SIMD Text Processing
Cyrillic to Latin transcoding

In this project we will be exploring the application of SIMD techniques into the large scale text data conversion. In Kazakhstan, Cyrillic alphabet has been used since old soviet times, but nowadays government is in the phase of switching it back to Latin alphabet. This requires a lot of work to convert all the existing Cyrillic texts to Latin, especially the electronic data. There are some simple programs which can perform text transformation between Cyrillic and Latin in a non-parallel fashion. We assume that non-parallel method of text transformation will be very slow for some of the big data conversion. We would like to apply a parallel technique to transform the text using SIMD and compare the results of both of the methods in solving this problem.

As an example, specific patterns will be utilized for transcoding UTF-8 to UTF-16. We will use this idea for our project so we will do a research to find out some patterns to transform between Cyrillic Unicode and its corresponding Latin Unicode. However, finding a visible pattern could be a challenge. For patterns we find, we will be using bit-parallel logic and shifting operations to convert between the characters. In case there are no patterns found between any two of the above mentioned Unicode, we will create a look-up table for mapping by utilizing SIMD registers and then using this mapping table we should be able map the characters from one script to another. However, we may sacrifice some of the register space for this mapping.

Cyrillic	Latin (QazAqparat)
А а	A a
Ә ә	Ä ä
Б б	B b
В в	V vs
Г г	G g
Ғ ғ	Ğ ğ
Д д	D d
Е е	E e
Ё ё	Yo yo
Ж ж	J j
З з	Z z
И и	Ï ï
Й й	Y y
К к	K k
Қ қ	Q q
Л л	L l
М м	M m
Н н	N n
Ң ң	Ñ ñ
О о	O o
Ө ө	Ö ö

Cyrillic	Latin (QazAqparat)
П п	P p
Р р	R r
С с	S s
Т т	T t
У у	W w
Ұ ұ	U u
Ү ү	Ü ü
Ф ф	F f
Х х	X x
Һ һ	H h
Ц ц	C c
Ч ч	Ç ç
Ш ш	Ş ş
Щ щ	Şş şş
Ъ ъ	(″)
Ы ы	I ı
І і	İ i
Ь ь	(′)
Э э	É é
Ю ю	Yw yw
Я я	Ya ya

Kazakh in Cyrillic script	**Kazakh in Latin script**
Барлық адамдар тумысынан азат және қадір-қасиеті мен құқықтары тең болып дүниеге келеді. Адамдарға ақыл-парасат, ар-ождан берілген, сондықтан олар бір-бірімен туыстық, бауырмалдық қарым-қатынас жасаулары тиіс.	Barlıq adamdar tumasınan azat jäne qadir-qasiyeti men quqıqtarı teñ bolıp düniyege keledi. Adamdarğa aqıl-parasat, ar-ojdan berilgen, sondıqtan olar bir-birimen tuwıstıq, bawırmaldıq qarım-qatınas jasawları tiyis.

For Cyrillic, the Unicode range is between U+0400 - U+04FF. For Kazakh version of Cyrillic, the characters that are used fall in the range of U+0400 - U+044F. For example, Cyrillic Ж (zhe) Unicode is U+0416 will be converted to J in Latin whose Unicode is U+004A. There are cases where one Cyrillic Unicode maps to two Latin Unicode, this will be one of our challenges.

Cyrillic script	**Latin script**
Ж - U+0416	J - U+004A
Ғ - U+0492	Ğ - C4 9E (UTF-8) – U+011E

Conclusion:

The performance of this tool will be tested by processing Wikipedia Kazakh text which is written in pure Cyrillic and includes approximately 10 million words and compare the outcome with the non-parallel method by processing the same text.