2021. *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic* 6. 5049.
https://doi.org/10.3765/ptu.v6i1.5049

Phonologically motivated orthographic variation in Modern Uyghur: the voicing of *h*

Michael Fiddler*

Abstract. In this paper, I present data from three corpora of written Uyghur showing that the conventionally voiceless letter *h*, which occurs in words of Arab-Persian etymology, sometimes patterns as voiced in stem-final environments where it is a trigger for morphophonemic voicing assimilation in a following segment. Results indicate that when authors omit root-final *h* from the spelling, they tend to use voiced suffix-initial consonants, but when the *h* is written there is considerable variation both between and within authors and lexemes. No other phonological or functional factors were identified as being strong predictors of the variation. I interpret this as reflecting a probabilistic process of lenition or deletion of root-final /h/ in the adaptation of these loanwords that has diffused at different rates across the lexicon for different speakers.

Keywords. Uyghur; orthography; loanword phonology; phonetics; corpus

1. Introduction. As corpus resources and statistical modeling techniques have become increasingly available to linguists, research in corpus linguistics has revealed complex multi-factorial interactions that condition variation in matters such as syntax and word choice (e.g., Wulff & Gries 2019, Bresnan et al. 2007, Gries & David 2007). More recently, the emergence of social media as platforms for written interaction has led to stimulating new developments in sociolinguistic research on language variation in the realm of orthographic practice. Studies in this vein have also revealed complex systems of variation. The spelling variants typically reflect aspects of phonetics or phonology, but their use in orthographic practice is conditioned by a wide variety of factors. Eisenstein (2015), for example, finds grammatical, phonological, and demographic factors at play in conditioning English g-deletion (from *-ing* forms) and *-t/-d* deletions on Twitter. Ilbury (2019), also working with Twitter data, describes how gay British men employed AAVE features to develop a "Sassy Queen" persona. Community-of-practice approaches have been taken as well, e.g., Iorio (2010) and Stewart et al. (2017).

While much of this type of research has focused on varieties of English, some work has been done on other languages, including French (van Compernolle & Williams 2010), Japanese (Joyce et al. 2012), Nigerian pidgin (Heyd 2016), and Romanized Lebanese Arabic (Sullivan 2017). The Turkic family, however, has been largely untouched as far as I have found, with the exception of some work on vowel harmony—Mayer (in prep) uses corpus evidence to address gradient aspects of opacity in Uyghur, and Washington (in draft) makes brief reference to the use of internet searches to support claims of variation in vowel harmony patterns in Kazakh. The agglutinative morphology and extensive morphophonemic processes in Turkic languages present very different challenges for corpus approaches from what we see in Standard Average European languages like English, but they also offer exciting possibilities for exploring conditioned variation.

* Thanks to Connor Mayer for sharing the corpus data with me, to Stefan Th. Gries for advice on corpus work and statistics, and to Matthew K. Gordon for helpful discussion of the results. All errors remain mine. Author: Michael Fiddler, University of California, Santa Barbara (mfiddler@ucsb.edu).

The present study focuses on a single letter in the orthography of Modern Uyghur (ISO 639-3 uig; Turkic; China and Central Asia). A look at the letter *h* brings us into the middle of a long story of Turkic morphophonology in contact with the Arabic and Persian lexicons, filtered through the social practice of Modern Uyghur orthography (also derived from Arab-Persian culture), and viewed through the lens of corpus techniques. The phoneme /h/ entered Uyghur through loanwords from Arabic (often via Persian), corresponding to the Arabic phonemes /h/ and /ħ/ (Comrie 1997: 916). Uyghur /h/ is traditionally described as a voiceless phoneme (e.g., Tüzdi 2002:47, Nazarova & Niyaz 2013:13, and Zakir 2007:8). However, there is considerable variation in its phonetic realization. Hahn (1991:74) states that /h/ is realized phonetically as [χ] before consonants and either [h] or [ɦ] elsewhere. In fact, the variation is even more complex than that. Figure 1 shows spectrograms of the word *qebih* 'wicked' pronounced in isolation by three different female speakers with /h/ realized as [h], [χ], and nothing. The uvular fricative in (b) was clear to the ear, and the lowering and backing of the tongue for the uvular produces a back allophone of /i/, as reflected by the noticeably lower second formant (Gordon et al. 2002). These spectrograms are evidence that [χ] can occur even in word-final position, and that in some instances the phoneme /h/ may not be produced at all.

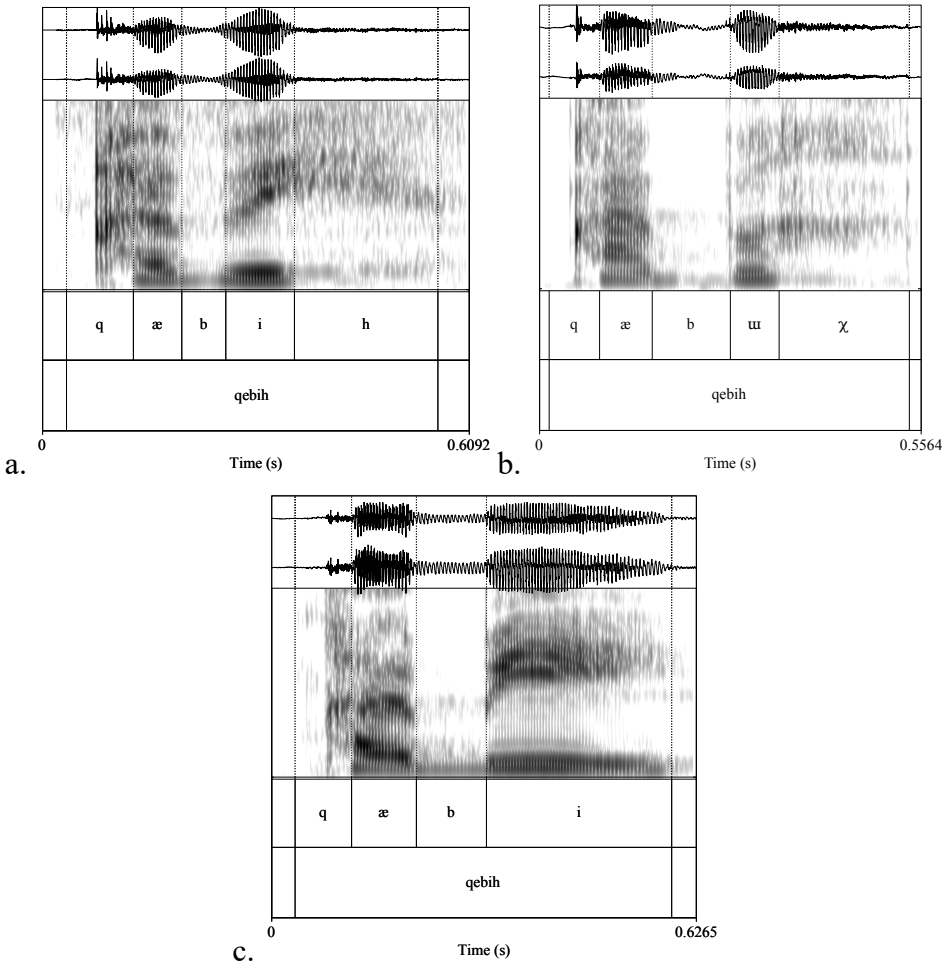

Figure 1. Uyghur /h/ as [h], [χ], and nothing in the word *qebih* 'wicked'

In addition to the allophonic variation, pharyngeal and glottal fricatives are notoriously gradient in terms of their acoustic properties. The boundary between, say, a slight bit of voiceless turbu-

lence or breathiness before or after a vowel to indicate an [h] and the absence of any realization is quite fuzzy, and the difference in voicing between [h] and [ɦ] can be similarly elusive.

However, while the phonetics are somewhat gradient, Uyghur orthography forces a binary decision between voiced and voiceless in many contexts. A number of verbal and nominal suffixes that begin with consonants undergo voicing assimilation[1] in which the suffix-initial consonant matches the voicing of the last segment in the stem. For example, as illustrated in (1), the ablative suffix *-Din* is written *-din* following stems ending in voiced sounds and *-tin* following stems ending in voiceless sounds.[2] This phonological alternation is reflected in the orthography; any time a writer produces a root ending in *h* with an ablative suffix, they must choose whether to write the first consonant of the suffix with a *t* or *d*.

(1)

Voiced stem-final consonant	*Assimilation to voiceless stem-final consonant*
yol 'road' + -din 'ABL' = yoldin	tarix 'history' + -din 'ABL' = tarixtin
yol 'road' + -da 'LOC' = yolda	tarix 'history' + -da 'LOC' = tarixta

This raises the question of how the binary phonological and orthographic categories map onto the gradient phonetic categories. Two quantitative questions emerge about words ending in *h*:

Question 1: a) How often do writers violate the conventional pattern of using a voiceless consonant letter in the suffix vs. a voiced letter?
b) How does this compare to words ending in other letters besides *h*?
c) What factors condition the variation?

Question 2: a) How often do writers omit the *h* from the spelling altogether?
b) What factors condition this choice?

2. Methods. Data for the study come from the archives of two Uyghur news sources, the Washington, D.C.-based *Radio Free Asia* (RFA) and the Kazakhstan-based *Uyghur Avazi* ("Uyghur Voice"), and the D.C.-based Uyghur American Association online discussion forum. The RFA corpus comprises ~9.2 million words from ~30,000 articles spanning 2008-2020, and the Uyghur Avazi corpus ~6 million words from ~14,000 articles spanning 2012-2020. The forum corpus contained over 24 million words from over 100,000 posts, but an unknown percentage of the posts were written in languages other than Uyghur. The contents of these three publicly accessible archives, including the article texts and metadata such as author and date, were collected with a custom web-scraper and compiled into a corpus. The RFA news was written in Uyghur Arabic script, the *Uyghur Avazi* news was written in a Latin script, and the forum posts were written in either Arabic script or a Latin script.

To address Question 1, an initial search was done in which all words containing a sequence of *h* followed by any of the relevant suffix variants (see Table 1) were extracted from the corpus. Many of these are verb suffixes, but it turns out there are essentially no *h*-final verbs in Uyghur. Like many languages, Uyghur tends not to borrow verbs from other languages, preferring borrowed noun+light verb constructions. However, the corpus search for verb suffixes with voice-assimilating consonants produced useful data of words ending in letters besides *h* (see below).

[1] I follow the convention in Turkic linguistics of using capital letters to indicate segments that alternate in assimilation or harmony processes, e.g., *D* represents an alveolar stop that surfaces as either [d] or [t]. It is common to analyze these as underspecified archiphonemes; Hahn (1991), on the other hand, argues that they are underlyingly voiced and become voiceless to match voiceless stem-final segments. My use of the capital letters is not intended to imply a stance on this issue but simply to indicate a segment with different surface realizations.

[2] All Uyghur forms in this paper are written in Latin script Uyghur orthography.

Suffix	Variants	Suffix	Variants
-DA 'LOC'	-da, -ta, -de, -te	-GU 'DESID'	-ghum, -qungiz, -güsi, -kümiz, etc.
-GA 'DAT'	-gha, -qa, -ge, -ke	-GUdek 'DESID.SIMUL'	-ghudek, -qudek, -güdek, -küdek
-Din 'ABL'	-din, -tin	-Di 'PST'	-di, -ti, -dim, -tim, etc.
-Dek 'SIMUL'	-dek, -tek	-Gili 'PURP'	-ghili, -qili, -gili, -kili
-DUr 'CAUS'	-dur, -tur, -dür, -tür	-Giche 'LIM'	-ghiche, -qiche, -giche, -kiche
-GAn 'PTCP'	-ghan, -qan, -gen, -ken[3]	-GAch 'SIMUL'	-ghach, -qach
-GAnliK 'NMLZ.PTCP'	-ghanliqi, -qanliqidin, -genlikini, -kenlikide, etc.	-GAchKA 'PTCP.REAS'	-ghachqa, -qachqa, -gechke, -kechke

Table 1. Suffix variants included in the corpus search

The expressions for *-GAnliK* and *-GU* were written to allow further suffixes added to the right, and *-Di* included the full paradigm of person/number combinations. All others were word-final. Corpus searches and statistical analysis were done with base R (R development core team 2020).

The initial search results were put through an extensive cleaning process. Each word was checked against a ~23,000-word dictionary list to remove any lexemes that ended in character strings that matched the regular expressions but were simply the end of the word, not suffixes, e.g., *puxta* 'strong' (not *pux-ta* '??? + ABL'). The root words were identified by removing the suffixes, and these were checked to flag any that were not in the dictionary list. All such roots were checked manually, and any that could be identified as known words that happened not to be in the dictionary list, recognizable morphological forms, obvious non-standard spelling variants, or phonological variants (e.g., with vowel raising) were kept; any that could not be verified were discarded. Some /x/-final words that were spelled with *h* on the *Uyghur Avazi* site were also cut.

To address the second part of Question 1, the initial search was repeated with all other letters in the alphabet in root-final position instead of *h*. The matches from this search (n ~6.1 million) were subjected to the same cleaning process as the *h* data, but due to time constraints manual checking was only done for items that occurred more than 250 times in the data. This did not affect the quality of the data included, it just meant more potentially useful data was discarded.

The initial search identified 66 *h*-final lexemes in the corpus. To address Question 2, a follow-up search was done to check for *h*-less variants of these 66 lexemes. The search expressions for this search included each specific root word spelled without the final *h*, checked first to be sure they were not just separate lexemes that differed only in that they did not have the final letter *h*. Because suffix assimilation was not crucial for the question of h-omission, the search expressions included bare forms of the roots, and also a few common non-assimilating noun suffixes, namely *-ning* 'GEN', *-ni* 'ACC', and *-lAr* 'PL.'

All matches were then coded for a variety of potentially predictive factors, including the phonological factors ROOT_LENGTH (in syllables), PRECEDING_VOWEL, SUFFIX_CONSONANT, and NEXT_LETTER (either the suffix consonant or, for unsuffixed forms, the first letter of the follow-

[3] All instances of *-ken* were eventually omitted from the data because they were computationally indistinguishable from cliticized forms of the evidential *iken*.

ing word), the regional/genre factor of which CORPUS the token came from, and the functional factor FREQUENCY (a continuous variable which was logged to reduce skewing). The phonological factors were essentially exploratory variables without theoretically motivated hypotheses as to how the factor would affect whether *h* is matched with a voiced or voiceless suffix consonant letter. For the other factors, some tentative hypotheses were posited. The forum corpus was expected to have more spelling variation in general than the news corpora, and possibly a higher percent of *h* matched with voiced suffix variants. The orthography used in the *Uyghur Avazi* news corpus conflates the phonemes /h/ and /x/ into a single letter. The association with the more saliently voiceless /x/ might motivate lower rates of voiced suffix variants with *h*-final roots in that corpus. Frequency could pull in either direction, as discussed in Bybee & Thompson (1997) and Diessel (2007), among others. On the one hand, high-frequency words are known to resist change, which would predict lower rates of voiced suffixes with *h* in this case. On the other hand, high-frequency words are also prone to phonetic reduction, which would in this case lead to the voicing or deletion of /h/ and increased use of voiced suffix consonants.

3. Results. Once the data was collected, the results were tabulated and statistical analyses were conducted to address the research questions. One initial observation is that the frequencies of the individual *h*-final lexemes were not evenly distributed; most were attested only a few times, and about 75% of the data came from the two most frequent lexemes, *allah* 'God' and *roh* 'soul.'

3.1. QUESTION 1: SUFFIX VOICING. For Question 1, the data was tabulated according to the outcome variable SUFFIX-VOICING, i.e., whether the suffix consonant letter was voiced or voiceless. For the 66 *h*-final lexemes, including variants spelled with the final *h* and without it, voiced letters in the suffixes were used 42.5% of the time (2020/4749). For only tokens spelled with the final *h*, the rate was 16.9% (552/3274).

The results for *h*-final stems were compared with the results for stems ending in other letters (Table 2). Since other stem-final letters may be either voiced or voiceless, the data is presented as either matching the expected pattern, i.e., voiced suffix letter with voiced stem-final letter and vice versa, or following the unexpected pattern, i.e., voiced suffix letter with voiceless stem-final letter and vice versa. The asymmetries are obvious: voiced suffix letters are matched with h-final stems (the unexpected pattern) over 40% of the time, but all other letters diverge from expectations less than 1% of the time. Within each group, there is a subset that violates expectations much more often: *h*-final stems spelled without the *h*, and stems ending in the voiced stops *b*, *d*, *g*, and *gh*[4], which are known to undergo final devoicing (Hahn 1991, Schwarz 1992).

		Unexpected pattern	Total	% unexpected
Stems ending in h:	Spelled with h	552	3274	16.9
	Spelled without h	1468	1475	99.5
	Total	2020	4749	42.5
Stems ending in other letters	b, d, g, gh	8260	20248	40.8
	All others	15741	3486372	0.45
	Total	24001	3506620	0.68

Table 2. Divergence from the expected patterns of morphophonemic voicing assimilation

[4] Phonetically, *gh* is a fricative, but in the phonology it patterns as the [+voice] counterpart of *q* and the [+back] counterpart of *g*. The other voiced obstruents which are not stops—*j*, *z*, and *zh*—behave similarly to all other letters in the corpus data in diverging from the expected patterns less than 1% of the time.

To examine the variation, two mixed-effects binary logistic regression model were fit to the data. Both models predicted SUFFIX-VOICING based on a set of fixed factors; the main difference was the random-effects structure. The first model included random intercepts for LEXEME and AUTHOR were included as well. The dataset was trimmed to include only LEXEMES and AUTHORS that contributed at least 10 tokens each, for a total of 659 total data points. The *Uyghur Avazi* news articles did not list authors, so trimming the data to use AUTHOR as a random effect excluded all the data from that corpus. Because this dataset included words spelled with and without the final *h*, all tokens were coded for an additional factor H, indicating whether the *h* was present or not. The fixed-effects structure included NEXT_LETTER, ROOT_LENGTH, PRECEDING_VOWEL, H, the log of FREQUENCY (FREQ.LOG), and their two-way interactions with CORPUS.

After a process of backward model selection, the final model included significant main effects for H, NEXT_LETTER, and PRECEDING_VOWEL. LEXEME was dropped from the random effects structure, but the random effect of AUTHOR was highly significant. The model's classification accuracy was 94% with a C-score of 0.9719. Table 3 shows the model summary for the final model, and Figure 2 shows effects plots for the three main effects in the model.

Formula: SUFF_VOICING ~ H + NEXT_LETTER + PREC_VOWEL + (1 | AUTHOR)

AIC	BIC	logLik	deviance	df.resid
332.4	354.9	-161.2	322.4	654

Scaled residuals:

Min	IQ	Median	3Q	Max
-4.6793	-0.1732	-0.0853	0.0064	6.4978

Random effects:

Groups	Name	Variance	Std.Dev.
AUTHOR	(Intercept)	9.886	3.144

Number of obs: 659, groups: AUTHOR, 30

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.6262	0.8092	-4.481	7.43e-06 ***
Hno	10.5952	1.8438	5.746	9.12e-09 ***
NEXT_LETTERG	1.4161	0.3644	3.886	0.000102 ***
PREC_VOWELo	-6.1989	1.8114	-3.422	0.000621 ***

Table 3. Model 1: model summary for binary logistic regression predicting SUFFIX-VOICING

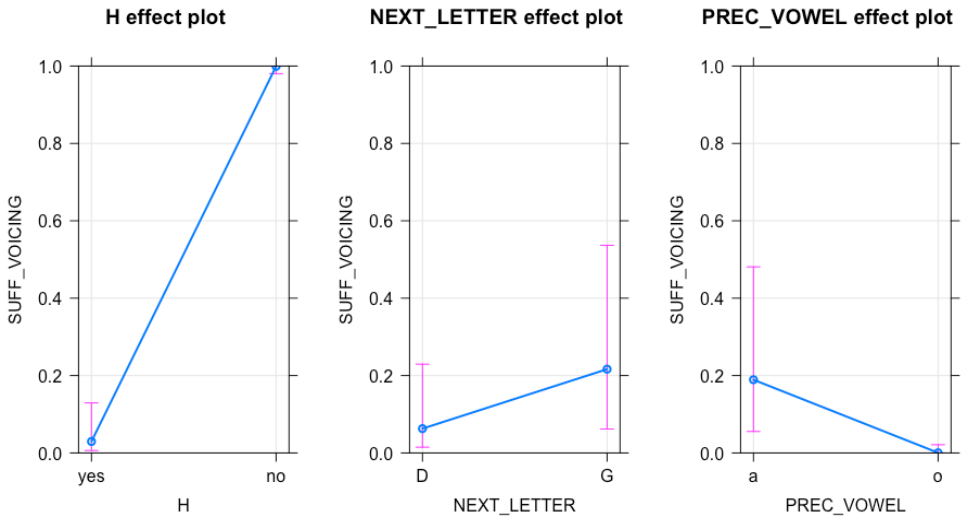

Figure 2. Effects of H, NEXT LETTER, and PRECEDING_VOWEL on SUFFIX_VOICING

Each plot in Figure 2 shows the model's predictions of how likely the suffix consonant is to be voiced (0.0 = very unlikely to be voiced, 1.0 = extremely likely to be voiced) based on the value of the factor on the x-axis. While all three factors had significant effects, the only one with a really notable effect size was H. As the leftmost plot shows, if a word was spelled without the final *h*, the model almost invariably predicted a voiced suffix consonant letter, and for words spelled with the *h* it predicted a voiceless suffix consonant. Looking at NEXT LETTER (center plot), the G suffix consonants were slightly more likely to be written as *g* or *gh* (rather than *k* or *q*) after a root-final *h* as compared to the odds of a D suffix consonant being written as *d* (rather than *t*). However, it made a relatively small difference in the model's predictions, and the confidence interval for /G/ was quite large. For PRECEDING_VOWEL (rightmost plot), an *a* before the root-final *h* predicted slightly higher odds of a voiced suffix letter as compared to an *o* before the *h*, but again the effect size was small and the confidence interval was large.

To illustrate the random effect of AUTHOR, the histogram in

Figure 3 shows the distribution of the percent voiced suffix letters used with *h*-final words by the authors in the dataset (raw data, not model predictions). The distribution is skewed left, which means that most authors used voiced suffix letters less than 25% of the time, but there is also considerable variation.

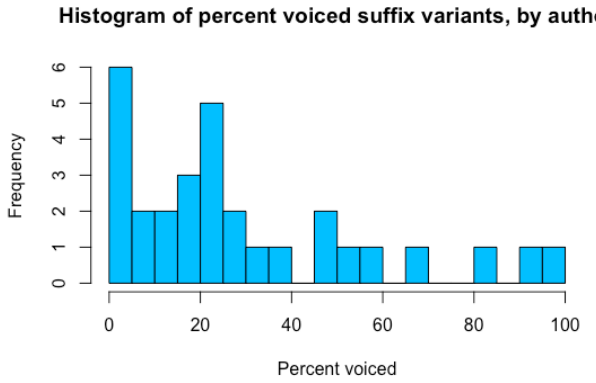

Figure 3. Raw data to illustrate the random effect of AUTHOR

In sum, what we learn from this model is that AUTHOR is a stronger random effect than LEXEME, and that the author's choice to write the *h* is the best fixed factor for predicting the spelling of the suffix consonant. There was little variation between the U.S.-based RFA news and the Uyghur American Association online forum, and the phonological factors had fairly little effect.

Because the criteria for including AUTHOR as a random effect meant that over 85% of the data had to be discarded, a second model was fit with the same fixed-effects structure but random intercepts only for LEXEME. This brought the size of the dataset back up to 4408 tokens. The interaction between H and CORPUS was not included in this model due to issues with complete separation. In the final model, the random effect of LEXEME was very significant, there was a significant interaction between PRECEDING_VOWEL and CORPUS, and there was a highly significant main effect of H. The model's classification accuracy was 88% with a C-score of 0.9240. The final model summary is shown in Table 4, and the effects plots are shown in Figure 4.

Formula:
SUFF_VOICING ~ H + PREC_VOWEL + CORPUS + (1 | MASTER_LEX) +
PREC_VOWEL:CORPUS

AIC	BIC	logLik	deviance	df.resid
2458.7	2509.8	-1221.3	2442.7	4400

Scaled residuals:

Min	IQ	Median	3Q	Max
-19.5396	-0.5830	-0.0434	0.0512	9.1928

Random effects:

Groups	Name	Variance	Std.Dev.
MASTER_LEX	(Intercept)	4.559	2.135

Number of obs: 4408, groups: MASTER_LEX, 16

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.27830	0.63484	-5.164	2.42e-07 ***
Hno	7.02413	0.47737	14.714	< 2e-16 ***
PREC_VOWELo	0.03047	2.29625	0.013	0.98941
CORPUSkaz	0.46702	0.68923	0.678	0.49803
CORPUSrfa	0.26368	0.15445	1.707	0.08777 .
PREC_VOWELo:CORPUSkaz	-3.55851	1.26576	-2.811	0.00493 **
PREC_VOWELo:CORPUSrfa	0.98001	1.02578	-0.955	0.33938

Table 4. Model 2: model summary of binary logistic regression predicting SUFFIX-VOICING

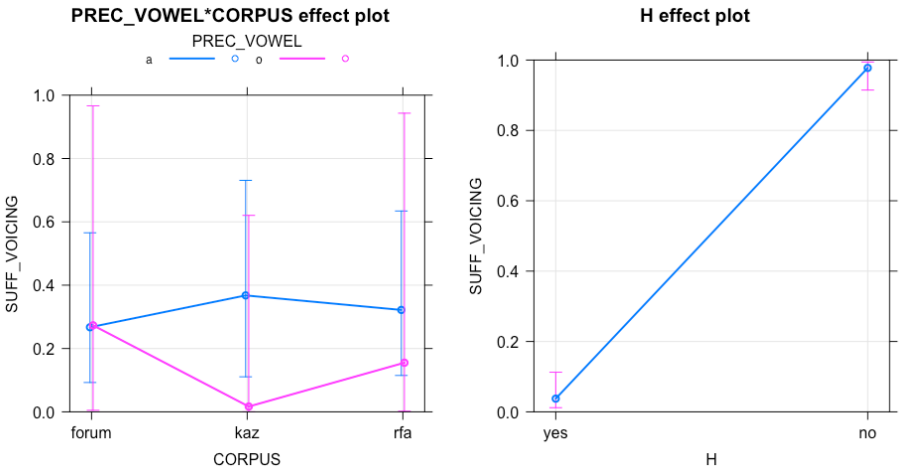

Figure 4. Model 2: effects plots of the interaction between PRECEDING_VOWEL and CORPUS and the main effect of H

As with the first model, it is H (Figure 4, right plot) which had the clearest effect on the choice of the suffix consonant letter. The model predicts that the suffix letter will be voiced whenever *h* is omitted, and voiceless when it is written. The interaction between PRECEDING_VOWEL and CORPUS (left plot) is more complicated. For the two news corpora, the model predicts slightly higher odds of a voiced suffix letter when there is an *a* before the *h*, but this trend does not hold in the forum corpus. However, the confidence intervals are very large, and much of the *a* vs. *o* effect is likely tied to specific lexemes (all data with *o* before the *h* come from the single word *roh* 'soul').

To illustrate the random effect of LEXEME, the histogram in

Figure 5 shows the distribution of the percent voiced suffix letters used with *h*-final words for the lexemes in the dataset (again raw data, not model predictions). Similar to AUTHOR, about half of the *h*-final lexemes were almost never paired with voiced suffix letters, but the other half often were.

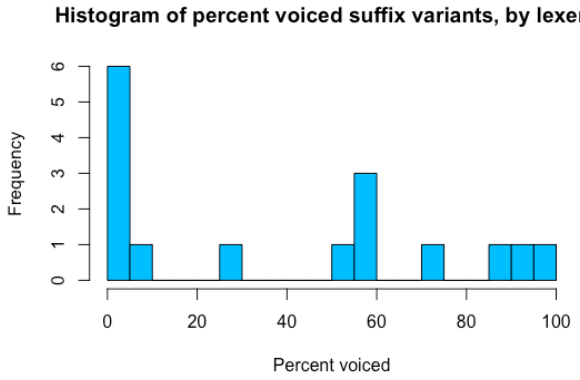

Figure 5. Raw data illustrating the random effect of LEXEME

The results from both models point to a strong influence of random factors and fairly little predictive power from phonological, functional, or regional/genre factors. Certain authors tend to use voiced suffix letters with *h*-final words more often, certain *h*-final words tend to be spelled with voiced suffix letters, and if the *h* is omitted from the spelling it is almost certain that a voiced suffix letter will be used. As a final illustration of this variation, Table 5 shows raw data

for the single lexeme *allah* 'God' from the five authors that contributed the most data in the forum and RFA news corpora.

Corpus	Author	Voiceless	Voiced	Total	% voiced
forum	A	18	25	43	58
	B	37	3	40	8
	C	30	0	30	0
	D	1	15	16	94
	E	8	6	14	43
RFA news	A	3	32	35	91
	B	23	5	28	18
	C	20	3	23	13
	D	17	0	17	0
	E	8	8	16	50

Table 5. Use of voiced vs. voiceless suffix consonants with the lexeme *allah* 'God' by the five authors that contributed the most data in each corpus

Two of the ten authors were consistent in always using voiceless suffix letters, but the other eight showed at least some variation in their spelling. Some mostly used voiceless letters, others mostly used voiced letters, and still others split pretty evenly.

We move on now to examine the question of how often and when *h* tends to be omitted from the spelling of these words.

3.2. QUESTION 2: OMISSION OF H. The tokens identified as a result of the initial search for *h*-final roots were tabulated along with the tokens from the follow-up search without the *h* (Table 6).

Spelling	n	%
With *h*	20902	63
Without *h*	12050	37
Total	32952	100

Table 6. Final *h*-omission

To test which factors might condition the omission of *h*, a mixed-effects binary logistic regression model was fit to the data. One additional factor was added, a binary variable SUFFIXED which indicated whether the word had a suffix or was just a bare root. The model predicted H-OMISSION based on the fixed factors of FREQ.LOG, ROOT_LENGTH, PRECEDING_VOWEL, SUFFIXED, and their two-way interactions with CORPUS, and also the non-interacting factor NEXT_LETTER. Random intercepts for LEXEME and AUTHOR were included as well (as before, this meant leaving out all data from the *Uyghur Avazi* news corpus). Discarding data from lexemes and authors that contributed less than 10 tokens, the dataset included 8754 tokens.

A backward selection process yielded a final model with both random factors, significant two-way interactions between CORPUS and FREQ.LOG, CORPUS and ROOT_LENGTH, CORPUS and PRECEDING_VOWEL, and CORPUS and SUFFIXED, and a significant main effect of NEXT_LETTER. The model failed to converge, but after restarting twice to run more iterations, the max|grad| was only 0.016052, which is reasonably close to convergence. The model's classification accuracy was 92% and the C-score was 0.9578. Figure 6 shows the effects plots for the significant interactions. (The model summary was too long to feasibly include here.)

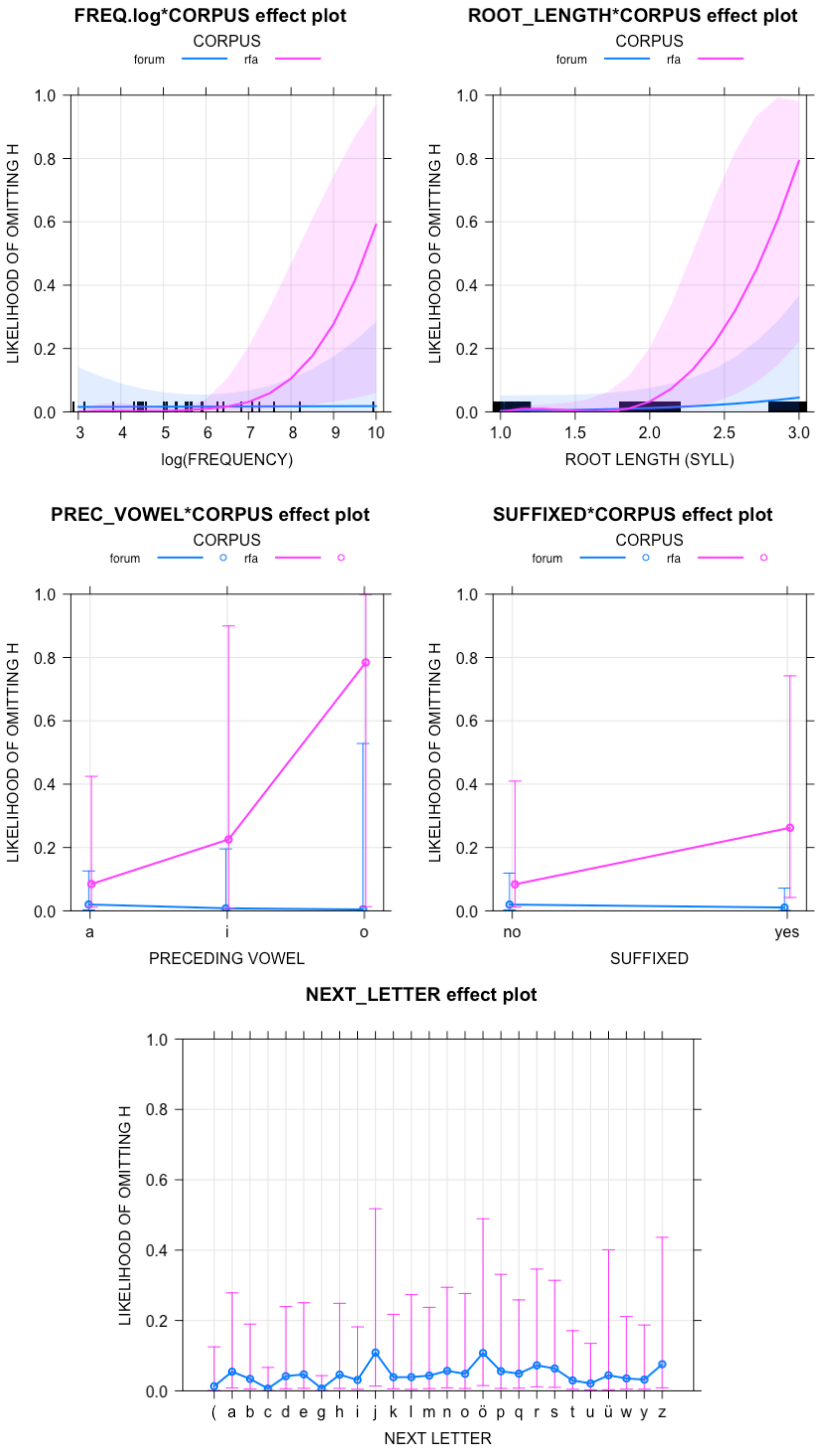

Figure 6. Model 3: effects plots of interactions between CORPUS and FREQ.LOG, CORPUS and ROOT_LENGTH, CORPUS and PRECEDING_VOWEL, and CORPUS and SUFFIXED, and a significant main effect of NEXT_LETTER

While these effects were all significant in the model, it is difficult to interpret them as indicating any coherent trends. In all the interactions, the model simply predicts that *h* is never omitted in data from the online forum. For the RFA corpus, the model predicts higher likelihood of *h*-

omission for words that are more frequent and longer. However, the increased odds of *h*-omission with higher frequency in the RFA corpus is likely due to the single lexeme *allah* 'God,' which is the only lexeme with very high frequency and happens to have *h* omitted more than other lexemes. Similarly, the prediction of 3-syllable words being spelled without *h* in the RFA corpus is due to the name *Abdullah*, which is spelled exclusively without an *h* in the RFA data. It is possible that these trends are significant, but given the limitations of the present dataset it is hard to separate them from lexeme-specific effects.

In the interaction between PRECEDING_VOWEL and CORPUS, the model predicts higher likelihood of *h*-omission when there is an *i* or especially an *o* before the *h*. However, the raw data shows the exact opposite trend, with the highest rates of *h*-omission when *a* precedes the *h* and lowest rates for *o*. Raw data obviously does not account for other factors in the model, but the enormous confidence intervals for *i* and *o* at least suggest the trend is shaky. The interaction between SUFFIXED and CORPUS seems solid—there is a slightly higher likelihood of omitting *h* in suffixed words in the RFA corpus—but that insight does not seem particularly useful. Finally, for NEXT_LETTER, there is some slight variation, but there are no coherent patterns and the confidence intervals are huge. In short, what we learn from this model is that there is considerable variation across individual authors and lexemes, and somewhat more variation in the RFA corpus than the forum, but no clearly interpretable trends from other variables.

4. Discussion At a very basic level, one interesting finding of this study is that there is indeed variation in spelling of *h*-final roots and the suffixes used with them even in the world of newswriting, where reporters are trained to follow standard conventions and their writing is edited carefully. It was surprising that the rates of variation in both h-omission and suffix consonant assimilation were comparable in the online forum data and news corpora. If these variables were indicators of informal writing or less careful attention to spelling, we would expect higher rates in web forum posts than in newswriting.

In addition, the striking difference between the rate of voiced suffix letter usage with h-final words (42.5%) versus words ending in all other letters (0.68%) suggests the influence of a real phonetic-phonological motivation, not just random typographical errors. It is likely that /h/ is not a very robust or salient consonant phonetically and is probably undergoing at least surface lenition—that is, becoming more sonorous by reducing the amount of constriction and frication and by allowing the vocal folds to continue vibrating—if not being deleted from underlying forms.

The next question, then, is whether the lenition or loss of /h/ is a general phenomenon or limited to word-final position. The morphophonology that provided a window into the process at the boundary between roots and suffixes cannot shed light on /h/ in other positions in the word. Even if there were prefixes that assimilated in voicing to root-initial segments, it would still not help with *h* in root-medial position. However, it is still possible to look at *h*-omission rates across different positions in roots. To this end, one final follow-up search was conducted. A convenience sample of words with *h* in word-initial, intervocalic, and pre-consonantal position was selected to compare with the previous results for root-final *h*. Most of the lexemes were represented only in bare form in the search expressions, but in a few cases common derivational or inflectional variants were included (e.g., *rehim-siz* 'merciless,' *méhmanxani-da* 'in the hotel'). All lexemes were checked to ensure that there was not a minimal pair differing only in the presence of the letter *h*. Spelling variants with *x* instead of *h* were included for pre-consonantal position, as [χ] is the expected allophone in that position. The results are presented in Table 7.

Position	Example	n lexemes	n tokens	% h omitted
Word-initial	hazir ‘now’	37	197327	0.3%
Intervocalic	bahar ‘spring’	47	88622	0.6%
Pre-consonantal	rehmet ‘thanks’	27	31408	2.2%*
Root-final	padishah ‘king’	59	32952	36.6%

*Spellings with *x* instead of *h* are counted as “not omitted”

Table 7. *h*-omission by position in the word

As Table 7 shows, *h* is not omitted with any notable frequency in any position except for word-final. This asymmetry is consistent with previous work on contextual variation of perceptual cues supporting phonemic contrasts (e.g., Steriade 1999), and more specifically with Bladon’s (1986) statements that [h] is perceptually non-salient in word-final position and that [h] is much more commonly attested in initial positions than final. The correspondence with perceptual motivation suggests that at a broad level the distribution of orthographic *h*-omission does in fact reflect the positions where /h/ is not produced or perceived in the spoken language.

In addition, looking back to the results of the model in Section 3.2, the fact that SUFFIXED had no significant effect on the omission of *h* shows the salience of bare noun forms even in this heavily suffixing language. In Hahn’s (1991) description, /h/ is realized with the more salient allophone [χ] in pre-consonantal position, but the corpus results suggest that this may only be true when the following consonant is part of the root. That is, for nouns where the root-final /h/ is not produced or is produced in a less salient form in the bare (nominative) form, the /h/ does not reappear and surface as [χ] when suffixes are added. Thus *rehmet* ‘thanks’ can surface as [ræχmæt], but *allah* + *-Din* ‘God + ABL’ is more likely to surface as [allahtin] or simply [alladin] rather than [allaχtin]. Diachronically there may have been a stage when the pre-consonantal [χ] allophone did indeed surface in derived environments, but at this point it seems that it only surfaces in non-derived environments. This illustrates a paradigm uniformity effect (Steriade 1999) where the phonetic properties of the bare root are extended to other forms in the paradigm even if a different surface realization could be allowed given the structure of those other forms.

The other possibility is that the final /h/ is simply no longer present in the phonemic representation for some speakers, despite being reinforced by standard orthographic practice. This seems to have happened at least for some personal names, as spellings like *Abdulla* are widely accepted as standard spellings, not just casual variants. However, the spectrograms in Figure 1 show that phonetic production varies at least between speakers, and the corpus results in Table 5 show that orthographic production varies even between authors, even newswriters, so it is difficult to make any broad statement of whether /h/ is in the underlying representation for any given word in the lexicon overall and even for any given word in an individual speaker’s lexicon.

These findings support the increasing trend in the field of looking at phonological phenomena as probabilistic rather than categorical. While phonetic data would be needed to confirm this, the adaptation of Arab-Persian loanwords in Uyghur is probably best viewed as a gradual process of diffusion rather than a uniform change, affecting different words at different rates for different speakers. Orthographic practice may eventually end up dropping the *h* from these words altogether. At any rate, the analyses conducted in this study were not able to identify clear trends.

Looking ahead to future work, one limitation of this study is that none of the data comes from texts produced in China, where the majority of the Uyghur population resides. To remedy this, I am planning to prepare a corpus of published works and possibly online communication from China-based sources. This study also did not take into account the possibility of regional variation, as information about the geographical origins of the authors of the texts were obvious-

ly not available, but that may be part of the picture too. Other future directions could include exploring the social implications or indexical meaning of orthographic variation involving *h* (and potentially other letters), and looking more systematically at the production of /h/ in audio data.

Abbreviations used in glossing

ABL	ablative	LOC	locative
ACC	accusative	PL	plural
CAUS	causative	PST	past
DAT	dative	PTCP	participle
DESID	desiderative	PTCP.REAS	participle of reason
DESID.SIMUL	desiderative simulative	PURP	purposive
GEN	genitive	SIMUL	simulative
LIM	limitative		

References

Bladon, Anthony. 1986. Phonetics for hearers. In Graham McGregor (ed.), *Language for Hearers*, 1-24. Pergamon Press, Oxford.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, 69-94. KNAW. http://hdl.handle.net/11858/00-001M-0000-0013-1A32-0

Bybee, Joan and Sandra Thompson. 1997. Three frequency effects in syntax. In *Annual Meeting of the Berkeley Linguistics Society* 23(1). 378-388.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2). 108-127. https://doi.org/10.1016/j.newideapsych.2007.02.002

Eisenstein, Jacob. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2). 161-188. https://doi.org/10.1111/josl.12119

Gordon, Matt, Paul Barthmaier, and Kathy Sands. 2002. A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*(2). 141-174. https://doi.org/10.1017/S0025100302001020

Gries, Stefan Th. and Caroline David. 2007. This is kind of/sort of interesting: variation in hedging in English. In Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen, & Jukka Tyrkkoö (eds.), *Towards multimedia in corpus linguistics*. Studies in variation, contacts and change in English 2, University of Helsinki. https://ufal.mff.cuni.cz/~cinkova/2015/docs/2007_STG-CVD_KindOfSortOf_MultMedCorpLing_0.pdf

Hahn, Reinhard F. 1991. *Spoken Uyghur*. Seattle: University of Washington Press.

Heyd, Theresa. 2016. Global varieties of English gone digital: Orthographic and semantic variation in digital Nigerian Pidgin. In Lauren Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Topics in English Linguistics (93). 101-122. De Gruyter Mouton. https://doi.org/10.1515/9783110490817

Ilbury, Christian. 2020. "Sassy queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, 24(2). 245-264. https://doi.org/10.1111/josl.12366

Iorio, Joshua B. 2010. *Explaining orthographic variation in a virtual community: Linguistic, social, and contextual factors*. Austin: University of Texas at Austin dissertation. http://hdl.handle.net/2152/ETD-UT-2010-05-727

Joyce, Terry, Hodošček, Bor. and Nishina, Kikuko. 2012. Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language & Literacy*, *15*(2). 254-278. https://doi.org/10.1075/wll.15.2.07joy

Mayer, Connor. (in prep). Gradient opacity in Uyghur backness harmony. https://linguistics.ucla.edu/people/grads/connormayer/papers/cmayer_uyghur_opacity_in_prep.pdf

Nazarova, Gulnisa and Niyaz, Kurban. 2013. *Uyghur: An Elementary Textbook*. Georgetown University Press.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwarz, H.G., 1992. *An Uyghur-English Dictionary*. Center for East Asian Studies, Western Washington University.

Steriade, Donca. 1999. Phonetics in phonology: the case of laryngeal neutralization. *UCLA Working Papers in Linguistics,* vol. 2, (October 1999).

Stewart, Ian, Chancellor, Stevie, De Choudhury, Munmun and Eisenstein, Jacob. 2017. December. # Anorexia,# anarexia,# anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE International Conference on Big Data (Big Data),* 4353-4361. IEEE. 10.1109/BigData.2017.8258465

Sullivan, Natalie. 2017. *Writing Arabizi: Orthographic variation in Romanized Lebanese Arabic on Twitter*. Austin: University of Texas at Austin dissertation. http://hdl.handle.net/2152/72420

Tüzdi, Sulayman. 2002. *Uyghur tili (fonëtika) [The Uyghur language (phonetics)]*. Ürümchi: People's Press of Xinjiang.

van Compernolle, Rémi A. and Williams, Lawrence. 2010. Orthographic variation in electronic French: The case of l'accent aigu. *The French Review*. 820-833. https://www.jstor.org/stable/40650541

Washington, Jonathan. In draft. Vowel harmony in Turkic languages. *Oxford Handbook of Vowel Harmony*.

Wulff, Stefanie. and Stefan Th. Gries. 2019. Particle placement in learner language. *Language Learning*, *69*(4). 873-910. https://doi.org/10.1111/lang.12354

Zakir, Hamit A. 2007. *Hazirqi zaman Uyghur tili: Introduction to Modern Uighur.* Urumqi: Xinjiang University Press.