

UNIVERSITY OF CALIFORNIA

Los Angeles

Evidence for base-driven alternation in
Tgdaya Seediq

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Arts in Linguistics

by

Jennifer Kuo

2020

© Copyright by

Jennifer Kuo

2020

ABSTRACT OF THE THESIS

Evidence for base-driven alternation in
Tgdaya Seediq

by

Jennifer Kuo

Master of Arts in Linguistics

University of California, Los Angeles, 2020

Professor Bruce Hayes, Chair

Standard approaches to learning the morphophonology of paradigms require positing URs from which surface contrasts can be derived. Often, URs are ‘**cobbled**’, in that they combine information from multiple forms of the paradigm, and do not necessarily correspond to any single surface form (Kenstowicz and Kisseberth, 1977: 33). In contrast, Albright (2002, et seq.) argues for a **single surface base approach**, where the UR must be based on the surface forms of one slot in the paradigm. In Tgdaya Seediq (Austronesian: Taiwan), all forms of a paradigm (non-suffixed vs. suffixed) suffer from some loss of contrasts, making it a good test case for comparing between the two theories of morphophonology. In this thesis, I conduct an empirical study of Seediq verbal alternations, and find support for Albright’s surface base hypothesis.

This thesis consists of three parts: First, based on a corpus on 340 Seediq verb paradigms, I describe the patterns of alternation in Seediq, and show how asymmetries in the data make it so that the suffixed forms of a paradigm are highly predictable from the non-suffixed forms. In the second part, I use a computational rule-based model to confirm this asymmetry, and

show that such an asymmetry is expected under Albright's single surface-base hypothesis, but unexpected under the traditional cobbled UR approach. Finally, based on these results, I develop a theoretical surface-base model of Seediq verbal alternations, which takes surface non-suffixed forms as inputs. The model uses Zuraw's (2000; 2010) DUAL LISTING framework to model speaker intuitions, and is implemented in Maximum Entropy Harmonic Grammar (Goldwater and Johnson, 2003), a stochastic variant of Optimality Theory (Smolensky, 1986; Prince and Smolensky, 1993).

The thesis of Jennifer Kuo is approved.

Claire Moore-Cantwell

Kie Zuraw

Bruce Hayes, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

1	Introduction	1
1.1	Two approaches to morphophonology	1
1.2	Phonological alternations in Seediq verbal paradigms	3
1.3	Data collection	4
1.4	Stress-driven vowel alternations	5
1.5	Final consonant alternations	8
1.6	Irregular alternations	10
1.7	Two approaches to morphophonology in Seediq	12
2	Statistical patterning of Seediq alternations	15
2.1	Vowel matching	16
2.2	Final consonant alternations	17
2.3	Final vowel alternations	19
2.4	Predictability from the suffixed form	21
2.5	Interim summary	25
3	Quantitative evaluation of stem and suffix bases	28
3.1	Model implementation	28
3.2	Model evaluation	30
3.3	Comparing stem and surface base models	31
3.4	Indirect evidence for historical reanalysis	33
3.5	The selection of a base form	35

4	Towards a theoretical surface-base model for Seediq	37
4.1	Input and GEN	38
4.2	Constraint set under a cobbled UR model	39
4.2.1	Constraints for final consonant alternation	40
4.2.2	Constraints for pretonic vowel neutralization	41
4.2.3	Dealing with saltatory alternations in Seediq	43
4.3	Model performance under a cobbled-UR model	46
4.4	Constraint set under a stem-base model	47
4.4.1	Constraints for pretonic vowel neutralization	47
4.4.2	Introducing Anticorrespondence	48
4.5	Model performance under a stem-base approach	49
4.6	Moving to a stochastic model	51
4.6.1	Predicting speaker intuitions	51
4.6.2	Accounting for stem-specific behavior	52
4.6.3	Constraint weights	53
4.6.4	Model performance	55
4.6.5	Frequency matching	56
4.7	Testing the productivity of Seediq alternations: current and future work	58
5	Conclusion	60
	Appendices	62
A	Summary of rule-based model (stem base)	62
B	Summary of rule-based model (suffix base)	64

C	Strict ranking OT model-constraints and constraint ranking	66
D	MaxEnt OT model-constraints and constraint weights	67

LIST OF FIGURES

2.1	How the reduced [u] of non-suffixed CVCuC is realised when stressed under suffixation	16
2.2	Rates of final consonant alternation	19
2.3	Final vowel alternations	20
2.4	Distribution of stressed vowels in non-monosyllabic suffixed forms	22
2.5	Realization of stressed vowels in monosyllabic suffixed forms	25
3.1	Performance of stem vs. suffix-base models	31
3.2	Model performance using real vs. simulated lexicon	34

LIST OF TABLES

1.1	Inflectional morphology of Seediq	4
2.1	Rates of final consonant alternation	18
2.2	Predictability of stem stressed vowels from suffixed forms in disyllabic verbs . .	24
2.3	Predictability of suffixed form stressed vowels from CVCuC stems	24
3.1	Derivation of ['talaŋ]→[tu'laman] in the stem-base model	30
3.2	Derivation of ['biciq]→[bu'ciqan] in the stem-base model	32
3.3	Derivation of [bu'ciqan]→['biciq] in the suffix-base model	33
4.1	Constraint weights in stem-base model	54
4.2	Example forms where observed outputs were assigned $p < 0.1$	55
4.3	Example forms where observed outputs, $0.1 < p < 0.3$	56
4.4	Predicted vs. observed rates of final alternation in lexicon vs. model	57
4.5	Predicted vs. observed rates of vowel matching in lexicon vs. model (stem V refers to the stressed vowel of the stem)	58

ACKNOWLEDGMENTS

I am grateful to Bruce Hayes, as well as Kie Zuraw and Claire Moore-Cantwell, for their advice, support, and many useful insights. I would also like to thank my Seediq consultants for their time and generosity, and for sharing their language with me. This work was supported by a UCLA GSRM grant.

CHAPTER 1

Introduction

1.1 Two approaches to morphophonology

The classical approach to analyzing the morphophonology of paradigms, laid out by Kenstowicz and Kisseberth (1977), involves setting up underlying forms (URs) which preserve as many contrastive phonological properties as possible. Often, when all forms of a paradigm are affected by neutralizing processes, the resulting UR must be ‘**cobbled**’, in the sense that it combines information from multiple forms of a paradigm. Consider, for example, the case of Tonkawa verbal paradigms, where verb roots display extensive morphophonemic alternations as illustrated in (1). In Tonkawa, different vowels of the verb stem surface depending on the phonological properties of its affixes. Crucially, for trisyllabic stems such as the ones in (1), no surface forms show all three vowels. Instead, URs must cobble together information about the first vowel from slots ‘A’ or ‘C’ of the paradigm, and information about the second and third vowels from other slots (such as slot ‘D’ of the paradigm) (Kenstowicz and Kisseberth, 1977: 33). Under this analysis, 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* (Kenstowicz and Kisseberth, 1977: p.16)

A (/C-stem-V/)	B (/V-stem-V/)	C (C-stem-C)	D (/V-stem-C/)	gloss	UR
notx	ntox	notxo	ntoxo	‘hoe’	/notoxo/
netl	ntal	netle	ntale	‘lick’	/netale/
picn	pcen	picna	pcena	‘cut’	/picena/

Albright (2002: et seq.) proposes an alternative approach, called the *single surface base hypothesis*, where the UR must be based on a single surface form in the paradigm. Specifically, 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 the surface allomorphs of slot ‘A’, ‘B’, ‘C’, or ‘D’ of the paradigm. Under this approach, the morphophonology will have less informational resources available, as no allomorph can perfectly predict all three vowels of a verb stem. On the other hand, the process of UR building is less complex and more restrictive.

In Tgdaya Seediq (henceforth Seediq), processes of vowel reduction and word-final consonant neutralization cause all forms of a paradigm to suffer from loss of contrasts, making it a good test case for comparing the two theories of morphophonological analysis. The current study presents evidence from Seediq in support of the Albrightian single surface base hypothesis. In particular, a survey of the Seediq lexicon reveals asymmetries which cannot be explained under the traditional UR analysis, but are predicted under a surface-base model.

The outline of the paper is as follows: in Section 1.2, I describe the patterns of alternation in Seediq verbal paradigms, and show how these patterns result in a loss of contrasts in all slots of the paradigm. In Chapter 2, a quantitative survey of a Seediq corpus reveals asymmetries such that the stem forms can predict the suffixed forms with much higher accuracy than the suffixed forms can predict the stem forms. In Chapter 3, I confirm this asymmetry through a computational model of surface-base learning. Based on these results,

I argue that Seediq has undergone restructuring in a direction which rendered the stem form of verb paradigms increasingly more informative than the other slots of the paradigm. Finally, in Chapter 4, I lay out a theoretical model of Seediq verbal alternations which takes surface forms as inputs. The model uses Zuraw’s (2000; 2010) DUAL LISTING framework to model speaker intuitions, and is implemented in Maximum Entropy Harmonic Grammar (Goldwater and Johnson, 2003), a stochastic variant of Optimality Theory (Smolensky, 1986; Prince and Smolensky, 1993).

1.2 Phonological alternations in Seediq verbal paradigms

Seediq is an Austronesian (Atayalic) language spoken in Central and Eastern Taiwan. The Tgdaya dialect, which is the focus of the current study, has roughly 1000 speakers aged around 40 and above (Kang, 1992).

The Seediq phoneme inventory is given in (2) and (3); the orthography that I will be using is given in italics. Where the orthography and IPA are different, phonetic transcription is given in brackets.¹

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 below in Table 1.1. Crucially, there are extensive vowel and consonant alternations between the non-suffixed and suffixed forms of a verb paradigm. Below, I describe these alternations in detail. Note that during elicitation of verb paradigms, described below in Section 1.3, all verbs were elicited with the /s-/, /-an/, /-un/, and /-i/ affixes. The patterns reported in the paper were found to be consistent across all pairs of non-suffixed and suffixed forms. As such, for simplicity, all examples (unless otherwise specified) will compare the bare stem forms to forms suffixed with /-an/ ‘locative focus, perfective’ (where the bare stem is representative

¹Note that the glottal stop is not written, but shows up between all vowel-vowel sequences. For example, *seediq* is phonetically [seʔediq]

of all non-suffixed slots of the paradigm).

	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	-i			

Table 1.1: Inflectional morphology of Seediq

(2) *Seediq consonant inventory*

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

(3) *Seediq vowel inventory*

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

1.3 Data collection

The alternations to be described in the rest of this chapter are based both on existing descriptive work by Yang (1976), as well as a corpus of 340 verbal paradigms. These 340 paradigms were drawn from two sources: (1) the Taiwan Aboriginal e-Dictionary (Mei-jin et al., 2014), and (2) fieldwork with three Seediq speakers, carried out by the author in Puli Township, Nantou, Taiwan. Data was collected over the course of three weeks in July 2019. The consultants were aged 69-78; note that there is a high rate of language attrition in Seediq communities, such that fluent speakers are all around age 40 and above, 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 range of Seediq speakers. All three consultants reported speaking both Mandarin and Seediq regularly at roughly equal rates.

184 paradigms were collected from the online dictionary, and the remaining 156 paradigms were collected directly from native speaker consultants. In addition, verb paradigms taken from the dictionary were confirmed with consultants, and omitted if my consultant(s) did not recognise the word; three verbs were omitted under this criterion. These forms are shown below in (4). In the case of (4a), the word was not recognised by my consultants. For (4b) and (4c), consultants disagreed with the dictionary on the suffixed form of the verbs. All three verbs were excluded from my analysis. 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	
		<i>dict.</i>	<i>consultant</i>
(a) ‘to hook’	'daquc	du'qut-an	NA
(b) ‘to increase’	uman	mal-an	man-an
(c) ‘to seal/close’	sepuy	supuy-an	supuw-an

1.4 Stress-driven vowel alternations

Consistent with Yang’s (1976) description, Seediq stress is always penultimate, such that suffixation shifts stress rightwards. This gives rise to alternations such as [**bun**uh~bu'**nu**han] ‘wear hat’. Crucially, stress interacts with processes of vowel neutralization, giving rise to alternations between the stem and suffixed forms of the paradigm.

Pretonically, all vowel contrasts are neutralised. First, onsetless pretonic vowels are deleted. This pattern was found for all 36 vowel-initial words in the data; as illustrated in (5), the stem’s initial vowel is deleted when stress shifts to the second syllable in the /an/-suffixed form. The vowel will assimilate to an adjacent stressed vowel if the two are separated by [ʔ] or [h] (see (6)); 35 verbs were found to match this description. Otherwise, vowels are reduced to [u] pretonically, as shown in (7). This last process of reduction to [u]

is by far the most common, occurring in 265 stems. All three patterns of pretonic vowel neutralization are exceptionless.

- (5) *Onsetless vowels delete* (36/36)
- (a) 'awak ~ 'wak-an 'lead (by a leash)
 - (b) 'eyah ~ 'yah-an 'come'
 - (c) 'uyas ~ 'yas-an 'sing'
- (6) *Vowel assimilation to stressed vowel* (35/35)
- (a) 'leʔiŋ ~ li'ʔiŋ-an 'hide (an object)'
 - (b) 'saʔis ~ si'ʔis-an 'sew'
- (7) *Vowel reduction to [u]* (265/265)
- (a) 'gedaŋ ~ gu'daŋ-an 'die'
 - (b) 'biciq ~ bu'ciq-an 'decrease'
 - (c) 'barah ~ bu'rah-an 'rare'
 - (d) 'burah ~ bu'rah-an 'new, create'

In addition, vowels are optionally deleted between nasals and stops, as shown in (8). This process was described by Yang (1976) to be obligatory, but speakers for the current study judged deletion to be optional, and accepted forms with or without vowel deletion.

Vowel deletion is only observed for two /an/-suffixed forms, but is more common in doubly suffixed forms, which arise when a stem is suffixed with a focus marker (e.g. /-an/) followed by the imperative /-i/. An example of this is shown in (9), where the final vowel of the stem for 'to cook' surfaces as [e] when stressed, but is deleted in the doubly suffixed form. Moreover, as demonstrated by the [hujedan~hun'dani] alternation, where vowel deletion occurs, the nasal will assimilate to the following stop, resulting in homorganic nasal-stop clusters.

- (8) *Optional vowel deletion between nasals and stops* (2/2)
- | STEM | SUFFIXED | |
|--------------|-----------------------------------|------------|
| (a) qu'nedis | qun'dis-an (~qun <u>u</u> dis-an) | 'lengthen' |
| (b) gu'natuk | gun'tuk-an (~gun <u>u</u> tuk-an) | 'peck' |

(9) *Vowel deletion in doubly suffixed forms*

STEM /STEM-an/ /STEM-an-i/
'haŋ**u**c hu'ŋ**e**dan hun'dani (~huŋ**u**dani) 'to cook'

Crucially, pretonic vowel neutralization always results in a loss of contrasts in the *suffixed* forms. For example, consider the two verbs in (7c-d). The two are distinctive in the isolation stem form, but become homophonous in the suffixed form due to reduction of the stem's initial vowel (indicated in boldface).

Post-tonically, there are similar but more restricted processes of vowel reduction, which result in a less drastic loss of vowel contrasts. First, /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 these alternations are shown below in (10). Moreover, with the exception of /uy/, diphthongs are prohibited word-finally. As a result, /ay/ and /aw/ are respectively monophthongised to [e] and [o] (11a-b), while /ey/ is monophthongised to [u] (11c).

Note that final [e] is only found as a result of monophthongization of /ay/. However, as will be discussed in Section 1.5, certain processes of final consonant alternation also result in stem-final [u] and [o].

Note that [o] has a limited distribution in the Seediq lexicon; [o] surfaces post-tonically as a result of word-final neutralization processes, but there are very few stems that surface with phonemic stressed [o] as in (10c). In the current data, only three such forms were found.

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

- (a) 'rem**u**x ~ ru'm**u**xan 'enter' (u~u, n=60)
(b) 'pem**u**x ~ pu'm**u**xan 'hold' (u~e, n=36)
(c) 'doʔ**u**s ~ doʔ**o**s-an 'refine' (metal)' (u~o, n=3)

(11) *Word-final monophthongization*

- (a) 'ra**e** ~ ru'ŋ**a**y-an 'play' (e~ay, n=7)
(b) 's**in**o ~ su'n**a**w-an 'to drink (alcohol)' (o~aw, n=1)
(c) 'de**n**u ~ du'ŋ**e**y-an 'dry (food)' (u~ey, n=11)

These post-tonic processes all result in a loss of contrasts in the stem forms, as it is not possible to predict how a final vowel will alternate from just looking at the isolation stem. For example, the stem-final vowels of (12a) and (12b) are contrastive in the suffixed form, but both reduced to [u] in the isolation stem form. Similarly, the stem-final syllables of (12c) and (12d) are contrastive in the suffixed form, but neutralised in the stem form due to monophthongization of /ey/.

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

	STEM	SUFFIXED	
(a)	'pem <u>ux</u>	pu'mexan	'hold'
(b)	'rem <u>ux</u>	ru'muxan	'enter'
(c)	'mal <u>u</u>	mu'ley-an	'able to'
(d)	ga'alu	gu'luw-an	'dote on'

1.5 Final consonant alternations

In addition to the above processes of pre- and post-tonic vowel reduction, Seediq has phonotactic constraints against word-final [p b m t d l g]. This results in various processes of word-final neutralization. First, /p, b, m, t, d, l/ are neutralised with other consonants as outlined in (13).

(13) *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 (13a), 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 (14a-c). In (14d-j), similar examples are provided for the other final consonant alternations.

Note that, as will be discussed further in Chapter 2, 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 forms (14h), while 32 forms are non-alternating (14g). In contrast, stem-final [c] almost always alternates with [t] or [d], and only surfaces as [c] in the suffixed form one time (14d).

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

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

All of these processes result in a loss of contrasts in the stem form. For example, (14a) and (14b) have contrastive stem-final consonants in the suffixed form. However, both surface with a final [k] in the isolation stem because /p/ neutralises to [k].

Stem-final /g/ shows more complicated patterns of alternation than those discussed above. As summarised in (15), /ag/ monophthongizes to [o] word-finally (15a), /eg, ug/ both monophthongize to [u] (15b), and /ig/ becomes [uy] (15c). These alternations are historically a result of /g/ weakening to [w] word-finally, followed by monophthongization of the resulting diphthong (Li, 1981). As with the other final consonant alternations, they result in neutralization of contrasts in the isolation stem form.

(15) *Alternation of final /g/*

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

1.6 Irregular alternations

The alternations discussed so far are all phonotactically driven. For example, pre-tonic vowel reduction is driven by restrictions on the distribution of vowels in non-stressed position, while post-tonic reduction (where /e,o/ reduce to [u]) is motivated by restrictions against non-peripheral vowels in post-tonic position. Final consonant alternations are driven by constraints against final [p, b, t, d, m, l, g].

However, Seediq also has a subset of irregular alternations; these are irregular in the sense that (i) they do not correspond to clear synchronic phonotactic motivations, (ii) lack generality and apply to very few lexical items. In the dataset used for this study, 27 such cases were found; these are listed below in (16), along with the expected suffixed form (given the non-suffixed stem form). The majority of these irregularities involves unexpected alternations in the post-tonic vowel of a stem (16a). In a small number of cases (n=5), stem-final open vowels are deleted in the suffixed form (16b). Three words ending in [an] or [un] do not alternate at all in the stem vs. suffixed forms (16c).

(16) *Verbs showing irregular alternations*

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

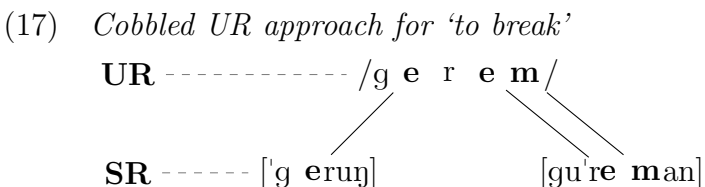
STEM	SUFFIXED	GLOSS	EXPECTED SUFFIXED
'huruc	hu'r idan	'come to a stop'	(hu'rudan, hu'redan, hu'rodan)
'raguh	ru'w ahan	'open'	(ru'guhan, ru'gehan, ru'gohan)
'tebas	tu'b esan	'sieve grains'	(tu'basan)
bu'leyaq	bulu'y eqan	'enjoy'	(bulu'yaqan)
se'yaŋ	su'y eqan	'get angry'	(su'yaqan)

	'ceyax	cu'yexan	'avoid'	(cu'yaxan)
	'nuqah	nu'qihan	'spin yarn'	(nu'qahan)
	'adis	'desan	'carry away'	('disan)
	'raqic	ru'qutan	'hook'	(ru'qitan)
	'patis	pu'tasan	'write'	(pu'tisan)
	'pehiq	pe'hejan	'destroy'	(pi'hijan)
(b)	<i>Irregular final vowel deletion (n=5)</i>			
	'hado	'hadan	'deliver'	(hu'dawan)
	'taqe	'tajan	'rely on'	(tu'neyan)
	'qene	'qenan	'extend'	(qu'neyan)
	'taqe	'tajan	'rely on'	(tujaan)
	'kesa	'kesan	'tell (someone)'	(kusaan)
(c)	<i>Irregular final consonant alternation (n=3)</i>			
	'qerac	qu'rapan	'grab'	(qu'racan, qu'ratan, qu'radan)
	'igin	'giman	'search'	('ginan, 'gilan)
	'murac	mu'rakan	'to scratch'	(mu'racan, mu'ratan, mu'radan)
(d)	<i>Non-alternating pairs (n=2)</i>			
	'caman	'caman	'pass the night'	(cu'manan, cu'malan)
	'tekan	'tekan	'pound (with pestle)'	(tu'kanan, tu'kalan)
(e)	<i>[n]-insertion (n=3)</i>			
	'apa	pa'an-an	'carry (on back)'	('pa-an)
	'qeya	qu'yan-an	'hang'	(qu'ya-an)
	'lawa	la'an-an	'invite'	(lu'wa-an)
(f)	<i>Other irregular alternations (n=3)</i>			
	'bege	'biqan	'give'	(bu'geyan)
	'uqun	qu'dalan	'eat'	('qulan, 'qunan)
	te'heyaq	ti'hiq-an	'play'	(tuhu'yaqan)

1.7 Two approaches to morphophonology in Seediq

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


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



Assuming, then, a cobbled UR approach, the majority of forms in Seediq can be derived using a series of phonotactically motivated markedness constraints. This is demonstrated for /gerem/ in (18). Highly ranked markedness constraint *m]_w rules out candidates (a) and (b), where final [m] surfaces faithfully. Post-tonic vowel reduction to [u] can be enforced by a positional licensing constraint, LICENSE(nonperipheral/stress), which limits non-peripheral vowel qualities to stressed syllables (Crosswhite, 2004). This constraint rules out candidates such as (c), where the vowel surfaces faithfully as [e]. Note that specific patterns of alternation observed (e.g. /m/ alternating with [ŋ]) results from the interaction of faithfulness

constraints; this will be discussed in more detail in Section 4.2.1.

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

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

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

In contrast, under the Albrightian surface-base approach to UR construction, the Seediq learner would designate either the isolation stem form or the suffixed form to be the privileged base. As discussed in Section 1.1, the base is constrained to a single slot of the paradigm. In the case of Seediq, this means that the base would have to be either the non-suffixed form for all verbs, or the suffixed form for all verbs. However, there cannot be a mix of non-suffixed and suffixed bases.

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

On the other hand, as noted in Section 1.1, UR discovery under the single surface base hypothesis is relatively easier. In addition, there is increasing evidence in support of the single surface-base hypothesis from various sources, including historical change in languages

like Korean (Kang, 2006) and Yiddish (Albright, 2010). This historical evidence is further supported by results of wug tests (Jun, 2010) and surveys of child errors (Kang, 2006) for Korean.

Both the cobbled UR approach and surface-base approach are able to account for the Seediq data (with relative strengths and advantages). Crucially, however, the two models make different predictions about mislearning (and analogical change), which has been employed since Kiparsky (1978) as a way of understanding the language-specific grammatical structure imposed by learners.

The cobbled UR approach predicts that when the learner has incomplete data (resulting in errors or reanalysis of paradigms), the UR will be determined solely on the basis of whatever surface forms happen to be available. For example, if a learner only hears the non-suffixed ['geruŋ], they might posit the UR /geruŋ/, and project the suffixed form [gu'ruŋan]. On the other hand, if they hear the suffixed form [gu'reman], they might posit the UR /gurem/, and project the isolation stem to be ['guruŋ]. In principle, this means that over time, reanalyses in both directions are plausible, and the resulting Seediq lexicon should reflect this.

In contrast, 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.

CHAPTER 2

Statistical patterning of Seediq alternations

In this chapter, 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, strong statistical tendencies in the data make it so that suffixed forms are highly predictable from stem forms, but the stem forms are not as predictable from the suffixed forms. This asymmetry supports the single surface-base approach.

In particular, 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. As such, if one cell in a paradigm is much more informative than the other, and this asymmetry cannot be attributed just to phonological neutralization processes (such as vowel reduction), it suggests that restructuring based on a single base form has happened.

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

Note that, as described in Section 1.2, Seediq verbal paradigms have prefixed forms. Moreover, the isolation stem forms and prefixed forms show the same patterns of neutralization and alternation, so there is no way to differentiate between stem and prefixed forms in terms of how suitable they are as bases. The data presented will use the ISOLATION STEM

form to represent all non-suffixed slots of a paradigm, but in principle, any non-suffixed slot of the paradigm 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 (Bybee, 2003; Albright, 2002; Albright and Hayes, 2003).

2.1 Vowel matching

As discussed in Section 1.4, /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 stressed vowel in the isolation stem.

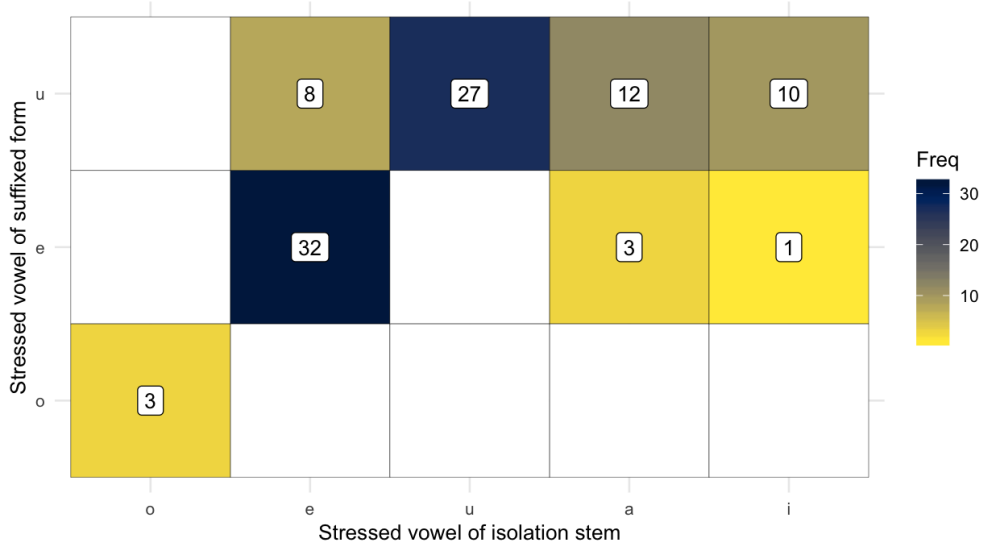


Figure 2.1: How the reduced [u] of non-suffixed CVCuC is realised when stressed under suffixation

Specifically, there is a tendency for VOWEL MATCHING, where the stressed vowel of the suffixed form ‘matches’ the stressed vowel of the isolation stem. This pattern is illustrated in Fig. 2.1, which shows the distribution of stressed vowels in CVCuC stems. 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 always surfaces as [u] in the suffixed forms (27/27, 100%). For [e], there was similarly a strong tendency for vowel matching, such that it was observed around 80% of the time (32/40). Otherwise, if the stem stressed vowel is /a/ or /i/, the reduced vowel is usually non-alternating, and surfaces as [u]. Consistent with these patterns, a Fisher’s exact test found that 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, whether the stem stressed vowel was [u] was significantly correlated with whether the suffixed stressed vowel was [u], $p < 0.001$ ($= 2.8 \times 10^{-8}$).

These regularities make it so that a speaker can predict, with relatively high accuracy, what a post-tonic [u] will surface as in the suffixed form. In other words, given a word of the form [ˈputus], the speaker can predict that the suffixed form will be [puˈtusan]. Given a word like [ˈpetus], the speaker could in principle predict that the suffixed form will most likely be [puˈtesan].

2.2 Final consonant alternations

Recall that word-finally, consonants /p, b, t, d, m, n/ are prohibited, resulting in the patterns of final consonant neutralization described in Section 1.5 and summarised below in (19). As a result of these alternations, when given just the isolation stems, it is not possible to perfectly predict whether final [c, k, n, ŋ] will alternate in the suffixed form.

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

	Cons.	Alternates?	Alternant	Example	Frequency
(a)	c	Yes	t	patic ~ putitan	18 (78%)
		Yes	d	patic ~ putidan	4 (17%)
		No		patic ~ putican	1 (4%)
(b)	n	Yes	l	patin ~ putilan	18 (75%)
		No		patin ~ putinan	6 (25%)
(c)	k	Yes	p	patik ~ putipan	6 (24%)
		Yes	b	patik ~ putiban	1 (4%)
		No		patik ~ putikan	18 (72%)
(d)	ŋ	Yes	ŋ	patiŋ ~ putiman	2 (6%)
		No		patiŋ ~ putiŋan	32 (94%)

Table 2.1: Rates of final consonant alternation

However, final consonants tend to either almost always or almost never alternate, as summarised in Table 2.1 and Fig. 2.2. Final [ŋ] almost never alternates with [m]; the hypothetical stem [ˈpatiŋ] will surface faithfully as [puˈtiŋ-an] 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] showing up faithfully as [k] in the suffixed form.

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

Because of these asymmetries in the alternation rate of final consonants, the final consonant of a suffixed form is actually highly predictable from the isolation stem. For example, a Seediq speaker knows that most [ŋ]-final stems (around 94%) will **not** show the [ŋ~m]

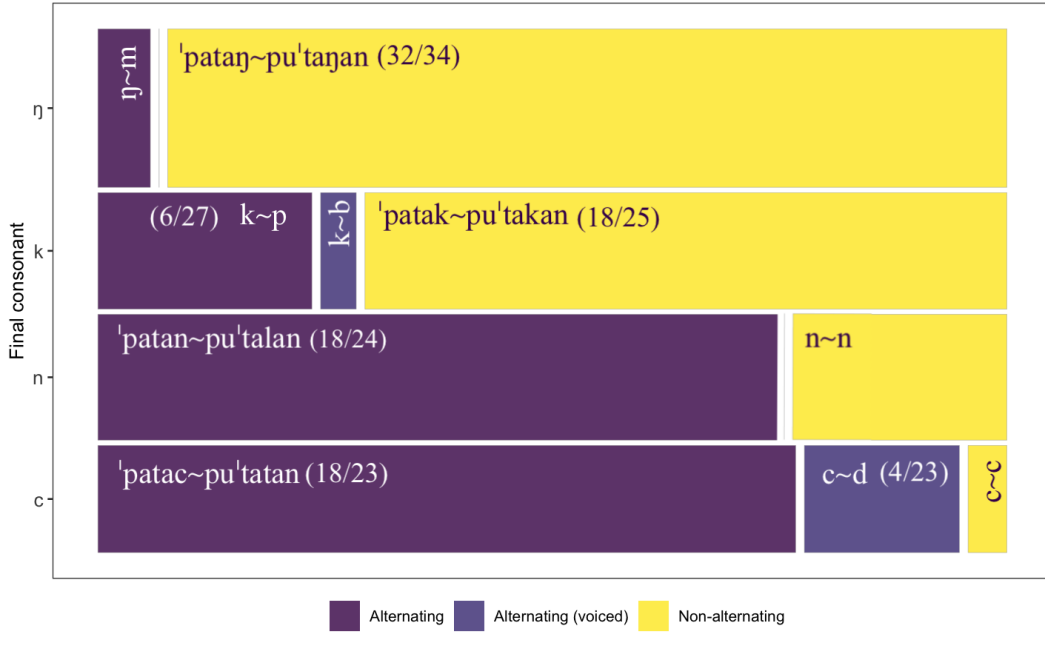


Figure 2.2: Rates of final consonant alternation

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.

2.3 Final vowel alternations

Stem-final vowels are expected to undergo the alternations summarized in (20). The patterns observed for these alternations are less clear (partially because of the relatively small number of available forms). As illustrated in Fig. 2.3, final [o] and [e] always alternate, and never surface faithfully as monophthongs; [o] tends to alternate with [ag] (9/12 cases, 75%), while final [e] tends to surface as [ay] in the suffixed form (7/10 cases, 70%). Final [u] shows more intermediate rates of alternation; about 39% of final [u]s are non-alternating (12/31); of the ones which do alternate, there is a slight preference for the suffixed form of surface with [ey]

(9/31, 29%).

Consistent with Yang's (1976) descriptions, final [i] never alternates. Final [a] generally also surfaces faithfully as [a] in the suffixed form. However, there are three forms where a final [n] is inserted, such as in the word 'qeya~qu'yan-an 'to carry'. Finally, in five irregular forms (discussed in (16)), open final vowels are deleted.

(20) *Final vowel alternations*

STEM		SUFFIXED
[u]	~	[u,ey,ug,eg]
[e]	~	[ay]
[o]	~	[ag, eg, aw]

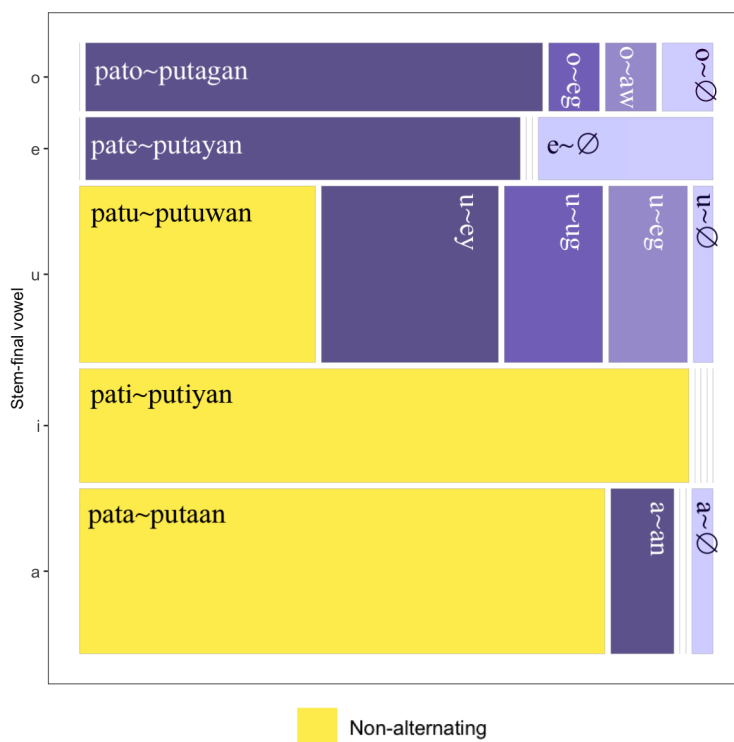


Figure 2.3: Final vowel alternations

Post-tonic vowel reduction and final consonant neutralization result in a loss of contrasts in the stem form. As a result, given just the stem form, it is impossible to perfectly predict the alternation of (i) [u] in post-tonic closed syllables, (ii) stem-final vowels, and (iii) final

consonants [c, n, k, ŋ, g]. However, the statistical patterns discussed so far suggest that these alternations are highly predictable from just the stem form.

2.4 Predictability from the suffixed form

Having established that stems can be used to predict suffixed forms with relatively 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. First, given a suffixed form, the realization of its stem-final consonants is completely predictable. For example, final /m/ will always be neutralised to [ŋ] in the non-suffixed stem. As such, as shown in (21a), when given a novel form [pu'tim-an], a Seediq speaker should in principle know with certainty that the isolation stem form will surface with a final [ŋ]. Similarly, final monophthongization is exceptionless, so given the suffixed form [pu'tay-an], speakers can predict that the stem form will surface with final [e] (21b). And, speakers should know that because of post-tonic mid-vowel reduction, a suffixed form such as [pu'tes-an] will surface with a reduced final vowel [u] in the stem form (21c).

(21) *Predicting stem forms from suffixed forms*

- (a) [pu'tim-an] → {'patiŋ, 'pitiŋ, 'petiŋ, 'potiŋ, 'putiŋ}
- (b) [pu'tay-an] → {'pate, 'pite, 'pete, 'pote, 'pute}
- (c) [pu'tek-an] → {'patuk, 'pituk, 'petuk, 'potuk, 'putuk}

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. First, consider vowel reduction and assimilation. As a result of these processes, the stressed vowel in the non-suffixed stem becomes neutralised in the suffixed form. This means that, given a suffixed form such as [pu'tis-an], it is impossible to perfectly predict what [u] will surface as when stressed in the stem form.

Vowel deletion adds further difficulties. Deletion of pretonic onsetless vowels affects onsetless disyllabic stems, resulting in alternations such as [ˈuʏas~ˈyas-an], where the suffixed form surfaces with a monosyllabic stem. Phonemically monosyllabic stems, though rare, do exist (e.g. [ˈreq ~ ˈreqan] ‘to swallow’). As a result, when given a novel suffixed form such as [ˈtis-an], where the stem appears to be monosyllabic, speakers must also predict whether the non-suffixed stem is monosyllabic, or if there is an initial (onsetless) vowel.

In the case of post-tonic vowel reduction (Section 2.1), a correlation between the stressed vowels of the stem and suffixed forms made it possible to predict alternation of suffixed forms with relatively high accuracy. For pre-tonic vowel reduction, however, this correlation is weaker. This is demonstrated in Fig. 2.4 and Fig. 2.5.

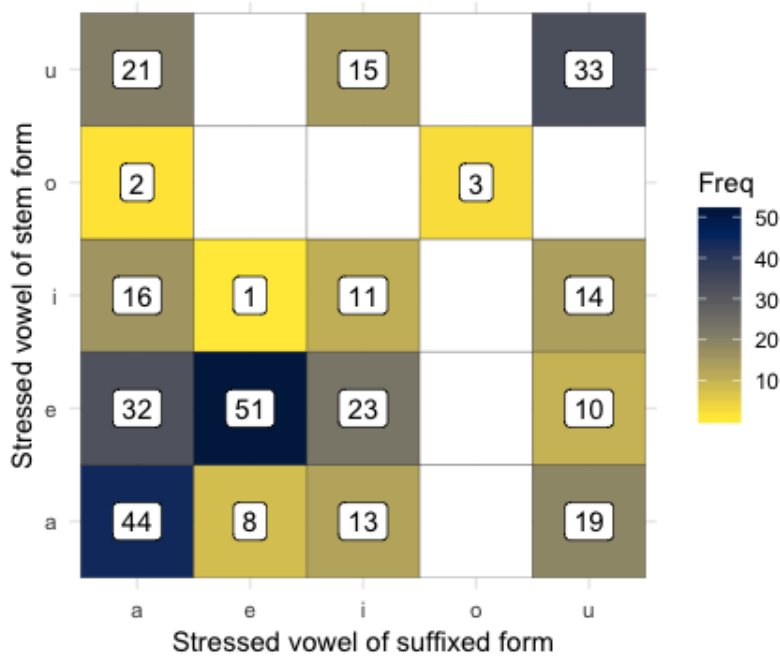


Figure 2.4: Distribution of stressed vowels in non-monosyllabic suffixed forms

Fig. 2.4 shows the distribution of stems which surface as *disyllabic* in the suffixed form. As seen in the figure, there is some predictability between the stressed vowel of the stems and the stressed vowel of the suffixed forms. However, these trends appear to be relatively weak compared to the patterns observed in Section 2.1, which looked only at contexts in

which post-tonic /e,o/ are reduced (i.e. stems of the form CVCuC, where the stem-final syllable is closed, and surfaces with the vowel [u]).

To get a general measure of the ‘predictability’ of the stem stressed vowel from the suffixed form’s stressed vowel, we can look each column of Fig. 2.4, and see how many forms are correctly predicted by selecting the largest number in each column. For example, looking at the first column of Fig. 2.4, 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 44 out of 115 (44+32+15+2+21) verbs where the suffixed form’s stressed vowel is [a]. In other words, if a speaker were given a suffixed form [pu'tak-an], they could apply this general pattern of predictability, and get the correct output 38% (44/115) 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, given a suffixed form such as [pu'tisan], the stem form is most likely to be [petis]. Applying this principle, we can correctly predict the stem stressed vowel for 37% (23/62) of relevant forms.

Using this method, I calculated the “predictability” of each column in Fig. 2.4 (i.e. predictability based on the identity of the suffixed form’s stressed vowel); these values are summarised below in Table 2.2, where the last column shows the proportion of forms correctly predicted. Based on these figures, predictability from some vowels (such as /i/, /a/, and /u/) is fairly low. Moreover, the last row of Table 2.2, which gives the total number of forms that can be correctly predicted, shows that even when picking the ‘best’ option based on statistical tendencies in the data, the correct vowel can be predicted only around 49% of the time.

As a comparison, Table 2.3 shows the predictability of the VOWEL MATCHING pattern discussed in Section 2.1. Here, values reflect predictability of the suffixed form’s stressed vowel based on the *non-suffixed stem’s* stressed vowel, in stems of the form CVCuC (with a reduced [u]). For example, if the isolation stem’s stressed vowel is [e], then we can predict with 80% (32/40) accuracy that the reduced [u] will surface as [e] when stressed in the

<i>Suff vow.</i>	<i>Predicted</i>	<i>Total</i>	<i>% correct</i>
/a/	71	115	38%
/e/	51	60	85%
/i/	23	62	37%
/o/	3	3	100%
/u/	33	76	43%
Total	181	316	49%

Table 2.2: Predictability of stem stressed vowels from suffixed forms in disyllabic verbs

<i>Stem vow.</i>	<i>Predicted</i>	<i>Total</i>	<i>% correct</i>
a	13	17	76%
e	32	42	76%
i	10	11	91%
o	3	3	100%
u	28	29	97%
Total	86	102	84%

Table 2.3: Predictability of suffixed form stressed vowels from CVCuC stems

suffixed form. This percentage is high for all vowels; the last row of Table 2.3 also shows that this VOWEL MATCHING principle correctly predicts the stressed vowel of the suffixed form 84% of the time. Overall, this comparison shows that in the contexts where vowel contrasts are neutralised, the stem form is much more informative than the suffixed form of a paradigm.

Fig. 2.5 shows the distribution of stems which surface as *monosyllabic* in the suffixed form. For these forms, the stem could either be monosyllabic in the isolation stem form (labeled \emptyset in Fig. 2.5), or surface with a stressed onsetless vowel. There is a general tendency for the stem to be disyllabic (monosyllabic stems are rare, with only four found in the current data), but once again, patterns of predictability are relatively weak.

The results so far suggest that the alternations which result from pretonic vowel neutralization are less predictable than the processes which result in loss of information in the stem form. Moreover, pretonic vowel neutralization affects a wider number of forms than the other neutralization processes; 336 verbs (the entire corpus, excluding 4 monosyllabic

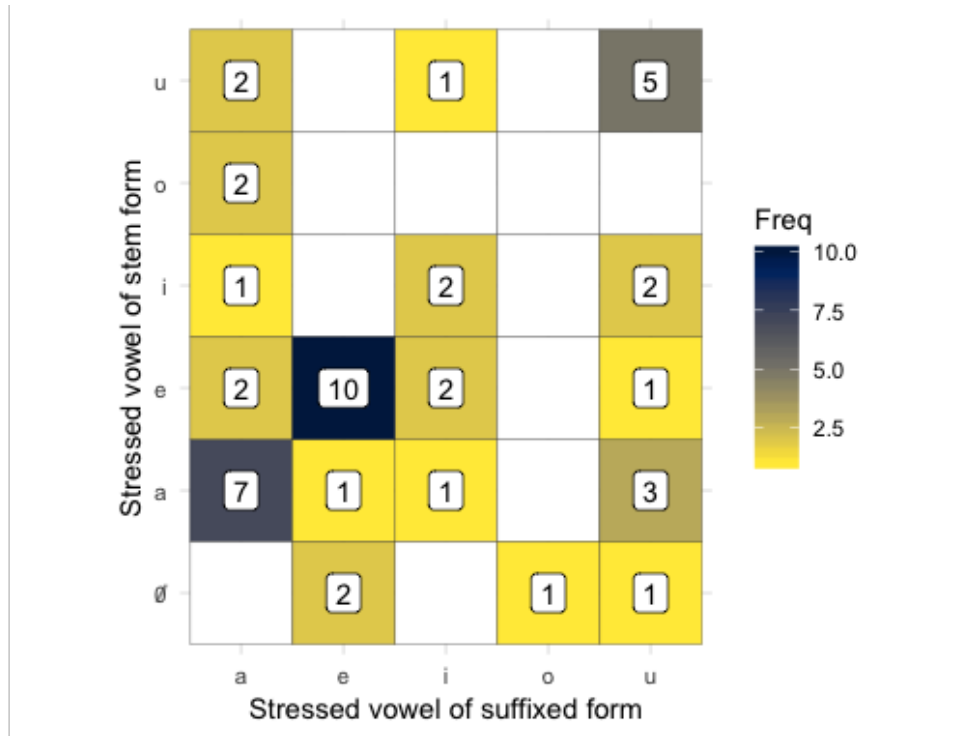


Figure 2.5: Realization of stressed vowels in monosyllabic suffixed forms

stems) are affected. In contrast, post-tonic vowel reduction affects 103 verbs, final consonant neutralization affects 101 verbs, and final monophthongization affects 97 forms.

2.5 Interim summary

In this chapter, comparison of the stem and suffixed forms revealed an asymmetry in how informative they are. Specifically, the stem forms can be used to predict the suffixed forms with much higher accuracy than the other way around. Notably, 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 neutralizing processes. Instead, the informativeness of the stem form is in part due to the very skewed rates of alternation in neutralised segments. In chapter 3, I confirm this asymmetry using models which quantify the relative informativeness of different base forms.

This asymmetry cannot be explained by a cobbled UR analysis; as discussed in Section 1.7, 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 relatively more symmetrical distribution of segments in underlying forms. However, as discussed in the beginning of this chapter, 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 the stem-final [ŋ]-[m] alternation. Historically, this dispreference may have been present as a weaker statistical tendency. If Seediq speakers have designated the stem to be the base, paradigms which showed the dispreferred [ŋ]-[m] alternation would gradually have been restructured, resulting in the very skewed rates of alternation that we see today. Although there is limited historical comparative data available for Seediq, I have found one example, elicited from my consultants, which suggests this type of reanalysis. As seen in (22), 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].

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

Restructuring of this variety is not unlikely; there are similar documented cases where a regular pattern rendered unpredictable by historical change is either partially or completely restructured. One notable case involves the Oceanic ‘thematic consonants’; for many Oceanic

languages, the loss of word-final consonants in intransitive verbs and nouns resulted in C/∅ alternations, where a consonant of unpredictable quality surfaces in a subset of suffixed forms (Hale, 1973; Blevins, 2008). Examples from Maori are shown in (23).¹

(23) *C/∅ alternations in Maori* (Hale, 1973: 414)

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

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

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

¹Note that only a subset of the possible thematic consonant are shown here; all Maori consonants, with the exception of /p/ and /w/, are attested alternants. The shape of the passive suffix is also conditioned by minimal word requirements.

CHAPTER 3

Quantitative evaluation of stem and suffix bases

In this chapter, a model of surface-base learning will be used to confirm the observed asymmetry between stem and suffixed forms. The model is not meant to be a theoretical one, but rather a way to quantify the predictability of different candidate base forms. Two surface-base models will be trained on the corpus of 340 verbs; one model will take the stem to be the base, and one will take the suffixed form as the base. Comparison of these two models will show that the stem-base model does a much better job of predicting the observed alternations in verb paradigms. Moreover, evaluation of the two models against a simulated lexicon will provide indirect evidence for historical reanalysis in a direction that supports Albright’s surface-base hypothesis.

3.1 Model implementation

The model employed here is based loosely off of the Minimal Generalization Learner (MGL, Albright and Hayes, 2003). Like the MGL, it is a rule-based model which takes surface forms as inputs, and attempts to derive the other slots of the paradigm using a series of rules. For Seediq, the model will take either the stem or suffixed form to be the base.

Note that the MGL iteratively learns rules from training data, resulting in a system which includes lexically-specific rules that may only apply to a single verb form. In contrast, in the current study rules were manually implemented.¹ Rules were also kept as general as

¹See Appendix A for the full list of rules under the stem-base model

possible, such that more specific rules were introduced only if needed to capture an irregular alternation. Although this could introduce a certain amount of subjectivity, the rules were designed to reflect the observed statistical patterns as closely as possible.

The key function of the MGL of interest to the current study is that it outputs a system of rules which vary in generality, and also has an explicit algorithm for rule evaluation. This allows us to quantify and measure the relative informativeness of different base forms.

In the current model, rules will be included for all observed alternations (including those of exceptional forms), regardless of how general they are. The resulting model will therefore consist of rules that vary in both **scope**-the number of forms in the input data that meet the rule’s structural description-and **hits**-the number of forms where rule application results in the correct output. For example, consider rules (24a-d), taken from the stem-base model. Pretonic vowel reduction (24a) and pretonic deletion of onsetless vowels (24b) are both exceptionless, but they differ in their *scope*; pretonic vowel reduction has a scope of 265 forms, while pretonic vowel deletion has a much smaller scope of 36. On the other hand, consider Rule (24c), which enforces the alternation of stem-final [ŋ] with [m] when projecting from stem to suffixed forms. Although this rule has a similar scope to Pretonic Vowel Deletion, it has a much smaller number of Hits (n=2).

(24) *Examples of rules in the stem-base model*

	Name	Rule	Example	p (H/S)	\hat{p}
(a)	Pret. VR	$\left[\begin{array}{l} +\text{syl} \\ -\text{stress} \end{array} \right] \rightarrow [u] / \#C_$	'patuk→pu'tukan	1.0 (265/265)	0.99
(b)	Pret. V-del.	$\left[\begin{array}{l} +\text{syl} \\ -\text{stress} \end{array} \right] \rightarrow \emptyset / \#_$	'awak→'wakan	1.0 (36/36)	0.95
(c)	ŋ-to-m	$[ŋ] \rightarrow [m] / _]_{\text{stem}} V$	'geruŋ→gu'reman	0.06 (2/34)	0.02
(d)	ruy-to-rig	$[ruy] \rightarrow [rig] / _]_{\text{stem}} V$	'baruŋ→bu'rigan	1.0 (3/3)	0.6

3.2 Model evaluation

In the model, each rule is evaluated in terms of its reliability as follows. First, a measure of **confidence** (p) is obtained by taking the ratio of Hits over Scope. Confidence reflects how reliable a rule is; an exceptionless rule like Pretonic Vowel Reduction is very reliable, and its confidence value of 1, shown in (24a), reflects this. In contrast, the rule for the [ŋ]-[m] alternation is not very reliable, and gets a low confidence value of 0.06.

Adjusted confidence (\hat{p}) is then calculated using lower confidence limit statistics (Mikheev, 1997; Albright and Hayes, 2003). Adjusted confidence penalises rules with small scope, to capture the intuition that the confidence of a rule is more well-grounded if there is more evidence for it (i.e. the Scope is high). For example, consider Rule (24d), which causes stem-final [ruy] to alternate with [rig] in the suffixed form. This rule, just like pretonic vowel reduction, has a confidence of 1. However, this value is based on just three forms, and is intuitively not as reliable. The rule’s much lower adjusted confidence of 0.6 reflects this.

Adjusted confidence values can then be used to evaluate how well each model captures the lexicon. To do this, each verb in the corpus is assigned a ‘score’, which is the product of the adjusted confidence of all the rules needed to derive the correct output form. For example, consider the verb [‘talaŋ~tu’laman] ‘run’. Under the stem-base model, the input [‘talaŋ] can be used to derive the observed output [tu’laman] using four rules, as summarised in Table 3.1. The score for [tu’laman] is therefore the product of the adjusted confidence values of these four rules; as shown in (25), this ends up being 0.019. Under this metric, the model that performs better (stem vs. suffix base) will assign higher scores to the lexicon.

Rule	Example	Hits/Scope	\hat{p}
Suffixation	'talaŋ →'talaman	1.00	0.99
PenultStress	'talaman →ta'laman	1.00	0.99
PretonicVR	ta'laman →tu'laman	1.00	0.99
ŋ-to-m	tu'laman →tu'laman	0.06	0.02

Table 3.1: Derivation of [‘talaŋ]→[tu'laman] in the stem-base model

$$\begin{aligned}
(25) \quad s('talaŋ) &= \hat{p}(\text{Suffixation}) \times \hat{p}(\text{PenultStress}) \times \hat{p}(\text{PretonicVR}) \times \hat{p}(\eta\text{-to-m}) \\
&= 0.99 \times 0.99 \times 0.99 \times 0.02 \\
&\approx 0.019
\end{aligned}$$

3.3 Comparing stem and surface base models

Fig. 3.1 below compares the stem and surface-base models on their performance in predicting the test data of 340 verb stems. From this figure, it is evidence that the stem-base model assigned much higher scores to the test data than the suffix-base model. In other words, the stem form of Seediq verb paradigms can be used to predict the suffixed forms of a paradigm with much higher accuracy than the other way around. This confirms the asymmetry observed in Chapter 2.

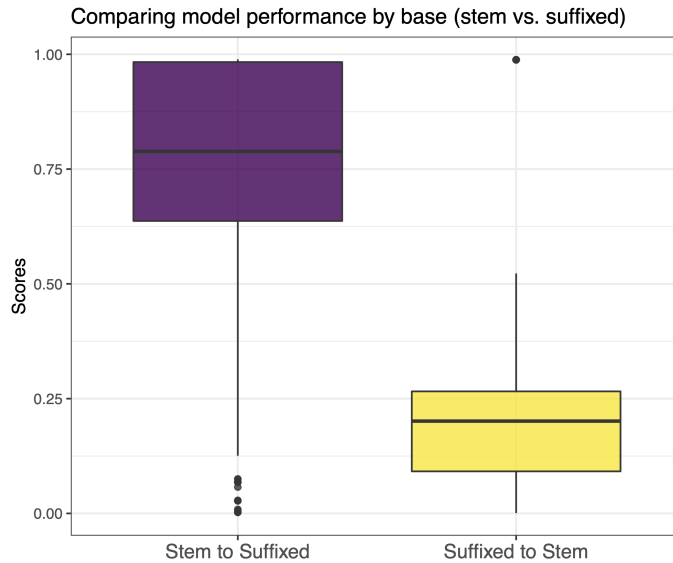


Figure 3.1: Performance of stem vs. suffix-base models

Note also that in the stem-base model, verbs which were assigned a very low scores (corresponding to the outliers in the bottom left corner of Fig. 3.1 either showed an irregular alternation, or a disfavored alternation; examples of these are shown below in (26), where (a) and (b) are both verbs showing irregular alternations, while (c) is a verb which shows a

regular but disfavored [ŋ]-[m] alternation.

(26) *Representative verbs assigned $p < 0.1$ in stem-base model*

STEM (input)	SUFFIXED (output)	SCORE
(a)	'adis	'des-an 0.02
(b)	'hado	'hadan 0.07
(c)	'talaŋ	tu'laman 0.05

In contrast, the stem-base model performs badly for almost all verb paradigms. This is primarily because, as discussed in Section 2.4, it is very hard to “undo” the effects of pretonic vowel neutralization; the patterns of predictability that allow us to predict what a vowel neutralised in the suffixed form will surface as in the stem form are weak. For example, consider the verb [ˈbiciq~buˈciqan] ‘few, less’. Under a stem-base model, the suffixed form can be derived using rules of stress assignment and pretonic vowel reduction, both of which are completely predictable and have high adjusted confidence; this is demonstrated below in Table 3.2.

In contrast, the suffix-base model, which takes [buˈciqan] as the input, must predict what the reduced [u] will surface as when stressed in the stem output. As seen in Table 3.3, a rule which changes /u/ to [u] when the following vowel is [i] has a very low adjusted confidence of 0.139. As a result, under this model, [ˈbiciq] gets a low score of 0.14.

Rule	Example	Hits/Scope	\hat{p}
Suffixation	'biciq → 'biciqan	1.00	0.994
PenultStress	'biciqan → bi'ciqan	1.00	0.994
Pretonic VR	bi'ciqan → bu'ciqan	1.00	0.993
score = $0.994 \times 0.994 \times 0.993 \approx \mathbf{0.98}$			

Table 3.2: Derivation of [ˈbiciq]→[buˈciqan] in the stem-base model

Rule	Example	Hits/Scope	\hat{p}
remove-suffix	bu'ciqan → bu'ciq	1.00	0.994
PenultStress	bu'ciq → buciq	1.00	0.994
/u/ → [i] / _C <i>0</i> i	'buciq → 'biciq	0.14	0.139
score = 0.994 × 0.994 × 0.139 ≈ 0.14			

Table 3.3: Derivation of [bu'ciqan] → ['biciq] in the suffix-base model

3.4 Indirect evidence for historical reanalysis

The asymmetry between stem and suffix bases by itself does not necessarily support the idea that historical re-analysis has taken place with the stem form as base.

In particular, the stem form is a good base in part because neutralised segments either almost always or almost never alternate. On one hand, this could be because historical reanalysis has exaggerated asymmetries in rates of final alternation, rendering the suffixed form predictable from the stem form. On the other hand, uneven rates of alternation may be an accidental effect of baseline preferences for certain sounds in the lexicon. For example, final [c] strongly prefers to alternate with [t]; this may be because there's a strong baseline phonotactic preference for [t] (relative to [c]). Similarly, the [ŋ]-[m] alternation is rare, but this could be because [m] is relatively low in frequency (and dispreferred) in the lexicon.

To test whether this is the case, the two surface-base models were tested against a simulated lexicon of 700 verb paradigms, in which the rates of alternation are determined by relative frequencies of sounds in the Seediq lexicon (regardless of which position in a word they occur in). For example, across the corpus of 340 paradigms, [ŋ] (n=104) is around 2.1 times more frequent than [m] (n=49). Corresponding to this, the [ŋ]-final forms in the simulated lexicon are 2.1 times more likely to *not* alternate (than to alternate with [m]).

To make this lexicon, 700 URs were first stochastically generated based on the distribution of sounds in the lexicon. These URs were then used to derive surface stem and suffixed forms based on the regular phonological rules described by Yang (1976); at these stage, the rules were assumed to be exceptionless. If the rates of alternation in the stem form are a result of

baseline phonotactic preferences, then the stem-base model should perform equally well on both the training data and the simulated lexicon.

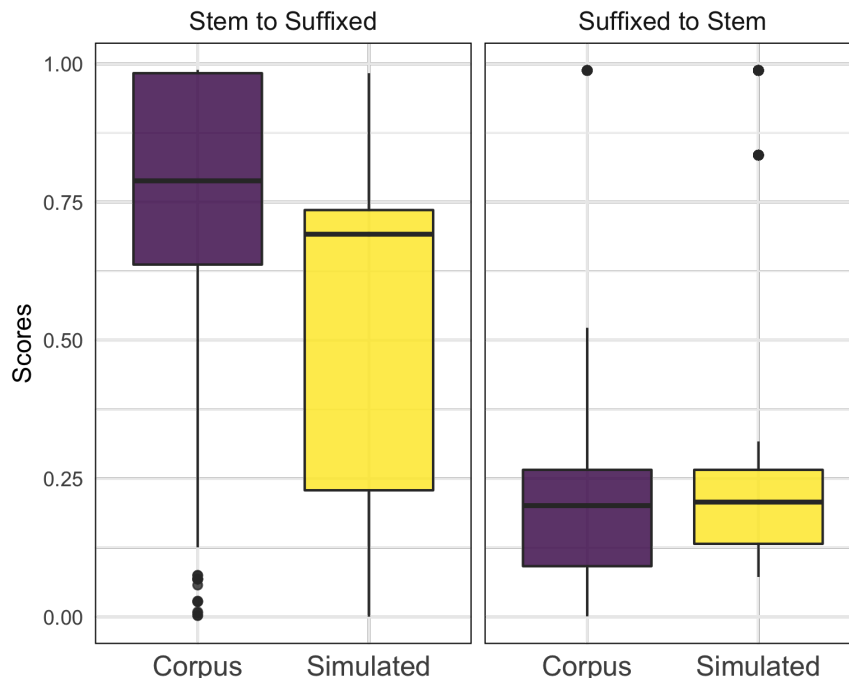


Figure 3.2: Model performance using real vs. simulated lexicon

Fig. 3.2 compares the performance of the stem- and suffix-base models on the real and simulated lexicons. We see that the suffix-base model shows equally bad performance on both lexicons. On the other hand, the stem-base model shows a clear asymmetry, and performs worse on the SIMULATED set. This suggests that the preference for alternation or non-alternation in the stem-base model cannot be explained away by gradient phonotactic preferences for certain segments. Instead, there appears to be an asymmetry in the distribution of sounds in exactly the places which show alternation. This is expected if Seediq verb paradigms have undergone reanalysis in a way which makes the stem forms a better predictor of the suffixed forms, as predicted by Albright’s single surface base hypothesis.

3.5 The selection of a base form

So far, evidence strongly suggests that speakers have designated a non-suffixed slot of the verbal paradigm to be the base. At this point, I briefly discuss potential reasons for why the non-suffixed form, rather than the suffixed form, was designated as the base form.

The literature on levelling suggests that frequency, morphological complexity, and phonological markedness may all influence the selection of a base form. However, in his formulation of the single surface base hypothesis, Albright (2002) posits a more restrictive criterion: the base should be the “most informative” member of the paradigm, in that it suffers from the fewest neutralizations and also affects the fewest lexical items.

In Seediq, non-suffixed forms do not intuitively suffer from fewer neutralizations than the suffixed forms: in the non-suffixed forms, both the final vowel and final consonant of the stem may be neutralised. In contrast, in the suffixed forms, contrast is only lost in the penultimate vowel of the stem (due to pretonic vowel neutralization). However, as discussed in Section 2.4, pretonic vowel neutralization affects a much larger number of lexical items than the other neutralization processes. As such, the suffixed forms could have already been less informative before any leveling happened, consistent with Albright’s criterion.

Moreover, in his reconstruction of Proto-Atayalic (which includes dialects of Seediq and Atayal), Li (1981) finds evidence that pre-tonic reduction occurred prior to all of the post-tonic neutralization processes: Post-tonic sound changes applied to different degrees in a subset of dialects of Seediq, while pretonic vowel neutralization affects all dialects of Seediq, and all but the two most conservative dialects of Atayal (Li, 1981: 239). From this, we could speculate that at some point after pretonic neutralization had taken place, the non-suffixed forms of the Seediq verb paradigm had become much more informative than the suffixed forms of the paradigm. In other words, pre-tonic vowel neutralization could have acted as a ‘tipping point’; it resulted in a lexicon where the stem forms were much more informative than the suffixed forms, and subsequent leveling processes would only have exaggerated

patterns of predictability from the stem form.

Modeling results from Figure 3.2 are also consistent with Albright's informativeness account. Notably, even when tested on the SIMULATED lexicon, the stem-base model performs much better than the suffix-base model. This suggests that, even before restructuring rendered caused verb paradigms to be increasingly predictable from the stem (or other non-suffixed slots of the paradigm), the stem form was already more informative than the suffixed form.

CHAPTER 4

Towards a theoretical surface-base model for Seediq

Having laid out the evidence in favor of a stem-base account of Seediq verbal paradigms, I now turn to the task of developing an explicit surface-base model to verbal morphophonology in Seediq. The model assumes Optimality Theory (Prince and Smolensky, 1993). In particular, I will use the framework of Maximum Entropy (MaxEnt) Harmonic Grammar (Goldwater and Johnson, 2003; Smolensky, 1986), which is a stochastic variant of OT.

A complete model of Seediq verbal phonology must account for how alternation is enforced under a surface-base approach. To address this issue, I first describe the constraints needed under a cobbled-UR model, and then outline the changes and potential challenges that are introduced in a stem-base model. To allow for a more direct comparison of the cobbled-UR and stem-base models, the constraint sets will first be introduced in strictly-ranked classical OT.

The model should also be able to characterise **native speakers intuitions** about how Seediq verbal alternations. In other words, it should make clear predictions about whether speakers will apply a given alternation to a novel stem, and also characterise well-formedness intuitions about novel stem-suffix pairs. To address this, I implement the model in MaxEnt. Because MaxEnt is stochastic, it assigns all candidate output forms a probability value; relative probabilities can then be interpreted as a prediction of native speakers' gradient intuitions about well-formedness.

Finally, assuming that well-formedness intuitions are gradient, the model should also be able to accommodate **stem-specific behavior**. Since lexical items in Seediq never show

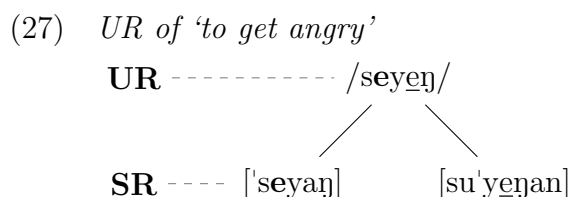
token variation in terms of whether they alternate, the model should have some way of enforcing the invariance of existing forms. To deal with the invariance of lexical items, I adopt the DUAL LISTING/ GENERATION MODEL of Zuraw (2000, 2010).

4.1 Input and GEN

For both the cobbled-UR and stem-base models, the inputs were based off of the 340 forms of my Seediq corpus, and the winning candidates were the attested forms. In the stem-base model, inputs are simply the 340 surface stem forms.

For the cobbled UR model, URs are formed by combining the most informative parts of the stem and suffixed forms, as detailed in Section 1.7. Both stem and suffixed URs were included as inputs, resulting in a total of 680 inputs.

In the case of irregular vowel alternations, the vowel taken to be in the UR is the one that is stressed in the surface form. For example, the verb [ˈseyɑŋ ~ suyeŋ-an] ‘to get angry’ shows an irregular vowel alternation, where the post-tonic [a] of the stem form surfaces as [e] when stressed in the suffixed form. As illustrated in (27), the cobbled UR takes /e/ to be the underlying vowel.



The candidate set included all applications and non-applications of attested alternations, regardless of how general the alternation is. Specifically, all combinatorial possibilities of alternation and non-alternation were included; candidates which show unattested alternations are assumed to be ruled out by undominated constraints, and therefore not included in the model.

For example, consider the verb [ˈheguc~huˈgetan] ‘to inhale’, whose candidates are shown in (28). Under the cobbled UR model, the stem has a UR /heget/. Given this UR, there are two possible sources of alternation: alternation of the post-tonic vowel /e/, and alternation of the final consonant /t/. As seen from the table, candidates such as [ˈhegut] are included to account for the alternation of post-tonic /e/ with [u]. However, candidates are also included for all attested irregular alternations involving post-tonic /e/. /e/-[a] alternation is irregular, but attested in the verb [ˈtebas ~ tuˈbesan] ‘to sieve grains’, while the /e/-[i] alternation is attested in the form [ˈadis~ˈdesan] ‘to carry away’. To account for these alternations, candidates such as [ˈhegic] and [ˈhegac] are included. For stem-final /t/, the /t/-[c] alternation is observed; candidates with a final [c] ([ˈheguc, ˈhegac, ˈhegic]) are included to account for this alternation.

(28) *Candidate set for ‘to smoke’ (/heget/ ˈheguc~huˈgetan)*

UR	Candidate	t~c	e~u	e~a	e~i
/heget/	ˈheget				
	ˈhegut		✓		
	ˈhegat			✓	
	ˈhegit				✓
	ˈhegec	✓			
	ˈheguc	✓	✓		
	ˈhegac	✓		✓	
	ˈhegic	✓			✓

4.2 Constraint set under a cobbled UR model


Under a cobbled-UR approach to analyzing Seediq verb alternations, the constraint set includes markedness constraints for each observed alternation, and relevant faithfulness constraints enforcing INPUT-OUTPUT correspondence (see Appendix B for the full constraint set and ranking). Note that in the model being described, markedness constraints were only

included if they were general enough to apply (i.e. assign violations) to at least three forms in the corpus. As a result, if an input form shows an alternation that can't be captured by a general enough constraint, the model will predict the wrong output for it; a more detailed discussion of this is found in Section 4.3. As will be seen, under a cobbled UR approach, phonotactically motivated markedness constraints and classical feature-based input-output correspondence constraints are mostly sufficient for capturing the regular patterns of alternation.

4.2.1 Constraints for final consonant alternation

Alternation of final consonants is enforced by positional markedness constraints. For example, consider the alternation of [d] with [c] in the verb ['harac~hu'radan] 'to build (with stone)', which has the UR /harad/. As shown in tableau (29) on the following page, alternation of /d/ with [c] results from ranking $*d]_w$, which prohibits final [d], above IDENT-IO[*delayed release*] and IDENT-IO[*voice*]. With this ranking, the faithful candidate (a) is eliminated by the higher ranked markedness constraint.¹

(29) *Tableau for /harad/ [harac]*

/harad/	$*d]_w$	ID[voice]	ID[del rel]
a. harad	*!		
 b. 'harac		*	*

The identity of the alternant simply results from the relative ranking of faithfulness constraints. This is demonstrated for the /d/-[c] alternation in tableau (30). Note that candidate (b), ['harat], whose faithfulness violations are a subset of the winning candidate, is ruled out by a highly ranked constraint $*t]_w$ (as both [t] and [d] are phonotactically illegal word-finally).

¹Although not shown here, other repair methods (e.g. deletion, vowel insertion) are ruled out by highly ranked MAX and DEP.

(30) *Constraint ranking for the /d/-[c] alternation*

/harad/	*t _w	*d _w	ID[cont]	ID[son]	ID[COR]	ID[voice]	ID[del rel.]
a. harad		*!					
b. 'harat	*!					*	
☞ c. 'harac						*	*
d. 'haras			*!			*	
e. 'haran				*!			
f. 'harap					*!		

4.2.2 Constraints for pretonic vowel neutralization

Pretonically, vowels either delete (if onsetless), assimilate to an stressed vowel (if separated by /h,ʔ/, or reduce to [u]. All three patterns can be captured by straightforward markedness constraints.

Pretonic vowel reduction to [u] is enforced by a positional licensing constraint (Crosswhite, 2004). The constraint LICENSE[u]/*pretonic* (henceforth LIC[u]/pret) is defined in (31), and essentially penalises non-[u] syllables in pretonic position. This is demonstrated below in tableau (32) for ['barig~bu'rigan].

The winning candidate violates IDENT[high] due to alternation of /a/ with [u]. However, the faithful candidate (a) fatally violates a higher ranked LIC[u]/pret, and is therefore eliminated. Candidate (c), which repairs the markedness violation by deleting the stem's initial syllable, is eliminated by ranking MAX-C above LIC[u]/pret.

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

(32) *Pretonic vowel reduction in ['barig~bu'rigan] /barig/*

/barig-an/	LIC[u]/pret	MAXC	ID[high]
a. ba'rigan	*!		
☞ b. bu'rigan			*
c. 'rigan		*!	

Seediq pretonic assimilation to stressed syllables can be enforced by an AGREE constraint (Bakovic, 2000); $\text{AGREE}(v(h)'v)$, defined below in (33), penalises non-identity between a pretonic vowel and a stressed vowel if the two are separated by [ʔ] or [h]. Note that although not discussed here in detail, this is a case of the widespread process of translaryngeal harmony (Steriade, 1987).

Tableau (35) below shows how this constraint can be used to derive the correct output for [leiq̣~li'ijan] /leiq̣/. The faithful candidate (a) is ruled out by $\text{AGREE}(v(h)'v)$; crucially, because $\text{AGREE}(v(h)'v) \gg \text{LIC}[u]/\text{pret}$, candidate (b), where the pretonic vowel reduces to [u], is also eliminated.

Candidates (c) and (d) both resolve the $\text{AGREE}(v(h)'v)$ violation, but candidate (c), which resolves the agree violation by changing the pretonic vowel, is the winning candidate. In contrast, candidate (d), where the stressed vowel assimilates to the pretonic vowel, is dis-preferred. To capture this generalization, we can bring in positional faithfulness constraints (Beckman, 1998); a positional IDENT constraint, which protects the stressed syllable, is defined in (34). In tableau (35), candidate (d) incurs a fatal violation of $\text{IDENT-}'\sigma[\text{high}]$, and is correctly eliminated.

(33) $\text{AGREE}(v(h)'v)$: A pretonic vowel must be identity to a following stressed vowel if separated by [h] or [ʔ].

(34) $\text{IDENT-}'\sigma(F)$: Output segments in a stressed syllable and their input correspondents must have identical specifications for the feature F (Beckman, 1998: 31).


(35) *Pretonic vowel assimilation in [leiq̣~li'ijan] /leiq̣/*

/leiq̣-an/	AGREEV	LIC[u]/pret	ID-' σ [hi]	ID[hi]
a. le'ijan	*!	*		
b. lu'ijan	*!			
☞ c. li'ijan		*		*
d. le'ejan		*	*!	*

Deletion of onsetless pretonic vowels can be enforced by the markedness constraint ONSET, defined in (36), which penalises onsetless vowels. This is demonstrated in tableau (37), for the word [ˈawak~ˈwakan] /awak/. Candidates (a) and (b) are eliminated because they violate the highly-ranked ONSET constraint. Candidates such as (d), which repairs the ONSET violation through epenthesis instead of deletion, are ruled out due to highly-ranked DEP constraints ².

(36) ONSET: Syllables must have onsets; incur a violation for each onset-less syllable (Prince and Smolensky, 1993).

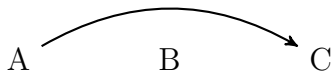
(37) *Onsetless vowel assimilation in [ˈawak~ˈwakan] /awak/*

/awak-an/	ONSET	DEPC	*[GLIDE]	LIC[u]/pret	MAXV	ID[hi]
a. aˈwakan	*!			*		
b. uˈwakan	*!					*
 c. ˈwakan					*	
d. tuˈwakan		*!				

4.2.3 Dealing with saltatory alternations in Seediq

The alternations discussed so far can all be dealt with in classical OT. However, a subset of alternations in Seediq involve saltation, which is known to be a problem for classical OT (Lubowicz, 2002; Ito and Mester, 2003). Saltation, as schematised in (38), describes cases where one sound A alternates with another sound C, but a phonetically and featurally ‘intermediate’ sound B doesn’t alternate.

(38) *Saltation*

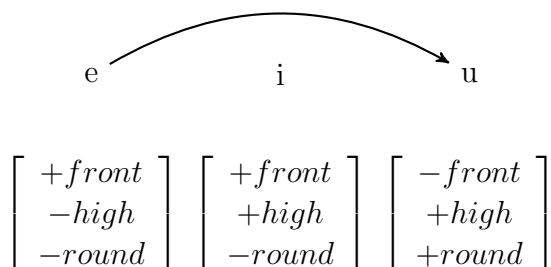


²Although not discussed here, onsetless vowels surface faithfully in stressed position, presumably due to higher-ranked positional faithfulness constraints

An example of saltation in Seediq is post-tonic vowel reduction. Post-tonically, because the mid vowels /e, o/ reduce to [u], only the peripheral vowels [i a u] are observed. This generalization can be captured by a markedness constraint LICENSE(nonperipheral/' σ). This constraint, defined in (39), penalises [e, o] in post-tonic position.

- (39) LICENSE(nonperipheral/' σ): Non-peripheral vowels can only appear in stressed syllables.


- (40) *Saltatory vowel reduction in Seediq*



However, LIC(non-periph)/' σ alone is not enough to enforce reduction of /e/ to [u]. Because /i/ is acceptable and non-alternating in post-tonic position, /e/ is expected to alternate with the featurally more similar /i/. Instead, as illustrated in (40), it reduces to the featurally more dissimilar [u]. As pointed out by Ito and Mester (2003), saltatory alternations of this variety cannot be derived in classical OT, and result in ranking contradictions.

This is demonstrated for the /e/-[u] alternation in (41), for the stem ['baluŋ] /baleŋ/ 'lay eggs'. Ranking LIC(non-periph)/' σ above faithfulness constraints correctly eliminates the faithful candidate (a). However, because a change of /e/ to [i] violates less faithfulness constraints than a change of /e/ to [u], this ranking will incorrectly predict the output to be (b). On the other hand, ranking faithfulness above LIC(non-periph)/' σ would prevent post-tonic /e/ from alternating, and incorrectly predict the winner to be candidate (a).

(41) *Post-tonic /e/-[u] alternation in classical OT*

/baleŋ/	LIC(non-periph)/' σ	ID[high]	ID[front]
a. 'baleŋ	*!		
 b. 'baliŋ		*	
c. 'baluŋ		*	*!

Aside from post-tonic vowel reduction, alternations involving final /g/ are also saltatory. Recall that word-final /ag, eg, ig/ respectively reduce to [o, u, uy] (refer to Section 1.5 for a more detailed discussion). These alternations are saltatory in that final /g/ is expected to alternate with the featurally more similar [k], which is licit word-finally. Word-final monophthongization of /ey/ to [u] is also saltatory, in that /ey/ is expected to reduce to [uy] (which, as discussed in Section 1.4, is allowed word-finally).

To deal with these saltatory alternations, I turn to the solution proposed by Hayes and White (2015), which is to use *MAP constraints. In this family of faithfulness constraints proposed by Zuraw (2013), *MAP(*a,b*) assesses a violation to a candidate if *a* is mapped to a corresponding *b*. Crucially, the corresponding segments *a* and *b* can differ more than one feature. Thus, there could be a *MAP constraint penalising the input-output pair /e/-[u], even though they differ along multiple features ([front], [high], and [round]). *MAP constraints have a default ranking constrained by the p-map (Steriade, 2001). However, Hayes and White (2015) propose that this ranking can be subverted given enough language-specific evidence, giving rise to saltatory alternations.

This is demonstrated in tableau (42) below, for the /e/-[u] alternation; the same approach can be straightforwardly extended to the other saltatory alternations. As with tableau (41) above, the high-ranking LIC(non-periph)/' σ constraint correctly eliminates the faithful candidate (a). In this case though, ranking *MAP(e,i) above *MAP(e,u) causes candidate (b), which shows the /e/-[i] alternation, to be correctly eliminated. The winner, candidate (c), has the observed /e/-[u] alternation.

(42) *Post-tonic /e/-[u] alternation using *MAP*

/bareŋ/	LIC(non-periph)/'σ	*MAP(e,i)	*MAP(e,u)
a. 'bareŋ	*!		
b. 'bariŋ		*!	
☞ c. 'baruŋ			*

4.3 Model performance under a cobbled-UR model

Overall, a cobbled-UR OT model of Seediq verbal alternation is able to capture most of the Seediq lexicon (barring irregularly alternating forms) using general markedness constraints and classical feature-based faithfulness constraints. In a subset of alternations, such as post-tonic /e/-[u], *MAP constraints are needed to get the correct alternant. In other words, feature-based faithfulness constraints are not sufficient even under a cobbled-UR approach. Crucially, however, all alternations have a clear phonotactic motivation, and can be explained using straightforward markedness constraints.

Note that under the restrictions I've set, where constraints must be general enough to apply to at least 3 forms, the cobbled UR model will inevitably make the incorrect prediction for verbs which show irregular alternations, and therefore cannot be straightforwardly derived under a cobbling analysis.

For example, consider the verb [ˈpatis~puˈtasan] ‘to write’, which has the UR /patas/. This verb shows an alternation of /a/ with [i]. As demonstrated in tableau (43), this alternation can be enforced by ranking a markedness constraint such as *as] (prohibiting final [as]) above faithfulness.³


(43) *Enforcing /a/-[i] alternation in ‘to write’*

/patas/	*as]	FAITH
a. patas	*!	
☞ b. patis		*

³In the tableau, FAITH represents feature-based constraints such as IDENT[high], Ident[back], etc.

However, as demonstrated in tableau (44), this ranking will predict the wrong output for regular, non-alternating forms such as [‘qamas~qu’mas-an] /qamas/ ‘to pickle’.

(44) *Constraint conflict in ‘to pickle’*

/qamas/	*as]	FAITH
a. qamas	*!	
 b. qamis		*

To resolve constraint ranking conflicts of this variety, I manually picked the ranking which would predict the correct output for more forms. For example, the ranking $*as] \gg FAITH$ predicts the right output for one verb (‘patis~pu’tasan), but results in the wrong output for six other [as]-final stems. In contrast, the opposite ranking of $FAITH \gg *as]$ gives us the right output for six verbs, and the wrong result for just one verb. As such, the final model adopts a ranking of $FAITH \gg *as]$.

Using this method, the cobbled-UR model correctly predicts 92% of input forms correctly, and misses 27 stem-suffix pairs. Note that all of the missed forms show irregular alternations (these were discussed above in Section 1.6).

4.4 Constraint set under a stem-base model

In this chapter, I will outline the constraints needed in a stem-base OT model of Seediq verbal paradigms, and point out deviations from cobbled-UR model discussed above. Crucially, whereas general markedness and faithfulness constraints were enough to enforce alternations under a cobbled UR model, this is not true for all alternations in a stem-base model.

4.4.1 Constraints for pretonic vowel neutralization

Pretonic vowel alternations can be analyzed with classical markedness and faithfulness constraints. Note that in the stem-base model, output-output (OO) faithfulness, rather than input-output (IO) faithfulness, is used to enforce non-alternation. With this minor modifi-

cation, the analyses above for pre-tonic vowel deletion, assimilation, and reduction can all be straightforwardly extended to the stem-base model.

For example, tableau (45) shows the analysis of pre-tonic vowel reduction for the input [ˈbarig]. Candidate (a), despite incurring no faithfulness violations, is ruled out by ranking LIC[u]/pret above faithfulness. Candidate (c), which resolves the LIC[u]/pret violation, is eliminated because MAXC ranks above IDENT[high]. Similarly, pretonic vowel assimilation is motivated by AGREE(v(h)'v), and deletion is motivated by ONSET.

(45) *Pretonic vowel reduction in [ˈbarig~buˈrigan] /barig/*

[ˈbarig]	LIC[u]/pret	MAXC	ID[high]
a. baˈrigan	*!		
☞ b. buˈrigan			*
c. ˈrigan		*!	

4.4.2 Introducing Anticorrespondence

While pretonic vowel alternations could be straightforwardly analyzed using general markedness constraints, this is not true of the other alternations, which cannot be related to or motivated by principles of general markedness.

For example, under the stem-base model, final [c] is expected to alternate with [t] in words like [ˈbirac~buratan]. However, [c] is allowed intervocally, as evidenced by stems like [ˈbiciq] ‘decrease.’ As such, markedness constraint such as *VcV (penalizing intervocally [c]) would incorrectly predict that VcV sequences are disfavored in the lexicon.


To deal with this issue, I turn to ANTICORRESPONDENCE constraints (Hayes, 1999). ANTICORRESPONDENCE, defined formally in (46), are markedness constraints which actively require that morphemes alternate in particular ways. For example, ANTICORR(c]_w, STEM, t, SUFFIX) would require final [c] in the isolation stem to alternate with [t] in the suffixed form. Note that, since all alternations discussed here will involve alternation between the base allomorph (i.e. the non-suffixed stem form) and the projected suffix form

allomorph, constraints such as $\text{ANTICORR}(c]_w, \text{STEM}, t, \text{SUFFIX})$ will simply be referred to as $\text{ANTICORR}(c, t)$, with the morphological context omitted.

- (46) $\text{ANTICORRESPONDENCE}(X, C, X', C')$: if morpheme μ appears with shape X in context C , it must appear with shape X' in a distinct context C' . In Seediq, C refers to the base (non-suffixed) allomorph, while C' is the projected suffix form allomorph.

Tableau (47) illustrates how $\text{ANTICORR}(c, t)$ can be used to enforce the $[c]$ - $[t]$ alternation in $['\text{birac} \sim \text{bu}'\text{ratan}]$. Simply ranking $\text{ANTICORR}(c, t)$ above faithfulness will correctly eliminate the non-alternating candidate (a), giving us the correct winning candidate (b).⁴

- (47) *Stem-base account of* $['\text{birac} \sim \text{bu}'\text{ratan}]$

$['\text{birac}]$	$\text{ANTICORR}(c, t)$	$\text{ID}[\text{del rel}]$
a. $\text{bu}'\text{racan}$	*!	
 b. $\text{bu}'\text{ratan}$		*

In the case of final $[c]$, where alternation with $[t]$ is preferred, the relevant ANTICORR constraint is ranked above faithfulness. For dispreferred alternations, on the other hand, ANTICORR will be ranked below faithfulness, preventing alternation.

4.5 Model performance under a stem-base approach

Recall that in a surface-base approach to Seediq, the base will not be able to perfectly predict outputs in the same way that cobbled URs can. For example, ranking $\text{ANTICORR}(c, t)$ above faithfulness correctly predicts the $[t]$ - $[c]$ alternation in 78% ($n=18$) of $[t]$ -final forms. However, this ranking will predict the wrong output for forms which are non-alternating ($n=1$, 4%), or show the $[t]$ - $[d]$ alternation (17%, $n=4$). Similarly, in the case of the $[\eta]$ - $[m]$ alternation, ranking faithfulness above $\text{ANTICORR}(\eta, m)$ generates the correct output for 32 non-alternating $[\eta]$ -final stems, and the wrong output for 2 alternating stems.

⁴Candidates with a non-reduced pretonic vowel, such as $[\text{bi}'\text{racan}]$, are ruled out by $\text{LIC}(\text{non-periph})/\prime\sigma$.

Because of this, the stem-base model inevitably performs worse than the cobbled UR model. Overall, it predicts the correct output form for 83% of stems (where 58 forms are missed by the model). Missed forms include both ones which show irregular alternations, and ones which show a dispreferred alternation (or non-alternation).

In contrast, as discussed in Section 4.3, the cobbled-UR model performs better, predicting the correct output for 92% of forms. In addition, the missed forms of the cobbled-UR model are a subset of the missed forms in the stem-base model; there are no verbs for which the cobbled UR predicts the wrong output, but the stem-base model predicts the correct output.

Examples of the forms which the stem-base model predicts the wrong output for are shown in (48). In forms such as (48a), the observed output shows a dispreferred alternation. For this specific case, the model predicts that ['geruŋ] will surface as [gu'reŋ-an] (because ANTICORR(ŋ,m) is ranked below faithfulness), but ['gerum] is observed. (48b) is an example of the opposite issue, where the observed form is non-alternating, but the model predicts an alternation. Specifically, the final [c] in [bu'cebac] is predicted to alternate with [t], but the observed suffixed form is non-alternating. Similar, the model predicts the wrong output in (48c) because a preferred alternation is not observed. In this case, the model predicts the output of ['egun] to be ['gelan], where the stem-final vowel shows an [u]-[e] alternation due to VOWEL MATCHING.

(48) *Examples of forms missed by the stem-base*

	INPUT	PREDICTED	OBSERVED
(a)	'geruŋ	gu'reŋan	gu'reman
(b)	bu'cebac	bucu'batan	bucu'bacan
(c)	'egun	'gelan	'gulan

The surface-base model also requires the introduction of ANTICORRESPONDENCE constraints, as alternation cannot be motivated by general markedness. In contrast, the cobbled-UR model enforced alternation using phonotactically motivated markedness constraints.

On the other hand, adopting anticorrespondence constraints does get us relatively far;

the stem-base model, which accounts for 83% percent of forms, performs only slightly less well than the cobbled-UR grammar (which has 92% accuracy), despite the latter having far greater access to informational resources. Most importantly, it is more consistent with the stem-suffix asymmetry observed in the data, and encodes the preferences that have guided restructuring in Seediq.

4.6 Moving to a stochastic model

In the following section, I discuss the motivations for adopting a stochastic model of Seediq verbal alternation, and outline how constraint weights are learned under this model.

4.6.1 Predicting speaker intuitions

A model of Seediq verbal alternations should be able to characterise the native speaker's **intuitions**. In particular, the model should make clear predictions on how a novel stem will alternate, and about the well-formedness of novel stem-suffix pairs.

The constraints and constraint rankings laid out above can be used to reflect speakers' general preference for alternation or non-alternation. However, the strictly-ranked OT model adopted so far predicts that speakers will apply alternations categorically, and have non-gradient judgements about the well-formedness of novel paradigms.

Contrary to this, a large body of work suggests that speaker intuitions about well-formedness are gradient (e.g. Frisch et al., 2004), and that speakers will apply alternations in a way which matches the frequency of alternations in the lexicon (e.g. Zuraw, 2000; Ernestus and Baayen, 2003; Hayes et al., 2009; Zuraw, 2010). As discussed in the beginning of Chapter 4, the stochastic nature of MaxEnt is able to capture this type of gradient behavior.

4.6.2 Accounting for stem-specific behavior

The model should also be able to accommodate **stem-specific behavior**. The available Seediq data suggest that variation in verb paradigms is primarily stem-by-stem variation; there is little, if any, token variation.

This invariance of existing forms must be encoded in the lexicon. To do this, I adopt the DUAL LISTING/GENERATION MODEL (Zuraw, 2000). Under this approach, both the stem and suffixed allomorphs of existing words are listed in the grammar. The grammar includes markedness constraints and output-output faithfulness constraints, whose relative weights give rise to speakers' gradient judgement of novel forms. However, highly weighted input-output faithfulness constraints protect listed forms from variation.⁵

First, consider the two verbs in (49), both of which have an [ŋ]-final stem. Because ANTICORR(ŋ,m) is in the grammar, there should be some probability that either form will show the [ŋ]-[m] alternation. However, for actual Seediq speakers, form (49a) always alternates, while (49b) never alternates. This is captured by listing the suffixed lexical entries /gureman/ and /buleŋan/ in the lexicon.

(49) *Listed forms for the [ŋ]-[m] alternation*

	EXAMPLE	LEXICAL ENTRY
(a)	<i>Alternating</i>	'geruŋ~gu'reman /gureman/
(b)	<i>Non-alternating</i>	'baluŋ~bu'leŋan /buleŋan/

Now, consider the verb ['geruŋ~gu'reman] from (49a). Because the [ŋ]-[m] alternation is dispreferred, we know that OO-FAITH should have high weight relative to ANTICORR(ŋ,m) (or, in a strict ranking model, OO-FAITH≫ANTICORR(ŋ,m)). Even if this is the case, because /gureman/ is listed as a lexical entry, highly weighted (or ranked) IO-FAITH causes the correct suffixed form to surface. This is demonstrated in tableau (50), where for simplicity

⁵Zuraw's (2010) model also employs USELISTED constraints, which require a listed entry to be employed. For the purposes of the present analysis, since stem-suffix pairs are not known to show token variation, USELISTED is assumed to be inviolable and highly weighted.

I use a strict ranking model. Candidate (b), which is faithful to the *surface stem allomorph* [ˈgeruŋ], is eliminated because it violates IO-FAITH (to the listed form /gureman/).

(50) *Using lexical listing to enforce alternation*

/gureman/ related to [ˈgeruŋ]	IO-FAITH	OO-FAITH	ANTICORR(ŋ,m)
☞ a. gu'reman		*	
b. gu'rejan	*!		*

4.6.3 Constraint weights

In my MaxEnt model, the input, candidate set, and constraint set are the same as the ones used in the strictly ranked stem-base model discussed above. To learn the weights of the markedness constraints motivating alternation, I first assume that that this part of learning is not influenced by IO-faithfulness to listed suffixed forms such as /gu'reman/ and /bu'lejan/ in (49a) above. In other words, weights were learned under a MaxEnt model which included only markedness, anticorrespondence, and OO-faithfulness constraints. This is the approach taken by Zuraw (2010) and Hayes et al. (2009), but there are in principle alternative approaches to learning constraint weights under a model which accounts for lexical specificity, such as constraint cloning (Becker, 2009). Weights were learned using Excel Solver (Fylstra et al., 1998), using the Conjugate Gradient Descent method.

Table 4.1 shows the resulting markedness and anticorrespondence constraint weights; note that only constraints which received a non-zero weight are shown here.⁶ In MaxEnt, constraints with higher weights are, intuitively speaking, stronger. In other words, unviolated markedness constraints are expected to have the highest weight, and constraints involving disfavored alternations are expected to have little to no weight.

Consistent with this, constraints such as AGREE(v'hv), which enforces the exceptionless pretonic assimilation of vowels, get a very high weight. Looking at the ANTICORRESPONDENCE constraints, we also see that in general, constraints associated with a high rate of al-

⁶For a full list of constraint weights, including for OO-faithfulness constraints, refer to Appendix C.

ternation are assigned a relatively higher weight. For example, $\text{ANTICORR}(\text{CeCuC}, \text{CuCeC})$ enforces vowel matching in isolation stems of the form CeCuC (which have a stressed [e] and reduced post-tonic vowel). In Section 2.1, a strong preference for this pattern of alternation was found (with alternation happening 80% of the time). Consistent with this, $\text{ANTICORR}(\text{CeCuC}, \text{CuCeC})$ has a relatively high weight of 5.62. In contrast, $\text{ANTICORR}(\eta, \text{m})$, which enforces the dispreferred $[\eta]$ - $[\text{m}]$ alternation, has a near-zero weight of 0.01.

General markedness			
(a)	ONSET	7.73	'awak~'wakan
(b)	*pretonic-VV	1.57	qe'epah~qu'pahan
(c)	Lic(u,pretonic)	3.05	'barah~bu'rahan
(d)	AGREE(v'hv)	7.23	'leiη~li'ijan
(e)	*HIATUS	2.58	'biki~bu'kiyan
Anticorrespondence			
(f)	$\text{ANTICORR}(\text{ruy}, \text{rig})$	6.89	'baruy~bu'rigan
(g)	$\text{ANTICORR}(\text{c}, \text{t})$	2.70	'birac~bu'ratan
(h)	$\text{ANTICORR}(\text{c}, \text{d})$	1.39	'haηuc~hu'ηedan
(i)	$\text{ANTICORR}(\text{k}, \text{p})$	1.82	'kayak~ku'yapan
(j)	$\text{ANTICORR}(\text{k}, \text{b})$	0.03	'eluk~'leban
(k)	$\text{ANTICORR}(\text{N}, \text{m})$	0.01	'talaη~tu'lam-an
(l)	$\text{ANTICORR}(\text{CeCuC}, \text{CuCeC})$	5.62	'pemux~pu'mexan
(m)	$\text{ANTICORR}(\text{CaCuC}, \text{CuCeC})$	2.59	'baluη~bu'leηan
(n)	$\text{ANTICORR}(\text{e}, \text{ay})$	4.95	'rage~ru'ηayan
(o)	$\text{ANTICORR}(\text{n}, \text{l})$	0.75	'tabun~tu'bulan
(p)	$\text{ANTICORR}(\text{o}, \text{aw})$	2.54	'sino~su'nawan
(q)	$\text{ANTICORR}(\text{o}, \text{ag})$	6.15	'baro~bu'ragan
(r)	$\text{ANTICORR}(\text{u}, \text{ey})$	0.35	'deηu~du'ηeyan

Table 4.1: Constraint weights in stem-base model

Note that the current model does *not* take into account the *amount* of data available for alternations that are exceptionless (i.e. in cases where there is no conflicting data). This means that markedness constraints which enforces exceptionless but very infrequent alternations will still end up with a very high weight. For example, $\text{ANTICORR}(\text{ruy}, \text{rig})$ enforces the alternation of stem-final $[\text{ruy}]$ with $[\text{rig}]$ in the suffixed form. This constraint is only applicable to three stems, but ends up getting a high weight of 6.89 because it is never

violated in winning candidates.

Intuitively, it seems as if an alternation that only applies to three forms should be less productive than something like the [c]-[t] alternation, which applies to many more forms. In other words, the inability to account for the *scope* of an alternation is a limitation of the current model. Although a full treatment of this issue is beyond the scope of the current work, a potential solution is to implement a bias term (Wilson, 2006), which penalises rules associated with less evidence.

4.6.4 Model performance

Overall, the model correctly assigns a probability of greater than 0.5 to around 71% of actually observed suffixed forms (243/340). Moreover, model performance is similar to the strictly ranked stem-base model described in Section 4.5, in that it predicts low probability for observed forms which either show an irregular alternation, or a dispreferred pattern of alternation.

For 31 verbs, the model assigned a very low probability ($p < 0.1$) to the observed output. Out of these, 25 were forms involving the irregular alternations discussed in Section 1.6. Table 4.2 shows representative examples of this variety; the third column shows the candidate that the model assigned the highest probability to. For example, the verb in (a) shows irregular stem-final [n]-insertion, while the verb in (b) shows both irregular vowel alternation and stem-final vowel deletion.

	INPUT	OBSERVED	p	PREDICTED	p
(a)	'apa	pa'an-an	0.029	'pa-an	0.83
(b)	'bege	'biq-an	0.057	bu'gay-an	0.73
(c)	'huruc	hu'rid-an	0.0001	hu'rut-an	0.75

Table 4.2: Example forms where observed outputs were assigned $p < 0.1$

The model assigned a probability of 0.1-0.3 ($0.1 < p < 0.3$) for 37 observed suffixed forms. As illustrated by the representative examples in Table 4.3, these cases primarily involved stem-

suffix pairs which showed a regular but dispreferred pattern of alternation. For example, in (a), the observed output [du'run-an] does not show the preferred [n]-[l] alternation. In contrast, in (b), the output [qu'rap-an] is assigned a low probability because it shows a dispreferred alternation ([k]-[p]). In (c), [ʔgul-an] gets a low probability because it doesn't show an [u]-[e] alternation (which is predicted by vowel matching pattern).

	INPUT	OBSERVED	<i>p</i>	PREDICTED	<i>p</i>
(a)	'durun	du'run-an	0.33	du'rul-an	0.67
(b)	'qerak	qu'rap-an	0.21	qu'rak-an	0.55
(c)	'egun	'gul-an	0.10	'gel-an	0.56

Table 4.3: Example forms where observed outputs, $0.1 < p < 0.3$

4.6.5 Frequency matching

As discussed in Section 4.6.1, the model should be able to learn the statistical distributions of the lexicon, and ‘match’ the frequency of alternations observed in the learning data. For example, stem-final [c] alternates with [t] 78% of the time in the Seediq lexicon. Corresponding to this, the model should predict that novel [c]-final stem forms will have around a 78% probability of showing the [c]-[t] alternation.

To confirm whether this is the case, I tested the model with novel input forms, using simplified candidate sets that isolated the effect of individual alternations. For example, to test whether the model matched rates of alternations involving final [c], I used the novel stem [ʔhunic]. As demonstrated in tableau (51), three input candidates were given to the model: [hu'nic-an] (non-alternating), [hi'nit-an] ([c]-[t] alternation), and [hu'nid-an] ([c]-[d] alternation).

(51) *Frequency matching for [c]-[t]-[d] alternation*

	$ID_{-IO[vc]}$	$ID_{-IO[del.rel]}$	$ANTICORR_{(c,t)}$	$ANTICORR_{(c,d)}$			
<i>'hunic/</i>							
<i>Weights</i>	0	0	2.7	1.39	\mathcal{H}	$e\mathcal{H}$	p
a. hu'nican			1	1	4.09	0.017	0.050
b. hu'nitan		1		1	1.39	0.248	0.748
c. hu'nidan	1	1	1		2.70	0.067	0.202

Using this method, the model's predicted probabilities for each alternation were obtained. Table 4.4 below compares predicted and observed rates of *final consonant* alternations; the model's predicted probability for each alternation is shown on the rightmost column. Overall, we see a close match between observed and predicted frequencies (correlation for the ten values given: $r=0.986$).

	Cons.	Pattern	Example	Rate	<i>prob.</i>
(a)	c	[c]-[t]	patic~putitan	78%	78.61%
		[c]-[d]	patic~putidan	17%	21.27%
		[c]-[c]	patic ~ putican	4%	0.02%
(b)	n	[n]-[l]	patin ~ putilan	75%	67.2%
		[n]-[n]	patin ~ putinan	25%	32.8%
(c)	k	[k]-[p]	patik ~ putipan	24%	26.4%
		[k]-[b]	patik ~ putiban	4%	4.6%
		[k]-[k]	patik ~ putikan	72%	69.0%
(d)	ŋ	[ŋ]-[m]	patiŋ ~ putiman	6%	6.1%
		[ŋ]-[ŋ]	patiŋ ~ putiŋan	94%	93.9%

Table 4.4: Predicted vs. observed rates of final alternation in lexicon vs. model

In Table 4.4, predicted and observed rates of VOWEL MATCHING are compared. Overall, as with the final consonant alternations, the model's predicted probabilities closely match rates of alternation (correlation for the eight values given: $r=0.96$). There is one slight mismatch: in the data, vowel matching is exceptionless when the stem stressed vowel is

[o], but the model predicts a lower probability of around 98%. In contrast, the model ‘exaggerates’ the vowel matching pattern for when the stem stressed vowel is [e], and assigns the vowel matching alternation a probability of around 85% (whereas the observed rate of alternation is 76%). This mismatch likely happened because the same ANTICORR constraint was used to enforce the [u]-[e] and [u]-[o] alternations.

	stem V	Pattern	Example	Observed Rate.	<i>prob.</i>
(a)	u	[u]-[u]	'putus~pu'tus-an	97%	99.92%
		[u]-[e]	'putus~pu'tes-an	3%	0.01%
(b)	e	[u]-[u]	'petus~pu'tus-an	24%	15.15%
		[u]-[e]	'petus~pu'tes-an	76%	84.85%
(c)	o	[u]-[u]	'potus~pu'tus-an	0%	1.66%
		[u]-[o]	'potus~pu'tos-an	100%	98.34%
(d)	a	[u]-[u]	'patus~pu'tus-an	76%	63.75%
		[u]-[e]	'patus~pu'tes-an	24.00%	36.25%

Table 4.5: Predicted vs. observed rates of vowel matching in lexicon vs. model
(stem V refers to the stressed vowel of the stem)

Overall, these comparisons suggest that the model successfully matched the statistical distributions of the lexicon it was trained on. Although not shown here, the model was able to match distributions of all other alternations (such as final vowel alternations).

4.7 Testing the productivity of Seediq alternations: current and future work

As discussed above in Section 4.6, MaxEnt models assign probabilities to all candidate output forms. In other words, the grammar of Seediq verbal alternations I have laid out so far makes empirical, testable predictions about how speakers will respond to novel forms. More specifically, the model predicts that, when given a stem form, the Seediq speaker should be able to actively ‘undo’ certain neutralizations, and apply frequent alternations even in the absence of a clear markedness motivation.

For example, (52) shows the model’s predicted probabilities of four different output forms for the novel input [ˈbekuŋ]. Given this input form, there are two potential alternations that may occur: (i) alternation of the final vowel [u] with [e], which is preferred under the principle of VOWEL MATCHING, and (ii) alternation of final [ŋ] with [m], which is statistically dispreferred in the lexicon. The model predicts gradient probabilities of output forms, depending on whether or not they show a specific alternation. Output (a), which shows the preferred vowel alternation but not the dispreferred [ŋ]-[m] alternation, is assigned the highest probability. On the other end of the spectrum, output (d), which has the [ŋ]-[m] alternation but fails to follow VOWEL MATCHING, has a near-zero probability.

(52) Example outputs for novel form:

<i>Input</i>		<i>Outputs</i>	<i>p</i>
ˈbekuŋ	(a)	buˈkeŋan	0.75
	(b)	buˈkuŋan	0.20
	(c)	buˈkeman	0.04
	(d)	buˈkuman	0.01

To test whether speakers will in fact behave as the model predicts, methods such as wug testing can be employed (Berko, 1958). In a pilot experiment with six subjects, I tested whether speakers could productively apply a subset of alternations to novel suffixed forms. The methodology used was a variant of the wug-test method; stimuli were not nonce-words, but rather ‘gapped forms’, or existing words in the Seediq lexicon which are never found in their suffixed forms (Pertsova and Kuznetsova, 2015). This method, though less common than wug-testing, was used out of respect for my Seediq consultants, who for cultural reasons did not want to work with nonce words. Early results from this pilot study suggest that speakers are able to productively extend some alternations, including the VOWEL MATCHING pattern. More extensive follow-up studies will be used to probe whether speakers do in fact productively apply the observed verb alternations, and whether they frequency-match as predicted under the current model.

CHAPTER 5

Conclusion

Based on a survey of 340 verb paradigms from the Seediq lexicon, the current study finds that Seediq paradigms show a striking asymmetry, whereby the stem forms (and more generally all non-suffixed slots of the paradigm) can be used to predict the suffixed forms with much higher probability than the other way around. I confirmed this asymmetry in a rule-based model of surface-base learning, and further show that the stem base performs substantially better on the real lexicon than on a simulated lexicon where rates of alternation were determined by baseline frequencies of sounds in Seediq.

This finding suggests that predictability from the stem form cannot be completely explained by baseline phonotactic preferences. Instead, I argue that there has been a gradual restructuring of Seediq verb paradigms based on the non-suffixed forms. This type of asymmetric restructuring is unexpected from a cobbled UR approach, where reanalysis from all allomorphs of a paradigm are possible. On the other hand, such an symmetry is the natural outcome under the single surface-based hypothesis.

Based on these empirical results, I offer an explicit constraint-based grammar of Seediq morphophonology where the input is the surface non-suffixed allomorph. In this grammar, ANTICORRESPONDENCE constraints are used to enforce alternations in the absence of a clear phonotactic motivation. I also adopt Zuraw's (2000, 2010) DUAL LISTING model, where stochastic constraint weighting models speakers' gradient preferences for alternation or non-alternation, while existing verbs are protected from variation through lexical listing. This grammar performs only slightly less well than a comparison grammar with cobbled URs,

despite the far greater informational resources to which the latter has access. Moreover, it codes the preferences that have presumably guided restructuring in Seediq.

Notably, the model makes clear empirical predictions about how Seediq speakers will respond to novel stem-suffix pairs. Future work should test these predictions, and confirm whether speakers can productively undo neutralizations of the stem form.

APPENDICES

A Summary of rule-based model (stem base)

*=applies to only one form.

Name	Rule	Example	p (H/S)	\hat{p}
stress assign.	$\sigma \rightarrow \left[\begin{array}{c} +\text{stress} \end{array} \right] / _(\sigma)]_w$	'bunuh→bu'nuhan	1 (340/340)	0.996
Pret. VR	$\left[\begin{array}{c} +\text{syl} \\ -\text{stress} \end{array} \right] \rightarrow [u] / \#C_$	'patuk→pu'tukan	1.0 (265/265)	0.99
Pret. V-del.	$\left[\begin{array}{c} +\text{syl} \\ -\text{stress} \end{array} \right] \rightarrow \emptyset / \#_$	'awak→'wakan	1.0 (35/35)	0.96
Pret. assim.	$\left[\begin{array}{c} +\text{syl} \\ -\text{stress} \end{array} \right] \rightarrow V_i / \#C_ \left[\begin{array}{c} -\text{LAB} \\ -\text{COR} \\ -\text{DORS} \end{array} \right] V_i$	'leiq→li'iq-an	1.0 (35/35)	0.96
Glide insertion	$\emptyset \rightarrow \left[\begin{array}{c} -\text{syl} \\ -\text{cons} \\ \alpha\text{high} \\ \beta\text{low} \end{array} \right] / \left[\begin{array}{c} +\text{syl} \\ -\text{low} \end{array} \right] _ \left[\begin{array}{c} +\text{syl} \end{array} \right]$	'leiq→li'iq-an	0.91 (32/35)	0.83
k-to-p	$[k] \rightarrow \left[\begin{array}{c} +\text{LAB} \\ -\text{DORS} \end{array} \right] / _]_{stem} V$	x→y	0.23 (6/26)	0.12
k-to-b	$[k] \rightarrow \left[\begin{array}{c} +\text{LAB} \\ -\text{DORS} \\ +\text{voice} \end{array} \right] / _]_{stem} V$	x→y	0.04 (1/26)	0.01
c-to-t	$[c] \rightarrow \left[\begin{array}{c} -\text{del rel} \\ -\text{strident} \end{array} \right] / _]_{stem} V$	x→y	0.74 (17/23)	0.62
c-to-d	$[c] \rightarrow \left[\begin{array}{c} -\text{del rel} \\ -\text{strident} \\ +\text{voice} \end{array} \right] / _]_{stem} V$	x→y	0.19 (4/23)	0.07
n-to-l	$[n] \rightarrow \left[\begin{array}{c} +\text{cont} \\ +\text{approx} \\ -\text{nasal} \\ +\text{lat} \end{array} \right] / _]_{stem} V$	x→y	0.62 (18/29)	0.49
u-to-ug	$\emptyset \rightarrow [g] / u_]_{stem} V$	x→y	0.28 (9/32)	0.18

u-to-ey	$u \rightarrow ey / _]_{stem} V$	$x \rightarrow y$	0.28 (9/32)	0.18
e-to-ay	$[e] \rightarrow [ay] / _]_{stem} V$	$x \rightarrow y$	0.64 (7/11)	0.42
an-to- \emptyset	$\left[\begin{array}{c} -front \\ +syl \end{array} \right] \left[\begin{array}{c} +COR \\ +nasal \end{array} \right] \rightarrow \emptyset / _]_{stem} V$	$x \rightarrow y$	0.24 (4/17)	0.10
fin vow del.	$\left[\begin{array}{c} +syl \\ -high \end{array} \right] \rightarrow \emptyset / _]_{stem} V$	$x \rightarrow y$	0.12 (6/52)	0.06
o-to-ag	$[o] \rightarrow [ag] / _]_{stem} V$	$x \rightarrow y$	0.83 (10/12)	0.65
ŋ-to-m	$[ŋ] \rightarrow \left[\begin{array}{c} +LAB \\ -DORS \end{array} \right] / _]_{stem} V$	'geruŋ→gu'reman	0.06 (2/34)	0.02
ruy-to-rig	$[ruy] \rightarrow [rig] / _]_{stem} V$	'baruŋ→bu'rigan	1.0 (3/3)	0.6
u-to-i	$[u] \rightarrow [i] / \#hur_c\#$	$x \rightarrow y$	1 (1/1)	0.14
Vow. match	$[u] \rightarrow \left[\begin{array}{c} -low \\ -high \\ \alpha front \\ \beta back \\ \gamma round \end{array} \right] / \left[\begin{array}{c} -low \\ -high \\ \alpha front \\ \beta back \\ \gamma round \end{array} \right] C_0_C_{stem} V$	'belux→bu'lexan	0.83 (91/109)	0.79
Irreg. [u]-to-[e]	$[u] \rightarrow [e] / \{a,i\}C_0_C_{stem} V$	'baluŋ→bu'lejan	0.12 (4/32)	0.05
Irreg. [i]-to-[e]	$[i] \rightarrow [e] / \{a,e\}C_0_C_{stem} V$	'adis→'desan	0.08 (2/26)	0.01
Irreg. [a]-to-[i]	$[a] \rightarrow [u] / \left[\begin{array}{c} +back \\ -syl \end{array} \right] _ C_{stem} V$	'nuqah→nu'qihan	0.18 (2/11)	0.04
o-to-aw*	$[o] \rightarrow [aw] / \#sin_]_{stem} V$	'sino→su'nawan	1 (1/1)	0.14
i-to-u*	$[i] \rightarrow [u] / \#raq_c]_{stem} V$	'raqic→ru'qutan	1 (1/1)	0.14
c-to-k*	$[c] \rightarrow [k] / \#mura_]_{stem} V$	'murac→mu'rakan	1 (1/1)	0.14
c-to-p*	$[c] \rightarrow [p] / \#qera_]_{stem} V$	'qerac→qu'rapan	1 (1/1)	0.14
uhu-to-u*	$[uhu] \rightarrow [u] / \#q_dey]_{stem} V$	'murac→mu'rapan	1 (1/1)	0.14
i-to-a*	$[i] \rightarrow [a] / \#pat_s]_{stem} V$	'patis→pu'tasan	1 (1/1)	0.14

B Summary of rule-based model (suffix base)

*=applies to only one form.

Name	Rule	Example	p (H/S)	\hat{p}
stress assign.	$\sigma \rightarrow [+\text{stress}] / _ (\sigma)_w$	bu'nuhan→'bunuh	1 (340/340)	0.996
Vow. match	$[u] \rightarrow \left[\begin{array}{c} \alpha\text{low} \\ \beta\text{high} \\ \gamma\text{front} \\ \delta\text{back} \\ \epsilon\text{round} \end{array} \right] / _ (C_0) \left[\begin{array}{c} \alpha\text{low} \\ \beta\text{high} \\ \gamma\text{front} \\ \delta\text{back} \\ \epsilon\text{round} \end{array} \right] C]_w$	bu'lexan→'belux	0.47 (159/340)	0.43
vow-insertion	$[u] \rightarrow [i] / _ (C_0)VC]_w$	tu'nunan→'tinun	0.68 (32/47)	0.58
u-to-a	$[u] \rightarrow [a] / _ (C_0)VC]_w$	gu'tukan→'gatak	0.21 (40/192)	0.17
u-to-e	$[u] \rightarrow [e] / _ (C_0)VC]_w$	bu'rasan→'beras	0.23 (42/186)	0.18
u-to-i	$[u] \rightarrow [i] / _ (C_0)VC]_w$	tu'nunan→'tinun	0.14 (42/304)	0.11
Vow. lengthening	$V_i \rightarrow V_i?V_i / _ (C_0)VC]_w$	gu'lekan→ge'ʔeguy	0.03 (7/264)	0.01
{e,u}g-to-u	$\{e,u\}g \rightarrow [u] / _]_w$	hu'yegan→heyu	0.9 (9/10)	0.71
l-to-n	$[l] \rightarrow \left[\begin{array}{c} -\text{lat} \\ -\text{cont} \\ +\text{nasal} \\ -\text{approx} \end{array} \right] / _]_w$	du'dulan→dudun	0.95 (18/19)	0.84
T-to-c	$\left[\begin{array}{c} +\text{COR} \\ -\text{son} \end{array} \right] \rightarrow \left[\begin{array}{c} +\text{strid} \\ +\text{del rel} \\ -\text{voice} \end{array} \right] / _]_w$	du'matan→damac	0.95(20/21)	0.85
m-to-ŋ	$m \rightarrow \left[\begin{array}{c} -\text{LAB} \\ +\text{DORS} \end{array} \right] / _]_w$	tu'laman→'talaŋ	0.5 (2/4)	0.15
P-to-k	$\left[\begin{array}{c} +\text{LAB} \\ -\text{son} \end{array} \right] \rightarrow \left[\begin{array}{c} -\text{LAB} \\ +\text{DORS} \\ -\text{voice} \end{array} \right] / _]_w$	ku'yapan→'kayak	0.88 (7/8)	0.65
aw/ag-to-o	$\{\text{aw,ag}\} \rightarrow [o] / _]_w$	bu'ragan→'baro	1 (11/11)	0.87
ay-to-e	$[\text{ay}] \rightarrow [e] / _]_w$	lu'hayan→'lahe	0.88 (7/8)	0.65
ey-to-u	$[\text{ey}] \rightarrow [u] / _]_w$	pu'heyan→'pahu	0.9 (9/10)	0.71

Final glide del	$\begin{bmatrix} -\text{syl} \\ -\text{cons} \\ +\text{son} \end{bmatrix} \rightarrow \emptyset / _]_w$	suqu'riyan→su'quri	1 (32/32)	0.95
Post. VR	$\begin{bmatrix} -\text{high} \\ -\text{low} \\ +\text{syl} \end{bmatrix} \rightarrow \begin{bmatrix} +\text{high} \\ +\text{back} \\ +\text{round} \end{bmatrix} / _]_w$	bu'leḡan→'baluḡ	0.92 (94/102)	0.88
ig-to-uy	rig] → [uy] / _] _w	bu'riḡan→'baruḡ	1 (3/3)	0.61
p-to-c*	[p] → [c] / #qera_] _w	qu'rapan→'qerac	1 (1/1)	0.14
k-to-c*	[k] → [c] / #mura_] _w	mu'rakan→'murac	1 (1/1)	0.14
m-to-n*	[m] → [n] / gi_] _w	'ḡiman→'iḡin	1 (1/1)	0.14
e-epenthesis	$\emptyset \rightarrow [e] / \#C \begin{bmatrix} -\text{front} \\ -\text{back} \end{bmatrix} C_]_w$	'ker-an→'kere	0.12 (3/25)	0.04
a-epenthesis*	$\emptyset \rightarrow [a] / \#\text{kes}_]_w$	'kes-an→'kesa	1 (1/1)	0.14
o-epenthesis*	$\emptyset \rightarrow [o] / \#\text{had}_]_w$	'had-an→'hado	1 (1/1)	0.14
u-to-o	[u] → [-high] / #_sa] _w	'sa-an→'osa	1 (2/2)	0.46
e-to-ya	[e] → [ya] / e(C ₀)_C] _w	ce'exan→'ceyax	0.60 (3/5)	0.27
an-epenthesis	$\emptyset \rightarrow [\text{an}] / \begin{bmatrix} +\text{syl} \\ -\text{LAB} \end{bmatrix} \begin{bmatrix} +\text{low} \\ -\text{back} \end{bmatrix} C_]_w$	'cam-an→'caman	0.14 (3/22)	0.04
a-to-i*	[a] → [i] / t_s] _w	pu'tasan→'patis	1 (1/1)	0.14
e-to-a*	[e] → [a] / b_s] _w	tu'besan→'tebas	1 (1/1)	0.14
u-to-i*	[u] → [i] / ruq_t] _w	ru'qutan→'raqic	1 (1/1)	0.14
e-to-i*	[w] → [i] / d_s] _w	'desan→'adis	1 (1/1)	0.14
i-to-u*	[i] → [u] / r_d] _w	hu'ridan→'huruc	1 (1/1)	0.14

C Strict ranking OT model-constraints and constraint ranking

★= Not necessary (but included to show
Faithfulness violations of a winning candidate)

Stratum #1

AGREE(v'hv)

*INITGLIDE

*HIATUS

ONSET

*pretonic.VV

ANTICORR(O.AG)

ANTICORR(RUY.RIG)

ANTICORR(C.T)

ANTICORR(E.AY)

ANTICORR(N.L)

ID[voice]

Stratum #2

matchVowel

MAXV

DEPC ★

IDENT-OO-low.fin

Stratum #3

LIC(u,pretonic)"

ID[high]/strs

DEP-g

Stratum #4

DEP-glide ★

Constraints which were tested, but **not necessary** in the ranking:

ANTICORR(A.AN)

ANTICORR(O.AW)

ANTICORR(K.P)

ANTICORR(N.M)

ANTICORR(AC.IC)

ANTICORR(IC.EC)

ANTICORR(U.UG)

ANTICORR(U.EY)

ANTICORR(CACuC.CuCEC)

ANTICORR(C.D)

D MaxEnt OT model-constraints and constraint weights

General markedness constraints

	<i>Constraint</i>	<i>weight</i>	<i>example</i>
(a)	ONSET	7.73	'awak~'wakan
(b)	*pretonic.VV	1.57	qe'epah~qu'pahan
(c)	Lic(u,pretonic)	3.05	'barah~bu'rahan
(d)	AGREE(v'hv)	7.23	'leij~li'ijan
(e)	*Hiatus	2.58	'biki~bu'kiyan

Anticorrespondence constraints

	<i>Constraint</i>	<i>weight</i>	<i>example</i>
(a)	ANTICORR(ruy,rig)	6.89	'baruy~bu'rigan
(b)	ANTICORR(c, t)	2.70	'birac~bu'ratan
(c)	ANTICORR(c, d)	1.39	'hajuc~hu'jedan
(d)	ANTICORR(k, p)	1.82	'kayak~ku'yapan
(e)	ANTICORR(k, b)	0.03	'eluk~'leban
(f)	ANTICORR(N, m)	0.01	'talaq~tu'lam-an
(g)	ANTICORR(CeCuC, CuCeC)	5.62	'pemux~pu'mexan
(h)	ANTICORR(CaCuC, CuCeC)	2.59	'baluq~bu'lejan
(i)	ANTICORR(e, ay)	4.95	'raje~ru'jayan
(j)	ANTICORR(n, l)	0.75	'tabun~tu'bulan
(k)	ANTICORR(o, aw)	2.54	'sino~su'nawan
(l)	ANTICORR(o,ag)	6.15	'baro~bu'ragan
(m)	ANTICORR(u, ey)	0.35	'deju~du'jeyan
(n)	ANTICORR(ae, \emptyset)	0.00	gene~qenan
(o)	ANTICORR(iC, eC)	0.00	adis~desan
(p)	ANTICORR(u,ug)	0.00	lihu~luhugan
(q)	ANTICORR(aC,eC)	0.00	tebas~tubesan

(r)	ANTICORR(aC, iC)	0.00	nuqah~nuqihan
(s)	ANTICORR(a,an)	0.00	qeya~quyanan

Faithfulness constraints

	<i>Constraint</i>	<i>weight</i>	<i>example</i>
(a)	IDENT-OO[DORS]	2.76	
(b)	IDENT-OO[round]/strs	0.64	
(c)	IDENT-OO[cont]	0.04	
(d)	IDENT-OO[round]	0.58	
(e)	MAX-V	1.76	
(f)	IDENT-OO[front]/strs	1.17	
(g)	IDENT-OO[high]/strs	1.54	
(h)	IDENT-OO[low]/strs	0.93	
(i)	IDENT-OO[LAB]	0.01	
(j)	DEP-g	0.83	
(k)	DEP-C	3.03	
(l)	IDENT-OO[high]	0.00	
(m)	IDENT-OO[front]	0.00	
(n)	IDENT-OO[low]	0.00	
(o)	IDENT-OO[voice]	0.00	
(p)	IDENT-OO[del. rel]	0.00	
(q)	IDENT-OO[lateral]	0.00	

Bibliography

- Albright, Adam (2010). Base-driven leveling in yiddish verb paradigms. *Natural Language & Linguistic Theory* 28(3). 475–537.
- Albright, Adam and Bruce Hayes (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.
- Albright, Adam C (2002). *The identification of bases in morphological paradigms*. PhD dissertation, University of California, Los Angeles.
- Bakovic, Eric (2000). *Harmony, dominance and control*. PhD dissertation, Rutgers, The State University of New Jersey.
- Becker, Michael (2009). *Phonological trends in the lexicon: the role of constraints*. University of Massachusetts, Amherst. PhD dissertation, PhD dissertation.
- Beckman, Jill N (1998). *Positional faithfulness*. PhD dissertation, University of Massachusetts Amherst.
- Berko, Jean (1958). The child's learning of english morphology. *Word* 14(2-3). 150–177.
- Blevins, Juliette (2008). Consonant epenthesis: natural and unnatural histories. *Language universals and language change* 79–109.
- Bybee, Joan (2003). *Phonology and language use*, volume 94. Cambridge University Press.
- Crosswhite, Katherine (2004). Vowel reduction. *Phonetically based phonology* 191–231.
- Ernestus, Mirjam and R Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language* 5–38.
- Frisch, Stefan A, Janet B Pierrehumbert, and Michael B Broe (2004). Similarity avoidance and the ocp. *Natural Language & Linguistic Theory* 22(1). 179–228.

- Fylstra, Daniel, Leon Lasdon, John Watson, and Allan Waren (1998). Design and use of the microsoft excel solver. *Interfaces* 28(5). 29–55.
- Goldwater, Sharon and Mark Johnson. Learning of constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111120, (2003).
- Greenhill, Simon J, Robert Blust, and Russell D Gray (2008). The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271–283.
- Hale, Kenneth (1973). Deep-surface canonical disparities in relation to analysis and change: An australian example. *Current trends in linguistics* 11(19731). 401–458.
- Hayes, Bruce. Phonological Restructuring in Yidin and its Theoretical Consequences. In Hermans, Ben and Marc van Oostendorp, editors, *The derivational residue in phonological optimality theory*, 175–205. John Benjamins Publishing Company, (1999). doi: 10.1075/la.28.09hay.
- Hayes, Bruce and James White (2015). Saltation and the p-map. *Phonology* 32(2). 267–302.
- Hayes, Bruce, Péter Siptár, Kie Zuraw, and Zsuzsa Londe (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*. 822–863.
- Holmer, Arthur (1996). *A parametric grammar of Seediq*. PhD dissertation, Lund University.
- Ito, Junko and 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.
- Kang, Chen (1992). *Taiwan gaoshanzu yuyan [Languages of the Gaoshan people of Taiwan]*. Central Minorities institute. Beijing.

- Kang, Yoonjung (2006). Neutralizations and variations in Korean verbal paradigms. *Harvard Studies in Korean Linguistics* 11. 183–196.
- Kenstowicz, Michael and Larry M Kisseberth. Topics in phonological theory, (1977).
- Kiparsky, Paul (1978). Analogical change as a problem for linguistic theory. *Studies in the Linguistic Sciences Urbana, Ill* 8(2). 77–96.
- Li, Paul Jen-kui (1981). Reconstruction of proto-atayalic phonology. *Bulletin of the Institute of History and Philology* 52.
- Lubowicz, Anna (2002). Derived environment effects in optimality theory. *Lingua* 112(4). 243–280.
- Mei-jin, Huang, Yu-yang Liu, and Xin-sheng Wu (2014). Taiwan aboriginal language e-dictionary. *Taiwan Journal of Indigenous Studies* 7(2). 73–118.
- Mikheev, Andrei (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23(3). 405–423.
- Pertsova, Katya and Julia Kuznetsova. Experimental evidence for lexical conservatism in Russian: Defective verbs revisited. In *Proceedings of FASL*, volume 24, (2015).
- Prince, Alan and Paul Smolensky (1993). Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in phonology* 3.
- Smolensky, Paul. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, (1986).
- Steriade, Donca. Locality conditions and feature geometry, (1987).
- Steriade, Donca (2001). The phonology of perceptibility effects: the p-map and its consequences for constraint organization. *Ms., UCLA* .

- 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*. PhD dissertation, University of California, Los Angeles.
- Zuraw, Kie (2010). A model of lexical variation and the grammar with application to tagalog nasal substitution. *NLLT* 28(2). 417–472.
- Zuraw, Kie. *map constraints. Master’s thesis, University of California Los Angeles, (2013).