The empirical consequences of data collection methods: A case study from Kazakh vowel harmony

Adam G. McCollum
University of California San Diego

Empirical data is crucial to all subdisciplines of linguistics. As a result, various subdisciplines have developed best-practices to ensure the integrity of linguistic research. This paper focuses on several methodological concerns from experimental and field research. The paper argues that fieldwork should be guided by best-practices from both, focusing on: stimulus ordering, register differences, and the effect of orthography. The paper describes and challenges results from a recent paper, Bowman & Lokshin (2014), which reports an unusual non-local interaction in Kazakh vowel harmony. Specifically, Bowman & Lokshin (2014) claims that two exceptional suffixes, the comitative and infinitive suffixes, exhibit what Mahanta (2012) calls "idiosyncratic transparency." Using data from colloquial and literary Kazakh, this paper argues that the data in Bowman & Lokshin (2014) are artefactual, and do not represent any known variety of Kazakh. The three methodological concerns discussed cross-cut both experimental and field methodologies, and the divergent results reported in Bowman & Lokshin (2014) serve to highlight their importance for linguistic fieldwork.

1. Introduction[*]

Modern linguistics has a variety of academic forebears, and the influence of each is evident in the methods used in linguistic research. Historically, linguistics has borrowed heavily from anthropology, employing fieldwork as a primary data collection strategy. Accordingly, best-practices for field research have developed throughout the last century. These often focus on how to work with speakers and communities, and how to elicit, record, and analyze data. Field methodologies emphasize the important interpersonal and intercultural issues that arise during fieldwork. The importance of these is manifest in the number of recent volumes devoted to fieldwork (descriptive, theoretical and documentary fieldwork; e.g. Abbi 2001; Newman & Ratliff 2001; Ladefoged 2003; Ameka et al. 2006; Gippert et al. 2006; Vaux et al. 2007; Bowern 2008; Chelliah & De Reuse 2011). On the other hand, decidedly experimental subfields in linguistics have emerged over the last half century, adopting methodologies from psychology and cognitive science.[1] The importance of experimental methods is similarly evident in its own growing body of literature (Cowart 1997; Sprouse 2007; Podesva & Sharma 2013; de Groot & Hagoort 2017). In high-resource languages, the use of varied data collection methods has become increasingly common. In lower-resource languages though, it is often challenging to find corpora, design experimental materials, and utilize the various technologies commonly found in contemporary linguistic research. In this paper, I argue that using a variety of data collection methods, even in understudied and lower-resource languages, greatly improves the quality of the resultant analysis.

---

[1]Modern linguistics certainly has other academic forebears, like philology, philosophy, and mathematics. These are not relevant for the paper, and so are not discussed.

I discuss the empirical consequences of several issues that cross-cut experimental and field research for the analysis of exceptionality in Kazakh vowel harmony. More specifically, I contend that principled stimulus ordering, register differences, and orthography all play a role in addressing the claims made in a recent paper on exceptionality in Kazakh (Bowman & Lokshin 2014). I argue that findings in Bowman & Lokshin (2014) are an artefact of their methodological choices, demonstrating the empirical and theoretical consequences of different data collection strategies.

The paper is organized as follows. In §2, I briefly discuss the overlap between experimental and field methodologies, laying out the three topics relevant to the discussion of Kazakh, stimulus ordering, register differences, and orthographic effects. In §3, I describe vowel harmony in Kazakh. There I also present the data reported in Bowman & Lokshin (2014; henceforth B&L). In §4, I briefly introduce the sociolinguistic factors relevant to the study, going on in §§5-6 to show that data from colloquial and literary Kazakh critically differ from B&L's description of the comitative suffix. §7 presents experimental results that demonstrate the consequences of stimulus ordering and suggests a possible explanation for our divergent findings. §§8-9 goes on to show that the realization of the infinitive suffix in both the colloquial and literary registers differs from the description in B&L. Finally, in §10, I discuss the descriptive and theoretical implications of the paper.

2. Experimental and field methodologies

Fundamentally, linguistic research follows the hypothetico-deductive method.[2] Given some amount of extant data, the linguist forms a hypothesis, which is then tested and recorded. The data recorded from the experiment, whether it be a casual elicitation session in the bush or an ultrasound study in a laboratory, are then used to re-evaluate the original hypothesis. Between hypothesis formation and actual testing, though, exists a planning phase, where the researcher determines which methods are most appropriate to address the question at hand. These may range from passive participation in a community event to visual masked priming. In each case the linguist chooses which data collection method is most appropriate to answer the question at hand (e.g. Yao & Scheepers 2011; Schütze & Sprouse 2013; Tonhauser & Matthewson 2015). In these ways, both the experimentalist and the fieldworker engage in the same larger program of hypothesis generation, testing, and evaluation.

There are numerous factors that inform the hypothetico-deductive method. Here I briefly touch on three methodological choices that relate to the Kazakh data to be discussed, stimulus ordering, register, and orthography. First, the importance of stimulus ordering is well-attested. In the early stages of fieldwork, this is often impossible, but in later stages, when specific hypotheses about the language under study are being considered, more controlled data collection methods become feasible. Specifically, the order of stimulus items has been shown to influence responses in both laboratory and field settings (Bock 1986; Snyder 2000; Bickel et al. 2007; Pickering & Ferreira 2008; Caballero 2010; Yu 2014). To reduce potential confounding effects, it has been common

---

[2]This conception of linguistics is rejected by some, particularly documentary linguists who view their role in a community as fundamentally different (see e.g. Gippert et al. 2006; Bowern 2008 for discussion). The reasons that underlie the more community-led, participatory approach are certainly valid. Even in community-based work, the curious linguist often wonders about what happens when *x*, or why does *y* happen. If these questions motivate questions, elicitation sessions, or corpora-based analyses, then the documentarian is engaged in the hypothetico-deductive method. For this reason, I also subsume community-led research under the same broad conception of linguistic research.

since Fisher (1935) to randomize stimuli. As stimulus-ordering effects (priming) has been shown to affect lexical access, syntactic structure, morpheme ordering, and tone, among other phenomena, ordering is relevant to most, if not all, linguistic research. In short, stimulus ordering is an important tool to help ensure the validity of one's data, both in the field and in the lab.

Second, register differences may affect linguistic patterning (Biber 1993, 1995, 2012; Face 2003). In essence, the context, including factors like formality, the modality of communication, and the specific interlocutors present change linguistic behavior. For instance, Face (2003) demonstrates that intonational contours in Catalan significantly differ in spontaneous speech and "lab speech" (see also Xu 2010). Thus, it is crucial to know which register is being elicited, as well as the relevant properties of the target register. In many cases, a particular result may only hold within a certain register and may not generalize to other varieties of the language.

Third and finally, orthography exerts a significant influence on linguistic performance (see Derwing 1992 for arguments on the influence of orthography on linguistic competence, too). Some experimental studies have argued that orthographic knowledge interacts with phonological knowledge (e.g. Damian & Bowers 2003; Perre et al. 2010); fieldworkers and sociolinguists have written a great deal about the effects of orthography on variation, identity, and language maintenance (Seifart 2006; Sebba 2007; Essegbey 2015). It is therefore important to consider the potential effects of orthography on the target register and target phenomenon.

In some cases though, it is difficult or impossible to avoid using orthographic representations. Most methods of data collection come with certain drawbacks, which are most effectively minimized by the use of multiple complementary methods. For instance, the effects of syntactic priming can be tested using orthographically-based methods, like self-paced reading, or by aural presentation of the target stimuli. In like manner, the fieldworker may want to test some phonological hypothesis using orthographic as well as pictorial prompts. The use of multiple methods allows the researcher to understand more fully the phenomenon in question, and also the differences that emerge from the various modalities employed during elicitation.

The three factors just discussed, stimulus ordering, register differences, and orthography, in addition to the general importance of multiple converging methods, will all factor into the discussion of Kazakh.

3. Locality and Kazakh vowel harmony

In this section I discuss vowel harmony in Kazakh. I first introduce the role of exceptionality in vowel harmony, laying out Mahanta's (2012) claim that all exceptions are local, as well as B&L's counterclaim from Kazakh. From there, I describe the general pattern of backness harmony in Kazakh, which lays the foundation for the findings reported in subsequent sections.

3.1. Locality and exceptions in vowel harmony

In vowel harmony, some vowel determines the realization of another. This dependency is often argued to be local, precluding long-distance effects in harmony (e.g. Gafos 1999; Baković 2000; Ní Chiosáin, & Padgett 2001). Consider the Turkish example below. Observe in (1a-b) that /a/ and /e/ regularly alternate for backness harmony in Turkish. In the plural suffix, /a/ occurs after back vowels while /e/ occurs after front vowels.[3] However, in exceptional suffixes, this dependency is

---

[3]The following glosses are used in the paper: 1= first person, 2= second person, 3= third person, ABL= ablative, ACC= accusative, ADJ= adjectival, AGT= agentive, COM= comitative, FORM= formal, GEN= genitive, INF= infinitive,

violated. One example is the polygon-forming suffix, /-gen/, which does not alternate based on the backness of the root (1c-e). Following back vowels, the polygon-forming suffix still surfaces with a front vowel, (6d-e). When /-gen/ occurs, the iterative spreading of root vowel backness is interrupted. When additional suffixes follow the exceptional polygon-forming suffix, in every case the polygon-forming suffix imposes its own backness on the subsequent vowel. In Turkish, /-gen/ *blocks* harmony, since it determines the realization of the PL suffix that follows it.

(1)   Exceptionality in Turkish (Clements & Sezer 1982)
       a.   dal-lar   'branch-PL'
       b.   el-ler   'hand-PL'
       c.   ytʃ-gen-ler   'three-PLGN-PL'
       d.   altɯ-gen-ler   *altɯ-gen-lar   'six-PLGN-PL'
       e.   tʃok-gen-ler   'many-PLGN-PL'

One could imagine, however, another kind of exception, where the morphological root controls the realization of PL regardless of what intervenes. In (1d), this kind of non-local interaction is shown. The possible form, *altɯ-gen-lar is ungrammatical in Turkish. In this type of exception, the backness of the root skips over /-gen/ to determine the proper allomorph of PL. In other words, the exceptional morpheme is skipped for harmony. In this scenario, the exceptional morpheme is *transparent*.[4] Blocking and transparency are depicted in Table 1 below using autosegmental association lines (Goldsmith 1976). In the blocking cell, all phonological interactions are local. The polygon-forming suffix does not undergo harmony, so locality requires that the following suffix agree in backness with the exceptional morpheme. In the transparency cell though, harmony is non-local, since the backness of the vowel preceding /-gen/ determines the backness of the vowel following the exceptional morpheme. Transparency is represented by crossing autosegmental lines, which is generally prohibited in autosegmental frameworks (see also Pulleyblank 1983; Archangeli & Pulleyblank 1994).

Blocking /-gen/	Transparent /-gen/ (unattested)

a l t ɯ – g e n – l e r
   | | |⌐‐‐‐
[+bk][+bk][-bk]
	a l t ɯ – g e n – l a r
	  | |‐‐‐‐‐╆‐‐‐
	[+bk][+bk][-bk]

Table 1: Blocking and (unattested) transparency in Turkish

Mahanta (2012) makes the strong claim that locality governs all exceptionality in harmony. Under her analysis, at a definitional level, all exceptional morphemes block harmony (cf. Finley 2010). In contravention of Mahanta's claim, B&L report that Kazakh possesses two exceptional suffixes that do not block harmony, but are, in fact, transparent. B&L describes two patterns of what Mahanta calls "idiosyncratic transparency" in Kazakh backness harmony, shown below. In (2a-b),

---

LOC= locative, NMZLR= nominalizer, NPST= non-past, PFV= perfective, PL= plural, PLGN= polygon, POSS= possessive, Q= question, S= singular. Suffixes are preceded by "-" and enclitics are preceded by "=."
[4]This does not preclude transparency, in general. There are many cases where a vowel's phonological status renders it transparent to harmony (van der Hulst & van der Weijer 1995; Vaux 2000; Svantesson et al. 2005). This type of exceptionality is distinct from the morphologically-conditioned exceptionality discussed here.

/i͡e/ regularly alternates with /aː/ for harmony. However, B&L argue that the comitative suffix, COM, does not undergo harmony, but still allows the root vowel's [back] feature to determine the realization of a following question enclitic. This interaction parallels the unattested variant of the Turkish data in (1d).

(2) Transparent COM
    a.   ki͡el-gi͡en=bi͡e   'come-PFV=Q'
    b.   qaːl-ʁaːn=baː   'stay-PFV=Q'
    c.   bəbi͡ek-pi͡en=bi͡e   'baby-COM=Q'
    d.   naːn-mi͡en=baː   'bread-COM=Q'

Likewise, in (3a-b), initial-syllable /u͡w/ regularly triggers back vowel suffixes in Kazakh. Yet, the infinitive suffix (INF), like COM, fails to undergo backness harmony, also allowing the root's backness to determine the backness of the next morpheme (3c-d).

(3) Transparent INF
    a.   tu͡w-də   'flag-ACC'
    b.   ru͡w-də   'clan-ACC'
    c.   ʒaːz-u͡w-də   'write-INF-ACC'
    d.   kɛr-u͡w-dɛ   'enter-INF-ACC'

The findings in B&L thus directly contradict the claims advanced in Mahanta (2012). This paper focuses on the empirical claims made in Bowman & Lokshin (2014), arguing against their description of COM and INF. I demonstrate that COM is not transparent but blocks harmony. Further, I show that INF is not even exceptional, but regularly undergoes harmony. As a result, I argue that B&L's claims should be regarded with caution, and that Kazakh does not instantiate a pattern of exceptional transparency. Rather, I suggest that two distinct registers, literary and colloquial Kazakh, stimulus-ordering, and orthographic effects all played a role in the surprising data described in B&L.

3.2. Kazakh vowel harmony

The Kazakh vowel inventory consists of at least the following nine phonemes, /ɛ ʏ i͡e ɵ ː æ ː a ː ɔ ː ə ɔ/, and potentially two more phonemes, /i͡j u͡w/ (McCollum & Chen submitted). The number of phonemes has been contested, and researchers have typically posited the nine phonemes above, excluding /i͡j/ and /u͡w/ (Dzhunisbekov 1972; Kirchner 1998; Muhamedowa 2015; Washington 2016). Using the features, [back], [high], [low], and [round], I assign contrastive features to the inventory in Table 2 below.

    In addition to these contrastive features, I use moras, μ, to differentiate the long and short vowels. The high and the low vowels are all long, and as a result, bimoraic. In contrast, the mid vowels contrast for length, and may either be monomoraic (short) or bimoraic (long). The length contrast seems to be emerging from what was likely a height contrast. The vowels described as high by previous writers are now produced as mid vowels and differ from the historical mid vowels

in that they are very short (Johanson 1998). The long vowels are over twice as long as the short vowels (Washington 2016; McCollum 2018; McCollum & Chen accepted).[5]

		[-back]		[+back]	
		[-round]	[+round]	[-round]	[+round]
[+high]		$\widehat{ij}$			$\widehat{uw}$
[-high, -low]	μ	ɛ	ʏ	ə	ɔ
	μμ	$\widehat{ie}$	ɵː		ɔː
[+low]		æː		aː	

Table 2: Feature chart for the Kazakh inventory

Typical for a Turkic language, Kazakh exhibits backness (or palatal) harmony. In most cases, the backness of the initial vowel determines the backness of all subsequent vowels (Balakaev 1962; Dzhunisbekov 1972, 1980; Kirchner 1992, 1998; Kara 2002; Muhamedowa 2015). This is true both within roots and in suffixes, as demonstrated below.

In (4a-j), observe that only /$\widehat{ie}$/, /ɛ/, and /ʏ/ may follow the front vowels, /æː $\widehat{ie}$ ɵː ɛ ʏ/. In (1k-r), only /aː/, /ə/, and /ɔ/ may follow the back vowels, /aː ɔː ə ɔ/. Observe also that the dorsal obstruents /k/ and /q/ are subject to the same co-occurrence restriction. The more posterior phoneme, /q/, occurs with [+back] vowels while the more anterior phoneme, /k/, occurs with [-back] vowels. There are some exceptions to this, but these exceptions are almost always foreign loans.

(4) Backness harmony within roots

[-back] roots   [+back] roots

a. æːr$\widehat{ie}$ŋ   'barely'   k. qaːlaː   'city'

b. æːlɛ   'yet'   l. qaːzə   'horse sausage'

c. k$\widehat{ie}$z$\widehat{ie}$k   'turn'   m. qəraːn   'hawk'

d. $\widehat{ie}$sɛk   'door'   n. qərəq   'forty'

e. tɵːb$\widehat{ie}$ 'hill'   o. bɔːlaːt   'steel'

f. kɵːsʏk   'desert carrot' p. qɔːzə   'lamb'

g. tɛz$\widehat{ie}$ 'knee'   q. qɔlaːq   'ear'

h. kɛsɛ   'person'   r. qɔlɔn   'colt'

i. tʏl$\widehat{ie}$k   'graduate'

j. ʒʏzʏk   'ring'

Backness harmony applies to suffixes, as well. In (5), only [-back] may follow [-back] roots. Moreover, in (6), only [+back] vowels may follow [+back] roots. Specifically, observe the

---

alternations for the locative and accusative suffixes. The locative suffix alternates between /-tie͡/ and /-taː/ in (5a-e) and (6a-e). The accusative suffix alternates between /-tɛ/ and /-tə/ in (5f-j) and (6f-j).[6]

(5)   Backness harmony in suffixes after [-bk] roots

a.	sæːt-ti͡e	'fortune-LOC'		f.	sæːt-tɛ	'fortune-ACC'
b.	i͡es-ti͡e	'memory-LOC'		g.	i͡es-tɛ	'memory-ACC'
c.	tɵːs-ti͡e	'chest-LOC'		h.	tɵːs-tɛ	'chest-ACC'
d.	tɛs-ti͡e	'tooth-LOC'		i.	tɛs-tɛ	'tooth-ACC'
e.	tʏs-ti͡e	'dream-LOC'		j.	tʏs-tɛ	'dream-ACC'

(6)   Backness harmony in suffixes after [+bk] roots

a.	taːs-taː	'stone-LOC'		f.	taːs-tə	'stone-ACC'
b.	qɔːs-taː	'hut-LOC'		g.	qɔːs-tə	'hut-ACC'
c.	təs-taː	'outside-LOC'		h.	təs-tə	'outside-ACC'
d.	qɔs-taː	'bird-LOC'		i.	qɔs-tə	'bird-ACC'
e.	tu͡w-daː	'flag-LOC'		j.	tu͡w-də	'flag-ACC'

In (7), backness harmony is iterative, affecting both short and long vowels alike. Note that the question enclitic undergoes harmony in these examples, which derive from the literary register. The differences between Q in the literary and colloquial registers will figure prominently in §4-6 (see also Muhamedowa 2015:282).

(7)   Iterative backness harmony

a.	i͡es-ti͡er = mi͡e	'memory-PL = Q
b.	i͡es-ɛ-m-ɛz = bi͡e	'memory-POSS-1-PL = Q'
c.	taːs-taːr = maː	'stone-PL = Q'
d.	taːs-ə-m-əz = baː	'stone-POSS-1-PL = Q'

Backness alternations are encoded orthographically in Kazakh. For instance, the plural suffix in (8) has two orthographic variants, <тер>and <тар>, which are used after front and back vowel stems, respectively. The question enclitic, which is separated from the stem by a space, marking its status as an enclitic, also exhibits orthographic variation according to backness harmony. In (8a&c), the question enclitic is written as <ме> after a front vowel stem and <ма> after a back vowel stem. Also, the initial consonant of the enclitic is determined by the sonority of the immediately preceding segment, with <ме> occurring after more sonorous and <бе> after less sonorous segments.

---

[6]Kazakh has a few relatively unproductive prefixes, which are typically of Persian origin. These prefixes do not undergo harmony, as in /bi͡ej-taːnəs/ [bi͡ejtaːnəs] 'PRV-acquaintance.'

(8)   Iterative backness harmony

	Phonology	Orthography	Gloss
a.	i͡es-ti͡er = mi͡e	естер ме?	'memory-PL = Q
b.	i͡es-ɛ-m-ɛz = bi͡e	есіміз бе?	'memory-POSS-1-PL = Q'
c.	taːs-taːr = maː	тастар ма?	'stone-PL = Q'
d.	taːs-ə-m-əz = baː	тасымыз ба?	'stone-POSS-1-PL = Q'

In addition to backness harmony, Kazakh exhibits rounding harmony, as is evident in some of the root-internal alternations shown in (9, see 9f,j&r). Rounding harmony is typically non-iterative (Balakaev 1962:102-103; Kirchner 1998:320-321; McCollum 2018; McCollum & Chen accepted). In contrast to backness harmony, rounding harmony is not encoded orthographically (compare [ʒʏzʏk] and <жүзік> 'ring'). Rounding harmony will not be further discussed in the paper.

3.3. Comitative suffix

The comitative suffix, /-mi͡en/, is one of the only invariant suffixes in the language (Krippes 1993; Kirchner 1998; Kara 2002; Muhamedowa 2015).[7] This suffix surfaces with the front vowel /i͡e/ regardless of the vowel that precedes it. The suffix onset surfaces as /m/ after sonorants, (9a-e), as /b/ after voiced obstruents, (9f), and as /p/ after voiceless obstruents (9g-j). Most importantly, COM is realized with a front vowel regardless of preceding vowel quality, (compare 9a-g with 9h-j).

(9)   Comitative suffix

	Phonology	Orthogaphy	Gloss
a.	aːptaː-mi͡en	аптамен	'week-COM'
b.	aːj-mi͡en	аймен	'moon-COM'
c.	naːr-mi͡en	нармен	'dromedary-COM'
d.	taːl-mi͡en	талмен	'willow-COM'
e.	naːn-mi͡en	нанмен	'bread-COM'
f.	qaːz-bi͡en	қазбен	'goose-COM'
g.	taːs-pi͡en	таспен	'stone-COM'
h.	sæːt-pi͡en	сәтпен	'fortune-COM'
i.	i͡es-pi͡en	еспен	'memory-COM'
j.	ɛs-pi͡en	іспен	'work-COM'

Recall from above that /i͡e/ regularly participates in harmony, both as a trigger and undergoer of harmony, as in /i͡es-ti͡er-ɛ-m-ɛz=bi͡e/ 'memory-PL-POSS-1-PL=Q' from (7b). Thus, it is not the

---

[7]B&L describe this affix as the comitative only, but the same affix functions as both instrumental and comitative in the language (Kirchner 1998:327). I use the term comitative to maintain consistency with B&L. The semantic distinctions between the comitative and instrumental are not relevant for the purposes of this paper, as they do not affect the phonological behavior of this suffix. Orthographically, the comitative is separated from the noun by a space while the instrumental is not.

feature specification of /ĩe/ that prevents harmony on COM. For vowels that are not exceptions to vowel harmony for featural reasons, Mahanta (2012) contends that these vowels ontologically block harmony in exceptional morphemes.

The data in (10) show the effect of COM on subsequent morphemes. B&L is the first to systematically investigate harmony on morphemes following the exceptional comitative suffix. They find that COM is transparent to harmony. Thus, the [back] feature of the preceding vowel determines the backness of the vowel following COM, as shown below. As far as I know, the question enclitic is the only morpheme that may follow COM.

(10)   Harmony after the comitative suffix (Bowman & Lokshin 2014:5)
         Phonology   Orthography   Gloss
    a.   naːn-mĩen = baː   нанмен бе   'bread-COM = Q'
    b.   bɵːbĩek-pĩen = bĩe   бөбекпен бе   'baby-COM = Q'

In summary, according to B&L, COM is exceptionally (idiosyncratically) transparent. First, this morpheme is exceptional because /ĩe/ is invariant in this morpheme, while in all other contexts participates in harmony. Second, /ĩe/ of COM is transparent because it does not spread its own backness feature but allows the backness of the preceding vowel to determine the backness of the following vowel.

3.4. Infinitive suffix

In addition to the comitative suffix, B&L report that the infinitive suffix is also invariant. Like COM, they find that INF is transparent to backness harmony.[8] The infinitive suffix is represented orthographically by <y>. Traditionally, this grapheme has been assumed to represent a regularly alternating high round vowel (Balakaev 1962; Dzhunisbekov 1972, 1980; Vajda 1994; Kirchner 1998; cf. Kara 2002). Thus, after front vowel stems, this grapheme is reported to represent surface [ʏw] while after back vowel stems the <y> of INF represents surface [ɔw], demonstrated in (11). Unlike the other vowels that alternate for harmony, this alteration is not encoded orthographically. Orthographic <y> is used to represent both [ʏw] and [ɔw]. This fact will play a role in the discussion of B&L's findings later, in §9.2.

(11)   Reported harmonization of INF (Vajda 1994:626)
         Phonology   Orthography   Gloss
    a.   dĩe-ʏw   деу   'say-INF'
    b.   aːw-ɔw   ауу   'overturn-INF'

In contrast, B&L finds only variable evidence for the harmonization of INF. One of the two speakers they consulted showed a clear difference in INF based on the preceding vowel. The other speaker, however showed almost no effect of preceding vowel on the backness of INF. INF shows clear differences in F2, the main acoustic correlate of backness for Speaker 1 below (left) but not

---

[8]B&L call this morpheme the infinitive suffix, which I follow for clarity. This affix functions more like a gerundial rather than an infinitive suffix.

for Speaker 2 (right). Vowel plots for each speaker's surface vowel inventory with /u/ 'INF' after front and back vowels are shown below:
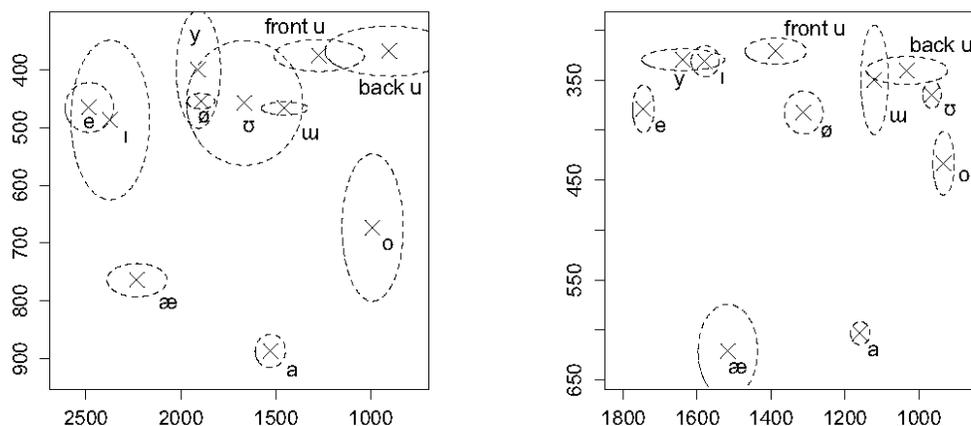

Figure 1: F1-F2 vowel plots for the two speakers consulted in Bowman & Lokshin (2014:4). The vowel plot for Speaker 1 is on the left, and the plot for Speaker 2 is on the right (Note their /y/ = my /ɤ/ and their /ʊ/ = my /ɔ/, among other differences in transcription)

B&L tentatively concludes that INF is transparent but phonetically affected by backness harmony, preserving an underlying [+back] feature (see also Kara 2002:9). Like COM, they argue that INF is transparent because affixes following INF bear the backness of the vowel preceding INF, as demonstrated below. In (12a-b) the accusative suffix surfaces with a front vowel, agreeing with the initial vowel rather INF. In (12c) the vowel of the accusative suffix agrees with both the initial stem and INF, as both are back vowels.

(12)   Backness harmony after INF (Bowman & Lokshin 2014:2)
        Phonology   Orthography   Gloss
   a.   ʒɤz-u͡w-dɛ   жүзуді   'swim-INF-ACC'[9]
   b.   kɛr-u͡w-dɛ   кіруді   'enter-INF-ACC'
   c.   ʒaːb-u͡w-də   жабуды   'close-INF-ACC'

It is also relevant to recall from (3) that <y> may occur in initial syllables too. When the grapheme <y> occurs in an initial syllable, it represents the high back vowel, /u͡w/. This vowel triggers [+back] suffixes, and is a regular trigger for harmony, as shown below. Thus, orthographic <y> represents a non-alternating [+back] vowel in initial syllables but an alternating vowel elsewhere.

(13)   Backness harmony after initial /u͡w/
        Phonology   Orthography   Gloss
   a.   su͡w-daː   суда   'water-LOC'
   b.   qu͡w-laːr   құлар   'crafty.person-PL'
   c.   tu͡w-də   туды   'flag-ACC'
   d.   bu͡w-dəŋ   будың   'steam-GEN'

---

[9]B&L (2014:2) incorrectly glosses this word as 'take-INF-POSS.3.'

Table 3 schematizes B&L's claims. Both COM and INF are invariant, surfacing as /i͡e/ and /u͡w/, respectively, regardless of preceding vowel backness. However, morphemes that follow these invariant suffixes, like Q and ACC, undergo long-distance harmony from the vowel preceding invariant COM or INF. The transparency of COM and INF is represented by the crossing autosegmental association lines below (Goldsmith 1976).

COM					
after [-back] stem			after [+back] stem		
COM	Q	Schema	COM	Q	Schema
i͡e	i͡e	Stem-COM-Q |___---|--- [-bk] [-bk]	i͡e	aː	Stem-COM-Q |___---|--- [+bk] [-bk]

INF					
after [-back] stem			after [+back] stem		
INF	ACC	Schema	INF	ACC	Schema
u͡w	ɛ	Stem-INF-ACC |___--|-- [-bk] [+bk]	u͡w	ə	Stem-INF-ACC |___--|-- [+bk] [+bk]

<div align="center">Table 3: Schematization of B&L's claims</div>

4. Two relevant influences on Kazakh phonology

Before addressing the phonological behavior of COM, Q, and INF in Kazakh, it is first important to note the influence of Russian and register differences in Kazakh. §4.1 describes some of the effects Russian has had on the Kazakh language, and §4.2 describes some of the distinguishing features and domains of the literary and colloquial registers in the language.

4.1. Russian influences on Kazakh phonology

From the eighteenth century through the Soviet Era (1917-1991), Russian influence in Kazakhstan monotonically increased. In fact, throughout the Soviet era and until the late 1990s Kazakhs did not constitute a majority in their own republic. Kazakhs were 81.7% of the population in 1897, but in 1989, only one century later they constituted only 40.1% of the population in the Kazakh Republic (Dave 2007). In addition to Slavic peoples, the Soviet Union moved Germans, Koreans, and peoples from the Caucasus to Kazakhstan in large numbers. The resultant diversity necessitated a bilingual population, which in conjunction with Russification policies also reduced the domains of usage for Kazakh (see Dave 1996, 2004, 2007; Fierman 1998; Grenoble 2003:196-197). As Dave (1996) notes, the majority of urban Kazakhs speak Russian as their first language. Though this tendency toward Russian is changing since Kazakh independence, the tendency is still quite pervasive.

There are several ways in which Russian exerts a significant effect on Kazakh phonology. First, Russian-dominant Kazakh speakers are far more likely to produce disharmonic words. They are less likely to produce words with suffix alternations than non-Russian-dominant speakers. Second, Russian-dominant speakers have more trouble with Kazakh phonemes that do not exist in

Russian, like /q/, /ə/, and /ʏ/. These troublesome sounds fall into two groups. First, some sounds are represented with orthographic characters not present in Russian orthography. The sounds /q/ and /ʏ/ fall into this category, since they are represented by orthographic <қ> and <ү> in Kazakh. The voiceless uvular stop is often produced as a velar, either /k/ or /x/, and the front round vowel /ʏ/ is often produced like /u/ by Russian-dominant speakers. The second class of troublesome sounds share the same grapheme but represent a different phoneme. Most significant among these sounds is /ə/. This Kazakh phoneme is represented orthographically as <ы>, but that same grapheme represents a high tense vowel in Russian /ɨ/ (or [i] under other analyses). The conflicting status of <y> can also be problematic, as we will see later. This grapheme in Russian always represents a high back vowel, but in Kazakh this grapheme represents a high back vowel in initial syllables but alternating [ʏw]~[ɔw] non-initially. Ignoring vowel quality alternations due to stress, Russian represents each phonemic contrast in all positions. Generally, Kazakh orthography represents backness harmony, but the one exception to this is <y>, which represents a back vowel in initial syllables but an alternating vowel in non-initial syllables.

The effects of Russian are more common in the speech of Kazakhs from northern and central Kazakhstan, due to the high percentage of Russians living in those areas (Fierman 1998:173-175). Kazakhs from other regions often comment that Kazakh in northern Kazakhstan speak almost entirely in Russian. Kazakh in central and northern Kazakhstan are more likely to be educated in Russian and are more likely to grow up with Russian neighbors and friends than Kazakh residing in more southerly regions of the country.

4.2. Literary and colloquial Kazakh

Colloquial and literary Kazakh exhibit some significant differences. Prior to the 19th century, there are few records of the Kazakh language. Jankowski (2012:26) notes "there is a great difference between written and spoken Kazakh, and it must have been so ever since the first Kazakh texts appeared." That being said, there is a more significant oral literary tradition, consisting of oral epics, music, poems and poetic dueling. During the 20th century, under the influence of Russian Kazakh literature began to emerge. However, Jankowski (2012:30) correctly observes that "modern Kazakh literature only has a minimal effect on spoken language." Children are taught the literary language in school, hear it spoken on the news and at formal events, but many Kazakhs do not command the literary register.

Literary Kazakh is differentiated from colloquial Kazakh in a number of ways. Perhaps most noticeable is the relative lack of Russian code-switching in the literary language. While very little of the Kazakh lexicon has been unaffected by Russian, in the literary register there is a conscious effort to purify the language of these foreign influences (Jankowski 2012:25-31). Grammatically, literary Kazakh is marked by increased morphological and syntactic complexity (e.g. Muhamedowa 2015:47-48) and exhibit distinct phonological patterns. One phonological difference between the two registers is rounding harmony. In literary Kazakh, it is far more pervasive, occurring more frequently and extending its influence further throughout the word than in colloquial Kazakh where it is variable and typically non-iterative (Balakaev 1962:102-103; Abuov 1994).

5. The comitative suffix in colloquial Kazakh

I conducted fieldwork on colloquial Kazakh in June 2014. Over fifteen hours of colloquial data were gathered though semi-formal conversational elicitation using the target language as the contact language. Data was collected from thirteen speakers (9 females, 4 males) residing in and around Taldykorgan, Kazakhstan. Data from two speakers were excluded because Kazakh was not their dominant language. Speakers ranged in age from 19 to 46, with a mean age of 33.5 years. Ten speakers were born in Kazakhstan, while one speaker was born in Mongolia. Among the 10 speakers from Kazakhstan, 7 were from southeastern Kazakhstan, and the 3 remaining speakers came from north-central, eastern, and southern Kazakhstan. Speakers also varied by educational achievement. Three speakers had master's degrees, one had a terminal bachelor's degree, eight had terminal high school diplomas, and one had completed some high school. Of the 11 speakers, only one exhibited significant influence from Russian. This speaker remarked on multiple occasions that they could not remember the Kazakh word for some item, or that their Kazakh was not as good as it should be. This speaker was educated in Russian and speaks a mix of Russian and Kazakh at home. The other 10 speakers were either educated entirely in Kazakh or grew up in a small village where Kazakh was the primary language used in both the home and community. The data were recorded to a Zoom H4N recorder at a sampling rate of 44.1 kHz with a Shure unidirectional microphone. The fieldwork data presented throughout the paper were normalized (Lobanov 1971) to facilitate more appropriate across-speaker comparisons. The normalized units for F1 and F2 are (z).[10]

5.1. The comitative suffix

During data collection, the comitative suffix occurred 381 times, 218 times after front vowel stems and 163 times after back vowel stems. Table 4 presents mean and standard deviations for F1 and F2 of COM after front and back vowels. Table 4 also compares F1-F2 of COM with non-initial (i.e. alternating) /aː/ and /i͡e/. Regardless of stem backness, F2 of COM always approximates F2 of non-initial /i͡e/. In fact, mean F2 of COM is higher than that of alternating /i͡e/. In other words, COM is more peripheral than /i͡e/ that surfaces due to backness harmony.

	Mean F1 (z)	SD	Mean F2 (z)	SD
Alternating /aː/	1.17	0.55	0.07	0.34
Alternating /i͡e/	-0.39	0.45	1.08	0.31
COM after [-back]	-0.20	0.54	1.28	0.49
COM after [+back]	-0.08	0.56	1.11	0.64

Table 4: Mean F1 and F2 (z-score) with SD of alternating /aː/, /i͡e/, and COM

The data from Table 4 are plotted in Figure 2 below. Compare the realization of /aː/ and /i͡e/ in harmonic affixes (n=652 and 846, respectively) to the realization of COM after front and back vowel roots. It is clear that COM does not alternate between /aː/ and /i͡e/.

---

[10]Lobanov's normalization method involves finding the center of a speaker's acoustic vowel space. Using the center of the vowel space as a reference point, F1 and F2 are z-values derived from the statistical z-distribution (or alternatively, normal distribution) measured from the center of the speaker's acoustic vowel space.
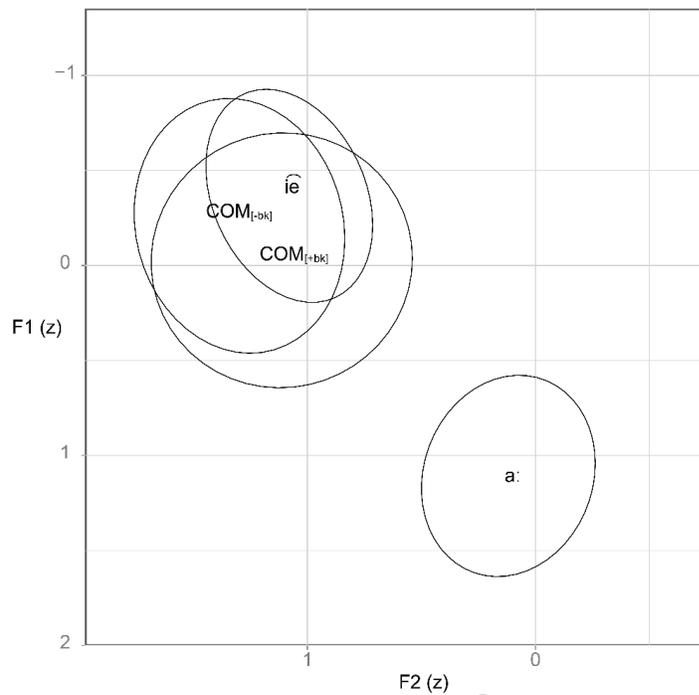
Figure 2: F1-F2 of COM in front and back vowel contexts, compared to /ie͡/ and /aː/ in alternating suffixes (in z-scores, with 1 SD ellipses)

The invariance of COM is readily attested in the descriptive literature (Balakaev 1962:157-159; Kirchner 1998:327; Kara 2002:33-34; Muhamedowa 2015). Thus, the result above is unsurprising, but serves to establish more concretely that COM in colloquial Kazakh does not alternate for backness (see also Userbaeva 2005; Niyazgalieva & Turganalieva 2013).

5.2. The realization of the question enclitic following the comitative suffix

As discussed in §5.1, the comitative suffix is invariant for backness. The most pressing issue, though, relates to the realization of the question enclitic after COM since only Q may follow COM. Traditionally, Q is treated as an alternating suffix, whose vowel varies between /aː/ and /ie͡/, depending on the backness of the stem (Balakaev 1962: 413-415; Kirchner 1998:321; Kara 2002:36-37). The traditional description of Q is demonstrated in (14) below. In (14a-f) Q is realized with /aː/ after back vowels, but with /ie͡/ after front vowels (14g-k). Note also that the alternation of the initial consonant of Q resembles that of COM in (9).

(14)   Traditional description of Q

	Phonology	Orthography	Gloss
a.	aːptaː＝maː	апта ма	'week＝Q'
b.	ɔːj＝maː	ой ма	'idea＝Q'
c.	tu͡w＝maː	ту ма	'flag＝Q'
d.	qəz＝baː	қыз ба	'girl＝Q'
e.	qɔs＝paː	құс па	'bird＝Q'
f.	aːt＝paː	ат па	'horse＝Q'
g.	ʒi͡ebi͡e＝mi͡e	жебе ме	'arrow＝Q'
h.	tɛl＝mi͡e	тіл ме	'tongue＝Q'
i.	tæːn＝bi͡e	тән бе	'body＝Q'
j.	tɛs＝pi͡e	тіс пе	'tooth＝Q'
k.	sʏt＝pi͡e	сүт пе	'milk＝Q'

During fieldwork, 35 tokens of the question enclitic were recorded. Of those, 20 occurred after front vowel stems. Of the 20 tokens following a front vowel, only one token of Q surfaced as a front vowel. This particular instance involved a mother instructing her son how to complete a map task derived from the HCRC Map Task Corpus (Anderson et al. 1991). It seems plausible that the mother was taking on the role of teacher, and as a result, switching to a more formal register. In literary Kazakh, Q alternates for harmony, but as seen in Table 5, Q in colloquial Kazakh is invariantly [+bk]. Elsewhere, this speaker's productions of Q were always [+bk]. Despite the small sample size, the acoustic realization of Q is clear. The question enclitic is produced with a [+bk] vowel in the colloquial language. Table 5 and Figure 3 compare the realization of Q after front and back vowels to non-initial (i.e. alternating) /aː/ and /i͡e/.

	Mean F1 (z)	SD	Mean F2 (z)	SD
Alternating /i͡e/	-0.39	0.45	1.08	0.31
Alternating /aː/	1.17	0.55	0.07	0.34
Q after [-back]	0.58	0.72	-0.24	0.44
Q after [＋back]	0.68	1.22	-0.18	0.43

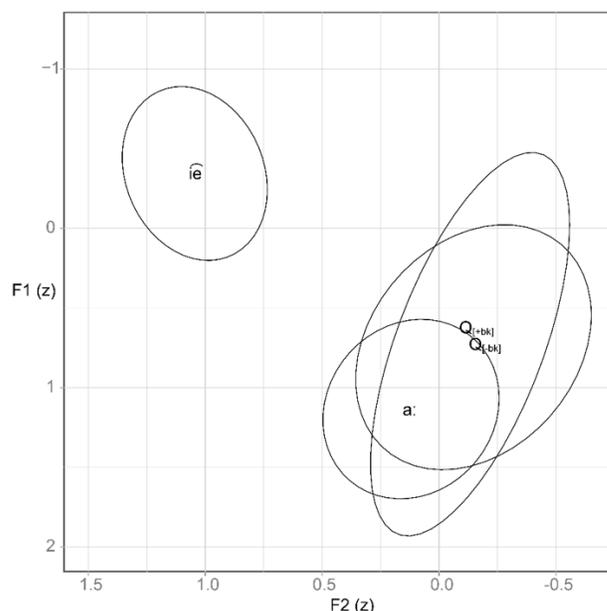Table 5: Mean and SD of alternating /aː/, /i͡e/, and Q

Figure 3: F1-F2 of Q in front and back vowel contexts, compared to /i͡e/ and /aː/ in alternating suffixes (in z-scores, with 1 SD ellipses)

Interestingly, Q exhibited a large amount of variation in F1, with many tokens approximating F1 of mid vowels rather than the low vowel /aː/ predicted by previous descriptions. In a number of related languages, including neighboring Kyrgyz (Hebert & Poppe 1963), Turkish (Lewis 1967; Underhill 1976) and Uyghur (Hahn 1991), the question enclitic is a high vowel, as opposed to the non-high vowel in Kazakh. Most importantly, though, Q does not alternate for backness harmony.

To make the contrast between previous descriptions and fieldwork data clear, several examples are presented in (15) below. In each example Q is underlined. In (15a-b) previous descriptions and fieldwork data correspond, since the root vowel is [+back]. However, when the root vowel is [-back], Q from fieldwork data is consistently disharmonic, in contrast to previous descriptions.

(15)   Realization of Q from fieldwork data compared to previous descriptions

	Fieldwork realization	Predicted realization based on previous work	Gloss
a.	bɔl naːjzaː = <u>maː</u>	bɔl naːjzaː = <u>maː</u>	'this spear = Q'
b.	baːr-aː-səŋ = <u>baː</u>	baːr-aː-səŋ = <u>baː</u>	'go-NPST-2S = Q'
c.	ɔːl tөːbi͡e = <u>maː</u>	ɔːl tөːbi͡e = <u>mi͡e</u>	'3S hill = Q'
d.	kөːpɣr-di͡en өːt-tɛŋ = <u>baː</u>	kөːpɣr-di͡en өːt-tɛŋ = <u>bi͡e</u>	'bridge-ABL cross-2S = Q'

Regardless of preceding vowel backness, Q is realized with a [+back] vowel in colloquial Kazakh, which is corroborated by Muhamedowa (2015:282-283) who notes the same invariance of Q. This is significant because Q is the only morpheme that may follow COM. Since Q is also invariant, colloquial Kazakh does not demonstrate the putative transparency of COM.

Before moving on, it should be noted that no tokens of COM+Q occurred during fieldwork. This construction occurs almost exclusively in literary texts, and even in those contexts, it is rare. While there is no direct evidence from colloquial Kazakh that COM is not transparent, the broader

invariance of Q in colloquial Kazakh suggests that the findings in B&L do not conform to the phonology of the colloquial language.

The next section shows that COM is similarly invariant while Q undergoes harmony in the literary register. The next section also demonstrates that Q in literary Kazakh undergoes harmony, in contrast to the colloquial data in this section. However, contra B&L, COM is not transparent, but blocks harmony in literary Kazakh. Using data from both the colloquial and literary registers, these two sections present a very different picture of exceptional COM.

6. The comitative suffix in literary Kazakh

Muhamedowa (2015) distinguishes between written and spoken Kazakh, noting that written Kazakh encodes an alternation on the question enclitic not present in the spoken language.[11] In §6.1, I present orthographic data from the Almaty Corpus of Kazakh (Madieva & Umatova 2015) and the Kazakh Language Corpus (Makhambetov et al. 2013), which show that Q agrees in backness with invariant COM rather than the stem vowel preceding COM. In other words, COM is not transparent in these written corpora. In §6.2, I go on to show from audio data in the Kazakh New Testament (kkitap.net) that COM is not transparent in spoken literary Kazakh, either. Whereas §5 conjectured that COM is not transparent in colloquial Kazakh, this section shows that COM is definitely not transparent in literary Kazakh.

6.1. Corpus data

The Almaty Corpus of Kazakh (Madieva & Umatova 2015) contains approximately 20 million morphologically tagged words from scientific, literary, and popular texts. When the corpus was queried for tokens of COM, 15,053 tokens from 461 documents were found. Of those, all were spelled with <e>, corresponding to phonemic /i͡e/. None were spelled with <a>, corresponding to phonemic /aː/. This result further supports the claim that COM does not alternate for backness harmony.

When the corpus was queried for tokens of Q, 7,127 tokens were found. Of those, 4,037 were written with <ма, ба, па> and 3,090 were written with <ме, бе, пе>. This indicates that, as noted throughout the descriptive literature, written Kazakh encodes a backness alternation for Q. The corpus was then queried for strings of COM followed by Q, returning only 7 instances of this morphological concatenation from 6 documents. Crucially, every instance of COM+Q was written <мен бе>, with graphemes representing front vowels. In short, COM blocks harmony on Q in the written language. Results from the Almaty Corpus of Kazakh are shown in Table 6.

Morpheme(s)	Allomorph	Token Count	Document Count
COM	<мен> /mi͡en/	15,053	461

---

[11]The present dissonance between the orthographic practices and the phonology of colloquial Kazakh is treated as a manifestation of register differences, where the orthographic conventions follow the literary register. However, this is not the only possibility. One possibility is language change. Perhaps this morpheme used to alternate, but has, after the development of an orthography, ceased to undergo harmony. Morpheme-specific patterns may arise when harmony evolves and decays (Anderson 1998), suggesting the plausibility of language change as a motivating factor in the misalignment of orthography to phonology in colloquial Kazakh.

Q	<ма> /maː/	2,296	125
	<ба> /baː/	1,118	63
	<па> /paː/	623	50
	<ме> /mie̯/	1,262	81
	<бе> /bie̯/	1,013	66
	<пе> /pie̯/	815	72
COM + Q	<мен бе> /mie̯n bie̯/	7	6

Table 6: Results from the Almaty Corpus of Kazakh

Given the rarity of COM+Q, I queried a second corpus, the Kazakh Language Corpus (Makhambetov et al. 2103). The Kazakh Language Corpus is a much larger corpus, containing over 135 million words, but lacks a graphical user interface comparable to the Almaty Language Corpus. When this larger corpus was queried for strings containing COM+Q, 77 tokens were found. All 77 tokens were written as <мен бе>, <бен бе>, or <пен бе>. None were written with <ба> following COM, e.g. <мен ба>.

From these two corpora of written Kazakh we see that COM is invariant while Q undergoes alternations in the written language. More importantly, when strings of COM+Q were queried, Q was always written <бе>, in accordance with the invariant [-back] feature of COM. There is thus no evidence for transparency in Kazakh orthography. Instead, Kazakh orthography treats COM as a blocker of backness harmony.

6.2. Kazakh New Testament

One corpus of searchable spoken Kazakh is available at present, the Kazakh New Testament (kkitap.net). Given that COM does not vary in colloquial Kazakh or in the Kazakh orthography, I did not cull acoustic data for COM. I did, however, cull 18 tokens of Q from this corpus. Nine tokens followed [+back] vowels and nine tokens followed [-back] vowels.

In the previous section, data was z-score normalized (Lobanov 1971) to facilitate across-speaker comparison. However, the data from the Kazakh New Testament came from only one speaker, so raw Hertz (Hz) values are presented. Mean F1 and F2 with standard deviations are shown in Table 7. The data from this audio corpus, like the orthographic corpora above, are clear. Q alternates according to the backness of the stem. Average F2 after [-back] stems is 2019 Hz, while average F2 after [+back] is 1190 Hz. There is an additional F1 difference between Q after these two stems due to the fact that the low vowel /aː/ alternates with the mid vowel /ie̯/ for harmony.

	F1 (SD)	F2 (SD)
Q after [-back]	445 (55)	2019 (156)
Q after [+back]	653 (80)	1190 (103)

Table 7: Mean F1-F2 (Hz) of Q in the Kazakh New Testament (n=18)

I then searched the corpus for instances of COM+Q. Only five instances of COM followed by Q were found in the corpus. Of these five, four followed the front vowel stem, /ie̯rk/ 'will' while one

followed the back vowel stem, /rūw͡χ/ 'spirit.' The relevant forms found in the text are shown in (16).

(16)   a.   адамның еркімен бе?

        aːdaːm-nəŋ   i͡erk-ɛ-mi͡en=bi͡e   Matthew 21:25; Mark 11:30; Luke 20:4

        'human-GEN   will-POSS-COM=Q'

        "by the will of humans?"

    b.   көктің еркімен бе?

        kɵːk-tɛŋ   i͡erk-ɛ-mi͡en=bi͡e   Matthew 21:25

        'heaven-GEN   will-POSS-COM=Q'

        "by the will of heaven?"

    c.   ілтипаттылық рухымен бе?

        ɛltij͡paːt-tə-ləq   rūw͡χ-ə-mi͡en=bi͡e   1 Corinthians 4:21

        'care-ADJ-NMLZR   spirit-POSS-COM=Q

        "with a caring spirit?"

As above, F1 and F2 were measured at the midpoint of each vowel. Observe in Figure 4 below the realization of Q in the word, /rūw͡χ-ə-mi͡en=bi͡e/ 'spirit-POSS-COM=Q. Here, Q is a front vowel, mirroring the regular alternation of Q found elsewhere in Kazakh New Testament.
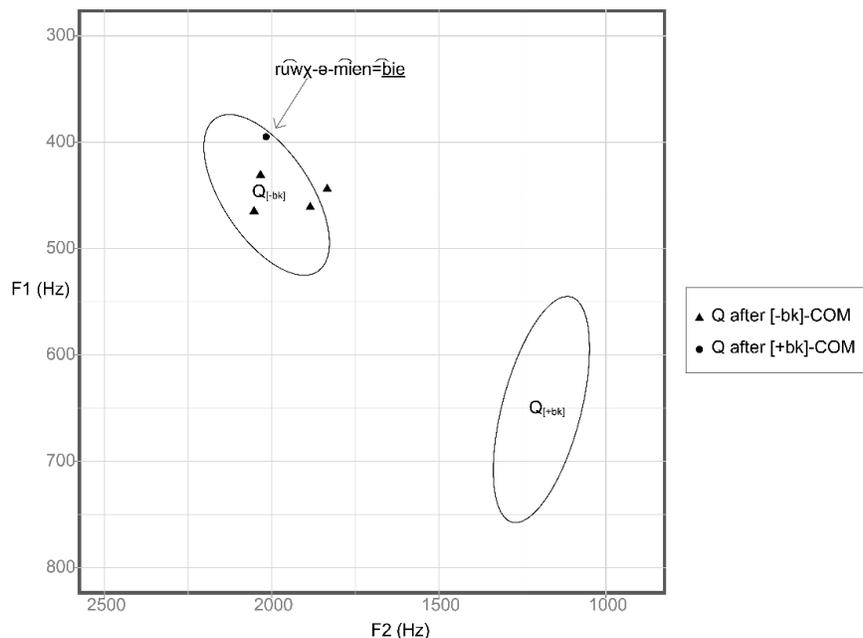

Figure 4: F1-F2 of Q after COM, compared to Q after [±bk] roots (in Hz, with 1 SD ellipses)

In summary, Q alternates according to the backness of the preceding vowel in this audio corpus. Further, when Q immediately follows COM, it is realized as a front vowel. In other words, COM is not transparent in the Kazakh New Testament. More broadly, this section has shown a difference in the application of vowel harmony between the colloquial and literary registers of the language. The question enclitic undergoes harmony in the literary register but does not in the colloquial register (Muhamedowa 2015).

The realizations of COM and Q in colloquial and literary Kazakh are compared to the findings from B&L in Table 8. The claims in B&L do not match the data for either colloquial or literary Kazakh. B&L and the results reported above agree that COM does not undergo harmony, but beyond that there is significant divergence. B&L reports that COM is transparent, allowing the backness of the preceding morpheme to dictate the backness of following Q. In both the written and acoustic data from literary Kazakh though, COM blocks harmony, forcing a following question enclitic to surface as [-back]. In the colloquial data from §5, both COM and Q are invariant. COM is always [-back] and Q is always [+back]. As for Q, B&L finds that Q undergoes harmony, like in literary Kazakh, in contrast to the pattern found in the colloquial language. In my data, neither of these morphemes undergo harmony in the colloquial register. Local harmony is exhibited in the literary register, with invariant COM dictating that following Q must be [-back]. In B&L, though, harmony is long-distance, skipping invariant COM and INF to target following morphemes, as shown by the crossing lines in the autosegmental schema below.

	after [-back] stem			after [+back] stem		
	COM	Q	Schema	COM	Q	Schema
Colloquial (no harmony)	i͡e	aː	Stem-COM-Q | | | [-bk] [-bk] [+bk]	i͡e	aː	Stem-COM-Q | | | [+bk] [-bk] [+bk]
Literary (blocks harmony)	i͡e	i͡e	Stem-COM-Q | |⁻⁻ [-bk] [-bk]	i͡e	i͡e	Stem-COM-Q | |⁻⁻ [+bk] [-bk]
B&L (transparent to harmony)	i͡e	i͡e	Stem-COM-Q |⁻⁻⁻|⁻⁻⁻ [-bk] [-bk]	i͡e	aː	Stem-COM-Q |⁻⁻⁻|⁻⁻⁻ [+bk] [-bk]

Table 8: Data from colloquial and literary Kazakh compared with B&L

I can think of three plausible explanations for the surprising data in B&L. One, their data may represent a dialectical difference between the speakers they consulted and those I worked with. Two, their data may come from a register that is neither colloquial nor literary, or three, their data may be an artefact of their data collection practices. I briefly address these three possibilities in order.

As to a potential dialectical difference, this would be surprising for several reasons. First, I have consulted speakers from central and northwestern Kazakhstan (where the speakers they worked with are from), and none of them produced the pattern described in B&L. Additionally, previous work on Kazakh dialects has reported only small differences between the dialects spoken in Kazakhstan, which are mostly lexical in nature (Amanzholov 1959; Kirchner 1998:330-331; Grenoble 2003:150). Lastly, I asked several speakers of other dialects if they have ever encountered data congruent with that reported in B&L and they said no. Further, each person

responded by saying that forms like (5a), /na:n-mi͡en=ba:/ 'bread-COM=Q', are ungrammatical in literary Kazakh.

Second, if these differences derive from a distinct register that is neither colloquial nor literary, it is unclear what kind of register this would be. If it relates to formal elicitation, then it should be possible to design a formal elicitation session to attempt to replicate their results. If formally elicited data match their results, then we could conclude that the data in B&L represent a potential elicitation register. If, however, formally elicited data does not match their findings, then we should conclude that a register difference is probably not involved in this discrepancy. The next section presents results from an experimental study that show these data do not derive from "lab speech" or some equivalent register used during formal elicitation.

Third, if these differences in the behavior of COM derive from the particular methods they employed to collect data, then we should expect to be able to generate their pattern of data only using certain methods. In the following section I attempt to replicate their results using two different elicitation strategies. I show that the ordering of stimuli corresponds to a large difference in vowel alternations for four speakers. If this result holds more generally, then their finding may, in fact, be an experimental artefact, and not representative of any known variety of Kazakh.

7. Stimulus ordering and the comitative suffix

The previous two sections described a register difference in Kazakh. The question enclitic alternates for harmony in literary Kazakh but not in the colloquial register. Interestingly, the pattern of data reported in B&L does not conform to either register. In this section, I explore the role of data collection methods on empirical results. I show that different stimulus presentation methods produce divergent results. At a very general level, the results reported in this section indicate the crucial role of stimulus presentation. On the other hand, the results from this section may offer an explanation for the surprising results found in B&L.

7.1. Participants

I recruited four Kazakhs residing in San Diego, CA to participate in the experimental study. Three of the participants were from southern Kazakhstan and one was from central Kazakhstan. All participants were in their 20's and spoke Kazakh and Russian, as well as some English.

7.2. Procedures and stimuli

The four elicitation sessions took place in quiet rooms near the campus of UC San Diego. Each speaker was presented a noun in its unmarked (i.e. NOM) form using Kazakh orthography. The speaker was then requested to produce this word in each of the seven pedagogical case endings (NOM, GEN, DAT, ACC, LOC, ABL, and COM) for both singular and plural numbers, with and without the question enclitic. Speakers were not explicitly asked to produce all case-inflected forms as quickly as possible, but all speakers completed the template for each lexical item very quickly. Twelve monosyllabic nouns were used as stimuli, half with [+back] and half with [-back] vowels. For each nominal root, 28 words (7 cases x 2 numbers x 2 question-related forms) were produced, resulting in a total of 336 words per speaker. The list of lexical items elicited is presented in (17). In both conditions described below the ordering of the 12 lexical items below was randomized.

(17)   Stimuli for formal elicitation

[-back] words

a.	k͡iɛn	кең	'mine'
b.	ɛn	iң	'den'
c.	k͡iɛz	кез	'time'
d.	sɛz	ciз	'2P.FORM'
e.	͡iɛs	ec	'memory'
f.	ɛs	ic	'work'

[+back] words

g.	naːn	нан	'bread'
h.	sən	сын	'test'
i.	qaːz	қаз	'goose'
j.	qəz	қыз	'girl'
k.	aːs	ac	'meal'
l.	əs	əc	'ash'

7.3. Condition 1: Ordered list

Each speaker was randomly assigned to one of two conditions. In the first condition, a template was provided using a sample word written with each of the seven case endings in singular and plural forms, with and without the question enclitic. An IPA-based version of this template is presented below in Table 9. Each speaker was given the template in Kazakh orthography and asked to familiarize themselves with it. Note that the template provided used a [-back] word, so as not to indicate the realization of COM in a back vowel context. If a [+back] stem had been used, each speaker would have seen an orthographic representation of a [+back] stem followed by invariant COM and a following [-back] Q (e.g. <атпен бе> /aːt-p͡iɛn=b͡iɛ/ 'horse-COM=Q').

   Crucially, the list shown in Table 9 uses a common ordering of cases found in pedagogical materials, where invariant COM follows all of the alternating cases (Rysbaeva 2000:27). Given that speakers were asked to produce a very grammar-focused, fairly unnatural task in a university setting, I expected participants to speak in a higher register, even with the rapidity with which they completed the task. Based on the results above, if a higher register was used, then all affixes except COM should alternate for harmony.

/͡iɛt/ 'meat'	SG	SG + Q	PL	PL + Q
NOM	͡iɛt	͡iɛt=p͡iɛ	͡iɛt-t͡iɛr	͡iɛt-t͡iɛr=m͡iɛ
GEN	͡iɛt-tɛŋ	͡iɛt-tɛŋ=b͡iɛ	͡iɛt-t͡iɛr-dɛŋ	͡iɛt-t͡iɛr-dɛŋ=b͡iɛ
DAT	͡iɛt-k͡iɛ	͡iɛt-k͡iɛ=m͡iɛ	͡iɛt-t͡iɛr-g͡iɛ	͡iɛt-t͡iɛr-g͡iɛ=m͡iɛ
ACC	͡iɛt-tɛ	͡iɛt-tɛ=m͡iɛ	͡iɛt-t͡iɛr-dɛ	͡iɛt-t͡iɛr-dɛ=m͡iɛ
LOC	͡iɛt-t͡iɛ	͡iɛt-t͡iɛ=m͡iɛ	͡iɛt-t͡iɛr-d͡iɛ	͡iɛt-t͡iɛr-d͡iɛ=m͡iɛ
ABL	͡iɛt-t͡iɛn	͡iɛt-t͡iɛn=b͡iɛ	͡iɛt-t͡iɛr-d͡iɛn	͡iɛt-t͡iɛr-d͡iɛn=b͡iɛ
COM	͡iɛt-p͡iɛn	͡iɛt-p͡iɛn=b͡iɛ	͡iɛt-t͡iɛr-m͡iɛn	͡iɛt-t͡iɛr-m͡iɛn=b͡iɛ

Table 9: Elicitation template for Condition 1

Speakers were not instructed how to order their productions of stimulus items, so one speaker inflected the target lexeme by-rows, producing all nominative-inflected forms first, then genitive and so on. A second speaker, however, inflected each lexeme by-columns, producing all singular non-questions, then singular questions, and so on. The number of stimulus items preceding

COM+Q varies some between these two speakers, but in each case a number of other forms precede COM, introducing a circumstance amenable to priming.

Predicted results are shown in Table 10 below. Table 10 replicates the three different COM+Q patterns found in the previous two sections. First, if speakers produce stimuli in the colloquial register, then Q should surface as [+back] regardless of stem backness. However, if speakers produce literary Kazakh, then Q should be realized in accordance with the backness of the preceding vowel. Thus, after COM, Q should always be realized as [-back]. If, however, participants produce the B&L pattern of transparency, then Q should be realized in accordance with the backness of the vowel preceding COM.

	after [-back] stem			after [+back] stem		
	COM	Q	Schema	COM	Q	Schema

Colloquial (no harmony)	i͡e	aː	Stem-COM-Q
			  | | |
			[-bk] [-bk] [+bk]
				i͡e	aː	Stem-COM-Q
						  | | |
						[+bk] [-bk] [+bk]

Literary (blocks harmony)	i͡e	i͡e	Stem-COM-Q
			  | |⟍
			[-bk] [-bk]
				i͡e	i͡e	Stem-COM-Q
						  | |⟍
						[+bk] [-bk]

B&L (transparent to harmony)	i͡e	i͡e	Stem-COM-Q
			  |⟍⟍⟍⟋
			[-bk] [-bk]
				i͡e	aː	Stem-COM-Q
						  |⟍⟍⟍⟋
						[+bk] [-bk]

Table 10: Predicted output patterns

Results from Condition 1 are shown in Table 11. Each speaker produced a root-COM=Q sequence 24 times. Of those 24, 12 critical productions occurred after [+back] roots. Speaker 1 produced the B&L (transparent) pattern, 3 of 12 times during elicitation. Speaker 4, on the other hand, produced the B&L (transparent) pattern 10 of 12 times.

	[-back] root		[+back] root	
Possible forms (register)	root-COM = bi͡e i͡es-pi͡en = bi͡e (Literary/B&L)	root-COM = baː i͡es-pi͡en = baː (Colloquial)	root-COM = bie aːs-pi͡en = bi͡e (Literary)	root-COM = baː aːs-pi͡en = baː (Colloqiual/B&L)
Speaker 1	11	1	9	3
Speaker 4	12	0	2	10

Table 11: Results from Condition 1

The question enclitic occurred 12 times per lexeme without a preceding COM. In these contexts, the realization of Q can shed light on the register employed during elicitation. These data are shown in Table 11. Speaker 1 produced only token of [+back] /paː/ after a [-back] stem, and Speaker 4 did not produce any tokens of [+back] /paː/ after a [-back] stem. In other words, in 143 of the 144 tokens of Q in [+back] contexts, Q was realized in a manner consistent with the literary register.

Since Q was not invariantly [+bk], it is unlikely that productions like /aːs-pi͡en=baː/ 'meal-COM=Q' reflect the colloquial register.[12]

Possible forms (register)	[-back] root		[+back] root	
	root = pi͡e i͡es = pi͡e (Literary)	root = paː i͡es = paː (Colloquial)	root = pie aːs = pi͡e (Unattested)	root-COM-baː aːs = paː (Colloqiual/Literary)
Speaker 1	71	1	4	68
Speaker 4	72	0	0	72

Table 12: The realization of Q after roots

In Table 12, we see that harmony almost always applies in root=Q sequences, in accordance with the data presented from literary Kazakh in §6. In Table 11, though, root-COM=Q consistently accorded with literary Kazakh for [-back] vowels only. For [+back] vowels, 11 of 24 tokens did not match the literary data in §6. If, as I have just argued, switching registers during elicitation does not drive this deviation from the literary register, then what does?

If the results from Condition 1 were the product of elicitation generally, then we should be able to replicate those results with a fully randomized word list. Again, if the pattern of harmony reported in Tables 11-12 is due to a general elicitation register, then this should hold across a variety of elicitation methods. If, however, the results in this subsection derive from some other factor, like stimulus ordering, then we predict that results obtained using a fully randomized stimulus list might differ from those in Condition 1.

7.4. Condition 2: Fully randomized list

Experimental Condition 2 used the same list of words, but instead of using the ordered template from the previous section, a random list of forms was generated. Each speaker was presented a root from the list in (16). Beside the lexeme, a second stimulus was presented. The second stimulus consisted of a randomly ordered combination of case, number, and the presence or absence of the question enclitic from Table 10. For instance, given the root /aːs/ 'meal' beside the paradigm cell, PL+ABL (in Kazakh orthography, көпше түрі + шығыс септігі), a speaker would produce [aːs-taːr-daːn] 'meal-PL-ABL.' After producing each of the 28 cells in Table 10 the next stimulus root was presented alongside a different randomized list of paradigm cells.

If the ordering of the list in Condition 1 resulted in the idiosyncratic transparency reported in B&L, then this effect should disappear in Condition 2 since the lists used in Condition 2 were randomized. This is exactly the result obtained. Results from Condition 2 corroborate the ordering-based prediction. Speakers 2 and 3 produced every form in accordance with a literary pronunciation. No forms exhibited transparency. Further, no forms exhibited the general invariance of Q that occurs in colloquial speech. Instead, all forms were representative of the literary register found in the two orthographic corpora and the Kazakh New Testament in §6.

---

[12] One could imagine that speakers vacillated between the colloquial and literary registers during elicitation. Under this hypothesis, speakers tended toward the literary register in front vowel contexts, but toward the colloquial register in back vowel contexts. This possibility seems highly unlikely and is not further discussed.

	[-back] root		[+back] root	
Possible forms	root-COM-b$\widehat{ie}$ $\widehat{ie}$s-p$\widehat{ie}$n = b$\widehat{ie}$ (Literary)	root-COM-baː $\widehat{ie}$s-p$\widehat{ie}$n = baː (Colloquial)	root-COM-b$\widehat{ie}$ aːs-p$\widehat{ie}$n = b$\widehat{ie}$ (Literary)	root-COM-baː aːs-p$\widehat{ie}$n = baː (Colloquial/B&L)
Speaker 2	12	0	12	0
Speaker 3	12	0	12	0

Table 13: Results from Condition 2

After sessions with Speakers 2 and 3, I asked if it was possible to produce Q as invariantly [+back]. Each speaker said yes, in the colloquial language, but not in the written language.

7.5. Discussion of results

In the two previous subsections I have demonstrated that two different experimental procedures produced divergent results. In Condition 1, a paradigm-based ordered elicitation session resulted in data that matched the general pattern found in B&L. In Condition 2, however, a randomized list of stimuli resulted in data entirely congruent with the literary register. How should we account for the different results obtained in §§7.3-7.4?

Given that the results from the randomized list conforms to numerous descriptions of the literary language, and that randomization is known to reduce the likelihood of ordering effects (Fisher 1935; Bock 1986), it seems most reasonable to conjecture that the results obtained from Condition 1 are, at least in part, are factual.

Concretely, I speculate that productions like /aːs-p$\widehat{ie}$n=baː/ 'meal-COM=Q' result from priming. Since COM occurs at the bottom of the list, as is typical in pedagogical grammars of the language, each speaker produced COM+Q at the end of group of related stimuli. Moreover, the colloquial variant of Q is identical to the literary variant after [+back] vowels. In other words, after [+back] roots, literary and colloquial Kazakh converge, resulting in /baː/. Thus, the distinction between the literary register, which is clearly used elsewhere in the elicitation, and the colloquial register is blurred for each of the words preceding COM in the list. When a speaker reaches the end of the list, a colloquial variant of COM has been repeatedly primed through the ordering of items in the template shown in Table 9. Thus, the realization of forms like /aːs-pien=baː/ could be due to the order of the list combined with a tendency towards the colloquial variant.

Some additional evidence for priming comes from Speaker 1. After finishing the paradigm for the [-back] stimulus, /ɛn/ 'den', she produced four instances of [-back] Q, /bie, pie/ with the [+back] root /aːs/ 'meal': /aːs=p$\widehat{ie}$/, /aːs-tǝŋ=b$\widehat{ie}$/ /aːs-taːr-dǝŋ=b$\widehat{ie}$/ and /aːs-taːn=b$\widehat{ie}$/. Since there is no motivation to preferentially produce [-back] variants of Q in either literary or colloquial Kazakh, the fact that a [-back] stimulus immediately preceded /aːs/ 'meal' offers a plausible cause for these unexpected productions.

My speculative hypothesis depends on the interaction of priming and a tendency toward the colloquial register. Several pieces of evidence suggest that Kazakhs gravitate strongly towards the colloquial rather than literary register. First, Kazakhs had almost no written literary tradition before the 19th century (Grenoble 2003:149-151; Olcott 2006:106-109; Jankowski 2012:25-26). The prestige of Russian throughout the Soviet Union then impeded large-scaled development of the

burgeoning literary tradition. As a result, the young Kazakh literary tradition was subordinate to Russian until very recently. As Jankowski (2012:30-31) observes, the current literary situation in Kazakhstan has not actually changed that much (see also Smagulova 2014) Many bookstores do not carry Kazakh books and many Kazakhs read only in Russian. Second, until very recently Kazakhs did not constitute a majority in Kazakhstan (81.7% in 1897, but only 40.1% in 1989; Dave 2007), which in effect necessitated a bilingual population (Dave 1996, 2004, 2007; Fierman 1998). Russification policies also reduced the domains of usage for Kazakh, and as a result, Kazakh was often spoken at home but not in public (see Grenoble 2003:196-197).

Thus, for many Kazakhs, even in post-Soviet Kazakhstan, there is little engagement with a higher, literary register. This is evident in the sentiment expressed by many Kazakhs that Kazakh is a language for speaking but Russian is a language for reading and writing. For these historical and sociolinguistic reasons it is plausible that Kazakh speakers gravitate towards the colloquial register, even in formal elicitation. While almost all Kazakhs can read in Kazakh, I speculate that it is more difficult to maintain a literary register than revert to the colloquial register.

In essence, two forces are pitted against each. On one hand, the formal task employed encourages a higher register. On the other hand, the general tendency towards the colloquial register favors less formal speech. In addition to these two factors, when the ordering of stimulus items favors the colloquial register, then the likelihood of colloquial [+back] Q increases significantly.

In this section, I reported on an experiment with four speakers to further determine if the proposals in B&L actually derive from a priming effect. Evidence from the experimental study reported in this section suggests that their results are potentially explainable as an ordering effect. More generally, though, I have demonstrated that vastly divergent empirical results are obtainable from simple differences in data collection methods. The difference between idiosyncratic transparency and canonical blocking here may fall out from something as seemingly insignificant as stimulus presentation method.

The following two sections focus on the infinitive suffix, where I demonstrate that INF in colloquial and literary Kazakh both alternates for [back] and spreads [back] to subsequent affixes, making INF a regular participant in harmony.

8. The infinitive suffix in colloquial Kazakh

This section uses audio data from fieldwork to investigate the claim that INF is transparent to harmony. In (18), repeated from B&L's analysis in (12) above, INF surfaces as invariantly [+back], but allows the backness of the preceding morpheme to pass onto subsequent affixes. In (18), the accusative suffix alternates according to the backness of the root despite the invariant [+back] feature of INF.

(18)   Backness harmony after INF (Bowman & Lokshin 2014:2)

	Phonology	Orthography	Gloss
a.	ʒʏz-u͡w-dɛ	жүзуді	'swim-INF-ACC'[13]
b.	kɛr-u͡w-dɛ	кіруді	'enter-INF-ACC'
c.	ʒaːb-u͡w-də	жабуды	'close-INF-ACC'

---

[13]B&L (2014:2) incorrectly glosses this word as 'take-INF-POSS.3.'

I demonstrate in this section that INF is not transparent to harmony in colloquial Kazakh, but regularly alternates in accordance with the backness of the preceding vowel.

8.1 The infinitive suffix

During fieldwork, I recorded 93 tokens of the infinitive suffix, 45 tokens after front vowels, and 48 tokens after back vowels. Below I compare the realization of INF with initial-syllable /ʏ/ and /ɔ/. This choice was made because round vowels are severely limited in non-initial syllables. If INF regularly alternates, the surface realization of INF should approximate initial-syllable /ʏ/ after front vowel stems and initial-syllable /ɔ/ after back vowel stems (Zsiga 1997:234-235). In Figure 5 below, INF shows a bimodal distribution for F2, which is expected for a backness harmonic alternation.
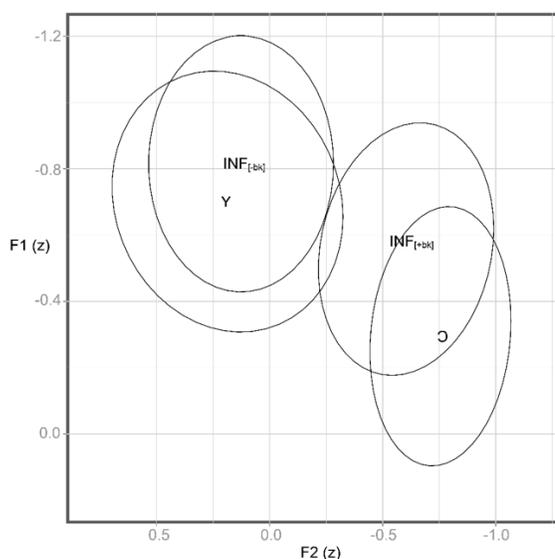
Figure 5: F1-F2 (z) of INF in front and back vowel contexts, compared to /ʏ/ and /ɔ/ in initial syllables (with 1 SD ellipses)

INF largely matches the realizations of /ʏ/ and /ɔ/ (n= 331 and 301 respectively). The distance between each allomorph of INF is slightly less than the distance between the two phonemes, but given that McCollum (2015:335) finds a 27% contraction of the vowel space in non-initial syllables, the fact that the allomorphs of INF are not quite as distinct as the initial-syllable productions of /ʏ/ and /ɔ/ is not surprising (see also McCollum & Chen accepted). Backness harmony in Kazakh peters out over the course of the word, so the distinction between front and back vowels is acoustically diminished in later syllables. This is further demonstrated in the density plot in Figure 6, where the F2 of initial-syllable /ɔ/ and /ʏ/ are compared with INF. The allomorphs of INF are represented with dashed lines while the initial-syllable realizations of /ʏ/ and /ɔ/ are represented with full lines. A clear bimodal distribution is evident, where INF[-back] and /ʏ/ group together while INF[+back] and /ɔ/ group together. In sum, even though INF is written with a single grapheme <y>, these data clearly show that INF alternates for backness harmony.

Since the alternation of INF may not be as perceptually salient as the /aː-/i͡e/ alternation, I tested the statistical significance of this alternation using a mixed effects model. The model included the following fixed effects: initial vowel backness, height, and rounding, and distance

from the initial vowel. Additionally, the model included speaker as a random effect. Using a likelihood ratio test between nested models to determine the significance of changes in F2, root backness was highly significant for predicting F2 of INF, ($\chi^2(1)= 99.48$, p < .001). The significance of root backness for F2 of INF further supports the claim that INF does, in fact, alternate for backness. In short, the backness of the root determines the backness of INF in colloquial speech. Descriptive statistics are presented in Table 14.

	F1 (SD)	F2 (SD)
INF[-back]	-0.83 (0.3)	0.13 (0.34)
INF[+ back]	-0.52 (0.34)	-0.63 (0.32)
ʏ	-0.75 (0.37)	0.24 (0.42)
ɔ	-0.72 (0.57)	-0.73 (0.28)

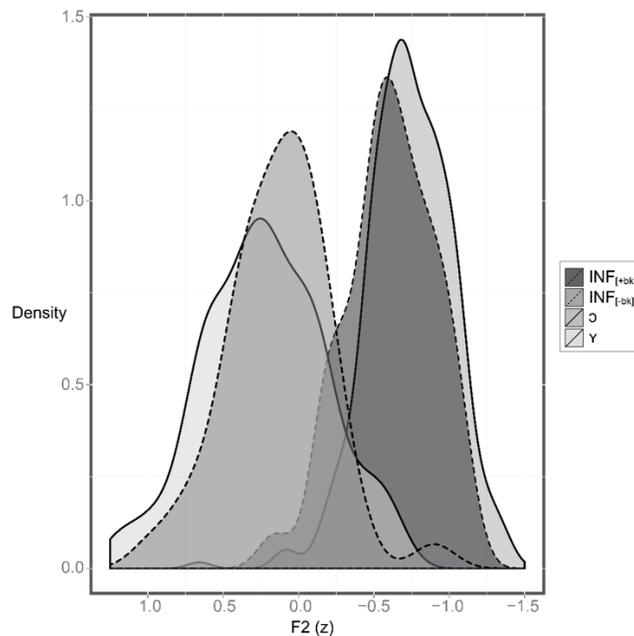Table 14: Mean F1-F2 (z) of INF compared to initial-syllable /ɔ/ and /ʏ/

Figure 6: F2 Density plot of INF in front and back vowel contexts, compared to /ʏ/ and /ɔ/ in initial syllables (in z-scores)

8.2. The realization of the agentive suffix following the infinitive suffix

Given that INF undergoes harmony, it is necessary to determine whether following affixes also undergo harmony in the colloquial register. To assess this, 93 tokens of the agentive suffix, /ʃɛ/~/ʃə/, immediately following INF were recorded during fieldwork (e.g. /qaːl-ɔw-ʃə/ 'remain-INF-AGT'). If AGT undergoes harmony, then we expect its surface realization to match those of front /ɛ/ and back /ə/ in non-initial (i.e. alternating) positions. Table 15, as well as Figures 7 and 8 confirm this prediction, showing that the acoustic realization of AGT approximates non-initial /ɛ/ and /ə/ (n= 435 and 277, respectively) in Kazakh.

	F1 (SD)	F2 (SD)
AGT[-back]	0.05 (0.54)	0.54 (0.31)
AGT[+back]	0.15 (0.44)	-0.11 (0.32)
ɛ	-0.22 (0.55)	0.57 (0.34)
ə	0.08 (0.57)	-0.10 (0.38)

Table 15: Mean F1-F2 (z) of AGT compared to alternating /ɛ/ and /ə/
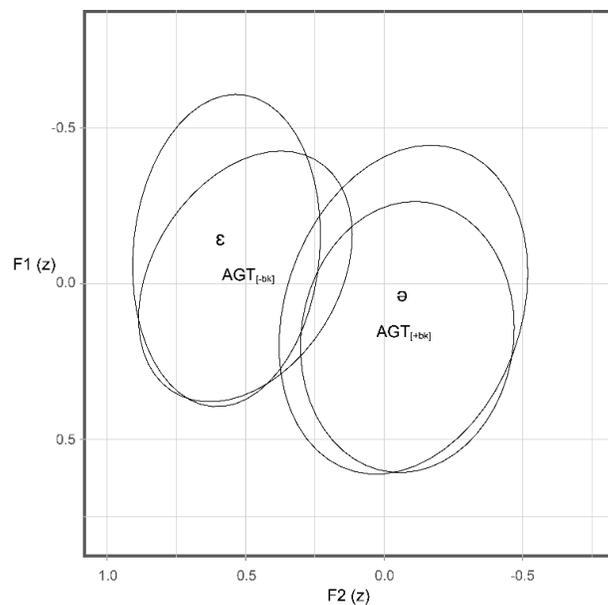
Figure 7: F1-F2 (z) of AGT after INF in front and back vowel contexts, compared to alternating /ɛ/ and /ə/ (with 1 SD ellipses)
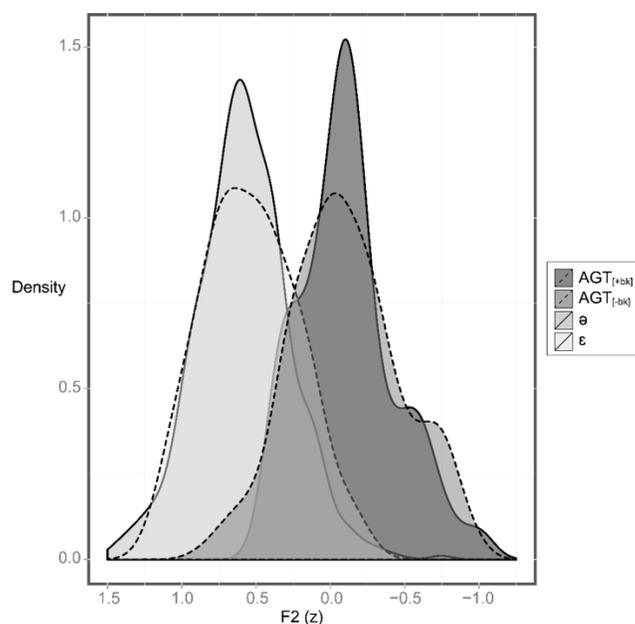
Figure 8: F2 Density plot of AGT after INF in front and back vowel contexts, compared to alternating /ɛ/ and /ə/ (in z-scores)

Impressionistically, AGT, as well as alternating /ɛ/ and /ə/, show more overlap in F2 than the /i͡e/-
/aː/ and /ʏ/-/ɔ/ alternations discussed above. Two forces produce this overlap. First, these two
phonemes are simply more similar to one another than the other harmonic pairings (see McCollum
& Chen accepted). Second, AGT, as well as other short vowel suffixes like ACC tend to occur
word-finally. As noted before, the vowel space shrinks in non-initial positions, resulting in more
significant F2 overlap for /ɛ/ and /ə/. Despite the contraction of the vowel space, root backness
was still highly significant for predicting F2 of AGT, (χ^2(1)=88.90, p < .001). Since the allomorphs
of AGT closely approximate the harmonic alternation between /ɛ/ and /ə/, I conclude that AGT
undergoes harmony. In short, both INF and AGT fully alternate for harmony. As a result, INF is a
regular suffix and not transparent in colloquial Kazakh. In the following section I examine acoustic
data from the Kazakh New Testament to demonstrate that INF in the literary register also regularly
undergoes harmony.

9. The infinitive suffix in literary Kazakh

9.1. The infinitive suffix

To assess the realization of INF in literary Kazakh, twenty tokens of INF after front and back
vowel roots in the Kazakh New Testament were culled. As above, F1 and F2 were measured at
vowel midpoint. This time I did not compare these realizations to initial tokens of /ʏ/ or /ɔ/. At this
point in the analysis, if we observe a significant alternation, irrespective of its approximation of
initial-syllable /ʏ/ and /ɔ/, we can reasonably conclude that INF undergoes alternations in both
colloquial and literary Kazakh. Descriptive statistics are shown in Table 16 and formant
frequencies are plotted in Figure 9. In both of these, F2 of INF is significantly higher after front
vowels, matching the results found in colloquial Kazakh.

	F1 (SD)	F2 (SD)
INF[-back]	332 (23)	1435 (135)
INF[+back]	398 (15)	959 (90)

Table 16: Mean F1-F2 (Hz) of INF in the Kazakh New Testament
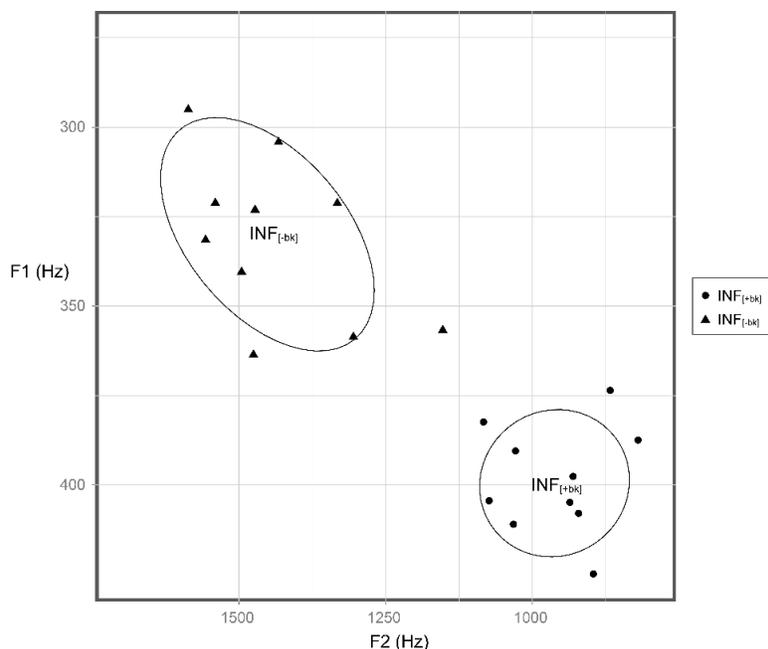
Figure 9: F1-F2 (Hz) of INF in front and back vowel contexts, in the Kazakh New Testament (with 1 SD ellipses)

All tokens were culled from one speaker, the narrator, in the text and so no normalization or random effects structure was used to assess statistical significance. Instead, a simpler t-test was conducted to determine the significance of root backness on F2 of INF. As in colloquial Kazakh, the effect was highly significant (t(19)= -9.3, p < .001). In sum, INF alternates for harmony in literary Kazakh.

9.2. The realization of the agentive suffix following the infinitive suffix

To determine whether or not INF spreads harmony onto following suffixes, twenty tokens of AGT immediately following INF were also culled. If F2 of AGT varies significantly based on the backness of the preceding vowel, then we can conclude that AGT undergoes backness harmony after INF in literary Kazakh. Since only an alternation is necessary, given the weight of evidence already put forth, no instances of regularly alternating /ɛ/ or /ə/ were measured for comparison. Descriptive statistics are shown in Table 17, and F1-F2 are plotted in Figure 10 below.

	F1 (SD)	F2 (SD)
AGT[-back]	376 (63)	1579 (109)
AGT[+back]	435 (37)	1305 (130)

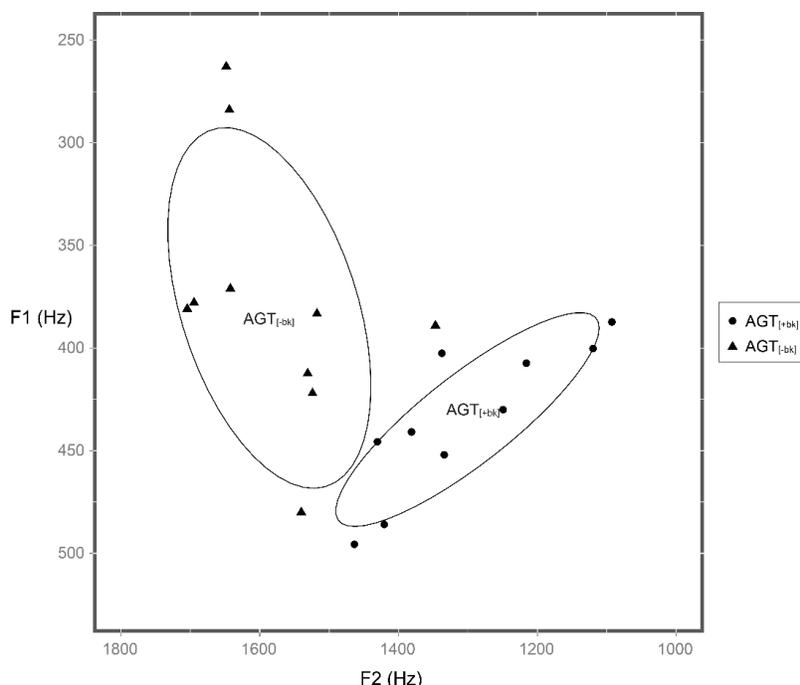Table 17: Mean F1-F2 (Hz) of AGT after INF in the Kazakh New Testament

Figure 10: F1-F2 (Hz) of AGT in front and back vowel contexts, in the Kazakh New Testament (with 1 SD ellipses)

The statistical significance of this F2 alternation was assessed using a t-test. As expected, F2 of AGT varies significantly based on the backness of the root (t(18)= -5.12, p < .001). When the t-values of AGT and INF are compared (-5.12 and -9.3, respectively), a more robust alternation is present in INF than in AGT. Again, this suggests a contraction of the vowel space due to the petering out of harmony throughout the word.

To summarize, both INF and AGT undergo backness harmony in colloquial and literary Kazakh. Findings from B&L are compared to colloquial and literary Kazakh in Table 18. In Table 8 above, we saw that the realization of Q varied by register, but this is not the case for affixes like INF and AGT. Instead, both colloquial and literary Kazakh accord with one another. Significantly, the data presented above suggest that the findings reported in B&L are not congruent with either register.

	after [-back] stem			after [+back] stem		
	INF	AGT	Schema	INF	AGT	Schema
Colloquial (no harmony)	YW	ε	Stem-INF-AGT |᠆᠆᠆᠆᠆ [-bk]	ɔw	ə	Stem-INF-AGT |᠆᠆᠆᠆᠆ [+bk]
Literary (blocks harmony)	YW	ε	Stem-INF-AGT |᠆᠆᠆ [-bk]	ɔw	ə	Stem-INF-AGT |᠆᠆᠆ [-bk]
B&L (transparent to harmony)	u͡w	ε	Stem-INF-AGT | | [-bk] [-bk]	u͡w	ə	Stem-INF-AGT | | [+bk] [-bk]

Table 18: Findings reported in B&L compared to colloquial and literary Kazakh

Recall also that in B&L, Speaker 2 exhibited significantly more variation between INF after [±back] roots, shown again in Figure 11. Observe that for Speaker 2 (right) the surface realization of INF after front vowels was a front vowel. It does not approximate initial-syllable /ʏ/ but exhibits an F2 that is characteristic of a front vowel. With the gradual petering out of backness harmony throughout the word in mind (McCollum 2015; McCollum & Chen accepted), it seems reasonable to conclude that for Speaker 2, INF does, in fact, alternate for harmony.
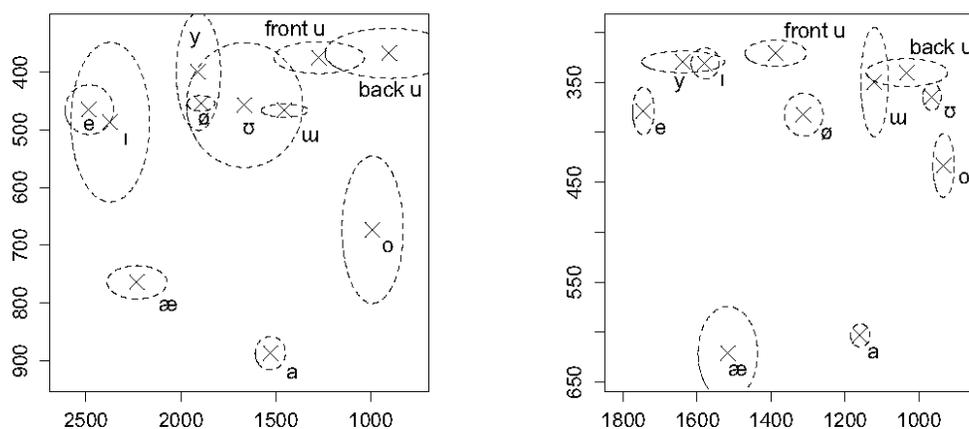

Figure 11: F1-F2 vowel plots for the two speakers consulted in B&L (2014:4). The vowel plot for Speaker 1 is on the left, and the plot for Speaker 2 is on the right.

If this is the case, then I only need to account for why INF for Speaker 1 failed to undergo harmony. The invariance of INF for Speaker 2 likely derives from an orthographic effect. B&L presented each stimulus orthographically. As noted earlier, in all other cases Kazakh orthography encodes backness harmony, but INF is always written with the grapheme, <y>, which in other contexts represents a back vowel. Further, if the words below were read assuming a one-to-one correspondence between grapheme and phone, this would result in transparency in backness harmony. In (19a-b), ACC is written as <дi> because the root is [-back]. In (19c), though, ACC is written as <ды> because the root is [+back]. If non-initial <y> is not treated as an alternating-vowel, then the orthographic representations below favor transparency for INF. Recall that <y> represents an alternating round vowel in Kazakh but non-alternating /u/ in Russian. If Russian influences these productions, then it's very possible that these were produced with transparency.

(19)   Backness harmony after INF (Bowman & Lokshin 2014:2)
           Phonology   Orthography   Gloss
     a.   ʒʏz-u͡w-dɛ   жүзуді   'swim-INF-ACC'[14]
     b.   kɛr-u͡w-dɛ   кіруді   'enter-INF-ACC'
     c.   ʒaːb-u͡w-də   жабуды   'close-INF-ACC'

B&L argues that INF is transparent for two reasons: one, it does not undergo categorical alternations for both speakers, and two, their consultants consciously identified both variants of INF as the same vowel. The data presented in this section as well as the interspeaker difference

---

[14]B&L (2014:2) incorrectly glosses this word as 'take-INF-POSS.3.'

between Speaker 1 and Speaker 2 in B&L suggest that phonetically, there is no reason to conclude that INF fails to undergo harmony generally in the language. As to their second point, it is not necessarily informative to know native speaker intuitions for this phenomenon. While native speaker intuitions may offer significant help (e.g. Sapir 1949), it is not the case that every intuition should guide linguistic analysis. For instance, given that both alternants of INF are represented by the same grapheme and B&L used orthographic prompts to elicit the data, it is possible that the speakers were answering an orthographic rather than a phonological question. Moreover, even if speakers are unaware of this alternation, it does not change the fact that it does alternate. Speakers are often unaware of labial harmony in Kazakh, because it is not represented orthographically and because it is gradient and non-iterative (McCollum 2018). Since the backness of INF depends on the backness of the root vowel, then we should conclude that it is a regularly alternating affix in the language, whether or not native speakers are aware of it.

I have argued that the use of orthography affected the surface realization of INF for Speaker 1. This does not generalize to all speakers, though. Note that INF alternates for Speaker 2, and moreover, that INF alternates in the Kazakh New Testament, although the narrator is reading from a script. One salient difference between these speakers and Speaker 1 is educational background. Speaker 1 was educated in Russian and is dominant in Russian (personal communication). Given the high percentage of Russians residing in north and central Kazakhstan, along with the general prestige of Russian, it is likely that Speaker 1 does not regularly use Kazakh orthography. In Russian, the grapheme <y> always represents a back vowel. However, in the Kazakh orthography, this grapheme represents an alternating vowel pair, /ɣw/ and /ɔw/, in non-initial syllables, and not just a single back vowel. It is entirely plausible then that Speaker 1 produced /uw/ simply as an orthographic effect (e.g. Derwing 1992; Damian & Bowers 2003; Perre et al. 2010). For speakers who were educated in Kazakh and who read and write regularly in Kazakh, the effect of orthography would likely be inconsequential. Yet, for speakers who read and write almost exclusively in Russian, the influence of orthography is presumably much more significant.

This section has further shown that INF regularly alternates for harmony. I have argued that the "idiosyncratic transparency" of INF in B&L likely results from their choice to prompt each target word with an orthographic representation. Kazakh orthography encodes all other backness alternates except that of INF. This fact, combined with the educational background of Speaker 1, make such an interpretation quite plausible. The next and final section summarizes the empirical findings and discusses the methodological contributions of the paper.

10. Summary and discussion

Empirically, the comitative suffix is not transparent in colloquial or literary Kazakh. The question enclitic, the only morpheme that may follow COM, does not alternate for harmony in the colloquial language. In the literary language, though, COM blocks harmony on Q, forcing the enclitic to surface as [-back]. Both registers differ from the data reported in B&L, though, where COM is transparent. Using the experimental results in §7, I suggested that this incongruence results from priming.

Moving on to consider INF in §§8-9, I showed that INF alternates for harmony in both registers, in contrast to the claims in B&L. Since INF alternates for harmony, the alternation of following AGT in the data presented in unsurprising. In short, INF is a regular undergoer of harmony. I suggested that the difference between the findings in B&L and those presented above depends on an orthographic effect. Overall, I have argued that neither COM nor INF are

transparent to harmony in Kazakh, showing that COM blocks harmony while INF regularly undergoes harmony. These results are summarized in Table 19 below.

COM						
	after [-back] stem			after [+back] stem		
	COM	Q	Schema	COM	Q	Schema
Colloquial (no harmony)	i͡e	aː	Stem-COM-Q | | | [-bk] [-bk] [+bk]	i͡e	aː	Stem-COM-Q | | | [+bk] [-bk] [+bk]
Literary (blocks harmony)	i͡e	i͡e	Stem-COM-Q | | [-bk] [-bk]	i͡e	i͡e	Stem-COM-Q | | [-bk] [-bk]
B&L (transparent to harmony)	i͡e	i͡e	Stem-COM-Q | [-bk] [-bk]	i͡e	aː	Stem-COM-Q | [+bk] [-bk]

INF						
	after [-back] stem			after [+back] stem		
	INF	AGT	Schema	INF	AGT	Schema
Colloquial (no harmony)	ɣw	ɛ	Stem-INF-AGT | [-bk]	ɔw	ə	Stem-INF-AGT | [+bk]
Literary (blocks harmony)	ɣw	ɛ	Stem-INF-AGT | [-bk]	ɔw	ə	Stem-INF-AGT | [-bk]
B&L (transparent to harmony)	u͡w	ɛ	Stem-INF-AGT | | [-bk] [-bk]	u͡w	ə	Stem-INF-AGT | | [+bk] [-bk]

Table 19: Findings reported in B&L compared to colloquial and literary Kazakh

Methodologically, I have argued that the pattern of data presented in Bowman & Lokshin (2014) result from the methodological choices used to collect their data. The following three issues were relevant: register differences, stimulus ordering, and orthography. The data described in Bowman & Lokshin (2014) are interpreted here as artefactual and demonstrate the importance of both experimental and field methodologies. From the experimental side, careful designs to avoid priming effects are necessary to ensure that data collected is representative of the language. Also, given the effect of orthographic representations on speech, it is important to consider the various possible outcomes of our choices. Further, using multiple methodologies can both provide converging evidence in favor of one's analysis and simultaneously safeguard against spurious results. The data presented in this paper come from multiple corpora as well as fieldwork data. By using multiple types of data from independent sources, I've provided robust evidence for the patterns described above. I have also argued, in line with general fieldwork manuals, that knowledge of the culture in which the language is spoken is an important part of field research. For research on Kazakh, it is crucial to know the linguistic ecology in Kazakhstan and the role that other languages, like Russian, might play during data collection.

At the theoretical level, locality has been shown to govern much vowel harmony in general, and in particular exceptionality in vowel harmony. These general findings are countered by Bowman & Lokshin (2014), though, who suggest that exceptional morphemes may exhibit "idiosyncratic transparency." At a formal level, such a result could undermine the assumptions of many theoretical models, including the autosegmental models used above. Crucially, the data from Kazakh, as reported in this paper, do not counter the descriptive and theoretical claims to-date. As far as we can tell, exceptionality in harmony is always governed by locality.

Finally, this paper has demonstrated the benefits of using a variety of methods to address a given research question. The combination of ethnographically-informed formal elicitation, production studies, and analysis of multiple corpora converge on a single analysis of exceptionality in Kazakh. When multiple data are brought to bear on a question, then we can be more confident that our contributions record the actual linguistic phenomenon under study.

References

Abbi, Anvita. 2001. *A manual of linguistic field work and structures of Indian languages*. Lincom Europa.

Amanzholov, Sarsen. 1959. *Voprosy dialektologii i istorii kazakhskogo yazyka.* National Instructional Institute in the name of Abai. Alma-Ata.

Ameka, Felix K., Alan Charles Dench, and Nicholas Evans, eds. 2006. *Catching language: The standing challenge of grammar writing*. Walter de Gruyter.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod and Stephen Isard. 1991. The HCRC map task corpus. *Language and speech* 34.4: 351-366.

Anderson, Gregory D.S. 1998. Historical aspects of Yakut (Saxa) phonology. *Turkic languages* 2: 1-32.

Archangeli, Diana, and Douglas Pulleyblank. 1994. *Grounded phonology*. MIT Press.

Baković, Eric 2000. *Harmony, dominance and control*. PhD dissertation, Rutgers University.

Balakaev, M. B. 1962. Sovremennij kazaxskij jazyk Fonetika i morfologiya. Nauka.

Biber, Douglas. 1993. Using register-diversified corpora for general language studies. *Computational linguistics* 19.2: 219-241.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory* 8.1: 9-37.

Bickel, Balthasar; Goma Banjade; Martin Gaenszle; Elena Lieven; Netra Prasad Paudyal; Ichchha Purna Rai; Manoj Rai; Novel Kishore Rai, and Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language* 83.1: 43-73.

Bock, J. Kathryn. 1986 Syntactic persistence in language production. *Cognitive psychology* 18.3: 355-387.

Bochnak, M. Ryan, and Lisa Matthewson, eds. 2015. *Methodologies in semantic fieldwork*. Oxford.

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. Palgrave Macmillan.

Bowman, Samuel R. and Benjamin Lokshin. 2014. Idiosyncratically Transparent Vowels in Kazakh. *Proceedings of the 2013 annual meeting on phonology*.

Caballero, Gabriela. 2010. Scope, phonology and morphology in an agglutinating language: Choguita Rarámuri Tarahumara variable suffix ordering. *Morphology* 20.1:165-204.

Chelliah, Shobhana L., and J. Willem. De Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. Springer.

Clements, George N. and Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. *The structure of phonological representations* 2: 213-255.

Cowart, Wayne. 1997. *Experimental syntax*. Sage.

Damian, Markus F., and Jeffrey S. Bowers. 2003. Effects of orthography on speech production in a form-preparation paradigm. *Journal of memory andlLanguage* 49.1: 119-132.

Dave, Bhavna. 1996. National revival in Kazakhstan: Language shift and identity change. *Post-Soviet Affairs* 12.1: 51-72.

Dave, Bhavna. 2004. Entitlement through numbers: nationality and language categories in the first post-Soviet census of Kazakhstan. *Nations and nationalism* 10.4: 439-459.

Dave, Bhavna. 2007. *Kazakhstan-ethnicity, language and power*. Routledge.

Derwing, Bruce L. 1992. Orthographic aspects of linguistic competence." *The linguistics of literacy*: 193-210.

Dzhunisbekov, A. 1972. *Glasnye kazakhskogo jazyka*. Alma-Ata: Nauka.

Dzhunisbekov, A. 1980. *Singarmonizm v kazakhskom jazyke*. Alma-Ata: Nauka.

Essegbey, James. 2015. "Is this my language?" Developing a writing system for an endangered language community. *Language documentation and endangerment in Africa*, Essegbey, James; Brent M. Henderson, and Fiona McLaughlin, eds., 153-176.

Face, Timothy L. 2003. Intonation in Spanish declaratives: differences between lab speech and spontaneous speech. *Catalan journal of linguistics* 2:115-131.

Fierman, William. 1998. Language and identity in Kazakhstan: Formulations in policy documents 1987–1997. *Communist and Post-Communist Studies* 31.2: 171-186.

Finley, Sara. 2010. Exceptions in vowel harmony are local. *Lingua* 120: 1549-1566.

Fisher, Ronald A. 1935. *The design of experiments*. Oliver and Boyd.

Gafos, Adamantios. 1999. *The articulatory basis of locality in phonology*. Garland Publishing.

Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel, eds. 2006. *Essentials of language documentation*. Walter de gruyter.

Goldsmith, John A. 1976. *Autosegmental phonology*. PhD dissertation, MIT.

Grenoble, Lenore. 2003. *Language policy in the Soviet Union*. Kluwer Academic Publishers.

de Groot, Annette, and Peter Hagoort, eds. 2017. *Research methods in psycholinguistics and the neurobiology of language: A practical guide*. John Wiley & Sons.

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

Hebert, Raymond, and Nicholas Poppe. 1963. *Kirghiz manual*. Uralic and Altaic Series vol. 33. Indiana University.

van der Hulst, Harry, and Jeroen van de Weijer. 1995. Vowel harmony. In *The handbook of phonological theory,* John A. Goldsmith, ed., 495-534.

Jankowski, Henryk. 2012. Kazakh in contact with Russian in modern Kazakhstan. *Turkic languages* 16.1: 25-67.

Johanson, Lars. 1998. The history of Turkic. *The Turkic languages,* Johanson and Csato, eds 81-125.

Kara, David Somfai. 2002. *Kazak*. Munich: Lincom Europa.

Kirchner, Mark. 1992. *Phonologie des Kasachischen: Untersuchungen anhand von Sprachaufnahmen aus der kasachischen Exilgruppe in Istanbul*. Harrassowitz Verlag.

Kirchner, Mark. 1998. Kazakh and Karakalpak. In *The Turkic languages*, Johanson and Csato, eds, 318-332. Routledge.

kkitap.net [website] 2010. Yeni Yaşam Yayınları New Life Publications. Istanbul.

Krippes, Karl. 1993. *Kazakh grammar with affix list*. Dunwoody Press.

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.

Lewis, Geoffrey. 1967. *Turkish grammar*. Oxford.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *JASA* 49: 606–608.

Madieva, G.B., and Zh. M. Umatova. 2015. Ob Almatinskom korpuse kazaxskogo jazyka. *Vestnik KazNU. Seriya filologischeskaya* 5:98-103.

Mahanta, Shakuntala. 2012. Locality in exceptions and derived environments in vowel harmony. *Natural language & linguistic theory* 30: 1109-1146.

Makhambetov, Olzhas; Aibek Makazhanov; Zhandos Yessenbaev; Bakhyt Matkarimo; Islam Sabyrgaliev, and Anuar Sharafudinov. 2013. Assembling the Kazakh Language Corpus. In *Proceedings of the 2013 conference on empirical eethods in natural language processing*, 1022–1031.

McCollum, Adam G. 2015. Labial Harmonic Shift in Kazakh: Mapping the Pathways and Motivations for Decay. In *Proceedings of the 41st annual meeting of the Berkeley Linguistics Society*, 329-351.

McCollum, Adam. G. 2018. Vowel dispersion and Kazakh labial harmony. *Phonology* 35.2: 287-326.

McCollum, Adam G. and Si Chen. accepted. Kazakh. *Journal of the International Phonetic Association*.

Menges, Karl. 1947. *Qaraqalpaq grammar*. New York: King's Crown Press.

Muhamedowa, Raihan. 2015. *Kazakh: A comprehensive grammar*. Routledge.

Newman, Paul, and Martha Ratliff, eds. 2001. *Linguistic fieldwork*. Cambridge.

Ní Chiosáin, Máire, and Jaye Padgett. 2001. Markedness, segment realization, and locality in spreading. *Segmental phonology in Optimality Theory: Constraints and representations*: 118-156.

Niyazgalieva, A. A. and G. G. Turganalieva. 2013. *Qazaq dialektologiyasy Oquw-adistemelik qural*. M. Otemisov atyndagi Batys Qazaqstan Memlekettik Universiteti. Oral, Qazaqstan.

Olcott, Martha Brill. 2006. *The Kazakhs*. 2nd edition. Hoover Press.

Perre, Laetitia; Chotiga Pattamadilok; Marie Montant, and Johannes C. Ziegler. 2010. Orthographic effects in spoken language: on-line activation or phonological restructuring? *Brain research* 1275: 73-80.

Pickering, Martin J., and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological bulletin* 134.3: 427-459.

Podesva, Robert J., and Devyani Sharma, eds. 2013. *Research methods in linguistics*. Cambridge University Press.

Pulleyblank, Douglas. 1983. *Tone in lexical phonology*. PhD dissertation, Massachusetts Institute of Technology.

Rysbaeva, G.K. 2000. Kazaxskij jazyk. Grammaticheskij spravochnik. Almaty: Sözdik-Slovar'.

Sapir, Edward. 1949. The psychological reality of phonemes. In Mandelbaum, D.G., ed. *Selected writings of Edward Sapir*, 46-60.

Schütze, Carson T., and Jon Sprouse. 2013. Judgment data. In *Research methods in linguistics*, Podesva, Robert J. and Devyani Sharma, eds., 27-50.

Sebba, Mark. 2007. *Spelling and society: The culture and politics of orthography around the world*. Cambridge.

Seifart, Frank. 2006. Orthography development. In *Essentials of language documentation*, Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel, eds., 275-299.

Smagulova, Juldyz. 2014. Early language socialization and language shift: Kazakh as baby talk. *Journal of sociolinguistics* 18.3: 370-387.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic inquiry* 31.3: 575-582.

Sprouse, Jon 2007. A program for experimental syntax. PhD dissertation, University of Maryland.

Svantesson, Jan-Olof; Anna Tsendina; Anastasia Karlsson, and Vivan Franzen. 2005. *The phonology of Mongolian*. Oxford.

Tonhauser, Judith, and Lisa Matthewson. 2015. Empirical evidence in research on meaning. unpublished manuscript. [http://ling.auf.net/lingbuzz/002595].

Underhill, Robert. 1976. *Turkish grammar*. MIT Press.

Userbaeva, G. 2005. *Bastawysh synypta esim sozderdi oqytuw*. M. O. Awezov atyndagy Ontustik Qazaqstan Memlekettik Universiteti.

Vajda, Edward. 1994. Kazakh phonology. *Opuscula altaica*, 603-650. Western Washington University.

Vaux, Bert. 2000. Disharmony and derived transparency in Uyghur vowel harmony. In *Proceedings of the North East Linguistic Society*, vol. 30, 672-698.

Vaux, Bert; Justin Cooper, and Emily Tucker. 2007. *Linguistic field methods*. Wipf and Stock Publishers

Washington, Jonathan North. 2016. *An investigation of vowel anteriority in three Turkic languages using ultrasound tongue imaging*. PhD dissertation, Indiana University.

Xu, Yi 2010. In defense of lab speech. *Journal of phonetics* 38.3: 329-336.

Yao, Bo, and Christoph Scheepers. 2011. Contextual modulation of reading rate for direct versus indirect speech quotations. *Cognition* 121.3: 447-453.

Yu, Kristine. 2014. The experimental state of mind in elicitation: illustrations from tonal fieldwork. *Language documentation & conservation* 8:738-777.

Zsiga, Elizabeth C. 1997. Features, gestures, and Igbo vowels: An approach to the phonology-phonetics interface. *Language* 73.2: 227-274.