] 'to break'. Given this paradigm, the learner can construct a UR /gerem/ which takes its initial vowel from the unsuffixed form, and its final vowel and consonant from the suffixed form; this is illustrated in (16).

(16) *Cobbled UR approach for 'to break'*

UR -------------- /g **e** r **e m** /

SR ------ [ˈg **e** ruŋ]   [guˈr **e m** an]
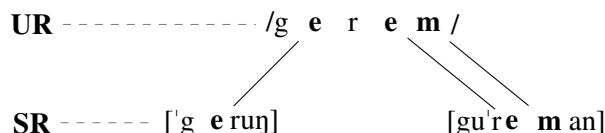
Assuming, then, a cobbled UR approach, the majority of forms in Seediq (excluding irregularly alternating ones) can be derived using phonotactically motivated markedness constraints. Example (17) below demonstrates how this can be done in Optimality Theory (Prince & Smolensky, 1993).

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, LIC(nonperiph/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] in unstressed position. For now, competing candidates are not shown, and the specific patterns of alternation observed (e.g. /m/→[ŋ]) are assumed to result from the interaction of faithfulness constraints.

(17) *Derivation of* [ˈgeruŋ] *under a cobbled UR approach*

/gerem/	*m]$_w$	LIC-NONPER	ID-LAB	ID[back]
a. ˈgerem	*!	*		
b. ˈgerum	*!			*
c. ˈgereŋ		*!	*	
☞ d. ˈgeruŋ			*	*

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 unsuffixed 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 unsuffixed form, suffers from neutralization. For example, if the isolation stem (unsuffixed) form were the base, the grammar would need to somehow 'undo' final consonant neutralization, which is impossible do 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 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 unsuffixed [ˈgeruŋ], and project the suffixed form [guˈruŋan]. On the other hand, if the learner hears the suffixed form [guˈreman], they might posit the UR /gurem/, and infer that the isolation stem is [ˈguruŋ]. 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 unsuffixed slots of a paradigm, but in principle, any unsuffixed 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, 2002b; 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 Section 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.
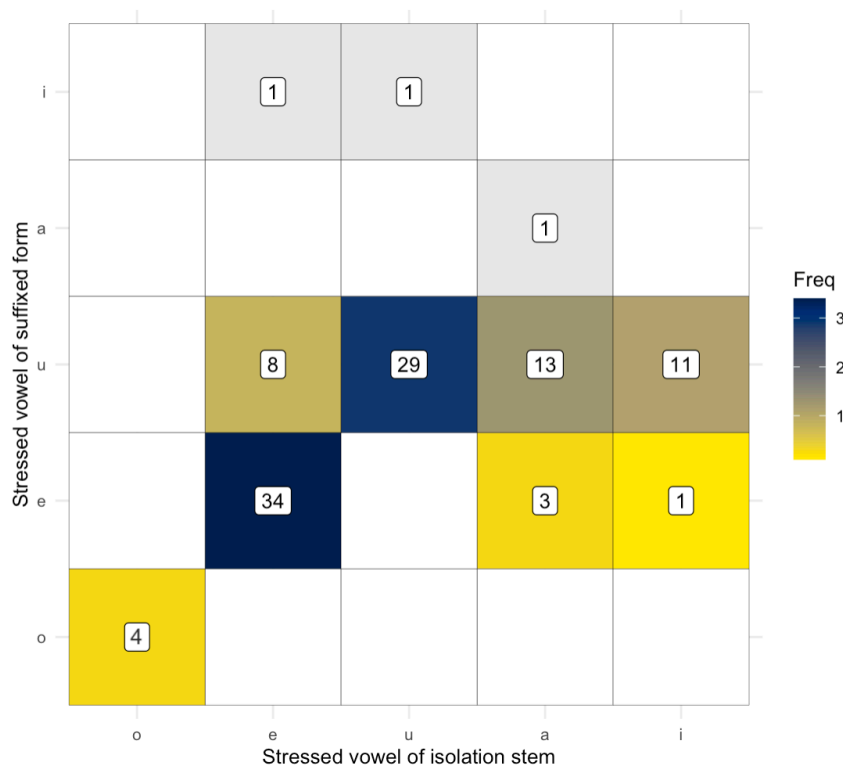

Figure 1: How the reduced [u] of unsuffixed CVCuC is realised when stressed under suffixation

Specifically, there is a tendency for VOWEL MATCHING, where the stressed vowel of the suffixed form

10

'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] alternates with [a] or [i] (instead of [e] or [o]). For example, the stem-suffix pair [ˈhuruc]∼[huˈrid-an] 'to stop suddenly' contains an irregular post-tonic [u]∼[i] alternation. These are shown in the top two rows of Fig. 1.

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], $p < 0.001$ ($= 2.8 \times 10^{-8}$).

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, with relatively high accuracy, "undo" the neutralization of post-tonic [u], and predict what it will surface as in the suffixed form.

To get a general measure of the 'predictability' of the stem stressed vowel from the suffixed form's stressed vowel, we can look at each column in Fig. 1, and select the majority variant, or the cell with the largest number of forms. For example, when the stem stressed vowel is [e], the post-tonic [u] is most likely to surface as [e]. This is is true for 34/(34+8+1) = 79% of relevant forms.

Table 4 summarizes the proportion of forms predicted by picking the majority variant for each column in Fig. 1. The column named MAJORITY shows the number of forms which follow the majority pattern, and the rightmost column shows the proportion of forms correctly predicted by picking his pattern. As a whole, by applying this principle and picking the majority variant in each vowel condition, we can correctly predict the stressed vowel of the suffixed form for 87% of CVCuC forms.

STEM	MAJORITY	TOTAL	% CORRECT
CoCuC	4	4	100%
CeCuC	34	43	79%
CuCuC	29	30	97%
CaCuC	14	17	82%
CiCuC	11	12	92%
Overall	**92**	**106**	**87%**

Table 4: Predictability of the suffixed form's stressed vowel from CVCuC isolation stems

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 Section 2.3 and summarised in (18). 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.

(18) STEM   SUFFIXED

    [c] ∼ [t, d, c]

    [k] ∼ [p, b, k]

    [n] ∼ [l, n]

    [ŋ] ∼ [m, ŋ]

However, final consonants tend to either almost always or almost never alternate, as summarised in Table 5. Final [ŋ] almost never alternates with [m]; the hypothetical stem [ˈpatiŋ] will surface as [puˈtiŋ-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.

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

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]).

FINAL C	MAJORITY	TOTAL	% CORRECT
c	16	21	76%
n	19	26	73%
k	19	27	70%
ŋ	34	36	94%
Overall	**88**	**110**	**80%**

Table 6: Predictability of the suffixed form's stressed vowel from CVCuC isolation stems

In general, given an unsuffixed form with a neutralized final consonant, the alternant that surfaces in the suffixed form is actually highly predictable. 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. Table 6 summarizes the proportions of forms that speakers will correctly guess by picking the majority alternant

for each neutralized final consonant. From here, we see that 80% of final consonant neutralizations can be accounted for using this strategy.

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 unsuffixed stem. When given a novel form [puˈtim-an], a Seediq speaker should know with certainty that the isolation stem form will end in [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 [puˈtis-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 results in alternations such as [ˈuyas]∼[ˈyas-an], where the suffixed form surfaces with a monosyllabic stem. As a result, for a novel suffixed form like [ˈtis-an], where the stem appears as monosyllabic, speakers must also predict whether the unsuffixed stem is monosyllabic (e.g. [ˈtis]∼[ˈtis-an]), or if there is an initial vowel ([ˈatis]∼[ˈtis-an]). Because underlyingly monosyllabic stems are rare (n=4 in the current corpus), I exclude them from the following discussion.

In the case of post-tonic vowel alternations (Section 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. 2.

Fig. 2 compares the stressed vowel of the suffixed form against the stressed vowel of the isolation stem; this figure shows the distribution of vowels across all disyllabic stems in the lexicon. There is some predictability between the stressed vowel of the suffixed form and the stressed vowel of the isolation stem forms. However, these trends appear to be relatively weak compared to the patterns observed in Section 3.1.1, which looked only at contexts in which post-tonic /e,o/ are reduced (i.e. stems of the form CVCuC).

Once again, we can get a general measure of the 'predictability' of the stem stressed vowel from the suffixed form's stressed vowel, by selecting the majority variant for each column in Fig. 2. For example, looking at the rightmost column of Fig. 2, 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 50 out of 125 (50+18+22+33+2) verbs where the suffixed form's stressed vowel is [a]. In other words, given a suffixed form [puˈtas-an], a speaker could pick the majority variant, and infer that the isolation stem form is [ˈpatas]. This choice predicts the correct output 40% (50/125) 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 [puˈtisan], the stem form is most likely to be [ˈpetis]. Applying this principle, we can correctly predict the stem stressed vowel for 39% (24/62) of relevant forms.

Table 7 summarizes the proportion of forms predicted using this method. 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
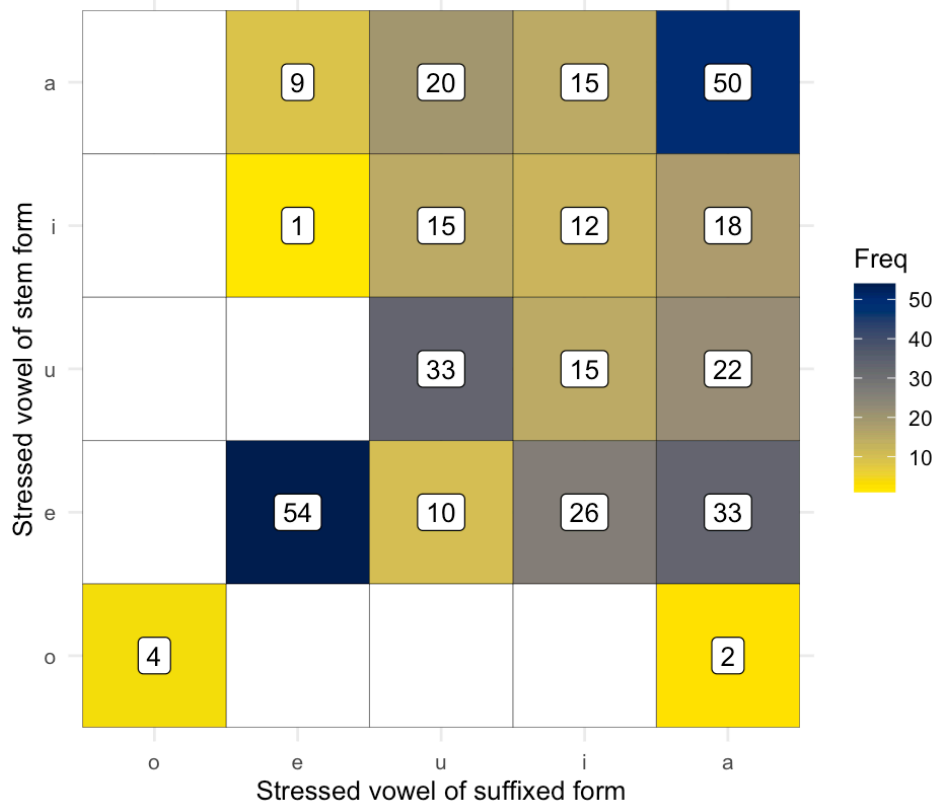
Figure 2: How the pretonic [u] of suffixed forms is realized when stressed in the isolation stem

only predicts the correct vowel 49% of the time.

Suff V	Predicted	Total	% correct
/o/	4	4	100%
/e/	54	64	84%
/u/	33	78	42%
/i/	26	68	38%
/a/	50	125	40%
Overall	**167**	**337**	**49%**

Table 7: Predictability of stem stressed vowels from suffixed forms in disyllabic verbs

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; 337 verbs (the entire corpus, excluding 4 monosyllabic stems) are affected. In contrast, post-tonic vowel reduction affects 106 verbs, and final consonant neutralization affects 110 verbs.

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. In this section, I explicitly demonstrates the asymmetry using the Minimal Generalization Learner algorithm (MGL; Albright, 2002b; Albright & Hayes, 2003). The MGL is a surface-base model which learns surface mappings between inflected forms. 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. I use the MGL to learn both a grammar that maps from isolation stem to suffixed forms, and one that maps from suffixed to stem forms. The two models are then assessed for how accurately they predict the Seediq lexicon.

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.

4.1   The MGL algorithm

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 & Hayes (2003, p. 123-128).

For example, in learning stem→suffixed mappings, the model could compare pairs the two inputs in (19a) to learn a rule of suffixation after non-continuant dorsals. Including an additional input [ˈberas]∼[buˈrasan] would result in the model learning a more general rule (19b). Eventually, consideration of a broader range of forms would result in a general suffixation rule $\emptyset \rightarrow$ **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.

(19)   *Rule generalization in the MGL (stem-to-suffix mapping)*

input	**rule**
a. ˈbeliŋ∼buˈliŋan, ˈbetaq∼buˈtaqan	$\emptyset \rightarrow$ an / [+DORSAL,-continuant]__
b. ˈbeliŋ∼buˈliŋan, ˈbetaq∼buˈtaqan, ˈberas∼buˈrasan	$\emptyset \rightarrow$ an / [-continuant]__

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 [ˈkayak]∼[kuˈyapan] 'to cut'. The general stem-to-suffixed mapping of $\emptyset$→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 the more general rule.

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, 2002b). 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, the stem→suffixed grammar learns three rules that apply to the input [ˈbirac]. These rules, given in (20), produce three output candidates of varying confidence.

(20)   *Examples of rules learned by the MGL in a stem→suffix mapping*

input	output	rule	confidence
ˈbira**c**	buˈra**c**an	$\emptyset \rightarrow$ an / __	0.65
	buˈra**t**an	c $\rightarrow$ tan / __	0.67
	buˈra**d**an	c $\rightarrow$ 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 341-word corpus. The informativeness of a paradigm slot was taken to be its accuracy in predicting the other slot. Specifically, for each word in a mapping, the generalization with the highest reliability was taken to be the model's prediction. These predictions were compared against to lexicon, to get the proportion of forms correctly predicted by the grammar.

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 & Wilson, 2008). A stem-suffix mapping is judged as being correctly predicted by the model only if both the segmental grammar and vowel-tier grammar predict the correct output.

4.3   Results: comparing accuracy of different mappings

Table 8 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.

Overall, the stem-to-suffix mapping is more more accurate, accounting for 72.7% of the lexicon (vs. 41.3% in the suffix-to-stem mapping). The discrepancy between the two mappings is largely because pretonic

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

Table 8: Model results: accuracy of mappings between Seediq stem and suffixed forms

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.

4.4 Predictions for reanalyses and novel forms

Table 9 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 MGL does not learn patterns that are observed in only one form. As a result, it does not learn the [k]-[b] alternation.

	rule	scope	hits	reliability	confidence
R1.	∅ →an	341	230	0.67	0.65
R2.	o→agan	13	11	0.85	0.75
R3.	n→l/i__	8	7	0.88	0.74
R4.	c→tan	23	17	0.74	0.67
R5.	ŋ→m	35	2	0.06	0.04
R6.	c→dan	23	4	0.17	0.13
R7.	k→pan	28	7	0.25	0.20

Table 9: Examples of rules learned in the segmental stem-base grammar

Overall, the model also learns the same vowel matching generalizations described in §3. Table 9 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.

17

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

Table 10: 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 Section 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 (i.e. [m] was already less frequent than [ŋ]). 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 (21), 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].

(21)　　ˈlauŋ∼luˈuŋan　(<*l-um-aum)　'to burn'
　　　　　　　(Li, 1981; Greenhill et al., 2008)

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 distributions of 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 & Albright, 2017; Ernestus & Baayen, 2003). For instance, when a Seediq speaker hears [ˈɡeruŋ], they would first posit its UR to be /ɡereŋ/ because Seediq stems with underlying /e/ in V1 position are most likely to have underlying /e/ in V2 position. From the UR /ɡereŋ/, they could then infer the suffixed form [ɡuˈreŋan].

Reanalyses should only arise in alternating environments. Given a disyllabic stem, this means that reanalyses from isolation stems, on the basis of lexical frequencies, should *only* 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 preserved because the corresponding stem form [ˈCeCaC] has a non-neutralized post-tonic vowel that is not vulnerable to reanalysis.

Put another way, renalysis from the stem would result in overrepresentation of URs with /e-e/, /o-o/, and /u-u/ (where /e-e/ is shorthand for /CeCeC/ stems, and likewise for the other vowel-vowel sequences). Vowel sequences such as /e-u/, /u-e/ will be underrepresented in URs, as these are the ones that are vulnerable to reanalysis. However, sequences such as /e-a/ and /e-i/ would not be underrepresented, but are instead expected to occur roughly at chance level. Finally, sequences like /i-i/ and /a-a/, which obey vowel matching but aren't expected as the output of reanalysis, should also occur roughly at chance level.

On the other hand, reanalyses from the suffixed form affects all stem-suffix pairs, as V1 is always neu-tralized in the suffixed form. Reanalysis from the suffixed form on the basis of lexical frequencies should therefore affect more vowel contexts. Table 11 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 an increase of the vowel sequences /a-a/, /e-e/, /e-i/, /o-o/, /u-u/ in URs. Ad-ditionally, sequences such as /a-u/ and /i-u/, which go against these statistical tendencies, are expected to be underrepresented.

V2	V1	%	UR vowels
a	a	0.38	/a-a/
e	e	0.83	/e-e/
i	e	0.37	/i-e/
o	o	0.75	/o-o/
u	u	0.42	/u-u/

Table 11: 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

Under the cobbled UR approach, where reanalyses from both stem and suffixed forms are possible, re-analyses should collectively result in vowel matching across most vowel contexts. In contrast, the surface-base approach predicts that reanalysis will happen from only the unsuffixed stem, and therefore increase vowel matching in more limited contexts. A subset of the vowel sequences predicted to be overrepresented or underrepresented under each approach are summarized in Table 12.

/V1-V2/	COBBLED UR	SURFACE-BASE
a-a	+	●
e-e	+	+
i-i	●	●
o-o	+	+
u-u	+	+
e-u	—	—
e-i	+	●
e-a	—	●
u-e	—	—
u-i	—	●
u-a	—	●

Table 12: Predicted UR vowel distributions in cobbled UR vs. surface-base approaches ('+'=overrepresented, '−'=underrepresented, '●'=at chance)

In order to determine whether certain vowel combinations are significantly more or less common than chance, I use a Monte Carlo procedure (Kessler, 2001; Martin, 2011) to approximate the distribution of the expected rate. Specifically, V1 and V2 of cobbled URs (constructed from the corpus of 341 stem-suffix pairs) were randomly recombined (vowel position is fixed, so that V1 is always sampled from initial vowels, and V2 from final vowels). After each such shuffling, the number of the target V1V2 pair under the new

permutation is recorded. This process was repeated 10,000 times, to give a reliable estimate of the expected (chance value) times the target V1V2 pair should occur.

Fig. 3 shows, for each V1V2 sequence, the 95% confidence interval for chance (found using the Monte Carlo procedure), plotted against the actual number of occurrences in the lexicon. If the actual value is smaller than the 95% confidence interval, then the target V1V2 sequence is underrepresented. Conversely, if the actual value is greater, than it is overrepresented. These values are also summarized and compared against the predictions of different UR structures in Table 13.
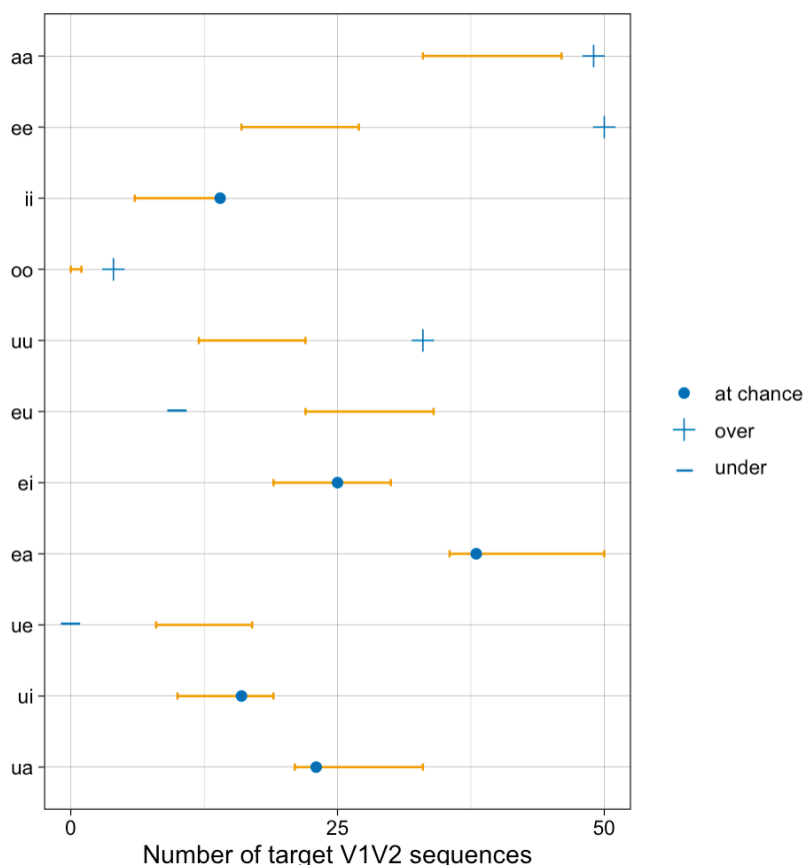
Figure 3: Comparing attested V1V2 pairs to Monte Carlo results.

Overall, in almost all cases, the results of the Monte Carlo simulation are exactly in line with the predictions of the surface-base approach. For example, /e-e/ stems are strongly overrepresented, and /e-u/ stems are underrepresented. In contrast, /e-i/ and /e-a/ sequences both occur at chance levels. This is the distribution that is expected to arise under reanalysis from the isolation stem, where [ˈCeCuC] stems are reanalyzed, while [ˈCeCaC] and [ˈCeCiC] stems are not. The COBBLED UR approach, on the other hand, would predict that sequences such as /e-a/ are also underrepresented, as they would be dispreferred under reanalysis from the suffixed form.

The only place where the results deviate from the predictions of the surface-base approach are the /a-a/ sequences, which are slightly overrepresented in the lexicon. In this case, the number of attested /a-a/ sequences is still relatively close to the distribution at chance level.

In summary, vowel-vowel sequences are over- and underrepresented specifically in environments that

/V1-V2/	COBBLED UR	SURFACE-BASE	RESULT
a-a	+	•	+
e-e	+	+	+
i-i	•	•	•
o-o	+	+	+
u-u	+	+	+
e-u	−	−	−
e-i	+	•	•
e-a	−	•	•
u-e	−	−	−
u-i	−	•	•
u-a	−	•	•

Table 13: Predicted UR vowel distributions in against Monte Carlo results ('+'=overrepresented, '−'=underrepresented, '•'=at chance)

would arise under reanalysis from the **surface stem**, but not from the suffixed form. Under a cobbled UR approach, these asymmetries would be arbitrary. On the other hand, these results provide indirect evidence that reanalyses are overwhelmingly from the stem, rather than from the suffixed form. This directionality falls out naturally from the surface-approach, if the isolation stem is the designated base.

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 Section 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.

The experimental methodology adopted was a modified version of a nonce-word task (i.e. wug test; Berko, 1958). Specifically, instead of nonce words, the current study uses 'gapped forms', or stems with no known suffixed forms. This was done out of respect for my participants, who raised concerns that the use of nonce words in experiments would interfere with ongoing language revitalization efforts.

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 & Baayen, 2003, and many others).

In the rest of this section, I summarize the predictions and results of the experiment. Detailed description of the experiment procedure can be found in Kuo (2020, 2022).

6.1 Predictions

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 14. 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.

Stimuli where V1 was /i,o/ were not included for several reasons. First, I limited the conditions tested to keep the experiment under two hours. In addition, there were also relatively few words where V1 was [i] or [o], and almost all of these were found to be non-suffixable during a pilot study.

Predicted responses, on the basis of the rule-based model in §4, are summarized in Table 14; 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, non-alternation should be preferred, while [u]∼[e] alternation is present as a minority pattern. 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	ˈdaruk (duˈruk-an) 'oil'
e	u	[u]∼[e] alternation	ˈkeruŋ (kuˈreŋ-an) 'wrinkles'
u	u	never alternate	ˈcuɡuk (cuˈɡuk-an) 'Bidens plant'
a	a		ˈsabak (suˈbak-an) 'dregs, pulp'
e	a	V2 never alternates	ˈrehak (ruˈhak-an) 'seed'
u	a		kuˈsuwak (kusuˈwak-an) 'yawn'

Table 14: Experimental conditions: vowel alternations

6.2 Results

Results for final vowel alternations are given in Fig. 4, which shows the proportion of response types by vowel condition. The distribution of vowels in the lexicon is given on the bottom column for reference. 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 consistent with predictions, final /a/ almost never alternates. There was one exception ([huˈrenaŋ∼huruneɲi]); in this case, an [a∼e] alternation resulted in vowel matching. On the right-hand column, in line with predictions, CeCuC stems prefer the [u]∼[e] alternation, and CuCuC stems never alternate.

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 novel [u∼a] alternation (e.g. [ˈdaruk∼duˈrak-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.
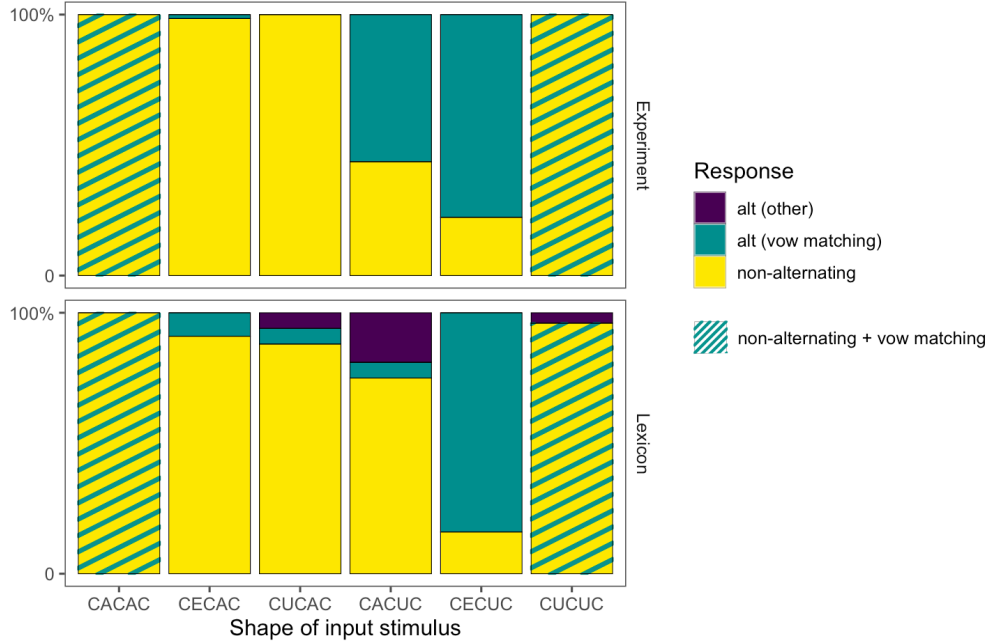
Figure 4: Vowel alternation rates in experiment vs. lexicon

6.3 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]. In Section 7.3, I explore this possibility, and model pattern over-extension as the result of a bias towards learning more general patterns.

Note that my 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 A constraint-based model of Seediq vowel alternations

§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 & Johnson, 2003), a stochastic variant of Optimality Theory (Smolensky, 1986; Prince & Smolensky, 1993).

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 log-likelihood. To learn model constraint weights, I use Excel's Solver add-in (Generalized Reduced-Gradient Algorithm; Fylstra et al., 1998).

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).

The rest of this section is organized as follows. Section 7.1 will introduce a base-driven analysis of Seediq vowel alternations, fit to the lexicon. In Section 7.3, I propose to model speakers' overextension of vowel matching in the experiment as the result of a bias towards simpler, less complex grammars. Specifically, overextension is modeled using a 'grammatical leakage' account (Martin, 2011).

7.1 A base-driven model of vowel alternations in the Seediq lexicon

This section introduces the constraints needed under a base-driven approach to Seediq post-tonic vowel alternations. The following section (7.2) will then present weights learned algorithmically in models trained on the lexicon.

7.1.1 Inputs and faithfulness

Following Albright (2002a,b, 2010, et seq.), I assume that the input to the model is the base form, in this case the isolation stem. Essentially, inputs are segmentally identical to isolation stems, but are under-specified for stress (as stress is non-phonemic). For example, given a stem-suffix pair like [ˈpukuc], the input to the model is /pukuc/. When deriving the suffixed form, candidates are in both an input-output (IO) correspondence relationship with these URs (e.g. /pukuc/), and an output-output (OO) correspondence relationship with the surface unsuffixed stems (e.g. [ˈpukuc]).

Additionally, inputs were schematized into 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. [ˈqene]) 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. [ˈatak]), ii) CVHVC stems (where H is [h] or [ʔ]),

and iii) CVCVC stems. Idealized training data was used for ease of interpretation, and so that results could be compared with experimental data.

The effects of IO-Faithfulness and OO-Faithfulness are demonstrated in tableau (22), which shows how suffixed forms are derived for the hypothetical input /petus/.[4] The constraints enforcing vowel alternation will be introduced in Section 7.1.3; for now, it is written as MatchV. Candidate (b), which undergoes vowel alternation to satisfy MatchV, violates faithfulness constraints, written in the tableau as IO-Faith and OO-Faith. However, because both faithfulness constraints have lower weight than the constraint which enforces vowel matching, candidate (b) is still preferred.

(22) *IO-faithfulness*

/petus-an/ ∼[ˈpetus]	P	$\mathcal{H}$	OO-MatchV	IO-Faith	OO-Faith
			7	2.5	2.5
a. puˈtus-an	0.12	7	1		
b. puˈtes-an	0.88	5		1	1

In this tableau, IO and OO-Faithfulness constraints have the same violation profile. In general, because all IO and OO-Faithfulness constraints have the exact same violation profile, I will omit IO-Faithfulnes from future tableau.

7.1.2 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] can be motivated by prominence alignment, where vowels reducing to less sonorous variants in non-prominent positions (Kenstowicz, 1994; Crosswhite, 2000). However, a sonority-driven reduction account does not explain why /i/ and /e/ reduce to [u]. In fact, Seediq vowel reduction to [u] is saltatory (White, 2013; Hayes & White, 2015), because /e/ reduces to [u] instead of the phonetically closer (and equally unmarked) [i]. Barnes (2002) speculates that the synchronic pretonic reduction pattern originated from a transparent process of pretonic vowels merged to a central high vowel. The pattern was obscured by subsequent phonetic rounding and backing of the reduced vowel to [u].

Saltatory alternations are problematic in parallel OT (Łubowicz, 2002; Ito & Mester, 2003; Hayes & White, 2015); even though pretonic neutralization to [u] is consistent with a prominence alignment account, it cannot be straightforwardly analyzed using standard markedness and faithfulness constraints. As such, I adopt a parsimonious constraint that describes the reduction pattern. This constraint, License[u]/pretonic (Lic[u]/pret), is defined in (23), and essentially penalizes non-[u] syllables in pretonic position. This is demonstrated in tableau (24) for [ˈbarah]∼[buˈrahan].

---

[4] Although it is not discussed here, both candidates (a) and (b) undergo pretonic reduction.

The faithful candidate (a) fatally violates the highly weighted L_IC[u]/pret, and is assigned near-zero probability. Candidate (b), which repairs the L_IC[u]/pret violation by incurring a violation of OO-I_DENT[high] (and other identity constraints), 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-weighed faithfulness constraints.

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

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

/barah-an/∼[ˈbarah]	P	$\mathcal{H}$	Lic[u]/pret	M_AX-V	M_AX-C	OO-ID[high]
			20	10	10	1
a. baˈrah-an	0	20	1			
b. buˈrah-an	1	1				1
c. ˈrah-an	0	20		1	1	

7.1.3   Constraints for post-tonic vowel alternations

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

I propose that speakers have learned both a specific constraint enforcing vowel matching in mid vowels, and a more general constraint enforcing vowel matching across all vowel categories. The choice to include the more general constraint is motivated by the experimental results, where speakers extended the vowel matching pattern, applying an innovative [u]∼[a] alternation to CaCuC stems (Section 6.2).

Both the specific and general constraints are formalized using prosodic correspondence (Crosswhite, 1998).[5] These are constraints which enforce identity between prosodic elements of related output forms, rather than linearly related segmental units. In particular, I use the two constraints N_UC-I_DENT-OO(M_ID-V) and N_UC-I_DENT-OO(V), to enforce vowel matching. These constraints, defined in (25), set up a correspondence relation between the stressed syllable nuclei of related output forms, and requires that these positions have the same vowel.

(25)   a.   N_UC-I_DENT-OO(M_IDV): For α, a stressed **mid-vowel** nucleus of the base, and β, a stressed nucleus of an output, where α corresponds to β, α and β must be the same.

   b.   N_UC-I_DENT-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.

I_DENT constraints typically reference feature specifications. For Seediq, the constraints enforce total identity. This approach builds on evidence that total identity is distinct from partial identity, and that constraints which enforce total identity are necessary (Coetzee & Pater, 2008; Gallagher & Coon, 2009).

---

[5]   In Kuo (2022), I go into a more detailed discussion of why Seediq vowel matching should be analyzed as prosodic correspondence.

Tableau (26) demonstrates how the two vowel matching constraints are used to derive outputs that match lexical frequencies.[6] The model learns a much higher weight for Nuc-ID(MiDV) than for Nuc-ID(V), reflecting how in the lexicon, vowel-matching alternation only happens for CeCuC and CoCuC stems. For the input /putus/, the faithful candidate satisfies Nuc-ID(V), and receives high probability. For /petus/, the high weight of Nuc-ID(MiDV) relative to competing faithfulness constraints causes candidate (e), which undergoes [u]~[e] alternation, to be preferred over the faithful candidate. Candidate (f) is ruled out because vowel alternation does not resolve violations of Nuc-ID(MiDV). For /patus/, candidate (j), which undergoes [u]~[a] alternation to resolve Nuc-ID(V) violations, is ruled out because the weight of Nuc-ID(V) is much lower than that of competing faithfulness constraints. As will be addressed in **??**, this differs from the experiment, where speakers did in fact apply the [u]~[a] alternation to around 50% of CaCuC forms.

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

	Obs	P	$\mathcal{H}$	Nuc-ID(MiDV) 5.18	Nuc-ID(V) 0.66	*P/i,u 3.89	ID[low] 0.99	ID[high] 5.98	ID[front] 2.20
/putus-an/~['putus]									
a. pu'tusan	1	0.93	3.89			1			
b. pu'tesan	0	0.01	8.84		1			1	1
c. pu'tosan	0	0.06	6.65		1			1	
/petus-an/~['petus]									
d. pu'tusan	0.20	0.18	9.68	1	1	1			
e. pu'tesan	0.80	0.81	8.15					1	1
f. pu'tosan	0	0.02	12.10	1	1			1	
/potus/~['potus]									
g. pu'tusan	0	0.04	9.68	1	1	1			
h. pu'tosan	1	0.96	6.40					1	
i. pu'tesan	0	0	13.84	1	1			1	1
/patus/~['patus]									
j. pu'tusan	0.76	0.90	4.09		1	1			
k. pu'tesan	0.18	0.01	8.26		1			1	1
l. pu'tosan	0	0.08	6.51		1			1	
m. pu'tasan	0.06	0	9.42				1	1	1

---

[6] The constraint *P/i,u will not be introduced until the following section, but is not crucial to this tableau.

7.1.4 Saltation in post-tonic vowel alternations

In tableau (26) above, the model generally over-predicts rates of [u]∼[e] alternation, while over-predicting rates of [u]∼[o] alternation. For the input /putus/, the model assigns the [u]∼[o] alternating candidate (c) a probability of 0.06, whereas in the lexicon this candidate is never observed. Similarly, for /patus/, the model assigns too much probability (P=0.08) to [puˈtosan], and too little probability to [puˈtesan] (P=0.01).

This issue arises because, similar to pretonic vowel reduction, post-tonic vowel alternations in Seediq are saltatory, in that post-tonic [u] prefers to alternate with [e] instead of the phonetically closer [o]. This results in a constraint weighting conflict; [u]∼[e] alternating candidates will always incur more faithfulness constraint violations than [u]∼[o] alternating forms, and therefore (when all else is held equal) be assigned lower probability.

To resolve this conflict, I use a constraint *[o], which simply penalizes surface [o]. Although this constraint is not rooted in a clear markedness motivation, it does capture fact that [o] is marginal in the Seediq lexicon, and occurs at a much lower frequency than the other vowels. An alternative approach to dealing with saltation, which Hayes & White (2015) adopt, is to introduce the *MAP family of faithfulness constraints Zuraw (2010, 2013). *MAP are less restrictive than classical faithfulness constraints, and can describe the correspondence between any two natural classes of sounds, even when the two classes differ in more than one feature. Using *MAP constraints, it is therefore possible to have both a constraint penalizing a change of [u]∼[e], and a constraint penalizing a change of [u]∼[o]. I explore this approach for Seediq in Kuo (2020).

The effect of *[o] is demonstrated for /putus/ and /patus/ in tableau (27), which is identical to tableau (26), other than the addition of *[o]. For both inputs, the [u]∼[o] alternating candidate receives much lower weight because they incur violations of *[o].

(27) *Tableau: Effect of *o on /putus/ and /patus/.*

		Obs	P	$\mathcal{H}$	Nuc-ID(MɪᴅV) 3.80	Nuc-ID(V) 0.69	*P/i,u 0.49	*o 3.86	ID[low] 0.67	ID[high] 0.32	ID[front] 3.04
/putus-an/∼[ˈputus]											
a.	puˈtesan	0	0	7.85	1	1				1	1
b.	puˈtusan	1	1	0.49			1				
c.	puˈtosan	0	0	8.35	1	1		1		1	
d.	puˈtasan	0	0	8.52	1	1			1	1	1
/patus-an/∼[ˈpatus]											
e.	puˈtusan	0.76	0.84	1.18		1	1				
f.	puˈtesan	0.18	0.06	3.85		1				1	1
g.	puˈtosan	0.00	0.04	4.17		1		1		1	
h.	puˈtasan	0.06	0.06	3.83					1	1	1

7.1.5 Explaining asymmetries in vowel alternation

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-ID(V).

There are multiple possible explanations for this; speakers could, for example, have learned a source-oriented generalization that only post-tonic [u] can alternate Becker & Gouskova (2016). However, source-oriented generalizations are difficult to capture in OT using standard markedness constraints, which target the output form. Instead, they require positing additional mechanisms. For example, highly ranked *directional* faithfulness constraints could be introduced to the grammar, to protect post-tonic /a/ and /i/ from alternating with [u] (but would not, for example, prevent /u/ from alternating with [a]) .

As an alternative, I analyze the directionality in vowel alternations 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. This approach captures the generalization that Seediq vowel alternations are prominence-aligning; pretonic vowels raise to the less sonorous [u], while stressed vowels prefer to be more sonorous. Typologically, a preference for sonorous vowels in foot-peaks (i.e. main-stressed syllables) is also well-attested, and has been observed in languages (e.g. Crosswhite & Jun (2001) on Zabiče Slovene and Chung (1983) on Chamorro.

Kenstowicz (1994) formalizes this preference for sonorous vowels in foot-peaks using a family of constraints, where less sonorous vowels are relatively more constrained from appearing in stressed positions. These constraints are given in (27); they penalize certain vowels in word-peaks, and are argued to follow a universal ranking hierarchy. In a weighted constraint model like MaxEnt, this means that a constraint like *P/ə should always have higher weight than *P/e,o. I will adopt this approach, and specifically use the constraint *P/i,u.[7] Other foot-peak constraints such as *P/a are assumed to be in the grammar, but are not active because of their relatively low weight.

(28) *Universal hierarchies and rankings for foot-peaks (Kenstowicz, 1994).*

    a. Hierarchy for foot peaks: a >e,o > i,u >ə

    b. Constraint formulation of*P/*x*: assign a violation to every vowel *x* that is in a foot peak (i.e. the nucleus of a stressed syllable)

    c. Constraint ranking: *P/ə ≫ *P/i,u ≫ *P/e,o ≫ *P/a

The effect of *P/i,u is demontrated in tableau (29), which compares the model's predictions of /patus/ and /putas/. Note that although *P/i,u is needed to explain experimental results, it has a very small effect in a lexicon-trained model. This is because, excluding some irregularities, only [u] alternates post-tonically. Consequently, there is little evidence to disambiguate between alternations which improve or reduce the stressed syllable's sonority.

Nevertheless, the model does learn a slight preference for [u]∼[a] alternation over [a]∼[u] alternation. For the input /patus/, candidate (d), which undergoes [u]∼[a] alternation, receives some probability (P=0.06) because alternation reduces violations of both Nuc-ID(V) and *P/i,u. In contrast, for the input /putas/, candidate (a), which undergoes [a]∼[u] alternation to resolve the violation against Nuc-ID(V), receives zero probability. The reason is that in this case, alternation actually increases violatons of *P/i,u.

---

[7] More sophisticated models of gradient constraints on syllable weight are explored in work like Flemming (2001) and Ryan (2011).

(29) *Effect of *P/i,u on vowel alternation*

					Nuc-ID(V)	*P/i,u	*[o]	ID[low]	ID[high]	ID[front]
					0.90	0.98	4.04	0.48	0.00	3.56
/patus/∼[ˈpatus]										
a.	puˈtusan	0.76	0.84	1.18	1	1				
b.	puˈtesan	0.18	0.06	3.85	1				1	1
c.	puˈtosan	0.00	0.04	4.17	1		1		1	
d.	puˈtasan	0.06	0.06	3.83				1	1	1
/putas/∼[ˈputas]										
e.	puˈtusan	0	0.00	5.02		1		1	1	1
f.	puˈtesan	0	0.02	4.95	1			1		1
g.	puˈtosan	0	0.02	4.94	1		1		1	
h.	puˈtasan	1	0.95	0.90	1					

Notably, this prosodic alignment predicts that sonority of the alternating vowel will affect how readily speakers extend vowel matching. For one, speakers should 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-ID(V). However, because alternation of [a] to [i] increases violations of *P/i,u, [a]∼[i] alternation should occur at a lower rate than [u]∼[a] alternation did for CaCuC stems. Predictions like this are hard to test using just corpus data, as post-tonic [i] and [a] generally never alternate in the lexicon. However, they could potentially be tested in future experimental work.

7.2 Model fit to lexicon and experimental results

The constraints presented in Section 7.1 were used to train a MaxEnt model. All markedness constraints were confirmed to be significant using likelihood ratio tests (Wasserman, 2004; Hayes et al., 2012). Additionally, all relevant faithfulness constraints were included even if they ended up testing non-significant.

The vowel alternation model has 7 constraints and a log-likelihood of -69.4 (null=-273.6). Fig. 5 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 ($r^2 = 0.97$). Additionally, the optimal weights assigned to the training data are shown in (30). From here, we see that the model has learned a much higher weight for Nuc-ID(MidV) relative to Nuc-ID(V); this matches the lexicon, where CeCuC and CoCuC forms undergo vowel-matching alternations, while other forms don't.

(30) *Constraint weights*

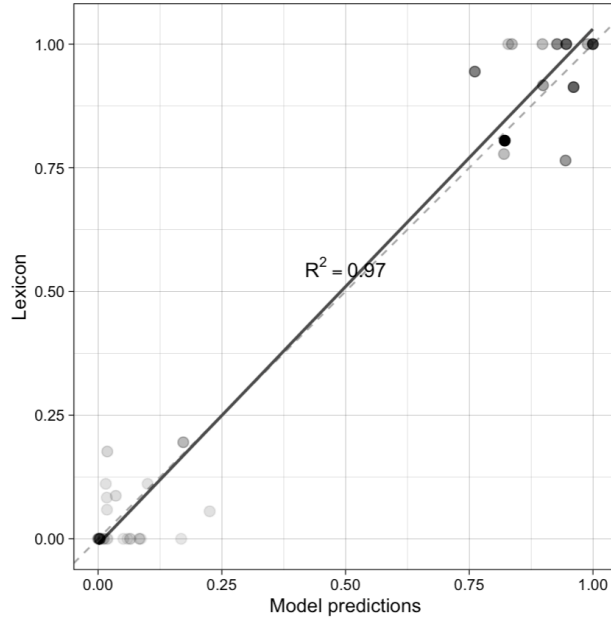| Nuc-ID(MidV)	5.33	*[o]	4.71
Nuc-ID(V)	0.17	ID[high]	2.10
*P/i,u	0.72	ID[low]	0.20
ID[front]	2.55		

Figure 5: Predicted vs. lexical probabilities of the vowel alternation model. Darkness of each point indicates frequency (black=highest freq). Solid lines are fitted regression lines; dotted line plots the line of best fit.

Although the model is able to closely match the lexicon, it does less well in matching the experimental results. In the experiment, speakers over-extended the vowel-matching alternation to CaCuC forms, resulting in alternations such as [ˈpatus]~[puˈtasan]. The model is not able to account for this; when fit to the epxerimental results, it severely under-predicts rates of [u]~[a] alternation. This is demonstrated in Fig. 6, which compares the fit of the model to the lexicon and to experimental results. As seen in Fig. 6b, the model mostly achieves a close to fit to experimental results, but makes the wrong predictions for /patus/. In the experiment, speakers produced the [u]~[a] alternating form around 56% of the time, but the model predicts that this candidate only occurs 6% of the time.

7.3   Explaining experimental results as a complexity bias

In the experimental results, speakers successfully learned generalizations which made alternations more predictable from a stem base. At the same time, speakers over-extended vowel matching beyond the environments observed in the lexicon. In this section, I discuss how the model above could be modified to learn over-extension of vowel matching.

In general, non-viridical learning may arise from different types analytic bias (Moreton, 2008). I propose specifically that the Seediq experimental results follow from a complexity bias, where learners preferentially learned the more general constraint Nuc-ID(V), over the more complex constraint Nuc-ID(MidV) (that is specific to only mid vowels). This analysis would be in line with a body of work suggesting that people preferentially learn simpler constraints (Pycha et al., 2003; Moreton & Pater, 2012)
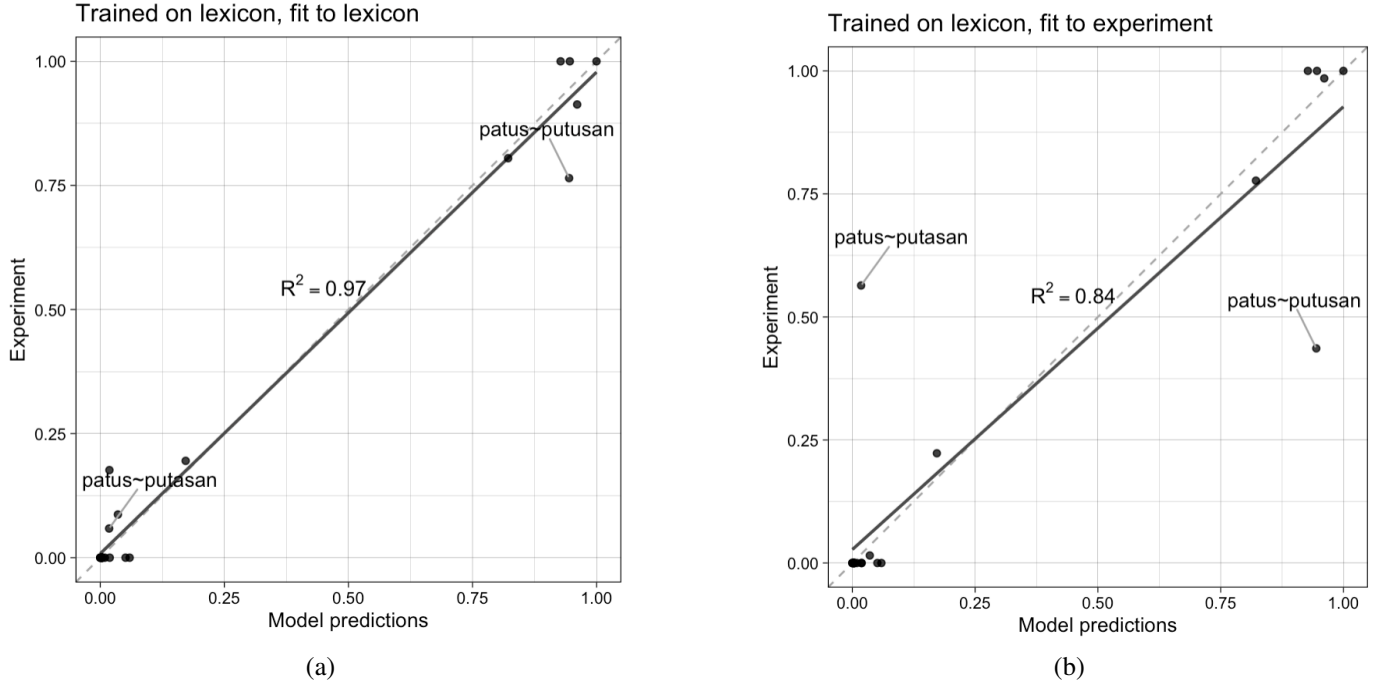
Figure 6: Model predictions plotted against lexicon vs. experimental results. Fitted regression lines are included.

7.3.1 Bias implementation

To implement a bias towards less complex, more uniform grammars, I adopt Martin's (2011) 'grammatical leakage' approach. Under this approach, the grammar contains a smoothing term that prefers to assign uniform weight to constraints (therefore penalizing complex grammars). For Seediq, because the grammar has both general and vowel-specific constraints for vowel matching, the smoothing term would cause the general matching constraint to be assigned some weight, resulting in an over-extension of vowel matching.

In MaxEnt, this smoothing term can be implemented as a Gaussian prior. In the current model, I implement the prior with a mean (μ) of zero and standard deviation (σ) of four for all constraints. In principle, μ can be any low value, as long as it is uniform across all the constraints. The value of σ was set by testing a range of values from 1-100, and finding the one that produced the best-fit model.

7.3.2 Model results and fit

Table 15 compares the weights learned in the BASELINE model (which has no smoothing term) against the BIASED model (which has the smoothing term). As expected, the BIASED model learns more uniform weights across-the-board. In particular, it learns a relatively higher weight for both the general vowel-matching constraint, and *P/i,u (which blocks alternations like ['putas]~[pu'tusan])

Overall, these differences strongly improve model fit to experimental results, as summarized in Table 16. In this table, the BASELINE model has no smoothing term, while the BIASED model has the smoothing term. In particular, adding a smoothing term significantly increases log-likelihood of the model fit to experimental results ($p = 8 \times 10^{-23}$, deg. freedom = 2).

	BASELINE	BIASED
Nuc-ID(MidV)	5.33	4.35
Nuc-ID(V)	0.17	1.94
*P/i,u	0.72	1.43
*[o]	4.71	2.36
ID[low]	0.20	0.33
ID[high]	2.10	2.37
ID[front]	2.55	2.32

Table 15: Constraint weights learned in baseline vs. biased models

	R^2	$\hat{L}$
BASELINE	0.84	-411.0
BIASED	0.88	-299.4

Table 16: Proportion of variance explained (R^2) and log-likelihood ($\hat{L}$) of model predictions fit to experimental results

The improvement in model fit is driven by the model's predictions for /patus/. Fig. 7 plots the predictions of the BASELINE and BIASED models against the experimental results. Note that these plots only show the vowel conditions that were tested in the experiment. As already discussed, the BASELINE model under-predicts the rate of [u]∼[a] alternation for /patus/ (P=0.02). In contrast, the BIASED model correctly predicts a higher rate of [u]∼[a] alternation (P=0.13).

Notably, the BIASED model still under-predicts rates of [u]∼[a] alternation compared to experimental results. Future work should consider alternative ways of modeling the preference for vowel matching. The current study implements a smoothing term where the μ and σ are constant across all constraints. An alternative approach, which could be fruitful for modeling Seediq, is to vary either σ or μ, so that some constraints are biased to have lower or higher weight than other constraints. This approach was pioneered by Wilson (2006), and has been utilized in work such as White (2013, 2017).
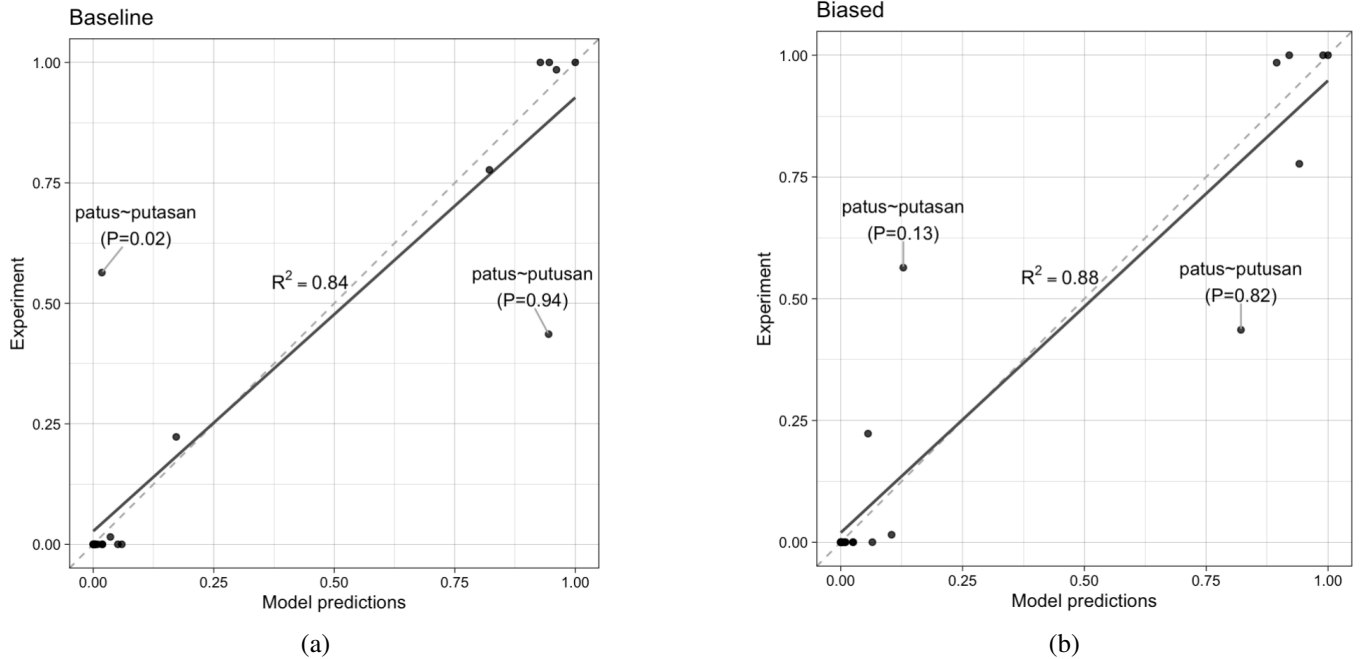
Figure 7: Predictions of models trained ont he lexicon, plotted against experimental results. Fitted regression lines are included.

8   Conclusion

Based on a survey of 341 Seediq verb paradigms, the current study finds that Seediq paradigms show a striking asymmetry, whereby the unsuffixed 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, 2002b).

This asymmetry is expected if there has been a gradual restructuring of Seediq verb paradigms based on the unsuffixed 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 an unsuffixed 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 to gapped forms that have no listed 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-ID(V) is used to enforce post-tonic vowel alternation. Additionally, I propose that speakers' non-viridical extension of vowel matching is rooted in a complexity bias, which could be implemented in my MaxEnt model as a Gaussian smoothing term.

References

Albright, Adam. 2002a. A restricted model of ur discovery: Evidence from lakhota. *Ms, University of California at Santa Cruz* .

Albright, Adam. 2010. Base-driven leveling in Yiddish verb paradigms. *NLLT* 28(3). 475–537.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.

Albright, Adam C. 2002b. *The identification of bases in morphological paradigms*: University of California, Los Angeles dissertation.

Barnes, Jonathan A. 2002. *Positional neutralization: A phonologization approach to typological patterns*: University of California, Berkeley dissertation.

Becker, Michael & Maria Gouskova. 2016. Source-Oriented Generalizations as Grammar Inference in Russian Vowel Deletion. *Linguistic Inquiry* 47(3). 391–425. doi:10.1162/LING_a_00217. `https://doi.org/10.1162/LING\_a\_00217`.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 84–125.

Berger, Adam, Stephen A Della Pietra & Vincent J Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22(1). 39–71.

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

Bybee, Joan. 2003. *Phonology and language use*, vol. 94. Cambridge University Press.

Chung, Sandra. 1983. Transderivational relationships in Chamorro phonology. *Language* 35–66.

Coetzee, Andries W & Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory* 26(2). 289–337. doi:10.1007/s11049-008-9039-z.

Council of Indigenous People. 2020. Aboriginal household registration statistics data analysis in oct 2020. `https://www.cip.gov.tw/portal/docDetail.html?CID=940F9579765AC6A0`.

Council of Indigenous Peoples. 2020. Online dictionary of aboriginal languages. `http://e-dictionary.apc.gov.tw/Index.htm`. Accessed: 2020-09-30.

Crosswhite, Katherine. 1998. Segmental vs. prosodic correspondence in Chamorro. *Phonology* 281–316.

Crosswhite, Katherine. 2004. Vowel reduction. *Phonetically based phonology* 191–231.

Crosswhite, Katherine & Alexander Jun. 2001. *Vowel reduction in optimality theory*. Psychology Press.

Crosswhite, Katherine M. 2000. Sonority-driven reduction. In *Annual meeting of the berkeley linguistics society*, vol. 26 1, 77–88.

Edwards, Jan, Mary E Beckman & Benjamin Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition .

Ernestus, Mirjam & R Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1). 5–38.

Flemming, Edward. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 7–44.

Fylstra, Daniel, Leon Lasdon, John Watson & Allan Waren. 1998. Design and use of the microsoft excel solver. *Interfaces* 28(5). 29–55.

Gallagher, Gillian & Jessica Coon. 2009. Distinguishing total and partial identity: Evidence from Chol. *Natural language & linguistic theory* 27(3). 545–582. doi:10.1007/s11049-009-9075-3.

Goldwater, Sharon & Mark Johnson. 2003. Learning ot constraint rankings using a maximum entropy model. In *Proceedings of the stockholm workshop on variation within optimality theory*, vol. 111120, .

Greenhill, Simon J, Robert Blust & Russell D Gray. 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271–283.

Hayes, Bruce, Péter Siptár, Kie Zuraw & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language.* 822–863.

Hayes, Bruce & James White. 2015. Saltation and the p-map. *Phonology* 32(2). 267–302. doi:10.1017/S0952675715000159.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *LI* 39(3). 379–440.

Hayes, Bruce, Colin Wilson & Anne Shisko. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 691–731.

Hoijer, Harry. 1946. Tonkawa. *Linguistic structures of Native America* 249.

Holmer, Arthur. 1996. *A parametric grammar of seediq*: Lund University dissertation.

Ito, Junko & Armin Mester. 2003. On the sources of opacity in OT: Coda processes in German. *The syllable in optimality theory* 271–303.

Jun, Jongho. 2010. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics* 19(2). 137–179.

Jun, Jongho & Adam Albright. 2017. Speakers' knowledge of alternations is asymmetrical: Evidence from seoul korean verb paradigms 1. *Journal of Linguistics* 53(3). 567–611.

Kang, Yoonjung. 2006. Neutralizations and variations in Korean verbal paradigms. *Harvard Studies in Korean Linguistics* 11. 183–196.

Kenstowicz, Michael. 1994. Sonority-driven stress. *Rutgers Optimality Archive* 33.

Kenstowicz, Michael & Larry M Kisseberth. 1977. *Topics in phonological theory*. New York: Academic Press.

Kessler, Brett. 2001. *The significance of word lists*. Cambridge, MA: MIT Press.

Kiparsky, Paul. 1978. Analogical change as a problem for linguistic theory. *Studies in the Linguistic Sciences Urbana, Ill* 8(2). 77–96.

Kuo, Jennifer. 2020. *Evidence for base-driven alternation in Tgdaya Seediq*: UCLA MA thesis.

Kuo, Jennifer. 2022. Evidence for prosodic correspondence in the vowel alternations of tgdaya seediq. *Phonological Data and Analysis* to appear.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. 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*, 884–891. Citeseer.

Li, Paul Jen-kui. 1981. Reconstruction of Proto-Atayalic phonology. *Bulletin of the Institute of History and Philology* 52.

Łubowicz, Anna. 2002. Derived environment effects in optimality theory. *Lingua* 112(4). 243–280.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87(4). 751–770. doi:10.1353/lan.2011.0096.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23(3). 405–423.

Moore-Cantwell, Claire. 2013. Over-and under-generalization in learning derivational morphology. In *Proceedings of nels*, vol. 42, .

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25(1). 83–127. doi:10.1017/S0952675708001413.

Moreton, Elliott & Joe Pater. 2012. Structure and substance in artificial-phonology learning, part i: Structure. *Language and linguistics compass* 6(11). 686–701. doi:10.1002/lnc3.363.

Noske, Roland. 2011. *A theory of syllabification and segmental alternation: with studies on the phonology of French, German, Tonkawa, and Yawelmani*, vol. 296. Walter de Gruyter.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive science* 33(6). 999–1035.

Pierrehumbert, Janet et al. 2003. Probabilistic phonology: Discrimination and robustness. *Probabilistic linguistics* 177–228.

Prince, Alan & Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in Phonology* 3.

Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of the 22nd west coast conference on formal linguistics*, vol. 22, 101–114. Cascadilla Press.

Ryan, Kevin M. 2011. Gradient syllable weight and weight universals in quantitative metrics. *Phonology* 413–454.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. Tech. rep. Colorado Univ at Boulder Dept of Computer Science.

Wasserman, Larry. 2004. *All of statistics: a concise course in statistical inference*. New York: Springer.

White, James. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a p-map bias. *Language* 93(1). 1–36.

White, James C. 2013. *Bias in phonological learning: Evidence from saltation*: UCLA dissertation.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science* 30(5). 945–982.

Yang, Hsiu-fang. 1976. The phonological structure of the paran dialect of Sediq. *Bulletin of the Institute of History and Philology Academia Sinica* 47(4). 611–706.

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*: UCLA dissertation.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to tagalog nasal substitution. *NLLT* 28(2). 417–472.

Zuraw, Kie. 2013. *map constraints. Ms, University of California, Los Angeles.