A large-scale corpus study of phonological opacity in Uyghur

Connor Mayer
Department of Language Science, UC Irvine
cjmayer@uci.edu

Abstract

This paper examines a case of phonological opacity in Uyghur resulting from an interaction between backness harmony and a vowel reduction process that converts harmonic vowels into transparent vowels. A large-scale corpus study shows that although opaque harmony with the underlying form of a reduced vowel is the dominant pattern, cases of surface-apparent harmony also occur. The rate of surface-apparent harmony varies across roots and is correlated with a number of factors, including root frequency. These data pose problems for standard accounts of opacity, which do not predict such variation. I propose an analysis where variation emerges from conflict between a paradigm uniformity constraint mandating that the harmonising behavior of a root remains consistent, and surface phonotactic constraints. This is implemented in a parallel model by scaling constraint violations according to certainty in a root's harmonic class. This aligns with past work suggesting some opacity is driven by paradigm uniformity.

This paper presents an exploratory written corpus study of a phonological pattern in Uyghur (Turkic: China) whereby a vowel reduction process converts harmonic vowels into transparent vowels, rendering the vowel harmony pattern opaque. Opaque patterns are of interest to phonological theory because of the challenges they pose for learning and for certain classes of phonological models. In particular, opacity has been a perpetual difficulty for strictly parallel phonological models such as classical Optimality Theory (Prince & Smolensky, 1993/2004), which do not straightforwardly predict its existence.

The paper has two primary goals. The first is to put empirical claims about opacity in Uyghur on stronger footing by presenting new empirical data. Uyghur provides a valuable opportunity to study an opaque pattern at scale using text data, because both vowel reduction and harmony are reflected orthographically. The data show that although opaque harmony is the majority pattern, there is variability in rates of opacity both within and between roots[1]: roots sometimes trigger surface-apparent harmony (McCarthy, 1999). Rates of opacity are correlated with factors like token frequency and the presence of certain derivational suffixes.

The second goal is to highlight challenges these data pose for standard theories of opacity, which do not predict such variability. I will outline an analysis that is compatible with the data, where observed variation emerges as the result of conflict between lexical knowledge of the harmonising class of a root, and sensitivity to surface phonotactic constraints. In addition to accounting for the various correlates of opacity observed in the corpus, this analysis can be implemented in a strictly parallel model using a constraint that mandates paradigm uniformity in the harmony system (Rebrus & Törkenczy, 2017, 2021; Rebrus et al., 2023). This aligns with previous proposals that lexical factors play a fundamental role in opaque phenomena (Łubowicz, 2003; Sanders, 2003; Mielke et al., 2003; Green, 2004; Pater, 2010; Nazarov, 2019).

[1] I use the term "root" in this paper to refer to monomorphemic roots as well as polymorphemic stems with only derivational suffixes. This is in part because the morphological transducer that will be described below typically treats these derived stems as single roots, and because derivational suffixes will often impose their own harmonic properties, while inflectional suffixes inherit the harmonic properties of the stem. I hope the reader forgives this slight abuse of terminology in the interests of avoiding awkward prose.

Section 1 provides background on phonological opacity. Section 2 describes the processes of backness harmony and vowel reduction in Uyghur, and how they interact to produce opacity. Section 3 presents the results of a large written corpus study looking at rates of opacity. Section 4 presents an analysis of the corpus data. Finally, Section 5 discusses implications for theories of opacity in general, some limitations of the work presented here, alternative analyses, and how future research might proceed.

1 Phonological opacity

Opacity is a type of structured phonological exceptionality. Kiparsky (1971, 1973) defines it as follows:

(1) Assume a phonological rule $\mathbb{R}$: A → B / C_D. $\mathbb{R}$ is *opaque* iff 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)

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 (1973) associated opacity of types (1a) and (1b) with *counterfeeding* and *counterbleeding* rule orders respectively. In 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.

More recently, interest in opacity 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 elsewhere.

A number of theoretical mechanisms have been proposed to handle these cases, including mechanisms that incorporate some degree of serialism into OT, such as sympathy (McCarthy, 1999), Stratal OT (Kiparsky, 2000; Bermúdez-Otero, 2003; Nazarov & Pater, 2017; Bermúdez-Otero, 2018), candidate chain theory (McCarthy, 2007), and serial markedness reduction (Jarosz, 2014), as well as purely parallel mechanisms, such as constraint conjunction (Kirchner, 1996), paradigm uniformity (Steriade, 2000), language-specific constraints (Pater, 2014), or indexed constraints (Pater, 2010; Nazarov, 2019, 2020, 2021). 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, the typologies of opacity enumerated in Baković (2007, 2011) and Baković and Blumenfeld (2019) show that these orderings are neither sufficient nor necessary conditions for opacity. They identify 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 accurate in broad strokes.

In light of the lack of a unified account of opacity from either serial or parallel theories, Baković (2007) 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 constitute exceptions to otherwise robust generalizations. Kiparsky (1971) supports this claim by presenting a number of cases of historical change where an opaque process is reanalyzed as a non-opaque one.

Subsequent research has presented evidence that opaque processes are learned as phonemic contrasts or lexicalized patterns rather than productive rules (e.g., Hooper/Bybee, 1976; Mielke et al., 2003; Sanders, 2003; Sumner, 2003; Zhang, 2019; Bowers, 2019), 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) as well as in behavioral experiments (Farris-Trimble & Tessier, 2019).

2 Opacity in Uyghur backness harmony

Uyghur is a southeastern Turkic language spoken by over 12 million people in the Xinjiang Uyghur Autonomous Region in the People's Republic of China, neighboring 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.

The opaque phenomenon under consideration arises from the interaction of two independent processes: backness harmony and vowel reduction. I will introduce these processes separately before demonstrating how their interaction leads to opacity.

The reader is referred to Mayer, McCollum, and Eziz (2022) for a more detailed description of Uyghur phonology.

2.1 Segments involved in backness harmony

Like most Turkic languages, Uyghur has backness harmony (e.g., Lindblad, 1990; Hahn, 1991a, 1991b; Engesæth et al., 2009/2010; Abdulla et al., 2010). It is most evident as alternations between suffix allomorphs, where, broadly speaking, segments in the suffix must agree in backness with the rightmost vowel of the roots they attach to. It may also be observed to a lesser extent in static root forms (particularly native Turkic roots), but extensive borrowing has led to many disharmonic roots.

	Front		Back	
	Unrounded	Round	Unrounded	Round
High	i	y̲		u̲
Mid	e	ø̲		o̲
Low	æ̲		ɑ̲	

Table 1: The Uyghur vowel system. Harmonising vowels are underlined.

	Front	Back
Voiceless	k	q
Voiced	ɡ	ʁ

Table 2: Harmonising Uyghur consonants

Segments that participate in backness harmony are shown in Tables 1 and 2. The underlined vowels in Table 1 serve as harmony triggers (that is, they determine the backness of suffixes attached to roots containing them), while the non-underlined vowels are transparent to harmony. The harmonising consonants may also serve as harmony triggers, though they tend to be weaker than vowels. This paper

will focus primarily on harmony driven by vowels. In addition to serving as triggers of harmony, the harmonising vowels and consonants both emerge as the outcome of harmony in harmonising suffixes.

2.2 A description of Uyghur backness harmony

The examples of harmony below include 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]). I assume that /A/ is unspecified for the feature [back], /D/ for [voice], and /G/ for both (Archangeli, 1988). Voicing alternations in the initial segment are caused by voice assimilation, and are orthogonal to harmony.

The basic characterization of backness harmony is that suffixes must agree in backness with the rightmost front /y ø æ/ or back /u o ɑ/ harmonising root vowel.

(2) *Simple front harmonising forms*
tyr-dæ 'type-LOC'
pæn-lær 'science-PL'
munbær-gæ 'podium-DAT'

(3) *Simple back harmonising forms*
pul-ʁɑ 'money-DAT'
top-lɑr 'ball-PL'
ætrɑp-tɑ 'surroundings-LOC'

The vowels /i e/ are *transparent* to harmony. They do not serve as harmony triggers, but allow the harmonic value of preceding segments to "pass through" them.

(4) *Front roots with transparent vowels*
mæstʃit-tæ 'mosque-LOC'
ymid-lær 'hope-PL'
mømin-gæ 'believer-DAT'

(5) *Back roots with transparent vowels*
student-lɑr 'student-PL'
uniwersitet-tɑ 'university-LOC'
ɑmil-ʁɑ 'element-DAT'

Roots without any harmonising segments typically take back suffixes, but some take front suffixes (see McCollum, 2021; Mayer, Major, & Yakup, 2022).

(6) *Neutral roots that take back suffixes*
sir-lɑr 'secret-PL'
din-ʁɑ 'religion-DAT'
hejt-tɑ 'festival-LOC'
peʔil-lɑr 'verb-PL'
tip-qɑ 'type-DAT'

(7) *Neutral roots that take front suffixes*
biz-gæ 'us-DAT'
bilim-gæ 'knowledge-DAT'
welisipit-lær 'bicycle-PL'

Back suffixes appear to be the unmarked class in Uyghur. There has been a general diachronic shift in the population of neutral roots towards back suffixes (Lindblad, 1990), and recent loanwords that lack harmonising segments typically take back suffixes (Mayer, McCollum, & Eziz, 2022).

2.3 Vowel reduction

The second process that contributes to opacity in the Uyghur harmony system is *vowel reduction* or *raising*, which raises the low vowels /ɑ æ/ to [i] in medial open syllables in derived environments.[2]

(8) */ɑ/ vowel reduction*

bɑlɑ	'child'	bɑli-ʁɑ	'child-DAT'
ɑpɑ	'mom'	ɑpi-si	'mom-3.POSS'
ɑŋlɑ-ʃ	'listen-GER'	ɑŋli-ʁɑn	'listen-PFV'
qɑrɑ-ŋ	'look-IMP'	qɑri-di	'look-3.SG.PST'

(9) */æ/ vowel reduction*

ɑpæt	'disaster'	ɑpit-i	'disaster-3.POSS'
ætæ	'tomorrow'	æti-gæ	'tomorrow-DAT'
søzlæ-ŋ	'talk-IMP'	søzli-gæn	'talk-PFV'
kytʃæ-ʃ	'strive-GER'	kytʃi-di	'strive-3.SG.PST'

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, as in /tɑksi/ 'taxi' or /æsli/ 'origin'. Certain roots resist raising categorically, particularly loanwords where the final vowel was long in the source language (Nazarova & Niyaz, 2013); in the current paper we focus on roots that undergo raising.

2.4 Opaque interactions between backness harmony and vowel reduction

Vowel reduction has the potential to introduce opaque behavior into the vowel harmony system. Consider, for example, the root /ɑʁinæ/ 'friend'. The final vowel undergoes raising when it occurs in a derived, word-medial open syllable:

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

What happens when the vowel in the suffix must harmonise with the final vowel in the root, like in the form /ɑʁinæ-DA/ 'friend-LOC'? There are two possibilities: *opaque harmony* according to the underlying form of the root and *surface-apparent harmony* according to the raised form of the root. We will set aside for a moment the question of which of these we actually see in Uyghur, and briefly explore some theoretical consequences of each realization.

A rule ordering where harmony precedes raising predicts the opaque form [ɑʁinidæ]:

(11) *Harmony precedes raising*

UR	/ɑʁinæ-DA/
Harmony	ɑʁinæ-dæ
Raising	ɑʁini-dæ
SR	[ɑʁini-dæ]

This opacity is precisely the kind that classical OT has difficulty accounting for: there is an explicit markedness violation (failure to harmonise), with no apparent motivation (cf. forms like /tɑksi-DA/ → [tɑksidɑ] 'taxi-LOC').

If raising instead precedes backness harmony, we would expect the form [ɑʁini-dɑ] with surface-apparent harmony:

[2] An analysis of Uyghur vowel reduction is beyond the scope of this paper, but past OT analyses of derived environment effects in vowel reduction have relied on Comparative Markedness (Mascaró, 2009; Khanjian, 2009) or local constraint conjunction (Łubowicz, 2002): the latter paper mentions Uyghur raising in passing (fn. 20) but does not provide a full analysis. McCollum (2020) and Mayer (2021b, Appendix E) provide analyses of vowel reduction in Uyghur, but neither attempts to account for derived environment effects.

(12) *Raising precedes harmony*

UR	/ɑʁinæ-DA/
Raising	ɑʁini-DA
Harmony	ɑ̲ʁini-dɑ̲
SR	[ɑʁini-dɑ]

This outcome can be predicted by both serial and parallel models.

2.5 Modeling opacity in serial and parallel models

The kind of opacity shown in (11) is straightforward to represent in serial rule-based models: the rule that drives harmony is simply ordered before the rule that drives raising. An analysis under such a model is fundamentally identical to the derivation in (11).

This pattern poses challenges for an analysis in classical OT. I will assume a simple markedness constraint that motivates vowel harmony, which is a combination of the local and non-local AGREE constraints used by Hayes et al. (2009):

(13) VAGREE: harmonising segments in a suffix must match the backness of the rightmost harmonising vowel in the stem.

The following constraints will drive raising:

(14) *UNREDUCED: Don't have low vowels in derived medial open syllables.

(15) ID[HEIGHT]: Don't change the height of segments in the input.

*UNREDUCED is shorthand for a more detailed analysis of the pressures that drive vowel reduction (for vowel reduction in general, see Crosswhite, 2001; de Lacy, 2002; for vowel reduction in Uyghur, see McCollum, 2020; Mayer, 2021b, App. E).

When relevant I will employ a constraint that prevents specified [back] values from being altered.

(16) ID[BACK]: Don't change the backness of segments in the input.

ID[BACK] prevents underlyingly specified vowels in roots and certain harmony-blocking suffixes from being altered. This constraint is not violated when a segment underspecified for backness in the input is assigned a backness value in the output, nor is it violated when /æ ɑ/ are raised to [i]: assuming that [i] is unspecified for backness, these processes violate DEP and MAX constraints respectively, which are low-ranked and omitted from the tableaux below.

/ɑʔilæ-lAr/	*UNREDUCED	VAGREE	ID[HEIGHT]
☹ a. ɑʔili-lær		*!	*
💣 b. ɑʔili-lɑr			*
c. ɑʔilæ-lær	*!		
d. ɑʔilæ-lɑr	*!	*	

Table 3: Failed tableau for [ɑʔili-lær] 'family-PL'. The sad face indicates the candidate that should have won, and the bomb indicates that candidate that did win.

These constraints allow Classical OT to derive only surface harmony, as shown in Table 3. Suppose that we want to derive opaque harmony for the suffixed form /ɑʔilæ-lAr/ 'family-PL'. The desired candidate [ɑʔili-lær] is harmonically bounded by the winning *[ɑʔili-lɑr], and so will never be the optimal candidate under any ranking.

An analysis using Stratal OT succeeds in capturing this opacity (Kiparsky, 2000; Bermúdez-Otero, 2003, 2018). Stratal OT divides the grammar into several strata (e.g., the stem, the word, the phrase) and assigns each of these levels a separate OT grammar with differing constraint rankings. The outputs of lower strata serve as the inputs to higher strata. We can capture the opaque pattern here by proposing that vowel harmony occurs at a lower stratum than vowel reduction, as in Table 4.

Word stratum

/ɑʔilæ-lAr/	ID[BACK]	ID[HEIGHT]	VAGREE	*UNREDUCED
a. ɑʔili-lær		*!	*	
b. ɑʔili-lɑr		*!		
☞ c. ɑʔilæ-lær				*
d. ɑʔilæ-lɑr			*!	*

Phrase stratum

/ɑʔilæ-lær/	ID[BACK]	*UNREDUCED	VAGREE	ID[HEIGHT]
☞ a. ɑʔili-lær			*	*
b. ɑʔili-lɑr	*!			*
c. ɑʔilæ-lær		*!		
d. ɑʔilæ-lɑr	*!	*	*	

Table 4: Tableaux for the derivation of [aʔili-lær] 'family-PL' at the word stratum (top) and phrase stratum (bottom). At the word stratum, the constraint driving raising is ranked below its corresponding faithfulness constraint, meaning harmony applies but raising does not. The output from the word stratum serves as the input to the tableau for the phrase stratum. At this stratum, the constraint driving raising is now ranked above its corresponding faithfulness constraint, meaning raising can apply.

I employ this formalism here because it is widely used in the contemporary literature, and because there is evidence in Uyghur that raising can apply at the level of the phrase, suggesting that it belongs to a higher stratum than backness harmony (specifically, it appears to be a post-lexical process; Kiparsky, 1982). For example, in phrases like <*Adil Hesenge berdi*> 'Adil gave it to Hesen', the dative <*-ge*> [-gæ] may raise to <*-gi*> [-gi] in rapid speech (Hahn, 1991b, p. 53).

To summarize, rule-based analyses predict opaque harmony straightforwardly, while strictly parallel analyses predict only surface-apparent harmony. Modifications to strictly parallel models that incorporate some degree of serialism, such as those listed in Section 2.1, also predict opaque harmony, though they differ in their attribution of the particular mechanism responsible for it. For example, while Stratal OT captures opaque patterns by positing multiple derivational strata with different constraint orderings, candidate chain theory (McCarthy, 2007) does so by evaluating candidate chains (roughly analogous to derivations) rather than candidates.

2.6 Past work on opacity in Uyghur

Which of these patterns do we observe in Uyghur? Pedagogical materials do not generally discuss these cases in any detail, since roots that can generate opaque harmony are a relatively small slice of the lexicon. Those that do suggest that opacity is the correct outcome (e.g., Hahn, 1991b, Section 4.3.5). Hahn describes this in terms of roots falling into a particular 'harmonic category', with vowel reduction processes 'disguising' the most salient clue to this category: the final vowel of the root.

As is typical for opaque phenemona (Kiparsky, 1971, 1973; Mielke et al., 2003), the rule ordering that produces opaque harmony reflects the relative diachronic development of each process. Backness harmony is an ancient property of Turkic languages (e.g., Clauson, 1972), while raising is a newer phenomenon in Uyghur: Chagatay, the closest direct ancestor to Uyghur, appears to have had no such rais-

ing process (Bodrogligeti, 2001). Opaque harmony thus maintains historical patterns of root backness at the cost of surface disharmony.

The exception to this is that certain derivational suffixes in Uyghur, such as the diminutive /-ʧæ/ and the adjectival suffix /-ɑnæ/ have been described as triggering surface-apparent harmony in raised forms (Hahn, 1991b; Vaux, 2000; Halle et al., 2000; Hall & Ozburn, 2018).[3]

(17) *Surface-apparent harmony in raised* /-ʧæ/

	/bɑʁ-ʧæ-DA/	→	[bɑ̱ʁʧidɑ̱]	'park-DIM-LOC'
cf.	/bɑʁ-ʧæ-m-DA/	→	[bɑʁʧæ̱mdæ̱]	'park-DIM-1.SG.POSS-LOC'

We will return to these suffixes below.

Uyghur speakers I have worked with agree that opaque harmony is the correct outcome. However, in addition to the suffixes above, speakers have identified certain forms where surface-apparent harmony is mandatory (e.g., /ærzɑn-i-GA/ → [ærzinigæ] 'cheap (ones)-3.POSS-DAT') or both surface-apparent and opaque harmony are acceptable (e.g., /æzɑn-i-GA/ → [æziniʁɑ]/[æzinigæ] 'call to prayer-3.POSS-DAT'). These forms demonstrate that although opaque harmony occurs in the vast majority of cases, we cannot always predict whether a root or stem will trigger surface-apparent or opaque harmony. These few elicited observations of variability were one of the motivations for the corpus study presented below.

Theoretical work on the interaction between vowel reduction and harmony has claimed that there is an asymmetry between vowels (Vaux, 2000; Halle et al., 2000; Hall & Ozburn, 2018): raised /æ/ is opaque and continues to behave as a front vowel trigger (with the suffixes above constituting notable exceptions), while raised /ɑ/ is transparent, behaving identically to underlying /i/. However, these claims have been based off only eight data points collected from a single speaker, and the empirical validity of these data is unclear (see Mayer, 2021b, Section 3.3.7). One of the main goals of this paper is to put claims about opacity in Uyghur on stronger empirical footing.

In addition to being an important empirical question, obtaining a better understanding of this pattern is valuable from a theoretical perspective: opaque patterns such as the one in (11) are not predicted to exist by many strictly parallel phonological models, and, indeed, the Uyghur pattern has been used to argue in favor of serial models (Vaux, 2000). The remainder of this paper will present a large-scale corpus study that examines the empirical facts of opacity in Uyghur and explore some of the theoretical implications of its results.

3 A corpus study of opacity in Uyghur backness harmony

In order to investigate the interaction of vowel reduction and backness harmony, I performed a corpus study using three large text corpora.[4]

Uyghur uses a number of different orthographies depending on where it is written: Perso-Arabic, Cyrillic, or Latin. In each of these, the alternations conditioned by the raising and harmony processes are represented orthographically.[5] Hence text corpora allow us to gather large-scale empirical data on their interaction.

The first corpus was generated from the Radio Free Asia (RFA; `https://www.rfa.org/uyghur/`) Uyghur language website. RFA is a US-sponsored non-profit news organization. The second was gen-

[3] /-ʧæ/-final stems are sometimes repaired to be internally harmonic: e.g., the root /bɑʁʧæ/ 'park' (lit. 'orchard-DIM') is frequently produced as [bɑʁʧɑ].

[4] The code used in this paper can be found at `https://github.com/connormayer/uyghur_corpora`.

[5] There is one specific exception to this: the editorial guidelines of Radio Free Asia, one of the websites from which the corpora were generated, require proper names to be written in their unraised forms, even when they undergo raising. For example, the word for 'American' is written as <*Amérikaliq*> even though it is pronounced [ɑmeriki̱-liq]. This results in a slight undercounting of raised forms in the following sections, but this does not affect the conclusions of this paper.

erated from the website of *Uyghur Awazi* (Uyghur Voice; `http://uyguravazi.kazgazeta.kz/`), an Uyghur-language newspaper published in Almaty, Kazakhstan. The third was generated from *Uyghur Akadémiyisi* (Uyghur Academy; `https://www.akademiye.org/ug/`), a legal research organization that publishes articles on Uyghur culture and politics.

Corpora were generated from the websites using *web scrapers*: 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 allow corpora to be generated from publicly available internet resources, in formats that are useful to researchers.

There are separate web scrapers for each of the websites, which are linked in the GitHub repository for this paper. These scrapers were written by undergraduate research assistants at UCLA and UCI in collaboration with the author.

A summary of the contents of each corpus is shown in Table 5.

Corpus	Article count	Word count	Date retrieved	Orthography
Uyghur Voice	24,080	4,197,550	January 2020	Latin/Cyrillic
RFA	45,777	7,986,484	July 2022	Perso-Arabic
Uyghur Academy	2,635	2,423,519	July 2022	Perso-Arabic

Table 5: Summary of corpora.[6]

In addition to the contents of each article, the scrapers retrieved the author, the date, and the URL.

3.1 Parsing the corpora

In order to extract information about the interaction between backness harmony and vowel reduction from the corpus, 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., 2019). This transducer is part of Apertium, a free and open-source rule-based machine translation platform (`https://www.apertium.org`).

The transducer maps from surface forms to underlying analyses that consist of roots plus morphological tags indicating the backness of any harmonising suffixes. For example, if the input is the surface form *qizingizgha* "to your daughter" the output analysis will be *qiz<n><px2sg><frm><dat-b>*. This indicates that the root is *qiz*, a noun <n>, and is suffixed with the 2nd person singular possessive marker in its formal form[7] *-ingiz* <px2sg><frm>, followed by the dative suffix in its back form *-gha* <dat-b>.

The output of the transducer was used to count the frequency of front or back suffixes for each root. Words for which the transducer was unable to produce a valid parse were excluded from analysis. Simple text processing comparing the parsed root and surface forms was used to detect whether vowel reduction occurred and to extract phonotactic properties of the roots and tokens.

Additional details of the transducer and data processing, including numerical validation, are presented in Appendix B.

[6]The earlier scrape date for the Uyghur Awazi corpus is due to a redesign that changed the structure of the site and pages, which requires modifications to the scraper. It was not possible to make these modifications in time for the submission of this paper.

[7]Uyghur has formal and informal versions of many 2nd person morphemes. The corresponding informal version of the 2nd person singular possessive marker in the above example would be *-ing*.

3.2 Quantitative results

3.2.1 Comparing harmonic and disharmonic roots

In this section I consider only tokens where (a) the rightmost two harmonising elements of the root are vowels; (b) the underlying final vowel in the root is either /æ/ or /ɑ/; (c) the final vowel undergoes raising; and (d) the raised vowel is followed by at least one harmonising suffix. The rightmost two harmonising vowels were chosen as the domain of analysis because the rightmost harmonising vowel in a stem almost invariably predicts suffix backness (see Mayer, 2021b, Ch. 4). When this vowel is reduced to transparent [i], the second-rightmost harmonising vowel has the potential to influence suffix backness. The effect of any preceding harmonising vowels on suffix backness appears to be negligible.

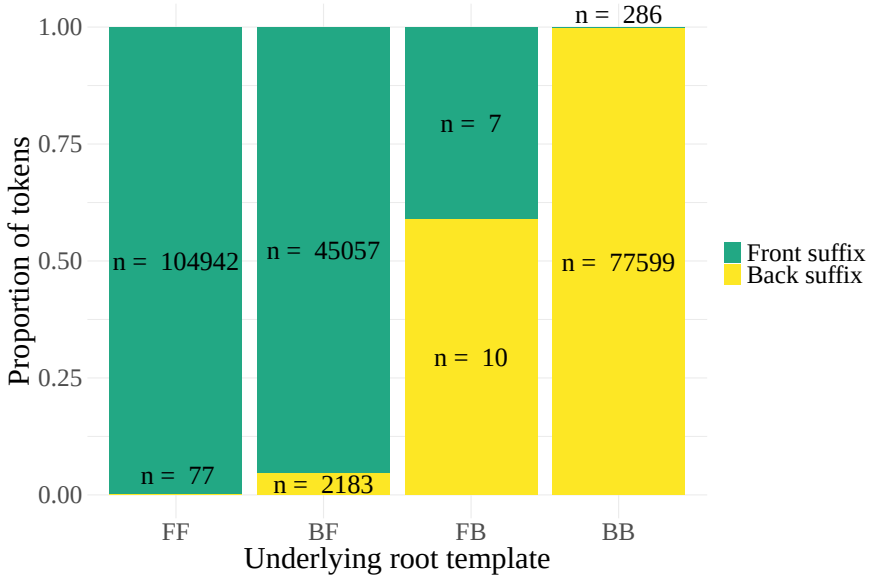

Figure 1: Suffix harmony choice in tokens where the final root vowel raises, broken down by root class. Token counts are overlaid on each category.

Fig. 1 breaks down parsed roots into four classes according to their rightmost two harmonising vowels: two back vowels (BB; 774 roots; e.g., /bɑlɑ/ 'child', /mɑʃinɑ/ 'vehicle, machine'), a back vowel followed by a front vowel (BF; 311 roots; e.g., /ɑdæt/ 'custom', /ɑʔilæ/ 'family'), a front vowel followed by a back vowel (FB; 7 roots; e.g., /ærzɑn/ 'cheap', /kæsipdɑʃ/ 'colleague'), and two front vowels (FF; 528 roots; e.g., /sypæt/ 'quality', /mæsilæ/ 'problem'). The BF and FB classes have the potential to produce opaque harmony.

The FB class has very few roots and tokens compared to the others. Roots with this shape are relatively uncommon, and those that exist tend not to undergo raising.[8] Fig. 2 breaks down BF roots and FB roots by their individual rates of front vs. back suffixes. In both cases we see that roots are typically categorical in whether they take back or front suffixes, while a smaller number (n = 101) show variation between the two (see Zuraw, 2016).

For example of such variation, consider the root <*idare*> /idɑræ/ 'office, bureau'. When used with the auxiliary verb <*qilmaq*> 'do', it can also mean 'to rule' or 'to govern'. This root has an overall frequency in the corpora of 1,122/million words and occurs in its raised form <*idari*> [idɑri] in 74% of tokens. The

[8]Three of these raising FB roots end in the derivational suffix /-dɑʃ/ '-mate', as in [xizmætdɑʃ] 'officemate'. This suffix may display idiosyncractic harmonising behavior in a similar way to the suffixes discussed in the next section, but there is insufficient data to determine this.

high frequency of the raised allomorph is likely due to the contexts in which this word tends to be used: since it's common to talk about an office relating to a person or entity, the root is often realized with the 3.POSS suffix [-si], as in <*ürümchi sayaset* ***idarisi***> 'Ürümchi Tourism Office'. When it occurs in its raised form with a harmonizing suffix attached, it displays opaque harmony in about 89% of cases (913 tokens) and surface-apparent harmony in about 11% of cases (113 tokens). The examples below show tokens of /idɑræ/ from the RFA corpus in its unsuffixed form (18), with opaque harmony (19), and with surface-apparent harmony (20).

(18) Unsuffixed token of 'idare'
<*... döletni qanun arqiliq* ***idare*** *qilish...*>
'... the rule of law..." (literally 'ruling the country by means of the law')

(19) Opaque token of raised 'idare'
<*1980-yillardin boyan merkiziy axbarat* ***idariside*** *ishligen.*>
"He has worked at the Central Intelligence Agency since the 1980s."

(20) Surface-apparent token of raised 'idare'
<*Gülnar xanim saqchi* ***idarisida*** *qandaq mu'amilige uchridi?*>
"What kind of treatment did Gülnar receive at the police station?"

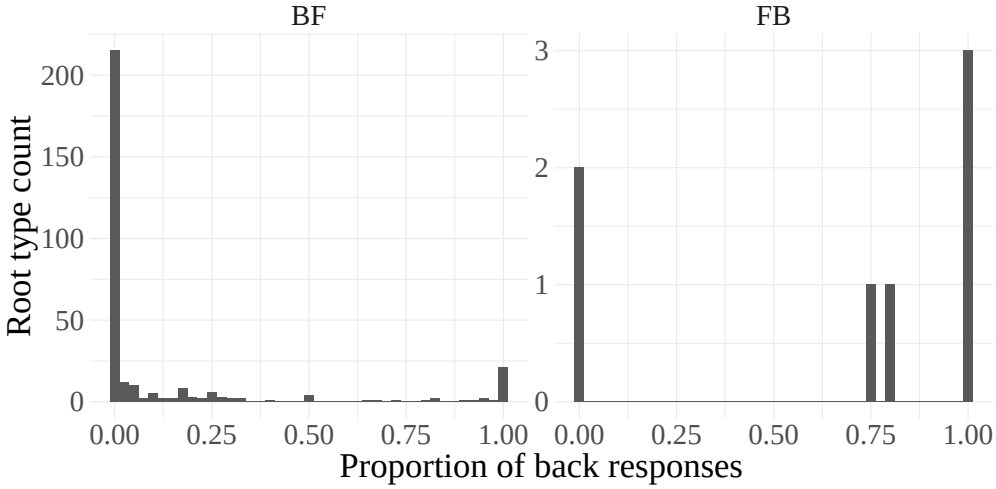

Figure 2: Histograms showing the distribution of rates of back suffix application in BF and FB roots. Note that for raised BF roots a back suffix constitutes surface-apparent harmony and a front suffix constitutes opaque harmony, while for raised FB roots it is the opposite.

Because the final two vowels in the BB and FF roots agree in backness, opaque harmony and surface harmony predict the same surface form. These roots almost categorically take the expected suffix forms.[9] The disharmonic FB roots and BF roots behave similarly to BB and FF roots, respectively, but both show higher rates of surface-apparent harmony. Chi-squared tests show significantly different rates of back suffix choice between BB and FB roots ($\chi^2 = 650.47$; df = 1; $p < 0.0001$) and between FF and BF roots ($\chi^2 = 4597.6$; df = 1; $p < 0.0001$). Thus the quality of the harmonising vowel preceding the raised vowel, and not just the underlying quality of the raised vowel, affects suffix choice: when the backness of the preceding vowel conflicts with the backness of the raised vowel, the suffix becomes more likely to agree with the preceding vowel.

[9]The small number of unexpected disharmonic suffixes could be the result of typos or misidentification of the quality of the underlying vowel.

3.2.2 Opacity in derivational suffixes

Recall that previous work on opacity in Uyghur has suggested that certain derivational suffixes like the dimunitive /-ʧæ/ and the adjectival suffix /-ɑnæ/ behave idiosyncratically, preferring surface-apparent harmony. Manual inspection of the corpus data revealed a similar pattern for the suffix /-nɑmæ/, meaning 'writings of' (e.g., /bɑburnɑmæ/ "The writings of Babur"). Fig. 3 breaks down harmony rates in the set of BF roots according to suffix.[10] The suffix /-ʧæ/ rarely displays opaque harmony, aligning with observations made in the literature. /-ɑnæ/ and /-nɑmæ/ display higher rates of opaque harmony, but not as high as the general population of BF roots.

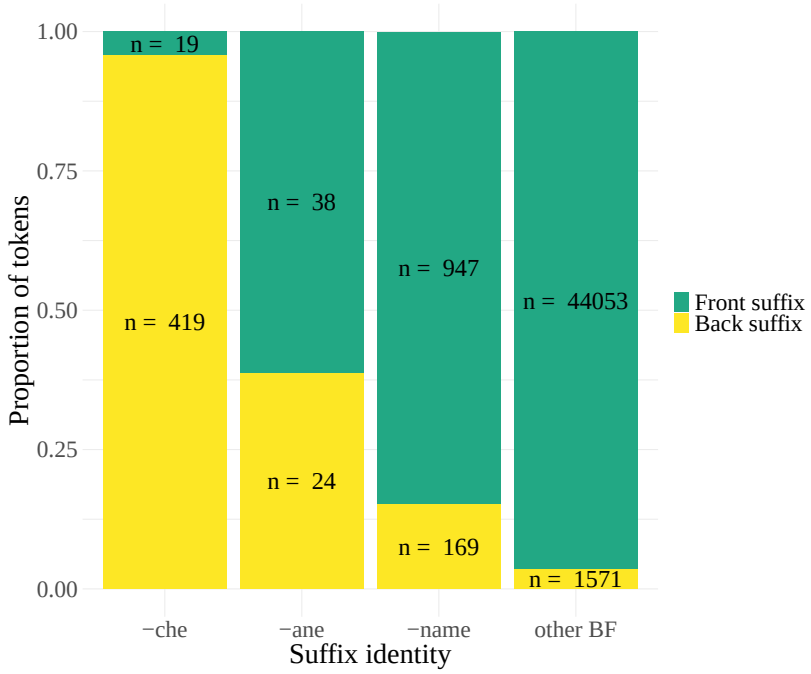

Figure 3: Suffix choice in raised BF roots broken down by root-final derivational suffix. 'Other BF' refers to BF roots that do not end in one of the three derivational suffixes. Token counts are overlaid on each category. The tokens of /-ʧæ/ included here all have a preceding B vowel, as in /bɑʁʧæ/ 'park'.

3.3 Predicting opacity

In order to identify potential factors that contribute to rates of opacity, I fit a Bayesian mixed-effects logistic regression model to the set of raised BF and FB tokens with at least one harmonising suffix attached (for discussion of the use of logistic regression in modeling categorical corpus data, see Speelman, 2014). This is a proper subset of the tokens described in the previous section, omitting the BB and FF roots. The model was fit in R (R Core Team, 2017) using the `brms` package (Bürkner, 2017).

Bayesian models treat the statistical parameters of the model as random variables and use sampling techniques to estimate the posterior distribution over parameter values given the observed data and prior beliefs about credible parameter values (see, e.g., Kruschke, 2014; Nicenboim & Vasishth, 2016). The model here uses the default, weakly-informative priors.

A Bayesian analysis is used for two reasons. First, the equivalent frequentist model often fails to converge with the random effects structure used here. Second, large sample sizes can lead to significant

[10] The transducer typically includes these suffixes as part of the root.

p-values for trivially small effects (e.g., Lin et al., 2013). I focus accordingly on reporting effect sizes: that is, the values of the coefficients of the fitted model. In addition to point estimates of these coefficients, Bayesian models also provide interpretable estimates of the range of credible values.[11]

The dependent variable was coded as either opaque harmony (1) or surface-apparent harmony (0). An opaque response was defined as a back suffix attached to a raised FB root, or a front suffix attached to a raised BF root.

The independent variables were selected based on previous work on Uyghur or other languages with similar vowel harmony systems:

- The **log token frequency of the root** in the corpora, normalized to count per million words. This was included as a predictor because frequency is often an important driver of phonological variability (e.g., Coetzee & Kawahara, 2012; Coetzee, 2016).

- The **proportion of tokens of the root that are raised**. This is defined as the number of tokens of a root containing the raised allomorph divided by the total number of tokens of that root.

  For example, the root /ɑpæt/ 'disaster' occurs 1719 times in the corpus. Of these tokens, 544 are in forms that exhibit raising (e.g., [ɑpit-i] 'disaster-3.POSS') and 1175 are in unraised forms (e.g., [ɑpæt-lær] 'disaster-PL'). Thus the proportion of raised tokens for this root is 544/1719 = 0.32.

  This variable was included based on the observation by Hahn (1991b) that raised forms obscure the harmonic class of a root (i.e., whether it takes front or back suffixes). That is, unraised tokens provide evidence of the harmonic class of the root via the identity of the final vowel, while raised tokens (particularly those with non-harmonising suffixes) do not.

- The **identity of the underlying raised vowel** (F or B). This allows us to test the proposal that raised /ɑ/ is more likely to display surface-apparent harmony than raised /æ/ (Vaux, 2000).

- The **distance between the rightmost two harmonising root vowels**, counted in segments. Previous work on Uyghur (Mayer, 2021a) and languages with similar harmony systems (Hayes et al., 2009; Rebrus & Törkenczy, 2017, 2021) suggest that the influence of a vocalic harmony trigger in the root decreases as greater numbers of transparent segments intervene between it and the following suffix. In a root like /ɑpæt/ 'disaster' this distance is 1 while in /pɑʔɑlijæt/ 'activity' it is 3.

- The **distance between the root and the first harmonising suffix**. This is calculated as the number of morphological tags between the final tag of the root and the tag of the first harmonising suffix. For example, /ɑpæt-i-GA/ 'disaster-3.POSS-DAT' has a distance of 1 (the non-harmonising 3.POSS intervenes between the root and the harmonising DAT) while /ɑpæt-GA/ 'disaster-DAT' has a distance of 0. Rebrus and Törkenczy (2017, 2021) found no influence of this factor on harmonising behavior in Hungarian.

- Whether the root ends in one of the three **derivational suffixes** discussed above. This was operationalized as three dummy-coded variables that took the value 1 if the root ends in /-ʧæ/, /-ɑnæ/, and /-nɑmæ/, respectively, and 0 if not.

Random intercepts were defined for:

[11] While confidence intervals provide a measure of dispersion for coefficient estimates in frequentist models, their interpretation is counterintuitive and frequently misunderstood (Hoekstra et al., 2014). The credible intervals produced by Bayesian models better align with intuitions about what such measures of dispersion should communicate.

- A nested effect of **author** within **corpus**. This controls for different rates of opacity across sources and individual writers.[12] A nested effect is used because authors are unique within corpora.
- **Root identity**. This controls for the idiosyncratic tendencies of roots that are not captured by the dependent variables.

Term	Mean estimate	95% CI
Intercept	3.46	[−0.89, 6.36]
*Log token count	0.32	[0.11, 0.51]
*Proportion raised	−2.25	[−3.77, −0.84]
*Final vowel (reference level BF)	−3.65	[−6.36, −1.02]
*Root vowel distance	0.80	[0.41, 1.23]
*Suffix distance	−1.29	[−1.45, −1.13]
*Has -ɑnæ	−4.62	[−6.77, −2.57]
*Has -nɑmæ	−2.12	[−3.42, −0.88]
*Has -ʧæ	−9.58	[−11.55, −7.73]
Corpus (standard deviation)	1.89	[0.16, 5.86]
Corpus:Author (standard deviation)	0.90	[0.64, 1.23]
Root (standard deviation)	2.36	[1.99, 2.78]

Table 6: Results from a mixed-effects logistic regression model whose coefficients were estimated using Bayesian inference. The 95% Credible Interval shows the central range in which 95% of the sample values occur. Credible intervals that does not contain zero are interpreted as a meaningful directional effect, and are marked with a '*' in the table.

The results shown in Table 6 show two common summary statistics for the coefficients[13] in Bayesian models: the mean, or most credible, value of each coefficient, and the 95% Credible Interval (95% CI): the range in which the central 95% of the sampled coefficient values fall given the model and dataset. Expressed in slightly different terms, the 95% CI tells us that the model estimates that there is a 95% chance that the true value of this parameter falls within this range, given the data. A 95% CI that does not include 0 is interpreted as meaningful, since it indicates that the directionality of the effect is highly credible. In addition, the 95% CI provides a measurement of uncertainty about the effect size. See Appendix C for more detail on the model.

These results suggest several frequency-related influences on opacity rates: more frequent roots are more likely to harmonise opaquely, but roots that frequently occur in their raised forms, where the underlying identity of the final vowel is obscured, are less likely to harmonise opaquely.

There are also phonological contributions. The model shows that underlying identity of the raised vowel is a significant predictor of opaque harmony, with underlying back vowels being less likely to harmonise opaquely (roughly aligning with the claims in Vaux, 2000). The distance between the rightmost

[12] Articles in the *Awazi* corpus are always published under the anonymous byline </*admin*>. The *Akadémiye* corpus is similar, with a small number of exceptions, resulting in a total of four unique author bylines. The RFA corpus provides greater attribution, but commonly uses partially anonymous bylines, such as <*muxbirimiz Erkin*> 'our reporter Erkin', resulting in a total of 154 unique author bylines. These conventions make consistent authorship attribution difficult. When coding the authorship variable used in the mixed effects logistic regression model, I use the listed author name when available. If no attribution is provided, the author is coded as 'None_<corpus_name>' to distinguish between unnamed authors in the different corpora.

[13] The coefficients here are the estimated change in the log odds of opaque harmony given a unit increase in the corresponding variable, where the log odds are defined as $\log \frac{P(\text{opaque harmony})}{1-P(\text{opaque harmony})}$. Positive values indicate an increased likelihood of opaque harmony, while negative values indicate a decreased likelihood. For example, the coefficient 0.32 for the log token count tells us that, for each unit increase in log frequency, the log odds of an opaque response relative to a surface-apparent response increase by 0.32, which corresponds to a 37% increase in the odds.

two harmonising vowels in the root is also positively correlated with rates of opacity: as the disharmonic vowel in the root becomes further from the suffix, its influence decreases.

Finally, there is also evidence for morphological influences on opacity rates. The three derivational suffixes described above each produce greater rates of surface-apparent harmony than seen in the general population of roots, and the specific rates vary between suffixes.

The number of non-harmonic suffixes directly between a root and the closest harmonic suffix was correlated with more surface-apparent harmony, which is unexpected given results from Rebrus and Törkenczy (2017, 2021). Because the majority of roots that can generate opacity are BF roots, this tendency might reflect an increased preference for the use of the default back suffix forms as distance between root and harmonising suffix increases.

4 An analysis of opacity in the corpus

Given the quantitative results described above, a descriptively adequate (Chomsky, 1965) model of opacity in Uyghur should be able to account for the following properties:

- The majority of raised tokens harmonise opaquely, but cases of surface-apparent harmony also exist.
- The rate of opaque harmony varies across roots. This variation is correlated with a number of phonological and morphological factors, as well as frequency.

In this section I will outline a simple parallel model that can account for these factors.

4.1 Phonological and lexical effects in backness harmony

Backness harmony is typically treated as driven by surface phonological constraints (e.g., van der Hulst, 2016): whether a root takes front or back suffix allomorphs depends on which variant will minimize resulting surface disharmony by some criteria, as in the simple model presented in Section 2.5. I will refer to this as the *phonological component* of backness harmony.

There are cases in Uyghur, however, where surface phonological properties are not sufficient to determine harmonising behavior (see Mayer, McCollum, & Eziz, 2022, for a more detailed description). For example, the pair of words /sir/ 'secret' and /bir/ 'one' are nearly identical: however, the former takes back suffixes (e.g., [sir-lɑr] 'secret-PL'), while the latter takes front suffixes (e.g., [bir-gæ] 'one-DAT'). Whether such roots take front or back suffixes cannot be predicted from their acoustic properties (Mayer, Major, & Yakup, 2022). Similarly, while the majority of roots that contain no harmonising vowels and a velar consonant /k ɡ/ take front suffixes (e.g., [kir-gæ] 'dirt-DAT'), a smaller number take back suffixes (e.g., [ɡips-qɑ] 'plaster-DAT').

Hayes (2016) uses the term *zones of variation* to describe similar roots in Hungarian. Because their harmonic class is at best partially predictable from phonological properties, speakers disproportionately rely on lexical knowledge: that /sir/ takes back suffixes while /bir/ takes front suffixes must simply be memorized as a fact about each root. I will refer to this as the *lexical component* of backness harmony. A consequence is that such roots typically display higher degrees of variability in suffix choice, particularly in wug tests where lexical knowledge is absent.

The effect of lexical information on backness harmony systems has been analyzed on the basis of *Harmonic Uniformity* (Rebrus & Törkenczy, 2017, 2021; Rebrus et al., 2023). This is a paradigm uniformity constraint (e.g., Steriade, 2000) which requires that the harmonic class of a root remain consistent across its extended paradigm. This constraint can override phonological processes that might otherwise

apply. Rebrus and colleagues have provided evidence for this constraint on the basis of a variety of phenomena in the Hungarian backness harmony system. Here we will focus on one particular consequence of Harmonic Uniformity: the harmonic class of a root should be consistent across its suffixed forms. That is, the same root should not take back forms of one harmonizing suffix and front forms of another.

Speakers simultaneously learn the lexical and phonological components of phonological systems (e.g., Zuraw, 2000, 2010). In the vast majority of cases in Uyghur and Hungarian backness harmony, these components favor the same suffix choices. In a smaller number of cases, such as the zones of variation described above, the phonological component is less informative and lexical knowledge plays a larger role. The opaque forms discussed in this paper are cases where the phonological and lexical components actively *conflict*: in a case like /ɑpæt-i-GA/ 'disaster-3.POSS-DAT', the phonological component favors [ɑpitiʁɑ] because it displays surface-apparent harmony, while the lexical component favors the disharmonic [ɑpitigæ] because it is consistent with the paradigmatic harmonising behavior of /ɑpæt/. In the following sections, I will demonstrate how Harmonic Uniformity can be used to predict the variable rates of opacity found in Uyghur.

4.2 Modeling gradience in opacity using Harmonic Uniformity

The models described below will use Maximum Entropy Harmonic Grammar (henceforth MaxEnt; Goldwater & Johnson, 2003), a generalization of Optimality Theory with numeric constraint weights (Pater, 2009). Higher weights indicate a greater penalty for constraint violation. MaxEnt uses these weights and violation profiles to compute probability distributions over output candidates. See Appendix D for more detail.

The phonological component of backness harmony will be modeled using variants of the simple VAGREE constraint introduced in Section 2.5. The constraints *UNRAISED and ID[HEIGHT], described in the same section, will be used to model vowel raising.

Lexical knowledge about the harmonic class of individual roots is modeled using the HARMONICUNIFORMITY constraint:

(21) HARMONICUNIFORMITY: the backness of a harmonising suffix must be identical to the harmonic class of the root.

I make a minor theoretical innovation here to allow this constraint to generate the variability in rates of opacity seen in Uyghur. Rebrus and collaborators divide roots into three harmonic classes: front roots, which consistently take front suffixes; back roots, which consistently take back suffixes; and vacillators, which take either. While this tripartite distinction is useful for capturing backness harmony patterns in broad strokes, it does not provide a mechanism to predict root-specific variation.

Instead, I propose that the violations of HARMONICUNIFORMITY are *scaled based on certainty in the harmonic class of a root*. That is, for roots where the harmonic class is certain, attaching a suffix that conflicts with its harmonic class will incur a large violation of HARMONICUNIFORMITY, while attaching a suffix that agrees with it will incur no penalty. For roots whose harmonic class is uncertain, violations of HARMONICUNIFORMITY will be similar between front and back suffixes, and thus phonological factors will play a greater role in deciding suffix backness. Under this conception, vacillating roots are those where certainty in the harmonic class is low.

It is natural to think of certainty in terms of a probability distribution over the harmonic classes FRONT and BACK given a root x. I notate this distribution as $P(\text{HC}|x)$, and occasionally use the abbreviated form $P(\text{HC})$ when the identity of the root is clear from context. Roots that categorically take front suffixes will have $P(\text{HC} = \text{FRONT}|x) \approx 1$ (and accordingly $P(\text{HC} = \text{BACK}|x) \approx 0$), roots that categorically take back suffixes will have $P(\text{HC} = \text{BACK}|x) \approx 1$ (and accordingly $P(\text{HC} = \text{FRONT}|x) \approx 0$), and maximally ambiguous roots will have $P(\text{HC} = \text{FRONT}|x) \approx P(\text{HC} = \text{BACK}|x)$. Concretely, both front and back suffix forms will

violate HARMONICUNIFORMITY, but the violations of each are scaled by certainty that the root falls into the opposite harmonic class: violations of front suffixed forms are scaled by $P(\text{HC} = \text{BACK})$ and violations of back suffixed forms are scaled by $P(\text{HC} = \text{FRONT})$.

Let's look at an example. Suppose the root /ɑʔilæ/ 'family' has $P(\text{HC} = \text{FRONT}) = 0.99$ and the root /ɑhɑlæ/ 'resident' has $P(\text{HC} = \text{FRONT}) = 0.7$. Tables 7 and 8 show the output of a simple MaxEnt model fit only to these data points, demonstrating how scaling the violations of HARMONICUNIFORMITY by these probabilities produces variability in whether harmony is opaque or surface-apparent. For /ɑʔilæ/, the certainty of harmonic class is so great that it overrides the violation of surface harmony. For /ɑhɑlæ/, where there is less certainty in class membership, we see variability.

/ɑʔilæ-lAr/	Obs. Freq.	Pred. Prob.	H	VAGREE $w = 11$	HARMONICUNIFORMITY $w = 31$
ɑʔili-lær	1	≈ 1	11.3	1	$P(\text{HC} = \text{BACK}) = 0.01$
ɑʔili-lɑr	0	≈ 0	30.7	0	$P(\text{HC} = \text{FRONT}) = 0.99$

Table 7: Tableau for the consistently opaque form /ɑʔilæ-lAr/ 'family-PL'.

/ɑhɑlæ-lAr/	Obs. Prob.	Pred. Prob.	H	VAGREE $w = 11$	HARMONICUNIFORMITY $w = 31$
ɑhɑli-lær	0.787	0.797	20.3	1	$P(\text{HC} = \text{BACK}) = 0.3$
ɑhɑli-lɑr	0.213	0.203	21.7	0	$P(\text{HC} = \text{FRONT}) = 0.7$

Table 8: Tableau for the variably opaque form /ɑhɑlæ-lAr/ 'resident-PL'.

Note that if $P(\text{HC} = \text{FRONT}) = P(\text{HC} = \text{BACK})$ or if the weight of HARMONICUNIFORMITY were 0, the output would be entirely determined by violations of VAGREE; conversely, if the weight of VAGREE were 0, the output would be determined entirely by lexical knowledge, and the predicted probability of [ɑhɑli-lær] in Table 8 would be higher.

4.3 Calculating $P(\textbf{HC}|x)$

The previous section showed how varying degrees of certainty in the harmonic class produce different rates of opacity across roots. We now turn to the question of how we determine $P(\text{HC}|x)$. That is, what properties are speakers sensitive to when determining the harmonic class of a root?

I propose a simple and rather coarse model that encodes some of the factors that may determine certainty in the harmonic class of a root based on the results of the corpus study. The key piece of evidence is, of course, the **distribution** of the root: that is, do we typically see this root with front suffixes or back suffixes? This can be thought of as the driving factor behind lexical knowledge of root harmonising class.

In addition, however, there are general properties of roots that fall into each harmonic class that can be used to infer the harmonic class of a root, even in the absence of clear distributional evidence: namely, the phonotactic properties of the root and its morphological composition, both mediated by frequency. These factors are enumerated below:

- In addition to suffixed forms, evidence for harmonic class comes from the **phonotactic** properties of the root. Some phonotactic properties of roots are highly predictive of harmonic class: if a root ends in a back vowel, you can be quite certain that it will belong to the class of back harmonisers, even if you have never encountered a suffixed form. Other properties are more weakly predictive, such as the presence of harmonising consonants in the root, as described briefly in Section 4.1.

- The **morphological composition** of the root is also important: certain derivational suffixes are more prone to surface-apparent harmony than others. This may relate to whether these derived forms are treated as roots in their own right, in which case opaque harmony might be expected, or as roots with disharmonic suffixes, in which case surface-apparent harmony may be preferred.

- **Prior biases**: Back harmony is the default class in Uyghur, and speakers may encode an overall preference for this class (Mayer, 2021a).

The frequency-based effects observed in the corpus study connect to each of these factors: frequent exposure to a root provides greater evidence of which suffixes it takes, as well as greater knowledge of its phonotactic properties; roots that typically show up in raised forms (particularly with non-harmonising suffixes) do not provide as much exposure to their final vowel, and accordingly are more prone to surface-true harmony; and the relative frequency of root and derived forms has been shown to predict morphological decomposability (e.g., Hay, 2001), although I do not pursue this idea further in this paper. Thus frequency plays an important role in this model, similar to other models of phonological variability (e.g., Coetzee & Kawahara, 2012; Coetzee, 2016).

The next section will present a modeling study that validates the claims made above.

4.4 Validating the model

To validate this proposal, I fit six simple MaxEnt models to the set of tokens from the corpora of roots whose final two harmonising segments were BB, BF, FB, and FF and which had at least one harmonising suffix. Note that this is a broader set of tokens than used in the statistical analysis in Section 3: it includes tokens in contexts that do not produce raising (such as /ɑpæt-lAr/ → [ɑpætlær] 'disaster-PL', tokens that categorically fail to raise in typical raising contexts (such as /dunjɑ-si/ → [dunjɑsi] 'world-3.POS'), and roots that are structurally ineligible for raising in any suffixed form (such as /nɑmærd/ 'disloyal', which will never raise due to its final complex coda, or /muʔællim/ 'teacher', which will never raise because its final harmonising vowel is not in the final syllable). This data set consisted of a total of 767,761 tokens. For simplicity, tokens for each root were aggregated based on whether they had front or back suffixes. This means the models do not consider the identity or number of suffixes, merely their backness.

The rationale behind choosing these particular six models is to deconstruct the various potential influences on backness harmony and gauge which factors play the greatest role in predicting suffix choice. The models are:

1. A **surface-oriented** model of harmony which contains the VAGREE constraint. This constraint mandates surface-apparent harmony. This model is only sensitive to the identity of the final surface harmonic vowel when determining suffix choice.

2. An **input-oriented** model that contains the VAGREEUNDERLYING constraint. This constraint mandates harmony with the underlying form of the final vowel. This model is only sensitive to the identity of the final underlying harmonic vowel when determining suffix choice.

3. A **lexical** model. This contains the HARMONICUNIFORMITY constraint defined above, with violations scaled according to $P(\text{HC}|x)$ for each root. This model is only sensitive to certainty in lexical harmonic class when determining suffix choice.

4. An **input-surface** model that combines the constraints in the input-oriented and surface-oriented models. This model allows independent contributions from the identity of both the underlying and surface final harmonising vowel when determining suffix choice.

5. A **lexical-surface model** that combines the constraints in the lexical and surface-oriented models. This model allows independent contributions from the identity of the surface final harmonising vowel and certainty in lexical harmonic class when determining suffix choice. This corresponds to the analysis presented above.

6. A **lexical-input-surface** model, which combines the constraints in the lexical, surface-oriented, and input-oriented models. This model allows independent contributions from the identity of both the underlying and surface final harmonising vowels, as well as the certainty in lexical harmonic class when determining suffix choice.

All of the models above include the ID[HEIGHT] and *UNRAISED constraints. A violation of ID[HEIGHT] was assessed for any form that exhibited raising. Violations of *UNRAISED were assessed when vowel raising did not occur in a context where it should have applied (i.e., a low vowel in a derived, word-medial open syllable). However, for roots that categorically resist raising (i.e., that never exhibited raising in the corpus), no violation was assessed.

In order to estimate $P(\text{HC} = B|x)$ for each root, I trained a mixed-effects logistic regression model to predict the likelihood of observing each root with a back suffix based on properties of the root.[14] The fixed effects in the model include a three-way interaction between final vowel identity, log root token frequency, and the proportion of raised tokens of the root, as well as the pairwise interactions of these variables. The model also includes three separate binary dummy-coded predictor variables corresponding to whether the root ends with the suffix /-ʧæ/, /-ɑnæ/, or /-nɑmæ/. Finally, the model includes a random intercept for root identity. This allows root-specific deviations from population-level trends to be encoded (i.e., idiosyncratic harmonising behavior of particular roots). I do not explicitly encode a bias towards back suffixes into the model. The separation of the fixed, population-level effects from the lexically-specific random effects is crucial for the model to be able to generalize to unseen roots. More will be said about this in the discussion. The dependent variable in the regression was the proportion of back suffixes taken by the root.[15] The coefficients of the model fit to the entire data set can be found in Appendix E.

Models were fit using the `maxent.ot` R package (Mayer et al., 2024). In order to prevent overfitting, k-fold cross validation was used, with $k = 10$. This means that instead of fitting the grammar to the entire data set, the tokens were randomly partitioned into ten subsets. One of the subsets (10% of the data) was held out and the model was trained on the remaining nine subsets (90% of the data). The trained model was then applied to predict the held-out subset and the log-likelihoods of the model applied to the training and test sets were recorded. This process was repeated ten times, with each subset being held out once. The same partitions were used for each model.

For the lexical models, the logistic regression model used to approximate $P(\text{HC} = B|x)$ was fit only to the training data in each case. It is important to note, however, that the predictor variables in the model, log root token frequency and the proportion of raised tokens, were calculated based on counts from the entire corpus. This includes unsuffixed tokens and tokens without harmonising suffixes, which are not included in the current analysis. This means that the values of these variables for each root were consistent across all folds, but the coefficients of the model that dictate how these properties are related to suffix backness differed depending on the training data. The models were configured to use root-specific random effects when they made predictions for roots found in the training data, but to make predictions using population-level effects when they encountered new roots.

[14] For computational tractability this model was not implemented in a Bayesian framework and did not include random intercepts for corpus/author.

[15] Specifically, the model was fit to an aggregated data set consisting root types with proportion of back suffixes as the dependent variable, weighted by the number of tokens.

L2 regularization was used to prevent overfitting (Goldwater & Johnson, 2003). For each model, the default value of each constraint, μ, was set to 0. σ, which determines how strongly deviations from the default value are penalized, was chosen based on a simple search over a range of values: 500, 200, 100, 50, 20, 10, 1, 0.5, 0.2, 0.1, 0.05, 0.02, 0.01, 0.001. The results below correspond to the values of σ for each model that produced the highest log-likelihood on the held-out data.

Table 9 shows for each model the mean and standard deviation of the log-likelihoods across the training and test sets for each of the ten folds, as well as the optimal value of σ. Higher log-likelihoods indicate better model fit and lower standard deviations indicate greater model stability across folds. Because of the large data set, the choice of σ was not particularly important so long as it was large enough to allow the weights to fit the data effectively. All simulations where $\sigma \geq 1$ produced similar results.

Model	Constraints	σ	Mean test LL	Test std. dev.	Mean train LL	Train std. dev.
Surface-oriented	3	10	−31,926	220	−290,574	347
Input-oriented	3	10	−16,570	127	−152,369	274
Lexical	3	10	−15,730	122	−144,397	258
Input-surface	4	20	−15,670	142	−144,272	281
Lexical-surface	**4**	**10**	**−15,097**	**142**	**−138,725**	**272**
Lexical-input-surface	5	20	−15,097	142	−138,725	272

Table 9: Mean log-likelihoods for the training and test sets across each of the ten folds and the optimal value of σ. The lexical-surface model obtains the best performance on the held-out test folds with the fewest number of constraints.

Of the six models tested, the lexical-surface model, which is sensitive to the lexical class of the root as well as the surface properties of the word, achieves the best performance with the fewest number of constraints. Three specific comparisons are particularly insightful: (1) that this model performs better than the lexical model indicates that there is a pressure towards surface-apparent harmony; (2) that it performs better than the input-surface oriented model indicates that suffix choice is sensitive to lexical factors such as root frequency and the proportion of raised tokens rather than simply underlying final vowel identity; and (3) that it performs identically to the lexical-input-surface model indicates that an additional constraint mandating harmony with the backness of the underlying harmonising vowel is unnecessary given the HARMONICUNIFORMITY constraint.

Although the improvement in log-likelihood between the input-surface and lexical-surface models is perhaps modest, it is important to keep in mind that the roots that have the potential to produce opaque harmony are a relatively small proportion of this data set, and the roots within that class that display variability in opacity are even fewer.

5 Discussion

Based on the results of the corpus and modeling studies, this paper suggests that the opacity observed in Uyghur backness harmony is mediated by both surface constraints on backness harmony as well as pressure for a root to display consistent harmonising behavior across its extended paradigm (Rebrus & Törkenczy, 2017, 2021; Rebrus et al., 2023). When certainty in the harmonic class of a root is high, this can override conflicts with phonological generalizations; conversely, when certainty in harmonic class is low, a greater influence of surface harmony constraints may be observed. This analysis is not only able to model root-specific rates of opacity, but also accounts for the relationship of these rates to factors such as frequency and morphological composition.

This account aligns with proposals that some cases of opacity are driven by paradigm uniformity

(Steriade, 2000), such as Canadian raising in certain dialects of English (Hayes, 2004), and several processes in Tiberian Hebrew (Green, 2004) and Polish (Sanders, 2003; Łubowicz, 2003). A clearer understanding of the relationship between paradigm uniformity and opacity may help to further unify at least some of the variegated types of opaque phenomena that have been observed to date (Baković, 2011; Baković & Blumenfeld, 2019).

An additional consequence of this treatment of opacity is that it allows the Uyghur data to be analyzed using a strictly parallel model. If a paradigm uniformity account is indeed correct, the inability of parallel models to represent this opaque phenomenon as a purely phonological process may be seen as a point in their favor, rather than a failure.

In the remainder of this section, I will touch on some concerns with the current model, alternative analyses, and possible directions for future research.

5.1 Generalization

One of the central goals of phonological theory is to account for generalization from attested data to novel forms, such as in wug test studies (Berko, 1958). Although the proposal made in this paper hinges on lexical knowledge of root harmonising class, it also predicts generalization to unattested roots based on the population-level properties of attested roots: in addition to learning the specific harmonic classes of attested items, speakers also learn the properties of roots that fall into each class. The claim made here is that, by relying on these generalizations about lexical classes, speakers can estimate the harmonic class of a new root even with only a single exposure. Indeed, the simple model of $P(\text{HC} = B|x)$ presented in the previous section predicts that an B-final root observed once in its unraised form will have a 0.9997 probability of being a back harmoniser. The estimated effects of frequency and the proportion of raised forms from the corpus data are subtle on forms with such obvious clues to their harmonic class.

This raises the question of how the claims in this paper might be tested experimentally if tokens with surface-apparent harmony are unlikely to be encountered without a large sample. I have two thoughts on this.

First, it may be the case that the model presented here overestimates the subtlety of these effects because it is trained on a relatively restricted genre of text: namely, newspaper and academic articles. Since such products are typically reviewed by an editor, surface-apparent forms, which are non-standard, are likely to be corrected prior to publication. It may be the case that these forms are more frequent in colloquial speech or writing, and therefore perhaps more amenable to simple wug-testing.

Second, even if this is not the case, there may be more nuanced ways of testing this phenomenon. Even if Uyghur speakers are highly consistent at producing opaque harmony on novel words, they might exhibit frequency-driven differences in language processing tasks (see Ellis, 2002, for an overview of such effects). For example, an unexpected suffix form on a highly frequent word might be more surprising or disruptive than one on a word they have only seen once. This effect might be tested using online behavioral measures such as eye-tracking, reading time, or EEG. It may also be the case that certainty in harmonic class is mediated by recency of exposure. One might test for this by asking participants to produce a suffixed form of a root immediately after seeing its unsuffixed form vs. seeing it in its raised form without a harmonic suffix, or to do a similar task in a priming study where they are primed with either raised or unraised forms of a root. If certainty in harmonic class is mediated by both recency and frequency, one might expect high-frequency roots to be more resistant to disruption via recent exposure to raised forms than low-frequency roots. I leave these as interesting areas for future research, in Uyghur or in other backness harmony languages where lexical factors are relevant.

5.2 Pathological roots

When considering root-specific behavior in the Hungarian vowel harmony system, Hayes and Londe (2006) note that an acceptable analysis must preclude the existence of pathological examples like a simple B root (e.g., the hypothetical /pɑb/) categorically taking front suffixes. The current model allows for this pathology, to the extent that it is possible to define $P(\text{HC} = \text{FRONT}|/\text{pɑb}/)$ in such a way that this behavior would be predicted.

Setting aside the question of how such a pathological root could arise in Uyghur in the first place, even in a case where there is sudden massive and unequivocal distributional evidence for /pɑb/ as a front harmoniser, the model predicts that its long-term survival would be uncertain for two reasons:

1. The phonological component exerts pressure for surface-harmony based on general patterns in the language.

2. A root like /pɑb/ would be highly atypical within the class of Uyghur front roots.

Both of these factors predict that /pɑb/ would be produced with back suffixes at least some of the time, particularly for speakers with low exposure to the root who rely more on knowledge of the general properties of front harmonisers than on root-specific knowledge. Over generations this variation would decrease $P(\text{HC} = \text{FRONT}|/\text{pɑb}/)$, resulting in a gradual drift towards back suffixation.

The model's predictions for the future of opacity in Uyghur are less clear. Although total surface-apparent harmony would be preferable from the perspective of the phonology, the zones of variation in the harmony system preclude a total shift away from lexical factors.

5.3 Alternative analyses using serial models

Analyses of opacity that use serial variants of classical OT grammars, such as Stratal OT, do not straightforwardly predict the kind of variability in rates of opacity seen in the corpus here. It is possible, however, to yield variation between opaque and transparent outcomes by introducing weighted constraints or variable constraint ranking into these models.[16] For example, Stratal OT can be made probabilistic by specifying a MaxEnt grammar for each stratum rather than a classical OT grammar (Nazarov & Pater, 2017). In terms of the Stratal OT analysis presented earlier in the paper in Table 4, if the weighting of ID[BACK] is allowed to vary probabilistically in the Phrase stratum as a function of individual lexical items, this model can produce opaque outcomes when ID[BACK] is weighted higher than VAGREE and transparent outcomes when it is ranked lower. Similarly, a Serial Markedness approach (Jarosz, 2014) could encode this variation by making the constraints that force extrinsic ordering of processes variably ranked (Jarosz, 2016; Prickett & Jarosz, 2021) in a way that is sensitive to lexical identity.

It is beyond the scope of this paper to consider these alternatives in any depth, but it will be important to compare their predictions against the account provided here. I will make two brief comments, however, on why the parallel account presented here may provide greater explanatory value.

First, in addition to the need for strata or serial constraint evaluation to model opacity and the variability therein, these serial analyses will still require some mechanism to encode the lexical aspects of harmony in Uyghur: namely, the need to memorize the harmonising class of roots that fall in zones of variation. The parallel analysis presented here is somewhat more parsimonious in that it attributes both phenomena to the same lexical factors.

Second, it is not clear to me how the relationship between root frequency and constraint weighting can be motivated in these models. To produce the necessary ranking of ID[BACK] in the probabilistic Stratal OT analysis, for example, requires that the weight of the constraint be scaled up for

[16] Thanks very much to an anonymous reviewer who pointed out these possible analyses.

high-frequency forms and scaled down for low-frequency forms. This runs counter to most accounts of frequency effects in phonology, where high-frequency forms are typically *less* faithful than their low-frequency counterparts (e.g., Coetzee & Kawahara, 2012; Coetzee, 2016). Some account for this discrepancy will be important. By contrast, the relationship between frequency and opacity is clear in the model presented in this paper: increased exposure to a root provides greater certainty in its harmonic class.

5.4 Limitations of the corpus

Because this is a corpus study, it is necessarily exploratory and does not constitute hypothesis confirmation (see Roettger, 2019, for more on this distinction). Exploratory analyses are most valuable as a tool for hypothesis generation. It will be important to test the predictions made by the model proposed here in controlled, experimental contexts, as described in the earlier section on generalization.

As well, because this study uses corpora where authorship cannot be uniquely determined, it does not tell us to what extent the variability we observe happens at the level of the individual or in aggregate across the population as the result of dialect variation. This is a particular concern because the corpora do not come from the same region: the *Awazi* corpus is written for Uyghurs living in Kazakhstan, while the other two corpora are written for Uyghurs living in China and the diaspora. The results of this study do, however, tell us where we might expect variation, either within or across speakers, or both. The sources of this variation can be carefully determined using more granular corpus or experimental studies.

5.5 A closing remark on corpus methodologies

In addition to its empirical and theoretical contributions, this paper demonstrates the value of taking a more holistic and comprehensive empirical approach to linguistic data collection and analysis. The internet has allowed for the proliferation of large amounts of textual data, even for relatively small languages such as Uyghur. Computational tools such as web scrapers and morphological transducers can allow us to marshal the complexity inherent in such large data sets, and provide access to new types of empirical data that allow us to supplement other data sources and measure phonological patterns writ large. Such tools will become increasingly important as phonological theory continues to develop.

Competing interests

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

I thank Bruce Hayes, Kie Zuraw, Adam McCollum, Tim Hunter, Eric Bakovic, Travis Major, Canaan Breiss, and the attendees of AMP 2020, mfm 2022, UCLA Phonology Seminar, and the Berkeley Phorum for their feedback at various stages of this project; Jonathan Washington for his help with the Apertium transducer; Tyler Carson, Daniela Zokaeim, and Rutvik Gandhasri for their work on the webscrapers; and Gülnar Eziz, Mahire Yakup, Gulnisa Nazarova, and other members of the Uyghur community for sharing their language and culture with me.

Funding statement

This work was supported by the Social Sciences and Humanities Research Council of Canada.

Data availability statement

The corpora and code used in the statistical analysis and MaxEnt model can be found at `https://github.com/connormayer/uyghur_corpora`. The modified transducer used to parse the corpus can be found at `https://github.com/connormayer/apertium-uig/tree/vowel_harmony`.

Ethical standards

The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Supplementary material

Supplementary material can be found at `TODO`.

A Abbreviations

1.SG.POSS = first-person singular possessive, 3.POSS = third-person possessive, 3.SG.PST = third-person singular past, ACC = accusative, DAT = dative, DIM = diminutive, GER = gerund, LOC = locative, PFV = perfective, PL = plural, PROG = progressive.

B The Uyghur morphological transducer

This appendix describes the morphological transducer used to analyze the corpora in this paper.

The transducer is a modified version of the *apertium-uig transducer* (Littell et al., 2018; Washington et al., 2019). This 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 roots 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 root must be explicitly encoded in the transducer.

The output of the first level then serves as input to the next level, which maps abstract morphophonological 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" (I use Latin orthography throughout this section rather than IPA, since it more closely reflects the input to the transducer).

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.

The transducer can also map in the oppposite direction: between surface forms and underlying analyses that consist of roots plus morphological tags. For example, if the input is the surface form <*qizingizgha*> "to your daughter" the output analysis will be *qiz<n><px2sg><frm><dat>*. This indicates that the root is *qiz*, a noun <n>, and is suffixed with the 2nd person singular possessive marker in its formal form <px2sg><frm> followed by the dative suffix <dat>.

B.1 Modifying the transducer

I modified this transducer to detect the harmonic quality of suffixes when mapping from surface to underlying forms.[17] While a form like *qizingizgha* maps to *qiz*<*n*><*px2sg*><*frm*><*dat*> under the original transducer, it maps to *qiz*<*n*><*px2sg*><*frm*><*dat-b*> under this modified system, 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* or *-ge*).

The modified transducer splits each tag corresponding to a harmonizing 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. In addition, the restrictions the phonological component of the transducer imposes on harmony have been lifted. The original transducer, for example, would reject a form like **at-ler*, "horses", for being disharmonic. The modified transducer will simply interpret this as an instance of the front form of the plural suffix.

In a few cases, multiple parses that correspond to identical surface parses were removed from the transducer for simplicity. Take the word *doktur* 'doctor' as an example. This noun may be parsed as a nominal (appropriate in cases like *doktur chong* 'the doctor is old'), copular form (appropriate in cases like *Adil doktur* 'Adil is a doctor'), and so on. Because such distinctions are not relevant for the current project, all but the nominal parse was removed.

Finally, the vowel raising processes described in Section 2.3 can obscure the harmonizing quality of suffixes: for example, the surface realization of /dost-lAr-m/ 'friend-PL-1.SG.POSS' is [dostlirim], which does not allow the backness of the plural morpheme to be determined. In such cases the modified transducer 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*>. This paper only uses tokens where the harmonic value of at least one suffix is unambiguous.

B.2 Interpreting and sanitizing the transducer output

Applying the transducer to the corpus produces one or more possible parses for each word that the transducer is able to recognize. The transducer was able to successfully analyze about 2.6 million of the 4.2 million words in the *Uyghur Awazi* corpus (61%), about 7 million of the 8 million words in the RFA corpus (87%), and about 1.8 million of the 2.4 million words in the Uyghur Academy Corpus (75%). The typical reason the transducer would fail to parse a word is because the root is not included in the transducer's list of valid roots: this is a key limitation of rule-based parsers. Tokens for which the transducer produced no parse were discarded.

Additional filtering was done to sanitize the data produced by the transducer. One challenge that arises is how to deal with multiple analyses in cases where they provide conflicting information about the root. The surface form *orgini*, for example, could correspond to underlying /orgɑn-i/ 'organ-3.POSS' or to /or-GAn-i/ 'harvest-PFV-3.POSS' (though the latter is unlikely because it corresponds to the front form of /-GAn/ attached to an umambiguously back root). I take a maximally conservative approach and discard all tokens for which such ambiguous parses are possible.

In addition, a number of suffixes are harmony blockers. Such suffixes, like the progressive suffix /-wɑt/, block harmony by failing to harmonize, and impose their own harmonic values on following

[17] `https://github.com/connormayer/apertium-uig/tree/vowel\_harmony`.

suffixes. Such suffixes are often historically derived from multi-word constructions (Mayer, McCollum, & Eziz, 2022).

(22) *Harmony blocking* /-wɑt/

	/søzlæ-wɑt-GAn/	→	[søzlæwɑtqɑn]	'speak-PROG-PFV'
cf.	/søzlæ-GAn/	→	[søzligæn]	'speak-PFV'

I discard tokens containing such suffixes.

Finally, a number of spurious parses were omitted based on manual inspection of the results. Particular attention was paid to tokens reflecting surface-apparent harmony to ensure these were not overcounted.

C Additional details on statistical models

The Bayesian models were fit using the default parameters of the `brm` function. Fig. 5 shows samples from the posterior for each coefficient.

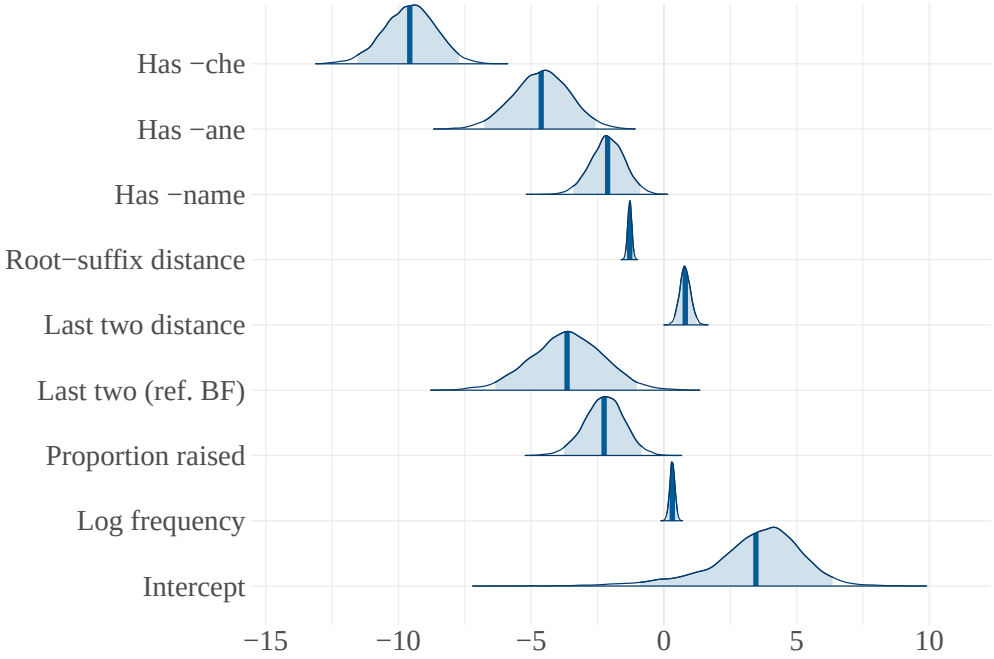

Figure 4: Plot of samples from the posterior for each model parameter. The dark areas are the mean values, and the shaded areas are the 95% Credible Intervals.

The DHARMa R package (Hartig, 2022) was used to apply standard residual diagnostics to the model (Fig. 5). The diagnostics show that the model satisfies tests for outliers, uniformity, and zero inflation (not pictured), but shows significant underdispersion. This means that the residual variance in the data is smaller than expected under the fitted model (see Fig. 6). The significance of this test tells us that, because of the large number of data points, we can be fairly certain that the model is underdispersed. However, this is not a concern for two reasons. First, the dispersion parameter is about 0.93, indicating that there is 7% lower variance in the observed data than predicted by the model. This is quite a small deviation. Second, underdispersion produces more conservative parameter estimates from the model (Hartig, 2022), so we may still be confident in the interpretation of the model presented in the paper.

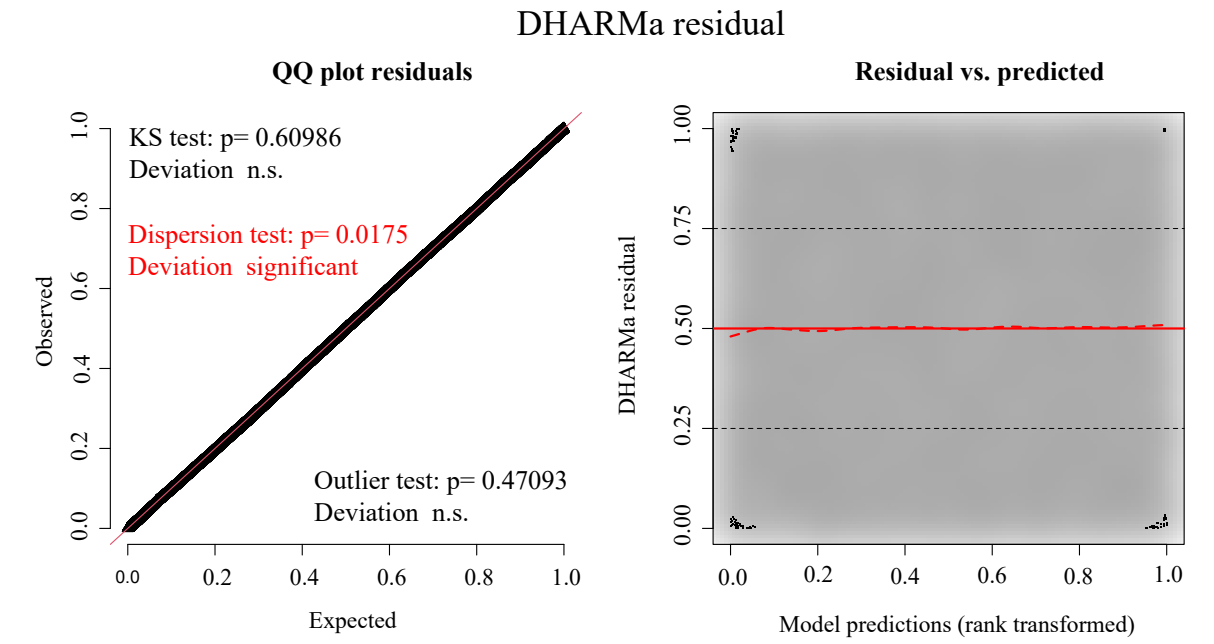

Figure 5: Output of DHARMa model checks run on the Bayesian logistic regression model.

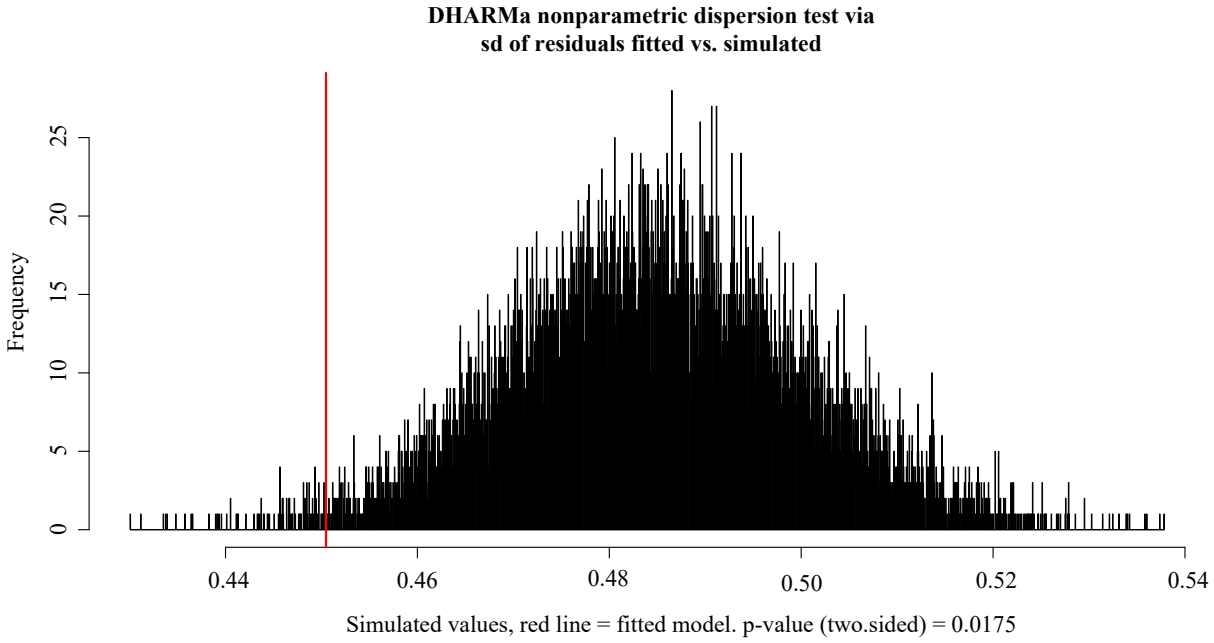

Figure 6: Histogram comparing simulated model residual standard deviations against the residual standard deviation from the fitted model.

D Maximum Entropy Optimality Theory

In a Maximum Entropy Optimality Theory 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 $C_i(x, y)$ returns the number of times an output candidate y for the input x violates the ith constraint. The *harmony* $H(x, y)$ of an output candidate y given the input x is:

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

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

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

where Ω is the set of all possible output candidates given the input x.

The likelihood of a data set under a MaxEnt model can be calculated by multiplying together the probabilities assigned to each token by the model (or summing them in log space). It is also straightforward to learn constraint weights that optimize fit to a dataset: see Hayes and Wilson (2008) for more details on this learning procedure.

E $P(\textbf{HC} = B|x)$ model coefficients

The coefficients of the model used to estimate $P(\text{HC} = B|x)$ in Section 4.4 are shown in Table 10. It is important to note that these are the coefficients learned when the model is fit to the entire data set. These will differ slightly from the coefficients learned during the k-fold cross validation process, though not substantially. p-values are not reported because this model is being used only for predictive purposes.

The coefficients correspond to the expected effects: the final vowel of the root has a substantial predictive effect on $P(\text{HC}|x)$, and this effect is strengthened the more frequent the root and weakened the more often it occurs in raised forms. As well, the three exceptional suffixes all increase $P(\text{HC} = F|x)$, which corresponds to surface-apparent harmony in those suffixed forms.

Term	Coefficient
Intercept	8.27
Final vowel (F)	-17.07
Log token count	0.66
Proportion raised	-0.08
Has *-ane*	10.02
Has *-name*	4.96
Has *-che*	8.01
Final vowel (F) : Log token count	-0.95
Final vowel (F) : Proportion raised	0.74
Log token count : Proportion raised	-0.13
Final vowel (F) : Log token count : Proportion raised	0.20
Root (standard deviation)	5.40

Table 10: Coefficients from the mixed-effects logistic regression model for approximating $P(\text{HC} = B|x)$ when it is fit to the entire data set.

References

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

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.

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.

Baković, E., & Blumenfeld, L. (2019). Rule interaction conversion operations. *Loquens*, *6*(2), e062.

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.

Bermúdez-Otero, R. (2003). The acquisition of phonological opacity. In J. Spenader, A. Eriksson, & Östen Dahl (Eds.), *Variation within optimality theory: Proceedings on the Stockholm workshop on 'Variation within Optimality Theory'* (pp. 25–36). Stockholm: Department of Linguistics, Stockholm University.

Bermúdez-Otero, R. (2018). Stratal phonology. In S. Hannahs & A. R. Bosch (Eds.), *The Routledge Handbook of Phonological Theory* (pp. 100–134). Abingdon: Routledge.

Bodrogligeti, A. J. E. (2001). *A grammar of Chagatay*. Muenchen, Germany: LINCOM EUROPA.

Bowers, D. (2019). The Nishnaabemwin restructuring controversy: New empirical evidence. *Phonology*, *36*(2), 187–224.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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.

Coetzee, A. W. (2016). A comprehensive model of phonological variation: grammatical and non-grammatical. *Phonology*, *33*(2), 211–246.

Coetzee, A. W., & Kawahara, S. (2012). Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, *31*, 47–89.

Crosswhite, K. (2001). *Vowel reduction in Optimality Theory*. New York/London: Routledge.

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

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.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. doi: 10.1017/S0272263102002024

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.

Farris-Trimble, A., & Tessier, A.-M. (2019). The effect of allophonic processes on word recognition: Eye-tracking evidence from canadian raising. *Language: Research Reports*, *93*(1), e136-e160.

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 Vari-*

ation within Optimality Theory (pp. 111–120). Stockholm: Stockholm University, Department of Linguistics.

Green, A. D. (2004). Opacity in Tiberian Hebrew: Morphology, not phonology. *ZAS Papers in Linguistics*, *37*, 37–70.

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.

Hartig, F. (2022). Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models [Computer software manual]. Retrieved from `http://florianhartig.github.io/DHARMa/` (R package version 0.4.6)

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, *39*(6), 1041–1070.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in phonological acquisition* (pp. 158–203). Cambridge: Cambridge University Press.

Hayes, B. (2016). Comparative phonotactics. In *Proceedings of the 50th meeting of the Chicago Linguistics Society* (pp. 265–285).

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.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164.

Hooper/Bybee, J. (1976). *An introduction to Natural Generative Phonology*. New York: Academic Press.

Jarosz, G. (2014). Serial markedness reduction. In J. Kingston, C. Moore-Cantwell, J. Pater, & R. Staubs (Eds.), *Proceedings of the 2013 Annual Meeting on Phonology*.

Jarosz, G. (2016). Learning opaque and transparent interactions in Harmonic Serialism. In G. Ólafur Hansson, A. Farris-Trimble, K. McMullin, & D. Pulleyblank (Eds.), *Proceedings of the 2015 Annual Meeting on Phonology*.

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.

Khanjian, H. (2009). Stress dependent vowel reduction. In I. Kwon, H. Pritchett, & J. Spence (Eds.), *Proceedings of the 35th Annual Meeting of the Berkeley Linguistics Society* (pp. 178–189). Berkeley, CA: Berkeley Linguistics Society.

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.

Kiparsky, P. (1982). Lexical morphology and phonology. In I.-S. Yang (Ed.), *Linguistics in the morning calm* (pp. 3–91). Seoul: Hanshin.

Kiparsky, P. (2000). Opacity and cyclicity. *The Linguistic Review*, *17*, 351–367.

Kirchner, R. (1996). Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry*, *27*(2), 341–350.
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).
Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
Lin, M., Jr., H. C. L., & Shmueli, G. (2013). Too big to fail: Large samples and the *p*-value problem. *Information Systems Research*, *24*, 906–917.
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.
Łubowicz, A. (2002). Derived environment effects in Optimality Theory. *Lingua*, *112*, 243–280.
Łubowicz, A. (2003). *Contrast preservation in phonological mappings* (Unpublished doctoral dissertation). University of Massachusetts Amherst.
Mascaró, J. (2009). Comparative markedness and derived environments. *Theoretical Linguistics*, *29*, 113–122.
Mayer, C. (2021a). Capturing gradience in long-distance phonology using probabilistic tier-based strictly local grammars. *Proceedings of the Society for Computation in Linguistics*, *4*(5).
Mayer, C. (2021b). *Issues in Uyghur backness harmony: Corpus, experimental, and computational studies* (Unpublished doctoral dissertation). University of California, Los Angeles.
Mayer, C., Major, T., & Yakup, M. (2022). Are neutral roots in Uyghur really neutral? Evaluating a covert phonemic contrast. In P. Jurgec et al. (Eds.), *Supplemental Proceedings of the 2021 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America.
Mayer, C., McCollum, A., & Eziz, G. (2022). Issues in Uyghur phonology. *Language and Linguistics Compass*, *16*(12). doi: 10.1111/lnc3.12478
Mayer, C., Tan, A., & Zuraw, K. R. (2024). Introducing maxent. ot: an R package for maximum entropy constraint grammars. *Phonological Data and Analysis*, *6*(4), 1–44.
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. (2020). Sonority-driven stress in Uyghur. In H. Baek, C. Takahashi, & A. H.-L. Yeung (Eds.), *Proceedings of the 2019 Annual Meeting on Phonology*.
McCollum, A. (2021). Transparency, locality, and contrast in Uyghur backness harmony. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *12*(10), 10.
Mielke, J., Hume, E., & Armstrong, M. (2003). Looking through opacity. *Theoretical Linguistics*, *29*(1–2).
Nazarov, A. (2019). Formalizing the connection between opaque and exceptionful generalizations. *Toronto Working Papers in Linguistics*, *41*(1).
Nazarov, A. (2020). Bedouin Arabic multiple opacity with indexed constraints in Parallel OT. In H. Baek, C. Takahashi, & A. H.-L. Yeung (Eds.), *Proceedings of the 2019 Annual Meeting on Phonology*.

Nazarov, A. (2021). Learnability of indexed constraint analyses of phonological opacity. In *Proceedings of the Society for Computation in Linguistics* (Vol. 4).

Nazarov, A., & Pater, J. (2017). Learning opacity in Stratal Maximum Entropy Grammar. *Phonology*, *34*, 299–324.

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

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistics Compass*, *10*(11), 591–613.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, *33*, 999–1035.

Pater, J. (2010). Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In S. Parker (Ed.), *Phonological argumentation: Essays on evidence and motivation* (pp. 123–154). London: Equinox.

Pater, J. (2014). Canadian raising with language-specific weighted constraints. *Language*, *90*, 230–240.

Prickett, B., & Jarosz, G. (2021). Modeling the acquisition of phonological interactions: Biases and generalization. In R. Bennett et al. (Eds.), *Proceedings of the 2020 Annual Meeting on Phonology.*

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.)

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rebrus, P., Szigetvári, P., & Törkenczy, M. (2023). How morphological is Hungarian vowel harmony? In N. Elkins, B. Hayes, J. Jo, & J.-L. Siah (Eds.), *Proceedings of the 2022 Annual Meeting on Phonology.*

Rebrus, P., & Törkenczy, M. (2017). Co-patterns, subpatterns and conflicting generalizations in Hungarian vowel harmony. In H. van der Hulst & A. Lipták (Eds.), *Approaches to Hungarian. Volume 15: Papers from the 2015 Leiden Conference.* Amsterdam: John Benjamins Publishing Company.

Rebrus, P., & Törkenczy, M. (2021). Harmonic Uniformity and Hungarian front/back harmony. *Acta Linguistica Academica*, *68*(1–2), 175–206.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology*, *10*(1).

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

Speelman, D. (2014). Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (Vol. 43, pp. 487–533).

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.

Sumner, M. (2003). *Testing the abstractness of phonological representations in Modern Hebrew weak verbs* (Unpublished doctoral dissertation). State University of New York at Stony Brook.

van der Hulst, H. (2016). Vowel harmony. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics.* Oxford: Oxford University Press.

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. (2019). Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019)*.

Zhang, J. (2019). Speakers treat transparent and opaque alternation patterns different – evidence from Chinese tone sandhi. In R. Stockwell, M. O'Leary, Z. Xu, & Z. Zhou (Eds.), *Proceedings of the 36th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project.

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

Zuraw, K. (2010). A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory*, *28*, 417–472.

Zuraw, K. (2016). Polarized variation. *Catalan Journal of Linguistics*, *15*, 145–171.