СЕКЦИЯ 2 | **РЕЧЕВЫЕ ТЕХНОЛОГИИ**

УДК 004.934.2

REGARDING THE IMPACT OF KAZAKH PHONETIC TRANSCRIPTION ON THE PERFORMANCE OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

Muslima Karabalayeva[1], Zhandos Yessenbayev[1], Zhanibek Kozhirbayev[1,2]

[1]National Laboratory Astana, Nazarbayev University, 53 Kabanbay batyr Ave., Astana 010000, Kazakhstan

[2]Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, 2 Satpayev Str., Astana 010008, Kazakhstan

muslima.karabalayeva@nu.edu.kz

Over the past decades automatic speech recognition has made remarkable advances, in both theoretical and practical aspects. Evolution of research in this field has been proceeding from the recognition of individual sounds and phonemes to the recognition of continuous and mixed speech, including tasks of automatic transcription of broadcast news and telephone conversations. Despite the high performance of continuous speech recognition systems, which makes up to 95%, the performance of phoneme recognition systems remains below 85%. However, phoneme recognition is widely used in a number of applications, such as spoken term detection, language identification, speaker identification and others.

The paper presents the results of the experiments on continuous Kazakh speech recognition using different phoneme sets and alternative phonetic transcriptions. This study was instigated by the fact that in modern Kazakh linguistics there is no common agreement about the phonetic system of the Kazakh language, while the list of phonemes and their number noticeably vary in different textbooks. Therefore, we aimed our experiments to study the impact of the phonetic system of the language, its orthoepic rules and the corresponding phonetic transcriptions on the performance of the phoneme recognition systems, which are the initial stage in the general systems of continuous speech recognition.

The following 6 systems of phonetic transcription have been considered and tested in our study. The first one is a project of the new Kazakh alphabet and a set of spelling rules proposed by Prof. A. Sharipbay. The second system is a set of orthoepic rules for the actual Kazakh Cyrillic alphabet, introduced by Kazakh linguists – the authors of the Kazakh "Orthoepical Dictionary". The third one of the systems considered is a phonetic system and a set of empirical transcription rules used by the authors of this work in their studies. The fourth variant is based on the actual Kazakh Cyrillic alphabet without taking into account any orthoepic rules, i.e. a transcription system in which one letter corresponds to one phoneme. The remaining two systems are variations or combinations of these systems mentioned above.

In total, three series of experiments were conducted: word-based recognition and two series of phone-based recognition. The latter two differ in test sets. Word-based experiments did not reveal any special differences in the recognition performance among the systems studied, which is due to the strong impact of the language model on the decoding process. On the contrary, phone-based experiments showed that: 1) the existing orthoepic rules are not fully adequate to the actual sounding of Kazakh speech; 2) the existing phonetic system of the Kazakh language can be optimized by removing some phonemes. The speech recognition system is implemented using the Kaldi platform.

Despite the fact that the present work is a preliminary study, on the whole, the presented experimental results make it possible to evaluate the adequacy of considered phonetic transcriptions to the actual sounding of Kazakh speech, and can be of particular interest in view of the forthcoming transformation of the Kazakh writing system.

Keywords: automatic speech recognition; phoneme recognition; phonetic transcription; Kazakh speech.

О ВЛИЯНИИ ФОНЕТИЧЕСКОЙ ТРАНСКРИПЦИИ КАЗАХСКОГО ЯЗЫКА НА КАЧЕСТВО АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Карабалаева Муслима Хизятовна, Есенбаев Жандос Аманбаевич[1], Кожирбаев Жанибек Мамбеткаримович[1,2]

[1]Национальная Лаборатория Астана, Назарбаев Университет, Астана, Казахстан

[2]Факультет информационных технологий, Евразийский университет имени Л.Н. Гумилева, Астана, Казахстан

muslima.karabalayeva@nu.edu.kz

Статья описывает результаты экспериментов по машинному распознаванию казахской речи с использованием разных систем фонетической транскрипции казахского языка. Эксперименты проведены на небольшом

речевом корпусе казахских предложений, озвученных дикторами в студийных условиях.

Целью экспериментов является сравнительный анализ нескольких систем фонетической транскрипции казахского языка, основанных на различном количестве фонем и использующих различные наборы орфоэпических правил. Результаты описанных экспериментов позволяют сделать выводы об адекватности зафиксированных орфоэпических правил реальному звучанию казахской речи.

Полученные результаты представляют особый интерес в связи с актуальной задачей реформирования казахской письменности, которое неизбежно затронет и фонетику, и орфоэпию казахского языка.

Ключевые слова: автоматическое распознавание речи; пофонемное распознавание; фонетическая транскрипция; казахская речь.

Over the past decades automatic speech recognition has made remarkable advances, in both theoretical and practical aspects. Evolution of research in this field has been proceeding from the recognition of individual sounds and phonemes to the recognition of continuous and mixed speech, including tasks of automatic transcription of broadcast news and telephone conversations. Despite the high performance of continuous speech recognition systems, which makes up to 95%, the performance of phoneme recognition systems remains below 85% (Lopes, Perdigão, 2011; Graves, Mohamed, Hinton, 2013). This is due to the fact that modern speech recognition systems are based on language models that are able to overcome the errors of the first stage of recognition, which is phonetic transcription. However, such systems have certain limitations imposed on the system's lexicon. Therefore, with no detraction from the merits of such systems, let us note that the task of recognizing phonemes by itself still remains relevant. This task has an important application value for speech recognition systems, which are constructed solely for the solution of this problem and which do not have those vocabulary restrictions mentioned above. Besides these continuous speech recognition systems, phoneme recognition is also widely used in a number of applications, such as spoken term detection (Schwarz, 2008), language identification (Matejka, 2009), speaker identification and others (Furui, 2005).

The present work is aimed to perform a comparative analysis of a few phoneme recognition systems for continuous Kazakh speech, which are based on different phoneme sets and different orthoepic rules, which, in their turn, generate alternative phonetic transcriptions necessary for constructing speech recognition systems.

This study was instigated by the fact that in modern Kazakh linguistics there is no common agreement about the phonetic system of the Kazakh language, while the list of phonemes and their number noticeably vary in different textbooks (Sharipbay, 2017, p. 20–22). Therefore, within the framework of our study, we considered several phonetic transcription systems proposed by different authors.

One of these is a project of the new Kazakh alphabet and a set of spelling rules proposed by Prof. A. Sharipbay in an attempt to introduce an optimal phonetic system as part of the transformation of the Kazakh writing system (Sharipbay, 2017). Another system is a set of orthoepic rules for the actual Kazakh Cyrillic alphabet, introduced by Kazakh linguists – the authors of the Kazakh "Orthoepical Dictionary" (Aitbaiuly et al., 2007). Yet another one of the systems considered is a phonetic system and a set of empirical transcription rules used by the authors of this work in their studies. There also was tested a variant on the basis of the actual Kazakh Cyrillic alphabet without taking into account any orthoepic rules, i.e. a transcription system in which one letter corresponds to one phoneme. The remaining two systems are variations or combinations of these systems mentioned above.

The experiments were conducted using high-quality speech signals recorded in unified studio conditions. This allowed us to avoid system errors, which could arise due to the diversity of acoustic conditions and the dissimilar quality of the sound recording equipment. The speech recognition system is deployed on the Kaldi platform (Povey et al., 2011), using a cluster for distributed computing managed by the batch-queuing system Univa Grid Engine, formerly the Sun Grid Engine (Univa Grid Engine, 2017).

The results of the described experiments make it possible to draw conclusions about the adequacy of the considered orthoepic rules to the actual sounding of Kazakh speech. The obtained results are of particular interest in view of the actual problem of the transformation of the Kazakh writing system, which will inevitably affect both phonetics and orthoepy of the Kazakh language.

1. Description of the speech corpus

The experiments have been conducted on two compact-sized speech corpora of the Kazakh sentences *kazspeech* (Makhambetov et al., 2013) and *enuspeech* (Tech. Report, 2014), recorded from native

speakers in a sound recording studio. The total duration of these audio files in the two corpora makes up 1832 minutes of speech (more than 30 hours). The whole set was divided into a training set, a development set, and a test set in a ratio of 81% / 10% / 9%, regarding the duration of audio files. At the same time, the ratio of male and female voices in each of the three sets has been kept approximately equal. The characteristics of the speech corpus are shown in Table 1. The last column contains the values of the OOV (out-of-vocabulary) rate, which is computed as the ratio of the number of unknown words in a set (i.e. words not found in the system's lexicon) to the total number of words in the set.

Table 1. **Characteristics of the speech corpus**

Set	Number of sentences	Duration (*hr*)	Number of speakers	Male speakers	Female speakers	OOV Rate
Train set	15937	24.8	179	76	103	0 %
Dev set	1875	2.92	21	9	12	19.67 %
Test set	1875	2.82	21	9	12	18.75 %

Each uttered sentence in the corpus has its corresponding orthographic transcription, which is the text of the sentence written in the actual Kazakh Cyrillic alphabet.

2. Description of the phonetic transcription systems

Each transcription rule is a single text line representing a rule of substitution. A rule of substitution consists of three parts: the substituted text, the equal sign, and the new text. For example:

```
Һ=Х
ЬЕ=ЙЕ
Ь=
```

Here it is possible to use regular expressions (Regular expression operations, 2017) and syntax of 'sed' scripts ('sed' manual, 2017). For example:

```
ГЫ$=ГІ
\(.\+\)Б$=\1П
```

If several transcription alternatives are available for the text being substituted, then the rule of substitution contains two or more equal signs. For example:

СБ=СП=ЗБ
ТБ=ТП=ДБ

A set of transcription rules is applied line by line to the input text, which is an orthographic transcription of the uttered sentences – a sequence of letters, while the output is its phonetic transcription – the sequence of phonemes. In this procedure, the order of the rules in the list matters, since earlier rules are applied earlier. And if you change the order of the rules in the list, the output transcription may also change.

The following 6 systems of phonetic transcription have been considered and tested in our study.

1.1. Grapheme = phoneme

This system of phonetic transcription is the most primitive, and it implies that the spelling of Kazakh words exactly matches their pronounce.

The phoneme list contains 42 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ё, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ, Ъ, Ы, І, Ь, Э, Ю, Я. This list coincides with the 42 letters of the actual Kazakh Cyrillic alphabet. Let us note that although Ъ (the hard sign) and Ь (the soft sign) are not formally phonemes, in our experiments they were modeled as separate phonetic units.

There are no transcription rules in this system.

1.2. Grapheme = phoneme, without Ё, Ъ, Ь

This system is similar to the previous one, but excludes three symbols: Ё, Ъ and Ь.

The letter Ё in the Kazakh language is found only in words adopted from Russian. In Russian, the usage of this letter in written texts is not mandatory. In the overwhelming majority of cases, Ё is substituted by Е in written texts, and this is exactly what happened to the orthographic transcriptions in our speech corpus. As a result, in the texts of the training set, the letter Ё occurs only three times (in words *шахтёрлер*, *актёрлердің* and *Гёте*), whereas in all other cases it is most likely written as Е.

The symbols Ъ and Ь are excluded from the system, since they do not denote any separate phoneme.

The phoneme list contains 39 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ,

Ы, І, Э, Ю, Я. This list coincides with the letters of the actual Kazakh Cyrillic alphabet, excluding symbols Ё, Ъ, Ь.

The transcription rules are:

Ё=Е

Ъ=

Ь=

This list contains only three transcription rules, meaning that the letter Ё should everywhere be replaced by Е, and the letters Ъ and Ь should be removed from the phonetic transcription: *шахтёрлер* [*шахтерлер*], *Ельцин* [*елцин*], *съезд* [*сезд*].

Here is a few sample transcribed words:

СУБЪЕКТІЛЕРІ → СУБЕКТІЛЕРІ

АНСАМБЛЬДІҢ → АНСАМБЛДІҢ

ФЕЛЬДФЕБЕЛЬ → ФЕЛДФЕБЕЛ

Sample source sentence:

ОСЫНДАЙ КҮДІКТІ ОПЕРАЦИЯЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЪЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

Sample transcribed sentence:

ОСЫНДАЙ КҮДІКТІ ОПЕРАЦИЯЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

1.3. Basic transcription

We gave the name of "basic transcription" to our own simple phonetic transcription system, which, firstly, takes into account the nascence of diphthongs, and, secondly, reduces the number of phonemes by joining rare phonemes into one class with other more frequent phonemes.

As used herein, the term "basic transcription" implies that this transcription system includes, in our opinion, the minimum necessary list of common transcription rules, applicable to the vast majority of words of the Kazakh language, both native and loan. And, conversely, this basic transcription does not include rules that describe nuances and complicated orthoepic cases. Also, the basic transcription does not consider cases when one word might have several alternative phonetic transcriptions.

For instance, if the letter Е is found at the beginning of the word, then it is always pronounced as the diphthong ЙЕ (*ел* [*йел*], *емес* [*йемес*]). In the same way, it is pronounced as the diphthong ЙЕ when found after a vowel, the hard sign or the soft sign (*ниет* [*нийет*], *пьеса* [*пйеса*]).

Likewise, the letter Ю, when found at the beginning of the word, after a vowel, the hard sign or the soft sign, is pronounced as ЙУ or ЙҮ (*Юрмала* [*йурмала*], *көбею* [*көбейү*]). When found after consonants, it gets close to the sound Ү (*жюри* [*жүри*], *эволюция* [*эволүция*]).

In the similar way, the letter Я at the beginning of the word, after vowels or the soft sign can be transcribed by the diphthong ЙА or ЙӘ (*қияр* [*қыйар*], *паэлья* [*паэлйә*]). After consonants, it gets close to the sound Ә (*лямбда* [*ләмбда*]).

Thus, we can remove the symbols Ю and Я from the phoneme list. Rarely used phoneme Һ can be joined into one class with the phoneme Х. Also, we can join the phoneme Щ into one class with the phoneme Ш, because Щ is the soft allophone of Ш.

In the basic transcription we introduced an additional symbol W to denote the short consonant У, which does not make a syllable and which can be considered as part of some diphthongs. For instance, it can be found between two vowels (*ауа* [*awa*]), at the beginning of words (*уақыт* [*wақыт*], *уәде* [*wәде*]), and also it often occurs at the beginning of words when the first letter is Ө or Ө (*он* [*wон*], *өйткені* [*wөйткені*]).

The phoneme list contains 36 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, W, Ұ, Ү, Ф, Х, Ц, Ч, Ш, Ы, І, Э.

The transcription rule list contains 139 rules. If necessary, the complete list of all transcription rules mentioned in the present paper can be provided upon request.

Sample transcribed words:

СУБЪЕКТІЛЕРІ → СУБЙЕКТІЛЕРІ

ОЙНАУҒА → WОЙНАWҒА

ОКТЯБРЬСК → WОКТӘБРСК

Sample transcribed sentence:

WОСЫНДАЙ КҮДІКТІ WОПЕРАЦИЙАЛАРДЫ ҚАРЖЫ МОНИТОРИНГІ СУБЙЕКТІЛЕРІ WОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМИТЕТКЕ ТАПСЫРАДЫ

1.4. Transcription based on the orthoepic dictionary

This system of phonetic transcription was based on the rules described in Chapter "Orthoepic rules or word sounding principals" of the Kazakh "Orthoepical dictionary" (Aitbaiuly et al., 2007, p. 770).

Unfortunately, it was not possible to formalize and apply in our system a complete list of all the rules described in the dictionary, since some of these rules take into account the origin of words. That is, orthoepic rules may differ depending on whether they are applied to original Kazakh words or words adopted from other languages (Russian, Arabic, Persian, etc.).

However, we tried to formalize and apply all the other rules described in the dictionary, those being the same (or having a priority use case) for all words, both native Kazakh and loanwords.

The phoneme list contains 39 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Һ, Ц, Ч, Ш, Щ, Ы, І, Э, Ю, Я. This list coincides with the letters of the actual Kazakh Cyrillic alphabet, excluding symbols Ё, Ъ, Ь.

The transcription rule list contains 252 rules.

Sample transcribed words:

ӨКІНУГЕ → УӨКҮНҮГЕ

ОКТЯБРЬСК → УОКТЯБІРСК

ШАЙБАНИ → ШӘЙБАНЫЙ

Sample transcribed sentence:

УОСҰНДАЙ КҮДҮКТІ УОПЕРАЦЫЙАЛАРДЫ ҚАРЖЫ МОНЫЙТОРІЙНГІ СУБЙЕКТІЛЕРІ УОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІЙТЕТКЕ ТАПСЫРАДЫ

1.5. Combined transcription rules

This system of phonetic transcription combines the two previous ones: at first we apply the rules according to the orthoepical dictionary (system 4), then we apply the rules of the basic transcription (system 3). This allows us, in addition to the Kazakh orthoepic rules, to take into consideration diphthongs (in those cases not considered in the orthoepical dictionary) and to reduce the number of phonemes to 36.

The phoneme list contains 36 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, И, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, W, Ұ, Ү, Ф, Х, Ц, Ч, Ш, Ы, І, Э.

The transcription rule list contains 380 rules.

Sample transcribed words:
ОКТЯБРЬСК → WОКТӘБІРСК
АҢҚИЫП → АҢҒЫЙЫП
ЖАПЫРАҚ → ЖАПЫРАҚ, ЖАПРАҚ
Sample transcribed sentence:
WОСҰНДАЙ КҮДҮКТІ WОПЕРАЦЫЙАЛАРДЫ ҚАРЖЫ МОНЫЙТОРІЙНГІ СУБЙЕКТІЛЕРІ WОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІЙТЕТКЕ ТАПСЫРАДЫ

1.6. Transcription proposed by Prof. A. Sharipbay

This system of phonetic transcription was based on the rules proposed and described in the monograph of Professor Altynbek Sharipbay (Sharipbay, 2017).

The automatic online transcriptor "Transcription of Kazakh Cyrillic writing into Latin" (Online Transcriptor, 2017) was used to generate phonetic transcriptions on the basis of these rules.

The phoneme list contains 31 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Ш, Ы, І.

The number of transcription rules is unknown.

Sample transcribed words:
ЦИРКУЛЯЦИЯСЫ → СІРКҮЛӘТСЫЙӘСЫ
ДУЭТ → ДҮЕТ
КОНЪЮНКТУРА → КОНЫЙҮНКТҰРА
Sample transcribed sentence:
ОСЫНДАЙ КҮДІКТІ ОПЕРАТСЫЙӘЛАРДЫ ҚАРЖЫ МОНІТОРІҢГІ СҮБЫЙЕКТІЛЕРІ ОЛАР АНЫҚТАЛҒАН СӘТТЕН БАСТАП КОМІТЕТКЕ ТАПСЫРАДЫ

3. The word-based experiment

We carried out a series of experiments on acoustic modeling and speech recognition over the described speech corpus (see Section 1) using the described phonetic transcription systems (see Section 2). The experiments were carried out on the Kaldi platform, using a cluster for distributed computing and a batch-queuing system.

At the first stage, word-based experiments were conducted, where the recognizable speech units were the words of the Kazakh language. Each uttered sentence was interpreted as a sequence of words separated by spaces.

The lexicon and the trigram language model (Federico, Bertoldi, Cettolo, 2008; The IRSTLM Toolkit, 2017) had been constructed from texts of the training set only, so the OOV rate was high (about 20%, see Table 1).

The experiments were carried out according to the standard Kaldi ASR algorithm (Povey et al., 2011; Yessenbayev, Karabalayeva, 2016), which was a sequential training of a monophone model, then several triphone models, and finally, an SGMM model.

To estimate the performance of a recognition system we use the value of WER (word error rate), which is computed as the ratio of erroneously recognized words to the total number of recognized words. WER is a common metric of the performance of experimental speech recognition models. The lower WER is, the better system performance is.

The results of the word-based experiment are shown in Table 2. The best results were obtained for the systems without transcription (systems 1, 2) and with the basic transcription (system 3). Here it can be noted that changing the phonetic transcription system does not lead to significant changes in the system performance on the test set (dispersion is about 0.5% on an absolute scale, about 1.1% relative to the lowest WER). As mentioned above, this is due to the fact that the language model alleviates and overcomes the errors that arise in the acoustic model of the system. Also, we can observe a big difference between the WER values on the training set and the test set. It is due to the high OOV rate in the test set, which again confirms the essential dependence of speech recognition systems on the lexicon and the language model.

Table 2. **The results of the word-based experiment, WER**

N	Phonetic transcription system	Train, %	Dev, %	Test, %
1	Grapheme = phoneme	0.50	40.28	36.88
2	Grapheme = phoneme, without Ё, Ъ, Ь	0.52	40.19	36.79
3	Basic transcription	0.54	40.21	36.93
4	Transcription based on the orthoepic dictionary	0.56	40.60	37.06
5	Combined transcription rules	0.59	40.71	37.21
6	Transcription proposed by Prof. A. Sharipbay	0.57	40.93	37.04

4. The phone-based experiment

At the next stage, independent phone-based experiments were conducted, where the recognizable speech units were phonemes. Each uttered sentence was interpreted as a sequence of phonemes separated by spaces.

As previously, the lexicon and the language model had been constructed from texts of the training set only. The experiments were carried out according to the standard Kaldi ASR algorithm.

The results of the phone-based experiment are shown in Table 3. As previously, the best results are obtained for the first three systems, with the leader being system 1. You can see that here, unlike the previous experiment, the WER values on the test and training sets are comparable, i.e. the dependence on the lexicon and the language model of the system is minimized. This factor makes the results of this experiment much more significant than the results of the previous word-based experiment (see Section 3). With regard to phonetic transcription systems, the error dispersion in the test set is notably larger than in the word-based experiment (about 2.9% on an absolute scale, about 17% relative to the lowest WER). The order of the error is also comparable with similar systems for English (Lopes, Perdigão, 2011; Graves, Mohamed, Hinton, 2013).

Table 3. **The results of the phone-based experiment, WER**

N	Phonetic transcription system	Train, %	Dev, %	Test, %
1	Grapheme = phoneme	12.15	18.46	16.78
2	Grapheme = phoneme, without Ё, Ъ, Ь	13.40	20.05	18.00
3	Basic transcription	13.35	19.48	17.94
4	Transcription based on the orthoepic dictionary	13.77	19.98	18.33
5	Combined transcription rules	14.67	21.63	19.43
6	Transcription proposed by Prof. A. Sharipbay	14.93	21.42	19.66

5. The test manual annotation experiment

Based on the experimental results presented in the previous section, we can already draw certain conclusions about the performance of the described systems of Kazakh phonetic transcription. However,

it is not entirely correct to compare these systems directly with each other, because they use different sets of phonemes, and therefore their test sets differ.

In order to compare the performance of the described phonetic transcription systems, we have chosen *a minimal phonetic alphabet* based on the phonetic system proposed by Prof. A. Sharipbay (Sharipbay, 2017, p. 104-105), since this alphabet contains the smallest number of phonemes among all considered phonetic systems.

For phonetic systems with a greater number of phonemes than in the minimal phonetic alphabet, we have defined the following rules of substitution in order to get rid of redundant phonemes in recognized texts:

```
Ц → ТС
Ч → ТШ
Э → Е
Һ → Х
Щ → Ш
Ю → Ү
Я → Ә
И → І
Ё → Е
Ъ →              (remove)
Ь →              (remove)
```

In addition, in the transcription systems (2) and (4) that contain the symbol `W` denoting the short consonant `У`, we have applied two more rules of substitution, in the order indicated:

```
У → Ұ
W → У
```

Using this minimal phonetic alphabet, we have annotated a small phonetically representative speech corpus *test_manual*, as an independent test set. This corpus contains 83 Kazakh sentences, uttered by three speakers (two males, one female). These 83 sentences consist of 534 words and 3220 letters. All the letters of the Kazakh Cyrillic alphabet are found in the corpus at least 11 times, except for the letters Ё (0 times), Щ (8 times), Ъ (6 times). All sentences of the corpus were transcribed manually, by careful listening to the uttered sentences and writing down their phonetic transcription.

The phoneme list contains 31 phonemes: А, Ә, Б, В, Г, Ғ, Д, Е, Ж, З, Й, К, Қ, Л, М, Н, Ң, О, Ө, П, Р, С, Т, У, Ұ, Ү, Ф, Х, Ш, Ы, І.

In this list, the symbol У is used only to denote the short consonant sound as part of diphthongs (just like the W symbol in the basic transcription system (2), see Section 2). Whereas the vowel У, which makes a syllable and can be found in loanwords, is transcribed either by the symbol Ұ or the symbol Y – depending on the hardness or softness of neighboring phonemes.

The symbol И is excluded from the system, since И is considered as an allophone of the phoneme I.

Besides that, we have excluded the symbol Һ (transcribed as X); the symbol Ц (transcribed as TC or C); the symbol Ч (transcribed as ТШ); the symbol Щ (transcribed as Ш); the symbol Э (transcribed as Е); the symbol Ю (transcribed as ЙҰ, or ЙY, or Y); the symbol Я (transcribed as ЙА, or ЙӘ, or Ә). The symbol Ё does not occur in the set.

Sample source sentences:

ОРЫСТЫҢ ТАМАША САТИРИК ЖАЗУШЫ МИХАИЛ МИХАЙЛОВИЧ ЗОЩЕНКО

СУБВЕНЦИЯ КӨЛЕМІ ЕКІ МИЛЛИАРД ТЕҢГЕ

РАДИЩЕВ ПЕТЕРБУРГТЕН МӘСКЕУГЕ САЯХАТ

Sample transcribed sentences:

УОРЫСТЫҢ ТАМАША САТІРІГ ЖАЗҰШЫ МІХАЙЫЛ МІХАЙЛАВІШ ЗОШЕНКА

СҰББВЕНСІЙӘ КӨЛЕМІ ЙЕКІ МІЛІАРТ ТЕҢГЕ

РАДЫЙШЕФ ПЕТӨРБУРКТЕН МӘСКЕУГЕ САЙАХАТ

The results of the test manual annotation experiment are shown in Table 4. The best result was shown by the system with basic transcription (3), and the second best is the system without transcription (1). At the same time, the inferiority of the system based on the orthoepic dictionary (4) and the system without Ё, Ъ, Ь (2) compared with the system 1 is not essential.

Table 4. **The results of the test manual annotation experiment, WER**

N	Phonetic transcription system	Test, %
1	Grapheme = phoneme	23.38
2	Grapheme = phoneme, without Ё, Ъ, Ь	23.68
3	Basic transcription	23.05
4	Transcription based on the orthoepic dictionary	23.48
5	Combined transcription rules	24.50
6	Transcription proposed by Prof. A. Sharipbay	25.03

6. Conclusions

The paper presents the results of the experiments on Kazakh speech recognition using different phoneme sets and alternative phonetic transcriptions. Based on the results obtained, it can be concluded that a system with basic transcription, which takes diphthongs into account and excludes rare loan phonemes, provides the most adequate model of real Kazakh speech. On the contrary, systems using transcriptions based on more complex orthoepic rules showed an inferior recognition performance, which, presumably, was due to some inconsistency of these rules with the actual sounding of Kazakh speech. Indeed, during the time that the actual Kazakh Cyrillic alphabet and the phonetic system on its basis are being used, native speakers have been able to adapt to foreign loan phonemes. Now they are able to utter them as required (e.g. in Russian), in an almost authentic manner. This conclusion is also supported by quite good recognition results obtained by the systems without using any orthoepic rules (in the "Grapheme = phoneme" experiments).

It can also be observed that reducing the number of phonemes to 36 (in the basic transcription) is absolutely harmless and even leads to an increase in the recognition performance. This seems to be due to that phonemes excluded from the phonetic system can be legitimately considered either as diphthongs or as allophones of other phonemes, more frequent in speech. At the same time, in the process of automatic recognition, we do not observe any negative confusion effects that could appear in consequence of classifying such phonemes to the wrong classes. On the contrary, such negative effects get reduced, and the recognition performance gets somewhat increased. This testifies to the correctness of the substitution rules that we use to remove the symbols Ё, Ю, Я, Һ, Щ, Ъ, Ь.

Further reducing the number of phonemes to 31 (in the minimal phonetic alphabet) leads to a decrease in the recognition performance. However, this can be explained by the complexity of the system of orthoepic rules, rather than by shortening the phoneme list. Let us also note that a certain portion of errors in the last experiment (with the manually annotated corpus) may result from the rough mapping of phonetic systems with a larger number of phonemes into the system with the minimal phonetic alphabet.

This work is a preliminary study. A more detailed study will require extensive analysis of the contribution of each single phoneme into the performance of continuous speech recognition. Only on the basis of such an analysis we shall be able to elaborate and introduce an adequate phonetic system with appropriate transcription rules. That is going to become the subject of our future work.

Acknowledgement

This work has been conducted under the targeted program O.0743 (0115PK02473) of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan.

REFERENCES

1. Lopes, C., Perdigão F. (2011). Phoneme recognition on the TIMIT database. Speech Technologies, 2011, pp. 285–302.

2. Graves, A., Mohamed, A.r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, 2013, pp. 6645–6649.

3. Schwarz, P. (2008). Phoneme recognition based on long temporal context. PhD thesis, Brno University of Technology, Faculty of Information Technology, 2008.

4. Matejka, P. (2009). Phonotactic and Acoustic Language Recognition. PhD thesis, Brno University of Technology, Faculty of Electrical Engineering and Communication, 2009.

5. Furui, S. (2005). 50 Years of Progress in Speech and Speaker Recognition Research. Proceedings of ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, 2005.

6. Sharipbay, A. (2017). Kazakh zhazuin latyn alipbiyine auistyru maseleleri. Monograph. Astana: L.N. Gumilyov Eurasian National University, 2017. – 136 p. – ISBN: 978-601-301-989-5.

7. Aitbaiuly, O., et al. (2007). Orthoepical dictionary. Ed. Aitbaiuly O., et al. – Almaty: Arys, 2007. – 800 p. – ISBN: 9965-17-473-3.

8. Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukas and Glembek, Ondrej and Goel, Nagendra and Hannemann, Mirko and Motlicek, Petr and Qian, Yanmin and Schwarz, Petr and Silovsky, Jan and Stemmer, Georg and Vesely, Karel. (2011). The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

9. Univa Grid Engine, official site. (2017). URL: www.univa.com [Access date: 15.07.2017].

10. Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., and Sharafudinov, A. (2013). Assembling the Kazakh Language Corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, October. Association for Computational Linguistics, pp. 1022–1031.

11. Technical Report (final). (2014). Sozdaniye akusticheskogo korpusa kazahskogo yazyka i utochneniye ego foneticheskogo stroya, predstavleniye kazahskih fonem v mezhdunarodnom foneticheskom alfavite. L.N. Gumilyov Eurasian National University; PI Sharipbay A.A.; exec.: Yessenbayev Zh.A. [et al.]. – Astana, 2014. – 68 p. – Reg No. 0112PK02344.

12. Regular expression operations, The Python Standard Library. (2017). URL: https://docs.python.org/2/library/re.html [Access date: 15.07.2017].

13. 'sed' manual. – GNU. (2017). URL: https://www.gnu.org/software/sed/manual/sed.txt [Access date: 15.07.2017].

14. Online Transcriptor. (2017). Transcription of Kazakh Cyrillic writing into Latin. URL: http://e-zerde.kz/latin/index22.php [Access date: 15.07.2017].

15. Yessenbayev, Zh., Karabalayeva, M. (2016) A baseline system for Kazakh broadcast news transcription. In the Proc. of the V International Scientific-Practical Conference on Informatization of Society. Astana, Kazakhstan, 2016, pp. 48–50.

16. Federico, Marcello and Bertoldi, Nicola and Cettolo, Mauro. (2008). IRSTLM: an open source toolkit for handling large scale language models. In the Proc. of Interspeech 2008, pp. 1618–1621.

17. The IRST Language Modeling (IRSTLM) Toolkit. (2017). URL: http://hlt-mt.fbk.eu/technologies/irstlm [Access date: 15.07.2017].