Language Specific Peculiarities Document for KAZAKH as Spoken in Kazakhstan

1. Dialects

Kazakh (also Kazak, Qazaq)[1] is a member of the Kypchak (Qypchaq, Qıpçaq) branch of the Turkic language family; its closest relatives being Karakalpak (Qaraqalpaq) and Nogai (Noğay), both of which are almost entirely mutually intelligible with Kazakh.

According to Amanjolov (1959), there are three main dialect areas for Kazakh as spoken in Kazakhstan: Northeastern, Southern, and Western (Cf. Аманжолов, 1959). The modern standard language is based on the Northeastern dialect. Speakers of standard Kazakh are found particularly among those who have been educated in Kazakh. While their native dialect may not be considered standard, they are often able to speak the standard dialect accurately. Nevertheless, many Kazakh speakers in Kazakhstan have received formal education only in Russian and consequently speak only their native variety of Kazakh.

Other sources (e.g., Ethnologue) assert that remarkably little dialectal variation occurs in Kazakh, despite Kazakhstan's status as the world's ninth-largest country in terms of area. With regard to the differences that do exist, some amount of further dialect levelling and mixing has been attested in recent years (Lewis et al., 2013).

For the purpose of this project, we collected data from speakers of the Northeastern and Southern dialects. These dialects are mutually intelligible (Appen Language Consultant, Personal Communication, 25 August 2013). Populations of speakers of each of these two dialects are significant; whereas, the Western dialect has relatively fewer speakers, posing potential risk to data collection from this dialect grouping.

Table 1. Dialect Regions and Main Cities (Cf. Аманжолов, 1959)

Dialect Region	Main cities (Population 250,000+)
Northeastern	Karagandy (450,000) Astana[2] (Capital; 400,000) Pavlodar (350,000) Oskemen (350,000) Semey (300,000)
Southern	Almaty[3] (1.3 million) Shymkent (450,000) Taraz (400,000) Türkestan (250,000)
Western (will not be collected)	Aktobe (300,000) Oral (250,000)

[1] Because of the relative familiarity of the English spelling, this document will adopt the spelling *Kazakh* henceforth when referring to the language, the fact that it breaks with other conventions used (namely the use of the grapheme <q> to represent the uvular stop) notwithstanding. In cases where particular place names are not standardized, the Kazakh government Romanization scheme detailed in column 4 of the table in §5.1 is followed.

[2] The last two decades have witnessed significant internal migration within Kazakhstan, especially to the cities of Almaty and Astana. As a result, the traditional dialects of these areas are not necessarily spoken by much of the population of these cities.

[3] Lewis et al. (2013).

Note that Kazakh is also spoken by ethnic Kazakhs in China and Mongolia, as well as communities in Turkey, Russia, and other nearby ex-Soviet states (considered to be diaspora communities). The varieties of Kazakh spoken in China and Mongolia are largely classifiable as belonging to the Northeastern dialect region, with some additional local nuances. Kazakh speakers from outside Kazakhstan will not be included in this data collection.

2. Deviation from native-speaker principle

Only native speakers of Kazakh from Kazakhstan are recruited in this project.

3. Special handling of spelling

Kazakhstan Kazakh is written in Cyrillic script and has reasonably standardized spelling. The spelling variation which does occur is largely related to pronunciation distinctions which are readily confused, e.g., /əj/ and /ǝj/. For example, и is used to represent both /əj/ (which could otherwise be written ый) and /ǝj/ (which would otherwise be written ій). While ій is never written instead of и, ый is occasionally encountered.

Kazakh contains a large number of frequently-occurring Russian loanwords. Nevertheless, the Kazakh alphabet contains the entire letter set of Russian Cyrillic, as well as nine additional letters which are specific to Kazakh. This permits Russian loanwords to be spelled in accordance with Russian orthography, which also has the benefit of avoiding creation of heterographs with true Russian words. In most speakers' idiolects, these loanwords are also generally given a pronunciation which closely resembles the native Russian pronunciation, with respect to phonological phenomena such as word-final devoicing, voicing assimilation processes and vowel reduction in unstressed syllables.

Similarly, in order to standardize strategies for dealing with loanwords from non-Slavic (or otherwise Cyrillic-scripted) languages, foreign proper nouns are represented with respect to their native, Latin-scripted orthography. Complementing this strategy, Kazakh suffixes are separated from the English word by a hyphen and written in full Kazakh Cyrillic. Where Latin-scripted Company names which are compounds occur, these words are bound by underscores. The use of more than one writing system has implications for Romanization (see §5 below).

As the default reference dictionary, we used Малбақов et al. (2006).

4. Description of character set used for orthographic transcription

Kazakh is written using a Cyrillic script composed of the 33-letter Russian alphabet, with 9 additional graphemes to represent specifically Kazakh phones. All 42 symbols are found in the Unicode range U+0400 – U+04FF.

Please note that for English and other European proper nouns that occur in the data, the standard Latin scripted character set (U+0041 – U+007a) is used (see §3 above and §5 below for more information).

Unicode	Kazakh upper case	Unicode	Kazakh lower case
0410	А	0430	а
04d8	Ә	04d9	ә
0411	Б	0431	б
0412	В	0432	в
0413	Г	0433	г
0492	Ғ	0493	ғ
0414	Д	0434	д
0415	Е	0435	е
0401	Ё	0451	ё
0416	Ж	0436	ж
0417	З	0437	з
0418	И	0438	и
0419	Й	0439	й
041a	К	043a	к
049a	Қ	049b	қ
041b	Л	043b	л
041c	М	043c	м
041d	Н	043d	н
04a2	Ң	04a3	ң
041e	О	043e	о
04e8	Ө	04e9	ө
041f	П	043f	п
0420	Р	0440	р
0421	С	0441	с
0422	Т	0442	т
0423	У	0443	у
04b0	Ұ	04b1	ұ
04ae	Ү	04af	ү
0424	Ф	0444	ф
0425	Х	0445	х
04ba	Һ	04bb	һ
0426	Ц	0446	ц
0427	Ч	0447	ч
0428	Ш	0448	ш
0429	Щ	0449	щ
042a	Ъ	044a	ъ
042b	Ы	044b	ы
0406	І	0456	і
042c	Ь	044c	ь
042d	Э	044d	э
042e	Ю	044e	ю
042f	Я	044f	я

Kazakh was previously written using Arabic script (before 1927) and Latin script (1927-1940). A variant of the former system is still used by Kazakh speakers in China (similar to the script used for the related Turkic language Uyghur), and a Latin script based on Turkish is used by Kazakh speakers in Turkey. The government of Kazakhstan plans to introduce a new Latin script in the near future, although previous such plans have been delayed (Novosti, R. 2012). Some government and news websites are already available in Romanized Kazakh (e.g., http://www.inform.kz/indexqaz.html, http://kazgazeta.kz/ , Wikipedia, cf. http://kk.wikipedia.org/wiki/?variant=kk-latn), although there is some variation in specific implementation of the orthography.

5. Description of Romanization scheme

The following table outlines Appen Butler Hill's Romanization scheme, which is fully reversible with the exception of Latin entries (see below). Note that the Kazakh Cyrillic script is fully readable and presents no particular issues. The Romanization scheme is primarily intended to be used as a reference for those unfamiliar with the Cyrillic script or for text-processing purposes which may be facilitated by use of the Romanization. The table also includes the Kazakh government's proposed Romanization scheme as discussed in the previous section.

Note that Latin-scripted text (U+0041 – U+007a) in the transcription will not be Romanized or otherwise converted by the following scheme. As a result, while this Romanization scheme is completely reversible for Cyrillic text, it is not completely reversible for data in a mixture of both Cyrillic and Roman scripts.

5.1 KAZAKH Romanization Scheme

Unicode	Kazakh	ABH Roman	Kazakh Gov't Latin
0410	А	A	A
0430	а	a	a
04d8	Ә	A1	Ä
04d9	ә	a1	ä
0411	Б	B	B
0431	б	b	b
0412	В	V	V
0432	в	v	v
0413	Г	G	G
0433	г	g	g
0492	Ғ	G1	Ğ
0493	ғ	g1	ğ
0414	Д	D	D
0434	д	d	d
0415	Е	E	E
0435	е	e	e
0401	Ё	Yo	Yo
0451	ё	yo	yo
0416	Ж	J	J

Unicode	Kazakh	ABH Roman	Kazakh Gov't Latin
0436	ж	j	j
0417	З	Z	Z
0437	з	z	z
0418	И	I1	Ї
0438	и	i1	ї
0419	Й	Y1	Y
0439	й	y1	y
041a	К	K	K
043a	к	k	k
049a	Қ	Q	Q
049b	қ	q	q
041b	Л	L	L
043b	л	l	l
041c	М	M	M
043c	м	m	m
041d	Н	N	N
043d	н	n	n
04a2	Ң	N1	Ñ
04a3	ң	n1	ñ / ŋ
041e	О	O	O
043e	о	o	o
04e8	Ө	O1	Ö
04e9	ө	o1	ö
041f	П	P	P
043f	п	p	p
0420	Р	R	R
0440	р	r	r
0421	С	S	S
0441	с	s	s
0422	Т	T	T
0442	т	t	t
0423	У	W	W
0443	у	w	w
04b0	Ұ	U	U
04b1	ұ	u	u
04ae	Ү	U1	Ü
04af	ү	u1	ü
0424	Ф	F	F
0444	ф	f	f
0425	Х	X	X
0445	х	x	x
04ba	Һ	H	H
04bb	һ	h	h

Unicode	Kazakh	ABH Roman	Kazakh Gov't Latin
0426	Ц	C1	C
0446	ц	c1	c
0427	Ч	Ch	Ç
0447	ч	ch	ç
0428	Ш	Sh1	Ş
0448	ш	sh1	ş
0429	Щ	Sch1	Şş
0449	щ	sch1	şş
042a	Ъ	"1	"
044a	ъ	"	"
042b	Ы	Y2	I
044b	ы	y2	ı
0406	І	I	İ
0456	і	i	i
042c	Ь	'1	'
044c	ь	'	'
042d	Э	E2	É
044d	э	e2	é
042e	Ю	Yw	Yu
044e	ю	yw	yu
042f	Я	Ya	Ya
044f	я	ya	ya

6. Description of method for word boundary detection

Word boundaries in Kazakh are generally indicated by the use of white space.

Rules regarding the use of comma, periods, dashes, and other common punctuation are similar to those of Russian; however, many people writing Kazakh do not put spaces after such punctuation. Even in texts where this is otherwise not the case, a person's initials (i.e., the first letters of their first name and patronymic) may be written together, e.g., Н.Ә. Назарбаев may be written Н.Ә.Назарбаев by less careful writers. Appen will enforce tokenization (following white space) in all such cases in order to avoid spelling variants due to different tokenization alone.

6.1. Initialisms and Acronyms

In Kazakh, initialisms (where the names of the letters forming an abbreviation are pronounced) and acronyms (where the letters forming an abbreviation are read as a new word) are both attested.

Some examples of initialisms include the following: А_Қ_Ш ("а-қы-шы") 'USA'; Б_Ұ_Ұ ("бә-ұ-ұ") 'U.N.'; Қ_Р ("қы-эр") 'Republic of Kazakhstan', and Р_Ф ("эр-эф") 'Russian Federation'.

Some examples of acronyms include the following: СЭС 'hydroelectric plant'; ЖИТС 'AIDS'; ҰҚШҰ 'Collective Security Treaty Organisation'. There are other types of abbreviations used, including ones like the following: ҚазМУ ("қазму") 'Al-Farabi Kazakh National University'; Еуроодақ ("еуро-одақ", = Еуропалық одағы) 'E.U.'; ЕурАзЭҚ ("еуразэқ", = Еуразиялық

экономикалық қауымдастығы) ‘Eurasian Economic Community’. There are also common foreign abbreviations of various types used in Kazakh. From English, ЮНЕСКО (“юнеско”) ‘UNESCO’; and from Russian: СССР (“(эс-)эс-эс-эр”) ‘USSR’, and СПИД (“спид”) ‘AIDS’. Note that many Russian abbreviations are used colloquially, as people are often not familiar with the Kazakh counterparts.

6.2 Hyphenation

There are many hyphenated words, including many place names (though most are now written together), and compound nouns. Many Kazakh place names are compounds, and are traditionally spelled with ‘-’: Қызыл-Орда, Талды-Қорған, and Алма-Ата. However, these are now mostly written together as one word: Қызылорда, Талдықорған, and Алматы. Quite a few Kazakh words (mostly compounds of various sorts) are spelled with ‘-’, e.g., бала-шаға ‘children’, қадір-қасиет ‘reputation’, and ар-намыс ‘conscience’.

7. Table containing all phonemes in the stipulated notation

For the phonemic transcription of words in this database we will use X-SAMPA symbols, which can be found at http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm. The total number of native phonemes is 34. There are 25 native consonants (including 2 glides/semi-vowels), and 9 vowels (6 monophthongs and 3 diphthongs). There are an additional 28 phonemes that occur almost exclusively in Russian loanwords.

KAZAKH PHONE CHART[4]

Phoneme			Typical Kazakh Correspondence		
IPA	SAMPA	Phonetic variants (IPA)	Orthographic	Unicode	Roman
CONSONANTS					
m	m		м	043C	m
p	p	p^h	п	043F	p
b	b	β, p	б	0431	b
f	f		ф	0444	f
v	v	f	в	0432	v
w	w		у	0443	w
n	n		н	043D	n
t	t	t^h, ts^j	т	0442	t
d	d	t, ð, dz^j	д	0434	d
s	s		с	0441	s
z	z		з	0437	z
ts	ts		ц	0446	c
r	4		р	0440	r
l	l	l, ɬ	л	043B	l

[4] ABH Language Consultant, personal communication; Vajda (1994).

Phoneme			Typical Kazakh Correspondence		
IPA	SAMPA	Phonetic variants (IPA)	Orthographic	Unicode	Roman
ʃ	S	ɕ, ʂ; tɕ, tʂ	ш	0448	sh1
ʒ	Z	ʑ, ʐ; dʑ, dʐ	ж	0436	j
ʃ.ʃ	S.S	ɕɕ, ʂʂ	щ	0449	sch1
j	j		й	0439	y1
ŋ	N	ŋ, ɴ	ң	04A3	n1
k	k	kʰ	к	043A	k
g	g	g, ɣ	г	0433	g
χ	X	χ , x	х	0445	x
q	q	qʰ, χ	қ	049B	q
ʁ	R	ɢ, ʁ	ғ	0493	g1
h	h	x, ∅	h	04BB	h
VOWELS					
i̯ə	i@\		е	0435	e
ə	@\		і	0456	i2
			и	0438	i1
			у	0443	w
			ю	044E	yu
ə	@		ы	044B	y2
			и	0438	i1
			у	0443	W
			ю	044E	yu
æ	{		ә	04D9	a1
			а	0430	a
ɣʉ	y}		ө	04E9	o1
ʉ	}		ү	04AF	u1
ɑ	A		а	0430	a
			я	044F	ya
u̯ʊ	uU		о	043E	o
ʊ	U		ұ	04B1	u
RUSSIAN-SPECIFIC PHONES					
x	x	x, χ	х	0445	x
ɛ	E		э	044D	e2
ɔ	O		о	043E	o
			ё	0451	yo
i	i		и	0418	i1

Phoneme			Typical Kazakh Correspondence		
IPA	SAMPA	Phonetic variants (IPA)	Orthographic	Unicode	Roman
u	u		у	0443	w
			ю	044E	jw
r	r		р	0440	r
mʲ	m'		м	043c	m
nʲ	n'		н	043d	n
pʲ	p'		п	043f	p
bʲ	b'		б	0431	b
tʲ	t'		т	0442	t
dʲ	d'		д	0434	d
kʲ	k'		к	043a	k
gʲ	g'		г	0433	g
fʲ	f'		ф	0444	f
vʲ	v'		в	0432	v
sʲ	s'		с	0441	s
zʲ	z'		з	0437	z
tʃ	tS'	tʂ, ɕ	ч	0447	ch
ʃː	S':	ɕɕ, ʂʂ	щ	0449	sch1
xʲ	x'		х	0445	x
rʲ	r'		р	0440	r
lʲ	l'		л	043b	l
ɨ	1		ы	044b	y2
ɪ	I		и	0438	i1
e	e		е	0435	e
a	a		а,о	0430, 043e	a,o
a:	a:		а	0430	a

7.1 Other symbols

Other Symbols	
ˈ	primary stress
.	syllable break
#	word boundary

8. Other Language Specific Items

8.1 Table of Digits

Numerals can be used before a noun or in place of a noun and inflect in the same way as nouns in Kazakh.[5]

Digits	Cyrillic	SAMPA
0	нөл ноль	nol nol'
1	бір	bir
2	екі	eki
3	үш	u1sh1
4	төрт	to1rt
5	бес	bes
6	алты	alty2
7	жеті	jeti
8	сегіз	segiz
9	тоғыз	tog1y2z

8.2 Other Numbers

Digits	Cyrillic	SAMPA
10	он	on
20	жиырма	ji1y2rma
30	отыз	oty2z
40	қырық	qy2ry2q
50	елу	elw
60	алпыс	alpy2s
70	жетпіс	jetpis
80	сексен	seksen
90	тоқсан	toqsan
100	жүз	ju1z
1,000	мың	my2n1
10,000	он мың	on my2n1
100,000	жүз мың	ju1z my2n1
10 million	он миллион	on mi1lli1on

[5] This table lists cardinal numbers. Ordinal numbers (which can also be used before or in place of nouns) are derived by adding the suffix *-(ı)нш(і)* and also inflect in the same way as nouns. Determiners (*e.g.,* demonstratives, possessives), most adjectives and 'collective numbers' (e.g., *екеу* 'the two, both' *бесеу* 'the five') also behave morphologically very similarly to nouns.

9. References

Аманжолов, С. (1959). *Вопросы диалектологии и истории казахского языка, часть первая*. Алма-Ата: Алма-Атинский государственный педагогический институт имени Абая.

Малбақов, М., Н. Оңғарбаева, and Т. Жанұзақов, eds. (2006). *Қазақ әдеби тілінің сөздігі : Он бес томдық*. Қазақстан Республикасы Білім және ғылым министрлігі, А. Байтұрсынұлы атындағы Тіл білімі институты. Алматы: Арыс.

Krippes, Karl A. (1994). *Kazakh (Qazaq-English dictionary)*. Kensington, MD: Dunwoody.

Novosti, R. (2012). Kazakhstan to Switch to Latin Script by 2025 – President. *Turkish Weekly*. Retrieved 18 July, 2013 from http://www.turkishweekly.net/news/145732/kazakhstan-to-switch-to-latin-script-by-2025-president.html

Omniglot (n.d.) *Kazakh*. Retrieved 18 July, 2013 from http://www.omniglot.com/writing/kazakh.htm

Pang, G-C & Cheng P. G. (2001). *Kazakhstan*. New York: Marshall Cavendish.

Schnitnikov, Boris N. (1997). *Kazakh-English dictionary*. Richmond, U.K.: Routledge.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). (2013). *Ethnologue: Languages of the World,* Seventeenth edition. Dallas, Texas: SIL International. Retrieved July 17, 2013 from http://www.ethnologue.com

UCLA Language Materials Project (n.d.) *Kazakh*. Retrieved July 17, 2013 from http://www.lmp.ucla.edu/profile.aspx?langid=60&menu=004

University of Cambridge Language Centre (n.d.). *Kazakh*. Retrieved July 17, 2013 from http://www.langcen.cam.ac.uk/resources/lang-ik/lang_ik.php?c=8

Vajda, Edward J. (1994). *Kazakh Phonology*. In Studies on East Asia, v. 19, Opuscula Altaica: Essays Presented in Honor of Henry Schwarz. Western Washington University, pp. 603-650.