CHAPTER 4

Gradient opacity in Uyghur backness harmony

4.1 Abstract

Opacity has long been an important issue for phonological theory, particularly in evaluating serial theories, such as SPE, against parallel theories, such as OT. This paper will examine a case of opacity in Uyghur (Turkic: China). In addition to exhibiting backness and rounding harmony, Uyghur has a vowel reduction process that neutralizes the front and back vowels /æ/ and /ɑ/ to the harmonically neutral vowel [i], potentially introducing opacity into the harmony system. Based on a combination of elicitation and a large-scale corpus study, I show that opaque harmony (harmonizing with underlying forms even when it produces surface disharmony) is the standard pattern in Uyghur. However, the same stems may also appear in tokens with surface-true harmony. An analysis of corpus frequencies shows that words whose unraised forms occur more frequently are more likely to display opaque harmony. I model this data using a variant of paradigm uniformity constraints, that, rather than requiring properties of the stem to be invariant across allomorphs, requires that all allomorphs of a stem take suffixes that harmonize with their base form, even when this violates surface harmony. The strength of a base is contingent on how frequently it is observed. I finish by discussing the implications of these results for theories of the representation and learning of opaque patterns.

4.2 Introduction

This paper examines a phonological pattern in Uyghur (Turkic: China) where a vowel raising process converts harmonic vowels into transparent vowels, thus rendering the harmony pattern opaque.

This paper has two primary goals: the first is to present new empirical data on this phenomenon based on a combination of elicitation and a large-scale corpus study. The second is to argue on the basis of this data and other properties of Uyghur that this opaque pattern is best modeled as a type of *paradigm uniformity* (Steriade, 2000), rather than as a type of procedural mapping from underlying to surface form, and to discuss the implications of this claim for theories of opacity in general.

In addition to this phonological analysis, I also hope to make a case for a more holistic and comprehensive approach to linguistic data collection and analysis. Although the interaction between vowel raising and backness harmony in Uyghur has been studied in the past (e.g., Vaux, 2000; Halle et al., 2000; Hall & Ozburn, 2018), this work has relied on a small number of data points which, I contend, incorrectly represent the two patterns and their interaction, at least in standard dialects of Uyghur. This paper demonstrates that modern computational tools can provide access to large amounts of data in a form amenable to phonological analysis, and that modern phonological theory, coupled with thorough knowledge of the language under study, can be used to marshal the complexity inherent in such large data sets.

4.3 Phonological opacity

Kiparsky (1971, 1973) defines phonological opacity as follows:

(45) Assume a phonological rule $\mathbb{R}$ of the form A $\rightarrow$ B / C_D. $\mathbb{R}$ is *opaque* if there are surface forms with either:

   a. *A* in the environment *C*_ *D* (underapplication opacity)

   b. *B* derived from *A* in environments other than *C*_ *D* (overapplication opacity)

In other words, opacity arises when either a conditioned alternation appears not to occur despite its conditions being met, or appears to occur when its conditions have not been met.

Kiparsky associated opacity of types (1a) and (1b) with *counterfeeding* and *counterbleeding* rule orders respectively. In the case of counterfeeding opacity, the structural conditions for rule $\mathbb{R}$

to apply are created by a different rule $\mathbb{P}$ that applies after $\mathbb{R}$: hence the necessary conditions are not met when $\mathbb{R}$ applies. Changing the rule ordering such that $\mathbb{P}$ applied before $\mathbb{R}$ would produce a *feeding* order where $\mathbb{R}$ applies transparently to the conditioning environment produced by $\mathbb{P}$.

In counterbleeding opacity, the conditions for $\mathbb{R}$ are met when it applies, but are subsequently altered by a different rule $\mathbb{Q}$ that applies after $\mathbb{R}$. Changing the rule ordering such that $\mathbb{Q}$ applies before $\mathbb{R}$ would produce a *bleeding* order where $\mathbb{R}$ transparently fails to apply because $\mathbb{Q}$ removes its conditioning environment.

Opacity has been a longstanding topic of interest in phonological theory since its definition by Kiparsky. In a large part, this interest has stemmed from debates on the merits of serial models such as *SPE*-style rules (e.g., Chomsky & Halle, 1968) vs. parallel models such as optimality theory (e.g., Prince & Smolensky, 1993/2004). Parallel models have difficulty correctly predicting cases of counterbleeding opacity, which generally produce faithfulness violations with no corresponding markedness repairs to motivate them. They also have difficulty with most types of counterfeeding opacity, which fail to repair a markedness violation whose repair is evident in other contexts. Additions such as sympathy (McCarthy, 1999) and candidate chain theory (McCarthy, 2007) have been incorporated into parallel models to handle opacity. The need for such bespoke mechanisms has been seen as a point in favor of serial models, which handle these cases of opacity without issue (e.g., Vaux, 2008).

Although counterfeeding and counterbleeding orderings are the best known configurations that result in opacity, Baković (2007, 2011) has shown that these orderings are neither sufficient nor necessary conditions for opacity as defined above. In particular, he identifies a number of cases of overapplication opacity that are not predicted by SPE-style rule ordering, and some which are *only* able to be described by parallel models. Thus the characterization of opacity as a unique challenge for parallel models is a simplification, though perhaps accurate in broad strokes.

In light of the lack of a unified account of opacity from either serial or parallel theories, Baković suggests that the field focus on Kiparsky's claim that opaque patterns are more difficult to learn than transparent ones. The basic motivation for this claim is that phonological processes that interact in an opaque fashion make generalization about those processes difficult: opaque forms

often appear on the surface to be counterexamples to generalizations that might robustly apply elsewhere. Kiparsky (1971) supports this claim by presenting a number of cases of historical change where an opaque process is reanalyzed as a transparent one. Subsequent research has presented evidence that opaque processes are learned as either phonemic contrasts or lexicalized patterns rather than productive rules (e.g., Hooper/Bybee, 1976; Mielke et al., 2003; Sanders, 2003), though evidence also exists that some opaque processes are applied productively in language games and other contexts (e.g., Donegan & Stampe, 1979; Al-Mozainy, 1981; Vaux, 2011).

4.4 Opacity in Uyghur backness harmony

Uyghur is a southeastern Turkic language spoken by approximately 10 million people in the Xinjiang Uyghur Autonomous Region in the People's republic of China, surrounding countries such as Kazakhstan and Kyrgyzstan, and various diasporic communities (Engesæth et al., 2009/2010; Nazarova & Niyaz, 2013). It has SOV word order with highly agglutinative morphology that is almost exclusively suffixing.

Opacity in Uyghur backness harmony arises from the interaction of two independent phonological processes: backness harmony and vowel reduction. I will introduce these processes separately before demonstrating how their interaction leads to opacity.

4.4.1 Segments involved in backness harmony

Like most Turkic languages, Uyghur has backness harmony. Borrowings from Persian, Russian, Arabic, and Chinese, many of which are quite old, have resulted in a high degree of stem-internal disharmony, though stems of Turkic origin tend to be harmonic. As a consequence, backness harmony is most evident as a morphophonological process, where, broadly speaking, segments in many suffixes must agree in backness with the stems they attach to (e.g., Lindblad, 1990; Hahn, 1991a, 1991b; Engesæth et al., 2009/2010; Abdulla et al., 2010).

The segments that directly participate in backness harmony are shown in Tables 4.1 and 4.2. The bolded vowels in Table 4.1 are serve as harmony triggers, while the non-bolded vowels are

	Front		Back	
	Unrounded	Round	Unrounded	Round
High	i	**y**		**u**
Mid	e	**ø**		**o**
Low	**æ**		**ɑ**	

Table 4.1: The Uyghur vowel system. Harmonizing vowels are in bold.

	Front	Back
Voiceless	k	q
Voiced	g	ʁ χ

Table 4.2: Harmonizing Uyghur consonants

transparent to harmony. Although these are the standard symbols used to transcribe these vowels, they are somewhat inaccurate representations of their phonetic realization. In particular, these vowels are generally produced less peripherally than their transcriptions would indicate. The vowel transcribed as /æ/ is acoustically intermediate between cardinal [æ] and [ɛ]. The vowels transcribed as /u/, /y/, /o/, and /ø/ are generally produced more closely to [ʊ], [ʏ], [ɔ], and [œ] respectively.

The transparent vowels, particularly /i/, display a much greater susceptibility to coarticulation than the harmonizing vowels. Hahn (1991b) describes no less than *fourteen* allophones of /i/, which range from [ɯ] to [ə] to [ɪ] to [i]. /e/ may surface as [ɛ], [e], or [i], with the latter a common allophone in initial syllables. Thus although these vowels are phonemically transcribed as non-low front vowels, the reader should keep in mind that their phonetic realization varies substantially.

Some researchers have proposed underlying back phonemic counterparts to /i/ and /e/, /ɯ/ and /ɤ/ (e.g., Hahn, 1991b; McCollum, 2019). Although allophonic variation is clear both subjectively and empirically, evidence of a phonemic contrast that is not motivated by parsimonious phonological analysis of the harmony system has been more difficult to obtain. Hahn, though in favor of a phonemic distinction, admits that these front and back counterparts "share the same set of allophones and are orthographically represented alike" (Hahn, 1991b, p. 34). He proposes a post-lexical fronting rule whereby underlying /ɯ/ and /ɤ/ are fronted to /i/ and /e/ in all contexts after

vowel harmony has applied. One challenge to this account is the total absence of homophones that differ in the backness of suffixes they take (i.e., underlying minimal pairs between /i/-/ɯ/ and /e/-/ɤ/).

Acoustic studies have found that acoustic measurements of vowel backness in stems containing only the vowels /i e/ are not a reliable predictor of whether such stems take front or back suffixes (Mayer & Major, 2018; McCollum, 2019; Mayer et al., in prep.). My anecdotal experience has been that Uyghur speakers often have intuitions about whether a token of /i/ is a "front *i*" or a "back *i*", but these intuitions do not always conform with the expected harmonizing behavior of stems. For example, one of my consultants reports that the final /i/ in /qojʧi/ 'shepherd' is front, but attaches back suffixes to this stem.

With these facts in mind, I assume the phonemic vowel inventory shown in Table 4.1, while acknowledging that it will be important to better understand the relationship between the phonetic realization of the transparent vowels and harmonic patterns with future phonetic study.

4.4.2 A description of Uyghur backness harmony

In the examples of harmony below, I use nouns with the locative suffix /-DA/ (surface forms: [-tɑ], [-dɑ], [-tæ], [-dæ]), the plural suffix /-lAr/ (surface forms: [-lɑr], [-lær]), or the dative suffix /-GA/ (surface forms: [-qɑ], [-ʁɑ], [-kæ], [-gæ]).[1] Voicing alternations in the initial segment are caused by voicing assimilation, and are orthogonal to harmony. All examples are attested tokens from the corpora described in Section 4.5.

The basic characterization of backness harmony is that suffixes must agree in backness with the final front (/æ ø y/) or back (/u o ɑ/) harmonizing stem vowel.

[1]The dative suffix may also surface as [-qæ] when attached to a stem with a front vowel that ends in a voiceless uvular, as in [χælq-qæ] 'people-DAT' (cf. [χælq-i-gæ] 'people-3.POS-DAT'). I consider this to be a case of place assimilation rather than harmony, though some previous work has ascribed greater significance to it (e.g., Pattillo, 2013).

(46) *Simple front harmonizing forms*

t**y**r-d**æ**/*-dɑ ‘type-LOC’

p**æ**n-l**æ**r/*-lɑr ‘science-PL’

munb**æ**r-**gæ**/*-ʁɑ ‘podium-DAT’

(47) *Simple back harmonizing forms*

p**u**l-**ʁɑ**/*-gæ ‘money-DAT’

t**o**p-**qɑ**/*-kæ ‘ball-DAT’

ætr**ɑ**p-t**ɑ**/*-tæ ‘surroundings-LOC’

The vowels /i e/ are *transparent* to harmony, meaning that they do not serve as harmony triggers for suffixes, but allow the harmonic value of preceding segments to “pass through” them.

(48) *Front stems with transparent vowels*

m**æ**stʃit-t**æ**/*-tɑ ‘mosque-LOC’

ymid-l**æ**r/*-lɑr ‘hope-PL’

m**ø**min-**gæ**/*-ʁɑ ‘believer-DAT’

(49) *Back stems with transparent vowels*

st**u**dent-l**ɑr**/*-lær ‘student-PL’

uniwersitet-t**ɑ**/*-tæ ‘university-LOC’

amil-**ʁɑ**/*-gæ ‘element-DAT’

If a stem contains no harmonizing vowels, the front dorsals /k g/ and back dorsals /q ʁ χ/ may serve as harmony triggers.

(50) *Stems with only front dorsals*

kishi-l**æ**r/*-lɑr ‘person-PL’

ne**g**iz-**gæ**/*-ʁɑ ‘basis-DAT’

(51) *Stems with only back dorsals*

qiz-l**ɑr**/*-lær ‘girl-PL’

ji**ʁ**in-d**ɑ**/*-dæ ‘meeting-LOC’

χiris-**ʁɑ**/*-gæ ‘grimace-DAT’

A small number of irregular stems containing front dorsals take back suffixes.

(52) *Front dorsal stems with back suffixes*

ingliz-l**ɑr**	‘English person-PL’
etnik-l**ɑr**	‘ethnic group-PL’
rentgen-**ʁɑ**	‘x.ray-DAT’
gips-**qɑ**	‘plaster-DAT’

Note that the opposite case, stems with back dorsals and no harmonizing vowels taking front suffixes, does not appear to occur.

In the absence of any harmonizing elements, stems are lexically listed for backness, with a strong statistical tendency towards back suffixes.

(53) *Neutral stems that take front suffixes*

biz-**gæ**/*-ʁɑ	‘us-DAT’
bilim-**gæ**/*-ʁɑ	‘knowledge-DAT’
welisipit-l**æ**r/*-lɑr	‘bicycle-PL’

(54) *Neutral stems that take back suffixes*

sir-l**ɑr**/-*lær	‘secret-PL’
din-**ʁɑ**/*-gæ	‘religion-DAT’
hejt-t**ɑ**/*-tæ	‘festival-LOC’
peʔil-l**ɑr**/*-lær	‘verb-PL’
tip-q**ɑ**/*-kæ	‘type-DAT’

In stems that contain a harmonizing vowel followed by a conflicting harmonizing dorsal, the vowel generally takes precedence. There is a small class of stems, however, that harmonize with the intervening uvulars.

(55) *Conflicting vowels and dorsals, vowel takes precedence*

m**æ**ntiq-q**æ**	‘logic-DAT’
æqil-**gæ**	‘intelligence-DAT’
r**ɑ**k-l**ɑ**r	‘shrimp-PL’
p**ɑ**kit-l**ɑ**r	‘fact-PL’

(56) *Conflicting front vowels and uvulars, uvular takes precedence*

tæsti**q-qa**	'approval-DAT'
tæʃwi**q**-l**ɑ**r	'publicity-PL'
tætqi**q**-l**ɑ**r	'research-PL'

Uyghur also has a process of *rounding harmony* that applies to high vowels. I do not discuss this here as it is not directly relevant to the opaque pattern under discussion.

4.4.3 Diachronic origins of Uyghur backness harmony

Uyghur, like Turkish, once had a corresponding back counterpart to /i/, /ɯ/ (the status of a back counterpart to /e/ is less clear). Thus no stems were harmonically neutral as is the case now, and front and back dorsals co-occurred with front and back vowels respectively. At some point in its history, Uyghur lost the distinction between /i/ and /ɯ/ (Hahn, 1991a), which complicated the harmony system. Lindblad (1990) shows that the most frequent stems that previously had /i/ continued to take front suffixes (e.g., [biz] 'we', [ilim] 'science', [itʃ-] 'drink'), the stems that previously had /ɯ/ continued to take back suffixes, and many less frequent stems that were underlyingly /i/ began to take the default back form of suffixes. Uyghur appears to be typologically unique in that the default harmony value is [+back], despite the transparent vowels being phonetically [–back]. In languages such as Mongolian and Finnish, which have similar transparent vowels, transparent stems generally behave as [–back] (Lindblad, 1990).

4.4.4 Vowel reduction

Uyghur has two independent, though similar, phonological vowel raising processes. The first, *vowel reduction*, raises the low vowels /ɑ æ/ to [i] in medial open syllables. The second, *umlauting* or *regressive vowel assimilation*, raises the same vowels to [e] in initial open syllables when the following vowel is [i] or [æ]. Because umlauting only targets initial syllables, it does not in general produce opacity in the harmony system (although it may render underlyingly harmonic stems harmonically neutral on the surface). Accordingly, this paper will focus on vowel reduction.

Vowel reduction raises the low vowels /ɑ æ/ to [i] in medial open syllables.

(57) */ɑ/ vowel reduction*

bɑl**ɑ**	'child'	bɑl**i**-ni	'child-ACC'
ɑp**ɑ**	'mom'	ɑp**i**-si	'mom-3.SG.POS'
ɑŋl**ɑ**-ʃ	'listen-GER'	ɑŋl**i**-di	'listen-3.SG.PAST'
qɑr**ɑ**-ʃ	'look-GER'	qɑr**i**-di	'look-3.SG.PAST'

(58) */æ/ vowel reduction*

ɑp**æ**t	'disaster'	ɑp**i**t-i	'disaster-3.SG.POS'
mew**æ**	'fruit'	mew**i**-si	'fruit-3.SG.POS'
søzl**æ**-ʃ	'talk-GER'	søzl**i**-di	'talk-3.SG.PAST'
kytʃ**æ**-ʃ	'strive-GER'	kytʃ**i**-di	'strive-3.SG.PAST'

This process generally applies only to derived environments. The stem /mɑqɑlæ/ 'article', for example, surfaces as [mɑqɑlæ] rather than *[mɑqilæ].

Note that the underlying form cannot in general be predicted from forms where vowel reduction could have applied, as many words have underlying /i/ in these positions.

(59) *Vowel reduction red herrings*

tɑksi	'taxi'	tɑksi-ni	'taxi-ACC'
æsli	'origin'	æsli-ni	'origin-ACC'
qeri-ʃ	'grow old-GER'	qeri-di	'grow old-3.SG.PAST'
ɑʁri-ʃ	'become ill-GER'	ɑʁri-di	'become ill-3.SG.PAST'

Even in derived environments, vowel reduction does not apply exceptionlessly. Certain roots resist raising categorically, though this appears to be more common when the potential raiser is /ɑ/.

(60) *Exceptions to vowel reduction with /ɑ/*

hɑwɑ	'weather'	hɑwɑ-si	'weather-3.POS'
dærjɑ	'river'	dærjɑ-si	'river-3.POSS'
mɑkɑn	'place'	mɑkɑn-i	'place-3.POSS'

(61) *Exceptions to vowel reduction with /æ/*

sæwæb	'reason'	sæwæb-i	'reason-3.POSS'
mɑqɑlæ	'academic article'	mɑqɑlæ-lær	'academic article-PL'
wæqæ	'accident'	wæqæ-gæ	'accident-DAT'

Raising has been claimed to occur less frequently in loanwords (Nazarova & Niyaz, 2013), and to be sensitive to vowel length distinctions and/or stress, with long or stressed vowels being less likely to raise, though there is scant phonetic evidence presented to support this (Hahn, 1991b).[2] Because of uncertainty about the nature of stress and/or vowel length in Uyghur, I will treat propensity to raise as a idiosyncratic property of particular words. Additional explanatory power may be gained by linking this to stress/vowel length once these phenomena are better understood.

4.4.5 Opaque interactions between backness harmony and vowel reduction

Vowel reduction has the potential to introduce opaque behavior as defined by Kiparsky into the vowel harmony system. Consider, for example, the root /aʁinæ/ 'friend'. When suffixes with appropriate forms are attached, the final vowel raises without exception:

(62) /ɑʁinæ-ni/ → [ɑʁini-ni]/*[ɑʁinæ-ni] 'friend-ACC'

What will happen for forms such as /aʁinæ-DA/, where the vowel in the suffix must harmonize with the final vowel in the stem?

If we assume a serial analysis where backness harmony precedes raising, we should expect to see the opaque form [aʁinidæ], shown in (63):

[2]The status of phonemic vowel length and stress in Uyghur is somewhat unclear. Hahn (1991b) claims that Uyghur has phonemic vowel length which is not represented orthographically, as well as lexical stress reflected by increases in pitch, duration, and intensity. A series of production and perception experiments in Yakup (2013) suggest that lexical stress does exist, but is reflected only by increases in duration: however, Uyghur speakers frequently disagreed as to which syllables were stressed in many words. Major and Mayer (2018) reproduce and expand on these results, suggesting that phrasal prosody is responsible for pitch contours that have previously been attributed to stress. All Uyghur speakers I have worked with have some intuition that certain vowels are longer than others.

(63) *Harmony precedes raising*

UR	/ɑʁinæ-DA/
Harmony	ɑʁin**æ**-d**æ**
Raising	ɑʁini-dæ
SR	[ɑʁini-dæ]

This is a case of underapplication, or counterbleeding, opacity. Backness harmony appears to fail to apply (or to apply incorrectly) to the surface form, resulting in a mismatch between the backness of the suffix and the final harmonizing vowel (cf. forms like /qojchi-DA/ → [q**o**jchid**ɑ**] 'shepherd-LOC').

If raising precedes backness harmony, we would instead expect the transparent form [ɑʁini-dɑ], shown in (64):

(64) *Raising precedes harmony*

UR	/ɑʁinæ-DA/
Raising	ɑʁini-DA
Harmony	**ɑ**ʁini-d**ɑ**
SR	[ɑʁini-dɑ]

Which do we see? Previous work on the interaction between vowel raising and harmony has claimed that there is an asymmetry between vowels (Vaux, 2000; Halle et al., 2000; Hall & Ozburn, 2018): raised /æ/ is opaque (i.e., serves as a harmony trigger), while raised /ɑ/ is transparent. This claim has been based on eight data points from Vaux (2000), which I reproduce below.

(65) *Data on /ɑ/ raising opacity from Vaux (2000)*

UR	**SR**	**Gloss**
/æswɑb-i-GA/	[**æ**swib-i-**gæ**]	'tool-3Pos-DAT'
/qæhwɑ-GA/	[q**æ**hwi-**gæ**]	'coffee-DAT'
/æmmɑ-lAr/	[**æ**mmi-l**æ**r]	'but-PL'
/ænʤɑn-i-GA/	[**æ**nʤin-i-**gæ**]	'Änjan-3Pos-DAT'

(66) *Data on /æ/ raising opacity from Vaux (2000)*

UR	**SR**	**Gloss**
/ɑdæm-i-GA/	[ɑdim-i-**gæ**]	'man-3Pos-DAT'
/ɑpæt-i-GA/	[ɑpit-i-**gæ**]	'disaster-3Pos-DAT'
/roʃæn-i-GA/	[roʃin-i-**gæ**]	'Roshän-3Pos-DAT'
/ɑʁinæ-lAr/	[ɑʁin-i-l**æ**r]	'friend-PL'

In the forms (65), the suffix harmonizes with the final surface vowel, with the raised /ɑ/ behaving like a transparent vowel. In the forms in (66), however, the suffix harmonizes with the underlying form of the raised /æ/, even though it is neutralized on the surface. This apparent asymmetry has been used to drive claims about how contrasts are represented in phonological inventories (Vaux, 2000; Halle et al., 2000; Hall & Ozburn, 2018).

There are reasons to be suspicious of these data, however. I asked five native Uyghur speakers to judge the forms in (65) and (66). They found all the forms in (65) and one form in (66) to be ungrammatical, for the following reasons:

- The final vowels of the stems /æswɑb/ 'tool', /ænʤɑn/ 'Änjan', and /roʃæn/ 'Rosän' never undergo raising. The surface forms provided by the consultants were [æsw**ɑ**b-i-**ʁɑ**], [ænʤ**ɑ**n-i-**ʁɑ**], and [roʃ**æ**n-i-**gæ**], respectively, with expected surface-true harmony.

- The Uyghur word for coffee is /qæhwæ/, not */qæhwɑ/. In addition, the final vowel does not raise, giving the surface form [qæhw**æ**-**gæ**].

- Three out of the four speakers rejected all forms of /æmmɑ-lAr/, which has the plural marker affixed to a conjunction. The one speaker who accepted it (under some duress) said they would say [æmm**a**-l**a**r], without raising.

It is possible that the forms from Vaux (2000) come from a different dialect of Uyghur, although at least one of my consultants felt that no native Uyghur speaker would produce them. In addition, data from the corpora are consistent with my consultants' judgements. Although this does not disprove the asymmetry between raised /ɑ/ and /æ/, it means that it must be validated by additional data. This will be addressed in the corpus study in Section 4.5.

4.4.6 Morpheme-specific opacity

Finally, certain morphemes in Uyghur have attracted attention in the literature because of they behave idiosyncratically with respect to the opaque process described above (Vaux, 2000; Halle et al., 2000). The best known of these is the diminutive suffix /-ʧæ/. The vowel in this suffix does not harmonize, and, unlike the examples ending in /æ/ above, becomes transparent to harmony when raised, even when it is the only harmonizing vowel in the stem.

(67) *Transparent raising of /-ʧæ/*

UR	**SR**	**Gloss**
/næj-ʧæ-DA/	[næjʧid**æ**]	'flute-DIM-LOC'
/kitɑb-ʧæ-DA/	[kitɑ**ɑ**pʧid**ɑ**]	'book-DIM-LOC'
/ziχ-ʧæ-GA/	[zi**χ**ʧiʁ**ɑ**]	'skewer-DIM-DAT'

My consultants accept the forms in (67).

4.5 A corpus study of opacity in Uyghur backness harmony

In order to investigate the interaction of vowel raising and backness harmony, I performed a large scale corpus study using two corpora from different geographical areas, and supplemented by judgments from my Uyghur consultants. The first corpus was generated from the website of *Uyghur Awazi* (Uyghur Voice), an Uyghur-language newspaper published in Almaty, Kazakhstan (`http://uyguravazi.kazgazeta.kz/`).

The second corpus was generated from the Radio Free Asia (RFA) Uyghur language website (`https://www.rfa.org/uyghur/`). RFA is a non-profit news organization that focuses on serving regions of Asia where government censorship of news reports is pervasive. The articles on the RFA Uyghur website are written in Central Uyghur, the standard Xinjiang dialect.

Corpora were generated from the the websites using *web scrapers*. A web scraper is software that, given a starting URL, instructions for how to navigate between pages, and instructions for which information to retrieve from each page, can download content from all pages on a site, or multiple sites. Such programs are useful for generating corpora from publicly available internet

resources, in formats that are useful to researchers.

There are separate web scrapers for the Uyghur Awazi[3] and the RFA[4] websites. These scrapers were written by the author in collaboration with undergraduate research assistants at UCLA.[5] They are freely available for use in research. Information on how to run the scrapers can be found at the links supplied in the footnotes.

The data presented in this paper were retrieved from the Uyghur Awazi and RFA websites in January 2020. They consist of a complete scrape of both sites. The Uyghur Awazi corpus contains approximately 14,000 articles with a total of about 6.1 million words. The RFA corpus consists of approximately 30,000 articles with a total of about 9.6 million words. The scrapers stored the results in comma-separated value files containing the article text in both the original orthography and converted to standard Uyghur Latin, the author, the date, the URL, and some additional metadata.

4.5.1 Uyghur orthography

Uyghur is spoken in a variety of countries and diasporic communities, and accordingly is represented by a variety of orthographies. Perso-Arabic script is typically used in China, while former Soviet countries tend to use Cyrillic script. Diasporic communities in the United States and elsewhere often use Latin orthography, of which there are several competing variants.

Despite this abundance of representations, all the official Uyghur orthographies map fairly closely onto pronunciation, which facilitates phonological research based on written corpora. In particular, the alternations conditioned by the raising and harmony processes described below are represented orthographically. The allophonic alternations in the transparent vowels discussed early are not represented, nor are other common processes like vowel devoicing.

[3] `https://github.com/connormayer/uyghur_tools/tree/master/uyghur_awazi_scraper`

[4] `https://github.com/yzgncx/RFA-Scraper`

[5] Thanks to Daniela Zokaeim and Tyler Carson for their work on these programs.

4.5.2 Parsing the corpora

In order to parse the corpus to retrieve information about the interaction between backness harmony and vowel raising, I modified an existing Uyghur morphological transducer to detect the backness of suffix forms (`https://github.com/apertium/apertium-uig`; Littell et al., 2018; Washington et al., to appear). This analyzer is part of Apertium, a free and open-source rule-based machine translation platform (`https://www.apertium.org`).

A morphological transducer maps between surface forms in various orthographies and underlying analyses that consist of roots plus morphological tags. This mapping may take place in either direction. For example, suppose we are mapping from surface forms to underlying analyses. If the input is the surface form *qizingizgha* "to your daughter" the output analysis will be *qiz<n><px2sg><frm><dat>* (I use Latin orthography throughout this section rather than IPA, since it more closely reflects the input to the parser). This indicates that the stem is *qiz*, a noun (<n>), and is suffixed with the 2nd person singular possessive affix (<px2sg>) in its formal form (<frm>) followed by the dative suffix (<dat>). The transducer can also carry out mapping in the opposite direction, from underlying analysis to surface form. In this case, the underlying input *qiz<n><px2sg><frm><dat>* produces the surface output *qizingizgha*.

I modified the uig-apertium transducer to detect the harmonic quality of suffixes when mapping from surface to underlying forms.[6] Under this modified system, the form *qizingizgha* presented above maps to *qiz<n><px2sg><frm><dat-b>*, indicating that the dative suffix surfaces in one of its back harmonizing forms (*-qa* or *-gha*) rather than one of its front harmonizing forms (*-ke*, *-ge*, or *-qe*). In addition, the restrictions the phonological component of the parser imposes on harmony have been lifted. The original parser, for example, would reject a form like *at-ler*, "horses", for being disharmonic. The modified parser will instead simply interpret this as an instance of the front form of the plural suffix.

The vowel raising processes described in Section 4.4.4 can obscure the harmonizing quality of suffixes: for example, the surface realization of /dost-lAr-m/ 'friend-PL-1SgPos' is [dostlirim],

[6]The code for this modified transducer can be found at `https://github.com/connormayer/apertium-uig/tree/vowel_harmony` .

which does not allow the backness of the plural morpheme to be determined. In such cases the modified parser does not attempt to guess the backness of the suffix (i.e., to report either <*pl-f*> or <*pl-b*>), but will instead remain agnostic, simply reporting <*pl*>.

I also performed some manual error correction on the transducer: primarily removing or modifying invalid stems.

Additional details of the parser are presented in Appendix A.

4.5.3 Interpreting the parser output

Applying the parser to the corpus will produce one or more possible parses for each word that the parser is able to recognize. We can use this output to calculate the frequency with which particular stems take front or back suffixes.

One challenge that arises is how to deal with multiple parses in cases where they provide conflicting information about the stem. The surface form *balilar*, for example, could correspond to underlying /bɑlɑ-lAr/ 'child-PL' or to /bɑl-i-lAr/ 'honey-3pos-PL.' Although the transducers used in the Apertium system can be augmented with contextual information to determine the most likely parse for ambiguous forms, this functionality is not yet available for Uyghur. I accordingly omit cases of ambiguity relating to stem identity from the results presented below.

A second challenge is how to treat forms that display conflicting harmony values among suffixes, such as the hypothetical form *qizlarge* "to the daughters", which contains the back form of the plural suffix /-lAr/ and the front form of the dative suffix /-KA/. I omit all tokens that produce at least one such parse. That being said, these tokens are exceedingly uncommon (about 500 tokens from 13.24 million words) and frequently result from highly improbable, though logically possible, parses. The suffixes applied to a stem almost always share the same backness values, unless one suffix is a harmony blocker.

Finally, I omit tokens containing suffixes like the continuous suffix /-(i)wɑt/ that block harmony by failing to harmonize, and impose their own harmony on following suffixes.

4.5.4 Quantitative results

The morphological parser was able to successfully parse 4.55 million of the 6.13 million words in the *Uyghur Awazi* corpus (74.3%) and 8.69 million of the 9.58 million words in the RFA corpus (90.6%), resulting in a total of 84.3% of all words across the corpora being successfully parsed. The poorer performance on the *Uyghur Awazi* corpus may be due to differences in the vocabulary of Uyghur spoken in Kazakhstan and in the content of the corpus. In addition, some articles contain sections of Kazakh text, which would fail to be parsed.

Of the stems that were successfully parsed, there were 215 that had the necessary structure to produce simple opacity. I divide these into two classes: BF stems ($n = 181$), whose final two harmonizing vowels are a back vowel followed by æ (e.g., /ɑdæt/ 'custom', /sijɑsæt/ 'politics'); and FB stems ($n = 34$), whose final two harmonizing vowels are a front vowel followed by *A* (e.g., /ætrɑp/ 'area', /æhwɑl/ 'condition'). For simplicity, I omit stems containing harmonizing dorsals, though transparent vowels may be found in the stems I consider.

Subsequent text processing comparing the stem identified by the parser and the surface form shows that 183 of these stems displayed vowel reduction. Of the BF stems, 177 (97%) exhibited vowel reduction, while of the FB candidates, only 6 (18%) did. This is in line with the general tendency for /æ/ to reduce more frequently than /ɑ/.

The rates of back responses for BF and FB candidates with and without raising are shown in Fig. 4.1. These figures show that opaque harmony (that is, harmony with the underlying form) is the norm for raised stems. Of the 183 stems represented in this graph, 96 display opaque raising without exception. 87 stems, however, vary in whether they display surface-true harmony or opaque harmony. The distribution of surface-true harmony across raised BF and FB stems are shown in Figs. 4.2 and 4.3 respectively.

I fit a linear regression model with proportion of raised tokens and overall frequency in corpus as predictors and rate of opaque harmony as the dependent variable. The proportion of raised tokens is defined as the number of raised tokens of a stem divided by the total number of occurrences. The model shows that proportion of raised tokens is a significant negative predictor of rate of opaque harmony (Fig. 4.4; $\beta = -0.12$, $t = -2.74$, $p < 0.01$), while the overall frequency is

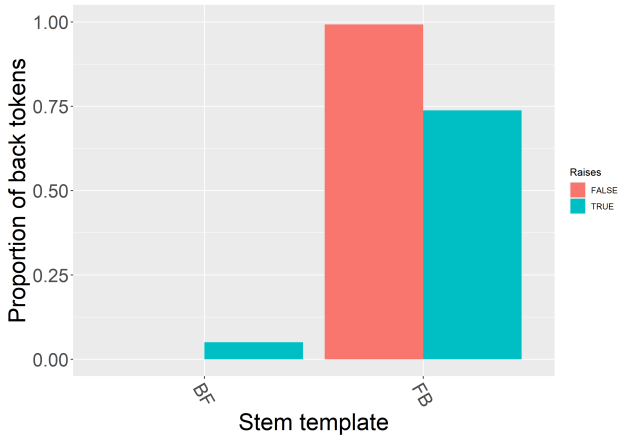

Figure 4.1: Proportion of back tokens for BF and FB stem types that undergo raising. Suffixes on unraised tokens always harmonize with the final vowel, while raised tokens vary in whether suffixes harmonize with the underlying form of the raised vowel, or the preceding vowel.

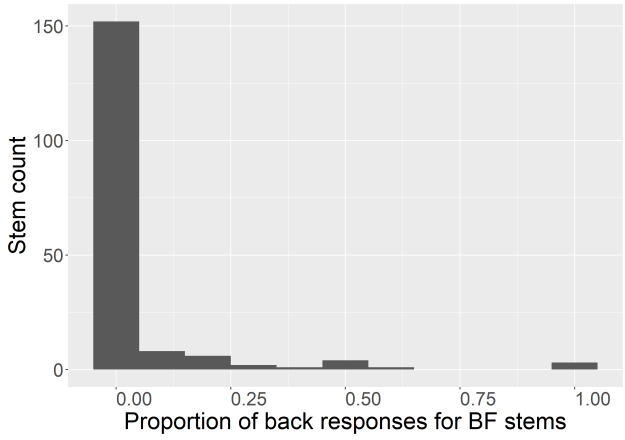

Figure 4.2: Histogram of back responses for BF stems in their raised forms.

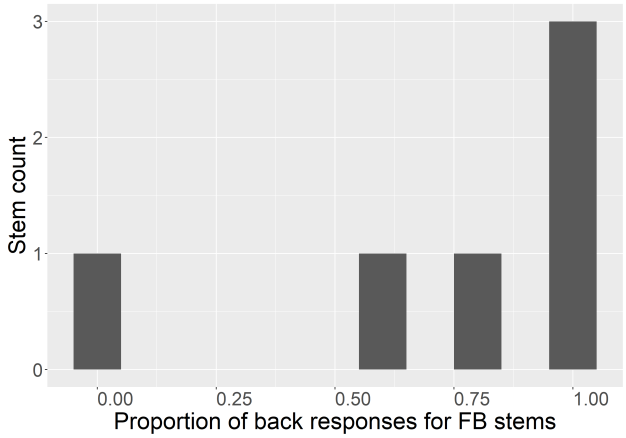

Figure 4.3: Histogram of back responses for FB stems in their raised forms.

a marginally significant predictor of the rate of opaque harmony (Fig 4.5; $\beta = 0.012$, $t = 1.88$, $p = 0.06$). In other words, the more frequently a stem occurs in its raised form, the more likely it is to display surface-true harmony.

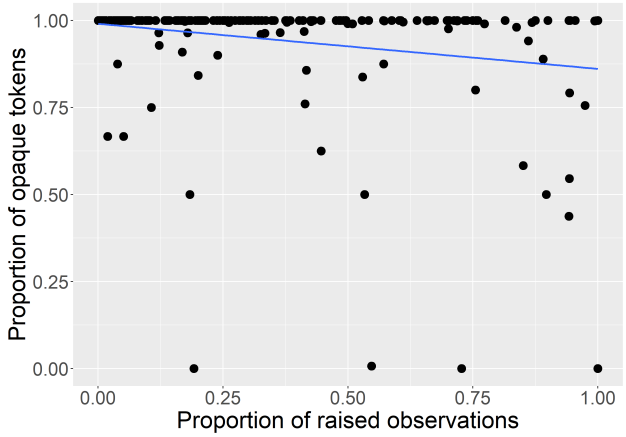

Figure 4.4: Proportion of harmony with the underlying form of the raised vowel (opaque behavior) plotted proportion of raised tokens.

4.5.5 Gradient opacity within a single stem

I will briefly illustrate examples of variation between opaque and surface-true harmony for a single stem: /sɑhɑbæ/ 'Companion to the Prophet'. This word occurs in its raised form [sɑhɑbi] in 95% of tokens (65/69) and displays opaque harmony (i.e., takes front suffixes) 43% of the time. The

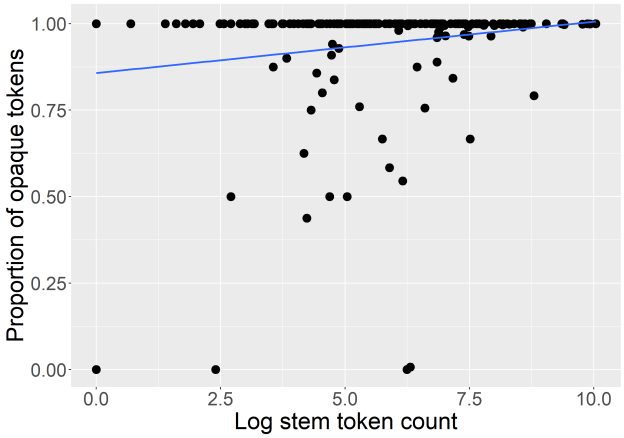

Figure 4.5: Proportion of harmony with the underlying form of the raised vowel (opaque behavior) plotted against overall frequency.

high frequency of raised tokens has to do with the contexts in which this word is typically used: it is almost always produced with the plural suffix /-lAr/.

Below are a few examples of this stem in its unsuffixed form and in its raised form with both opaque and surface-true harmony. All examples come from the RFA corpus. Note that these examples are presented in Latin Uyghur orthography, rather than IPA, where /æ/ is notated as 'e' and /ʁ/ as 'gh'. Translations are my own.

(68) *Unsuffixed tokens of 'sahabe'*

⟨⟨Xalid⟩⟩ dégen sözdin eqlimge islam tarixidiki meshhur ***sahabe****, qabil qomandan Xalid Ibni Welid keldi.*

The word 'Khalid' brought to my mind the powerful commander Khalid Ibn Walid, a well-known Companion in the history of Islam.

(69) *Opaque tokens of raised 'sahabe'*

Siyasetchiler üchün ereblerning wehhabiy bolghini bilen iranliqlarning ***sahabilerni*** *tillishi otturisida perq yoq.*

For politicians, there is no difference between the Arabs' having become Wahhabi and the Iranians' insulting of the Companions.

Andin ***sahabilerning*** *hijret qilishigha ruxset qilghan.*

Then (he) allowed the Companions to emigrate (make the hijrah).

(70) *Surface-true tokens of raised 'sahabe'*

Uni körgen ***sahabilar*** *kütülmigen bu ehwalni körüp ghezeblinidu.*

The Companions who saw him were shocked to see this unexpected situation.

⟨⟨*Adaletlik bir padishah bar*⟩⟩ *dégen sözi* ***sahabiligha*** *bergen bir kapalet idi.*

His words that 'there is a righteous king' were a promise given to his Companions.

4.5.6 Supplementing corpus results with elicitation

My Uyghur consultants provided insights on several forms that weren't in the corpus. There are cases of raised /ɑ/ being opaque, as in /ʃæytɑn-i-GA/ → [ʃæytin-i-ʁɑ] 'satan-3Ps-DAT'. There are also cases where it's transparent, as in /ærzɑn-i-GA/ → [ærzin-i-gæ] 'cheap-3Ps-DAT'. This is an interesting case because this is a somewhat idiosyncratic construction meaning '(to sell) for cheap': adjectives generally cannot take a possessive suffix. Finally, there are cases where both harmonizing forms may be acceptable, such as /æzɑn-i-GA/ → [æzin-i-ʁɑ] or [æzin-i-gæ] 'call to prayer-3Ps-DAT'. My consultants generally prefer the former, but says the latter also sounds fine.

4.6 Analysis of the corpus results

The quantitative results from the corpus show that (a) although opaque harmony is the norm for raised forms, the same stems may be found in tokens with surface-true harmony; and (b) the rate of surface-true harmony increases with the proportion of raised tokens observed.

On the basis of this data, I suggest that the opaque pattern in Uyghur is best modeled as *the maintenance of a relationship between surface realizations of a stem*, rather than a type of procedural mapping from underlying to surface form. This can be thought of as a type of paradigm uniformity (Steriade, 2000) where all the allomorphs of a stem must take suffixes of the same backness, even if this violates surface harmony.

Specifically, I suggest that this type of paradigm uniformity can be modeled as the requirement of harmony with a remote base (Stanton & Steriade, 2014). Following Albright (2002, 2010), who finds that in cases of morphological paradigm leveling the most phonologically predictive allomorph, I assume that there is strong pressure for the unraised form to serve as the remote base,

since raising neutralizes predictive phonological contrasts. However, the influence of this base on harmonic behavior across the paradigm is weighted as a function of exposure: the more frequently speakers are exposed to the unraised form, the stronger its effect as a remote base will be.

4.6.1 Phonological modeling

I will model the corpus results using Maximum Entropy Harmonic Grammar (henceforth MaxEnt; Smolensky, 1986; Goldwater & Johnson, 2003), a variant of Optimality Theory (Prince & Smolensky, 1993/2004). MaxEnt is able to generate probability distributions over output candidates for a given input, making it well-suited to capturing phonological gradience of the kind we see in Uyghur.

In a MaxEnt grammar, each constraint is associated with a real-valued weight that represents its strength. In a grammar with N constraints, the weight of the ith constraint can be notated w_i. The function $C_i(x)$ returns the number of times an output candidate x violates the ith constraint. The *harmony* of an output form x is:

$$H(x) = \sum_{i}^{N} w_i C_i(x)$$

where higher values of $H(x)$ are associated with more constraint violations. The probability of an output candidate x is

$$P(x) = \frac{\exp(-H(x))}{\sum_{y \in \Omega} \exp(-H(y))}$$

where Ω is the set of all possible output candidates.

Given a data set of observed frequencies and violation profiles for each candidate, the weights that maximize the likelihood of the data set can be calculated using standard numerical optimization techniques (see, e.g., Hayes & Wilson, 2008).

4.6.2 Constraints

I assume a set of constraints that motivate backness harmony and raising. For simplicity's sake I omit constraints related to consonant harmony here. The constraints that enforce vowel harmony

are:

- VAGREEBACK: Suffixes must agree in backness with the final back vowel of the stem.
- VAGREEFRONT: Suffixes must agree in backness with the final back front of the stem.

Note that these are local agree constraints: they can only be violated based on the final harmonizing vowel in the stem.

The constraint that motivates vowel reduction is:

- *UNREDUCED: Don't have low vowels in medial open syllables.

Vowel reduction violates several faithfulness constraints: ID(low), ID(high), and MAX(back), assuming [i] and [e] are unspecified for [front]. These constraints should have vowel-specific versions for [ɑ] vs [æ]. I omit these constraints here since I model only forms that undergo raising.

In a more comprehensive analysis the definition of these faithfulness constraints would need to be expanded to account for stems that do not raise, either by lexical indexation or some representation of lexically-specified stress patterns. I leave this aside for this paper, since the focus is the behavior of raising stems.

4.6.3 Capturing lexical variation using weighted bases

I assume that the input passed into GEN consists of the underlying form plus an associated *base weight*, w_b, that indicates the strength of the base's influence on realizations across the paradigm. For example, a base that raises but never displays surface-true harmony, like /ɑʔilæ/ 'family' might have a relatively high base weight, while a word like /sɑhɑbæ/ which displays surface-true harmony the majority of the time will be expected to have a lower base weight.

These weights are used in the calculation of violations of a constraint that motivates suffix paradigm uniformity.

- SPU($\mathscr{C}$): If the candidate allomorph of the stem does not match the base allomorph, calculate the number of violations of the proposed suffix form against the base allomorph using

the constraints in the set $\mathscr{C}$. The number of violations are multiplied by the base weight w_b to get the overall violation count for this constraint.

How do we determine the base weight for a stem? Results from the corpus study suggest that both how frequently a stem is observed in its unraised form and its overall frequency are related to how frequently it displays opaque harmony, which in turn suggests a higher stem weight. To approximate these qualities in a single numeric value, I use the natural log of the number of unraised tokens in the corpus as the base weight. This has several desirable properties: first, it is easily calculated; second, both high stem frequency and a high proportion of unraised tokens are positive correlated with the total number of unraised tokens.

Returning to the examples above, the number of unraised tokens of /ɑʔilæ/ is 2435, so its base weight $w_b = \ln(2435) = 7.8$, while the number of unraised tokens of /sɑhɑbæ/ is 4, so its base weight $w_b = \ln(4) = 1.39$.

4.6.4 Evaluating against corpus data

I fit a MaxEnt OT model using the constraints and method of calculate base weights described above to a subset of the corpus data consisting of stems whose final two harmonizing vowels fall into one of the following four templates: FF, BB, FF, and BB, where the final vowel will always be one of /æɑ/. As above, I only consider stems that undergo raising, and I omit stems with harmonizing consonants for the sake of simplicity. This data set contains 1397 stems corresponding to 231,274 tokens. I consider only suffixes that trigger raising and do not model suffix identity. The optimal model assigned a log likelihood of -8337 to the data.

I will illustrate the learned weights and the calculation of predicted probabilities of three prototypical forms: one with strongly opaque harmony, one with strongly transparent harmony, and one that is intermediate.

The tableau for calculating possible surface realizations of the input /ɑʔilæ-lAr/ 'family-PL' is shown in Table 4.3. The predicted frequencies by the model are a close approximation of the observed frequencies in the corpus. Because the subset of the corpus I examine contains only tokens that undergo raising, the weight for *UNRAISED is accordingly high, proving fatal to all candi-

/ɑʔilæ-lAr/ $w_b = 7.8$	Pred. Freq.	Obs. Freq.	H	VAGREEBACK $w = 1.11$	VAGREEFRONT $w = 3.56$	SPU(VAGREE) $w = 0.81$	*UNRAISED $w = 27.52$
ɑʔilæ-lær	0	0	27.52				1
ɑʔilæ-lɑr	0	0	31.08		1		1
ɑʔili-lær	0.995	1	1.11	1			
ɑʔili-lɑr	0.005	0	6.30			7.8	

Table 4.3: Tableau for the strongly opaque form /ɑʔilæ-lAr/.

dates that do not raise. The final two rows of Table 4.3 compare raised candidates that harmonize opaquely and transparently respectively. Although the opaque candidate [ɑʔili-lær] is surface-disharmonic and violates VAGREEBACK, this violation is overshadowed by the transparent candidate [ɑʔili-lɑr]'s violation of SPU(VAGREE), which, due to /ɑʔilæ/'s high base weight, is the worse of the two candidates.

Table 4.4 shows possible surface realizations of the input /sɑhɑbæ-lAr/ 'disciple-PL', which exhibits a high frequency of tokens exhibiting transparent frequency. Note that although these candidates have exactly the same violation profiles as /ɑʔilæ-lAr/'s candidates in Table 4.3, the weaker base weight for /sɑhɑbæ/ makes the transparent candidate [sɑhɑbi-lɑr] roughly as harmonic as the opaque candidate [sɑhɑbi-lær], and producing similar predicted frequencies.

/sɑhɑbæ-lAr/ $w_b = 1.4$	Pred. Freq.	Obs. Freq.	H	VAGREEBACK $w = 1.11$	VAGREEFRONT $w = 3.56$	SPU(VAGREE) $w = 0.81$	*UNRAISED $w = 27.52$
sɑhɑbæ-lær	0	0	27.52				1
sɑhɑbæ-lɑr	0	0	31.08		1		1
sɑhɑbi-lær	0.50	0.44	1.11	1.10			
sɑhɑbi-lɑr	0.50	0.56	1.12			1.4	

Table 4.4: Tableau for the strongly transparent form /sɑhɑbæ-lAr/.

Finally, Table 4.5 shows an intermediately opaque form /kæsipdɑʃ-i-lAr/ 'colleague-3Pos-PL' (the use of the possessive morpheme here is necessary to produce the environment for raising). The violation profiles for output candidates are similar to the previous two tableaux (modulo the differing backness of the final vowel), and the base weight, which is intermediate between those of

the previous two tableaux, accordingly produces an intermediate degree of opaque harmony.

/kæsipdɑʃ-i-lAr/ $w_b = 6.9$	Pred. Freq.	Obs. Freq.	*H*	VAGREEBACK $w = 1.11$	VAGREEFRONT $w = 3.56$	SPU(VAGREE) $w = 0.81$	*UNRAISED $w = 27.52$
kæsipdɑʃ-i-lær	0	0	28.63	1			1
kæsipdɑʃ-i-lɑr	0	0	27.52				1
kæsipdiʃ-i-lær	0.11	0.16	5.61			6.9	
kæsipdiʃ-i-lɑr	0.89	0.84	3.56		1		

Table 4.5: Tableau for the intermediate form /kæsipdɑʃ-i-lAr/.

4.6.5 Opacity in suffixes

The most robust cases of vowel reduction leading to transparent harmony in Uyghur are found in suffixes, like the diminutive /-ʧæ/ (Example 67). Such suffixes, that are both harmonically invariant and undergo raising, are generally uncommon in the language. Suffixes that raise tend to be harmonizing suffixes (e.g., the plural /-lAr/), while suffixes that do not harmonize also tend to not raise (e.g., the progressive /-wɑt/). The fact that /-ʧæ/ always behaves transparently when raised can be explained if it not part of the stem: there is no paradigmatic pressure for following suffixes to harmonize with its unraised form. This corpus corroborates this, with 95% of raised tokens of stems ending in a back vowel followed by /-ʧæ/ displaying transparent harmony.

It is instructive to compare /-ʧæ/ with a much less productive suffix /-jæ/, which derives country names from names of ethnicities.

(71) *Uses of /-jæ/*

finlɑnd	'Finnish person'	finlɑndijæ	'Finland'
rus	'Russian person'	rusijæ	'Russia'
ispɑn	'Spanish person'	ispɑnijæ	'Spain'

This suffix is not generally productive (even among names of countries), and may be considered somewhat analogous to English -ia, as in 'Bulgaria', 'Russia', 'Estonia', etc. Forms with a back vowel followed by /-jæ/, such as those shown above, show opaque harmony 97% of the time (e.g.,

/finlɑndijæ-DA/ → [finlɑndiji-dæ], not *[finlɑndiji-dɑ]).

The difference in harmonic behavior between these two similar suffixes can be accounted for by positing that certain derivational morphemes such as /-jæ/, by virtue of their lack of productivity, may be considered part of the base, while others such as /-ʧæ/, which are more productive, are not treated as part of the base, and thus do not exert pressure to harmonize with their underlying vowels in the face of raising. This predicts that less productive suffixes should generally exhibit higher degrees of opacity in their raised forms. I leave this as an area of exploration for future research.

4.6.6 Other empirical considerations

Several other properties of the Uyghur backness harmony system, and of backness harmony systems in general, support this analysis.

First, the Uyghur backness harmony system already requires a large degree of lexical specification. The phonetic and phonological properties of the class of stems containing only transparent segments are not effective predictors of suffix backness (Examples 53 and 54). In addition, subsets of the stems containing dorsals violated expected harmony patterns: some stems containing /k g/ take back suffixes (Example 52), and some stems with a front vowel followed by a uvular take back suffixes (Example 56). Thus harmony is already unpredictable to some extent from surface forms, and treating harmony as a partially morphological phenomenon does not introduce complications that are not already found in other areas of the system.

I also note that researchers like Kaun (2004) have suggested that vowel harmony serves to enhance the perceptibility of vowels in the stem by spreading their features through the word. If harmony in Uyghur were not generally opaque in the face of raising, this would mean that the same stem could take suffixes of differing backness in its different forms, which could potentially render harmony's perceptual benefits much less effective.

4.7 Discussion

Based on the results of the corpus study and general consideration of the properties of the Uyghur harmony system, this paper suggests that the opacity observed in Uyghur backness harmony is best considered as the maintenance of a relationship between surface forms of a stem, rather than a type of mapping from underlying to surface forms.

This might be thought of as a stronger version of grammatical gender. In languages with grammatical gender there are often strong phonological clues to which gender a noun is: in Spanish, for example, masculine nouns often end in -o and feminine nouns in -a. There are also frequent exceptions to these generalizations: for example, Spanish *mapa* 'map' is masculine despite ending in -a. Just as we should not expect a morphological property such as gender to change across the allomorphs of a word, we may similarly expect the harmonic value of a stem to remain constant, even when allomorphy renders its surface form disharmonic.

It's unclear whether this case of opacity is of relevance to the debate between serial vs. parallel theories, despite suggestions that it is (e.g., Vaux, 2000). It is unlikely that either serial or parallel theories can explain the kind of variation observed in the corpus without some kind of appeal to lexical specification: serial theories will fail to predict cases of surface-true harmony, and parallel theories will fail to predict opacity. Similarly, I suspect that both serial and parallel theories can be modified to capture this data given such an appeal. This paper has shown one way in which a parallel theory may explain this data. The question remains whether opaque patterns in general can be analyzed on the basis of lexical or morphological features, or if any can be unequivocally analyzed as strictly phonological processes.

Adopting Baković's suggestion, we might ask about the learnability and productivity of this pattern. Should this pattern under the analysis presented above be considered productive? This depends on how the idea of productivity is defined. As a mapping from underlying representation to surface form, it may not be: some knowledge of the morphological paradigm of a stem is necessary to compute the outputs. However, given sufficient knowledge of a stem's allomorphs, it would be surprising if speakers were not able to generalize effectively. This is an interesting area for future study using artificial grammar learning experiments (e.g., Moreton & Pater, 2012),

where participants could be presented with tokens of stems that vary in their rate of opacity, and tested to see if they generalize differently on the basis of these frequencies.

Similarly, we might wonder whether to expect a shift to transparent harmony as predicted by Kiparsky (1971). It is probably too early to tell at the time of this writing: raising is a relatively new process in Uyghur, and the phonological system may not have fully absorbed its consequences. The analysis presented above predicts that we should not see a shift towards surface harmony, but rather that backness harmony should be represented in speakers' minds as a morphological property with strong phonological correlates, rather than a strictly phonological process. It will be interesting and informative for phonological theory to see how the Uyghur backness harmony system evolves over subsequent generations.

An important limitation of the current study is that it relies on a specific genre of text: newspaper articles. This is a limitation for two reasons: the first is that it has a significant effect on word frequencies, as newspaper articles tend to be about specific topics. Thus it will be important to compare the relationships between opacity and the token frequency effects discussed in this paper on other corpora, and determine whether a more robust way of calculating base weight is possible. Although the log of the number of unraised tokens fits this corpus reasonably well, it both over- and under-predicts rates of opacity for certain stems. Second, newspapers generally use prestige varieties of the languages they are written in. My consultants, who are generally highly educated, see opaque harmony as the 'correct' behavior, despite their acknowledgment of some areas of variability. It will be valuable to look at opacity in more conversational corpora of Uyghur, where surface harmony could potentially be more prevalent. Although there are online Uyghur language message boards that use less formal language, these are frequently written using non-standard, phone-friendly Latin orthographies that eliminate the distinctions between some front and back vowels (e.g., /y/ and /u/ may both be written as *u*), making corpus study difficult.

Finally, this paper demonstrates the value of taking a broad empirical approach to a language. The internet has allowed for the proliferation of textual data, even for smaller languages. Computational tools such as morphological parsers, when applied with discretion and a knowledge of the language under study, can provide large amounts of empirical data that allow us to supplement other data sources and measure phonological patterns writ large. I anticipate that such tools will

become increasingly important as phonology continues to develop as a science.

Appendix A: The apertium-uig transducer

The apertium-uig transducer is implemented using finite-state transducers (FST): specifically, within the HFST framework (Helsinki Finite State Technology; Linden et al., 2011). A FST is a finite-state automaton (FSA) that contains two tapes: in this case, one corresponding to underlying analyses and one to surface forms. Each transition or arc in the transducer has a symbol corresponding to each tape. Either tape may be designated as the input. The transducer reads the input and takes the appropriate transitions between states. The symbols on the transitions corresponding to the output tape are written to an output buffer. If the transducer reaches a valid output state after consuming the entire input, then the contents of the output buffer are returned.

Any SPE-style system that uses sequences of rewrite rules to map from underlying analyses to surface forms can be implemented as a finite-state transducer (Johnson, 1972; Kaplan & Kay, 1994). In practice, this poses several problems, the most serious of which is that although the mapping from an underlying analysis to a surface form is deterministic given a set of rules, the inverse is not true in general. In fact, it is possible for a given surface form to correspond to a large, or even infinite, number of underlying analyses under certain rule systems. This quickly becomes intractable for any practical implementation of a morphological transducer. The *two-level morphology* framework (Koskenniemi, 1983, 1984, 1986; Beesley & Karttunen, 2003), which is implemented in HFST, was designed to mitigate these issues.

Two-level morphology divides the mapping between underlying analyses and surface forms into two stages. The first stage maps between a morphological analysis and an abstract morphophonological form, which allows a minimal representation of stems and suffixes. For example, the analysis *qiz*<*dat*> will map to *qiz*>$\{G\}\{A\}$ at this level, where > represents a morpheme boundary and $\{G\}$ and $\{A\}$ are essentially archiphonemes. It is this stage that solves the problem of overgeneration of underlying analyses: every valid underlying stem must be encoded at this level.

The output of the first level then serves as input to the next level, which maps abstract mor-

phophonological forms to surface forms. In this case, the phonological rules specified in the transducer will map $\{G\}$ to *gh* and $\{A\}$ to *a*, producing the surface form *qizgha* "to a girl."

In HFST, the first stage is implemented using the LEXC formalism, while the second is implemented using the TWOLC formalism. The rules specified at these levels are compiled into FSTs, which are then compose-intersected to form a single transducer. This transducer will only accept or propose underlying roots that are specified in the lexicon. Unfortunately, this introduces a degree of brittleness, since the transducer will not recognize any forms that are not present in the lexicon, and has no means by which to 'guess' the underlying form from the surface form unless augmented with additional tools.

.0.0.1 Modifying the transducer

The parser was modified to detect harmony by splitting each tag or tags corresponding to a haronizing suffix into three different forms corresponding to front variants (e.g., <*dat-f*>), back variants (e.g., <*dat-b*>), and ambiguous variants (e.g., <*dat*>). These tags are mapped to more restricted, though still abstract, morphological forms in the first stages. For example, <*dat-f*> will map to $\{Gf\}\{Af\}$, while <*dat-b*> will map to $\{Gb\}\{Ab\}$.

The second stage has been modified to map the newly introduced archiphonemes at the first stage to a restricted set of surface forms with corresponding backness. For example, it maps $\{Gf\}$ and $\{Af\}$ to only front allophones, and $\{Gb\}$ and $\{Ab\}$ to only back allophones. Several other complications relating to the interaction of harmony with other processes were also accounted for.

References

Abdulla, A., Ebeydulla, Y., & Raxman, A. (2010). *Hazirqi zaman uyghur tili [Modern Uyghur]*. Ürümchi: Xinjiang Xelq Neshriyati [Xinjiang People's Publishing House].

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Aksënova, A., Graf, T., & Moradi, S. (2016). Morphotactics as tier-based strictly local dependencies. *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 121-130.

Albright, A. (2002). *The identification of bases in morphological paradigms* (Unpublished doctoral dissertation). University of California, Los Angeles.

Albright, A. (2010). Base-driven leveling in Yiddish verb paradigms. *Natural Language and Linguistic Theory*, *28*, 475–537.

Al-Mozainy, H. Q. (1981). *Vowel alternations in a Bedouin Hijazi Arabic dialects* (Unpublished doctoral dissertation). University of Texas, Austin, Austin, TX.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology*, *5*(2), 183-207.

Baek, H. (2017). Computational representation of unbounded stress: tiers with structural features. *Proceedings of the 53rd Annual Meeting of the Chicago Linguistic Society*, To appear.

Baković, E. (2007). A revised typology of opaque generalisations. *Phonology*, *24*(2), 1–43.

Baković, E. (2011). Opacity and ordering. In J. A. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory* (2nd ed., pp. 40–67). London: Wiley-Blackwell.

Becker, L. (2016). *Vowel-consonant harmony in Uyghur.* (Ms, Leipzig University)

Beesley, K., & Karttunen, L. (2003). *Two-level rule compiler* (Retrieved from https://web.stanford.edu/ laurik/.book2software/twolc.pdf). Stanford University.

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

Braver, A. (2013). *Degrees of incompleteness in neutralization: Paradigm uniformity in phonetics with weighted constraints* (Unpublished doctoral dissertation). Rutgers University.

Burness, P., & McMullen, K. (2019). Efficient learning of output tier-based strictly 2-local functions. In *Proceedings of the 16th meeting on the mathematics of language* (pp. 78–90).

Association for Computational Linguistics.

Chandlee, J. (2014). *Strictly local phonological processes* (Unpublished doctoral dissertation). The University of Delaware.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Clauson, G. (1972). *An etymological dictionary of pre-thirteenth-century Turkish*. Oxford: Clarendon Press.

de Lacy, P. (2002). *The formal expression of markedness* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

de Santo, A., & Graf, T. (2017). *Structure sensitive tier projection: Applications and formal properties.* (Ms., Stony Brook University)

Donegan, P. J., & Stampe, D. (1979). The study of natural phonology. In D. A. Dinnsen (Ed.), *Current approaches to phonological theory* (pp. 126 – 173). Bloomington, IN: Indiana University Press.

Eilenberg, S. (1974). *Automata, languages, and machines*. Academic Press, Inc.

Engesæth, T., Yakup, M., & Dwyer, A. (2009/2010). *Teklimakandin salam: hazirqi zaman Uyghur tili qollanmisi / Greetings from the Teklimakan: a handbook of modern Uyghur*. Lawrence: University of Kansas Scholarworks.

Gafos, A. I. (1999). *The articulatory basis of locality in phonology* (Unpublished doctoral dissertation). John Hopkins University.

Gallagher, G. (2016). Vowel height allophony and dorsal place contrasts in Cochabamba Quechua. *Phonetica*, *73*, 101-119.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447-474.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, & Östen Dahl (Eds.), *Proceedings of the stockholm workshop on variation within optimality theory* (pp. 111–120). Stockholm: Stockholm University, Department of Linguistics.

Graf, T. (2010). Comparing incomparable frameworks: A model theoretic approach to phonology. *University of Pennsylvania Working Papers in Linguistics*, *16*. Retrieved from `http://repository.upenn.edu/pwpl/vol16/iss1/10`

Graf, T. (2017). The power of locality domains in phonology. *Phonology*, *34*, 385-405. Retrieved from https://dx.doi.org/10.1017/S0952675717000197 doi: 10.1017/S0952675717000197

Graf, T., & Mayer, C. (2018). Sanskrit n-Retroflexion is Input-Output Tier-Based Strictly Local. In *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 151–160). Brussels, Belgium: The Special Interest Group on Computational Morphology and Phonology.

Hahn, R. F. (1991a). Diachronic aspects of regular disharmony in modern Uyghur. In W. Boltz & M. Shapiro (Eds.), *Studies in the Historical Phonology of Asian Languages.* John Benjamins.

Hahn, R. F. (1991b). *Spoken Uyghur*. Seattle, WA: University of Washington Press.

Hall, D. C., & Ozburn, A. (2018). *When is derived [i] transparent? a subtractive approach to Uyghur vowel harmony.* (Talk presented at the 49th Northeast Linguistics Society Conference (NELS 49), Cornell University, 5-7 October)

Halle, M., Vaux, B., & Wolfe, A. (2000). On feature spreading and the representation of place of articulation. *Linguistic Inquiry*, *31*, 387-444.

Hansson, G. O. (2001). *Theoretical and typological issues in consonant harmony* (Unpublished doctoral dissertation). University of California, Berkeley.

Hayes, B., & Londe, Z. (2006). Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology*, *23*, 59–104.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Hayes, B., Zuraw, K., Siptar, P., & Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, *85*, 822–863.

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry*, *41*, 623-661.

Heinz, J. (2018). The computational nature of phonological generalizations. In *Phonological Typology.*

Heinz, J., Kasprzik, A., & Kötzing, T. (2012). Learning with lattice-structure hypothesis spaces. *Theoretical Computer Science*, *457*, 111-127.

Heinz, J., & Koirala, C. (2010). Maximum likelihood estimation of feature-based distributions. In *Proceedings of the 11th Meeting of the ACL-SIGMORPHON. ACL 2010* (pp. 28–37). Uppsala, Sweden: Association for Computational Linguistics.

Heinz, J., Rawal, C., & Tanner, H. G. (2011). Tier-based strictly local constraints for phonology. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers*, 58-64.

Hooper/Bybee, J. (1976). *An introduction to nature generative phonology*. New York: Academic Press.

Jardine, A. (2016). Computationally, tone is different. *Phonology*, *33*, 247-283.

Jardine, A., & Heinz, J. (2016). Learning tier-based strictly 2-local languages. *Transactions of the ACL*, *4*, 87-98.

Jardine, A., & McMullin, K. (2017). Efficient learning of tier-based strictly *k*-local languages. In F. Drewes, C. Martín-Vide, & B. Truthe (Eds.), *Language and automata theory and applications, 11th international conference* (p. 64-76). Springer.

Johnson, C. D. (1972). *Formal aspects of phonological decription*. The Hague: Mouton.

Kaplan, R., & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics*, *20*, 331-378.

Kaun, A. (2004). The typology of rounding harmony. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 87–116). Cambridge, MA: Cambridge University Press.

Kimper, W. A. (2011). *Competing triggers: Transparency and opacity in vowel harmony* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Kiparsky, P. (1971). Historical linguistics. In W. Dingwall (Ed.), *A survey of linguistic science* (pp. 576–642). College Park: University of Maryland Linguistics Program.

Kiparsky, P. (1973). Abstractness, opacity, and global rules. In O. Fujimura (Ed.), *Three dimensions of linguistic theory* (pp. 57–86). Tokyo: TEC.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production* (Publication 11). University of Helsinki, Department of General Linguistics, Helsinki.

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Coling'84* (pp. 178–181).

Koskenniemi, K. (1986). Compilation of automata from morphological two-level rules. In F. Karsson (Ed.), *Papers from the fifth scandinavian conference on computational linguistics* (pp. 143–149).

Lai, R. (2015). Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, *46*, 425–451.

Lindblad, V. M. (1990). *Neutralization in uyghur*. University of Washington.

Linden, K., Silfverberg, M., Axelson, E., Hardwick, S., & Pirinen, T. (2011). HFST – Framework for compiling and applying morphologies. In C. Mahlow & M. Pietrowski (Eds.), *Systems and frameworks for computational morphology* (Vol. 100, pp. 67–85). Communications in Computer and Information Science.

Littell, P., Tian, T., Xu, R., Sheikh, Z., Mortensen, D., Levin, L., ... Hovy, E. (2018). The ARIEL-CMU situation frame detection pipeline for LoReHLT16: a model translation approach. *Machine Translation*, *32*, 105–126.

Major, T., & Mayer, C. (2018). Towards a phonological model of Uyghur intonation. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of the 9th speech prosody international conference.*

Martin, A. (2005). *The effects of distance on lexical basis: Sibilant harmony in Navajo compounds* (Unpublished master's thesis). UCLA.

Mayer, C., & Major, T. (2018). A challenge for tier-based strict locality from Uyghur backness harmony. In A. Foret, G. Kobele, & S. Pogodalla (Eds.), *Formal Grammar 2018. FG 2018. Lecture Notes in Computer Science, vol 10950.* Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-57784-4_4

Mayer, C., Major, T., & Yakup, M. (2019). *Wug-testing Uyghur vowel harmony.* (Talk presented at the 27th Manchester Phonology Meeting. Manchester, England. May.)

Mayer, C., Major, T., & Yakup, M. (in prep.). Are neutral stems in Uyghur really neutral? acoustic and corpus evidence.

Mayer, C., & Nelson, M. (in press). Phonotactic learning with neural language models. In *Proceedings of the society for computation in linguistics 2020.*

McCarthy, J. J. (1999). Sympathy and phonological opacity. *Phonology*, *16*, 331–339.

McCarthy, J. J. (2007). *Hidden generalizations: Phonological opacity in optimality theory*. London: Equinox Publishing.

McCollum, A. (2019). *Transparency, locality, ad contrast in Uyghur backness harmony* (ms.) Rutgers.

McMullin, K. (2016). *Tier-based locality in long-distance phonotactics: Learnability and typology* (Unpublished doctoral dissertation). University of British Columbia.

McMullin, K., & Hansson, G. O. (2019). Inductive learning of locality relations in segmental phonology. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*, 14.

McNaughton, R., & Papert, S. (1971). *Counter-free automata*. MIT Press.

Mielke, J., Hume, E., & Armstrong, M. (2003). Looking through opacity. *Theoretical Linguistics*, *29*(1–2).

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning. part i: Structure, part ii: Substance. *Language and Linguistics Compass*, *6*, 686-701 and 702-718.

Nazarova, G., & Niyaz, K. (2013). *Uyghur: An elementary textbook*. Washington, DC: Georgetown University Press.

Pattillo, K. E. (2013). The typology of Uyghur harmony and consonants. *Rice Working Papers in Linguistics*, *4*.

Pin, J. E. (1986). *Varieties of formal languages* (R. E. Miller, Ed.). Plenum Publishing Co.

Prince, A. (1998). *Two lectures on optimality theory.* (Handout of paper presented at Phonology Forum 1998, Kobe University.)

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Cambridge, MA: Blackwell. (Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993.)

Rogers, J., Heinz, J., Fero, M., & Hurst, J. (2013). Cognitive and sub-regular complexity. In G. Morrill & M.-J. Nederhof (Eds.), *Formal grammar, volume 8036 of lecture notes in computer science* (p. 90-108). Springer.

Rogers, J., & Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*.

Rose, S., & Walker, R. (2011). A typology of consonant agreement as correspondance. *Language*, *80*, 475–531.

Sanders, R. N. (2003). *Opacity and sound change in the Polish lexicon* (Unpublished doctoral dissertation). UCSC.

Schützenberger, M. (1965). On finite monoids having only trivial subgroups. *Information and Control*, *8*, 190-194.

Simon, I. (1975). Piecewise testable events. In H. Brakhage (Ed.), *Automata theory and formal languages 2nd gi conference* (Vol. 33, p. 214-222). Berlin: Springer. Retrieved from `http://dx.doi.org/10.1007/3-540-07407-4_23` doi: 10.1007/3-540-07407-4_23

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & T. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 194–281). Cambridge, MA: MIT Press/Bradford Books.

Stanton, J., & Steriade, D. (2014). *Stress window and base faithfulness in English suffixal derivatives.* (Paper presented at the 22nd Manchester Phonology Meeting)

Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology interface. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology v* (pp. 313–334). Cambridge, MA: Cambridge University Press.

Sylak-Glassman, J. (2014). The effects of post-velar consonants on vowels in Ditidaht. In N. Weber, E. Sadlier-Brown, & E. Guntly (Eds.), *Papers for the international conference on salish and neighbouring languages 49, university of british columbia working papers in linguistics* (Vol. 37).

Vaux, B. (2000). Disharmony and derived transparency in Uyghur vowel harmony. *Proceedings of NELS 30*, 671-698.

Vaux, B. (2008). Why the phonological component must be serial and rule-based. In B. Vaux & A. Nevins (Eds.), *Rules, constraints, and phonological phenomena* (pp. 20–60). Oxford University Press.

Vaux, B. (2011). Language games. In J. A. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory* (2nd ed., pp. 722–750). London: Wiley-Blackwell.

Washington, J., Salimzianov, I., Tyers, F. M., Gökırmak, M., Ivanova, S., & Kuyrukçu, O. (to appear). Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019).*

Yakup, M. (2013). *Acoustic correlates of lexical stress in native speakers of Uyghur and L2 learners* (Unpublished doctoral dissertation). University of Kansas.

Yli-Jyrä, A. (2003). Describing syntax with star-free regular expressions. In *EACL.*

Yli-Jyrä, A. M., et al. (2013). On finite-state tonology with autosegmental representations. In *Proceedings of the 11th international conference on finite state methods and natural language processing.*

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*, 153–201.

Zuraw, K. (2000). *Patterned exceptions in phonology* (Unpublished doctoral dissertation). University of California, Los Angeles.

Zymet, J. (2014). Distance-based decay in long-distance phonological processes. In *Proceedings of the 32nd west coast conference on formal linguistics.* Cascadilla Press.