ISSN 2007-9737

Resurrection: The Khazar Language Reconstruction Using Computer Science Technologies

Elina Makipova[1], Iskander Akhmetov[1,2], Alexander Gelbukh[*,3]

[1] KIMEP University, College of Humanities and Education,
Kazakhstan

[2] Insitute of Information and Computational Technologies, Almaty,
Kazakhstan

[3] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

elina.makipova@kimep.kz, i.akhmetov@ipic.kz,
gelbukh@cic.ipn.mx

Abstract. Decrypting or reconstructing extinct languages is challenging, especially when the objective is to reconstruct a language with no or very few texts left, such as the Khazar language or early Slavic and Ugric languages. In this paper, we lay out the historical perspective of the Khazar people, their language, and contemporary descendant ethnic groups, namely the Chuvash and Tatar people. Then we discuss ways Computer Science can help researchers in language reconstruction and decryption. Finally, we pilot an approach to find Khazar/Bulgar word candidates in Chuvash and Tatar languages by (1) normalizing the words of two languages and (2) comparing them, accounting for the semantic concepts to solve the homonymy problem, and (3) excluding common Turkic words and borrowings from the Russian language.

Keywords. Khazar, language reconstruction, extinct languages, historical linguistics.

1 Introduction

Nowadays, there are more than thousands of different languages that can disappear.

Of the approximately 6,000 existing languages in the world, more than 200 have become extinct during the last three generations, 538 are critically endangered, 502 are severely endangered, 632 are definitely endangered, and 607 are unsafe [14]. However, some people think it is unimportant, because languages are much easier nowadays than before, so there is no need to learn and study them.

Moreover, it can seem unnecessary because no one speaks these languages, so there is no need to recognize them. However, reconstructing or decrypting an extinct language can significantly benefit by filling up the gaps in our historical knowledge and linguistics.

For instance, the language of ancient Egyptians can seem useless because people in Egypt do not speak this language anymore and use Arabic instead. Nevertheless, there are many scriptures in the Pyramids of Giza which scientists decrypt to learn more about the history, life, and tradition of the people living millennia before in the region.

The Voynich manuscript decryption is another example of the ancient language deciphering task which is not been solved to date [21]. However, what can we do about an ancient language that left no written artifacts to decrypt?

One of the approaches from historical linguistics is called Language Reconstruction, when we use a language or a set of languages

doi: 10.13053/CyS-28-1-4902

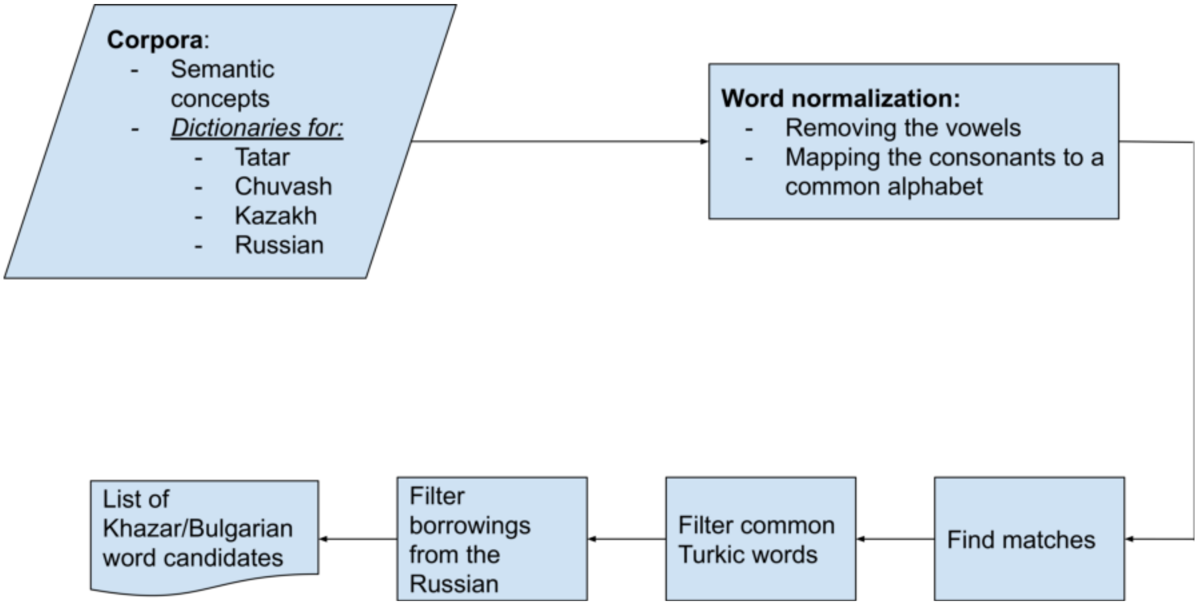
Corpora:
- Semantic concepts
- *Dictionaries for:*
- Tatar
- Chuvash
- Kazakh
- Russian
Word normalization:
- Removing the vowels
- Mapping the consonants to a common alphabet
Find matches
Filter common Turkic words
Filter borrowings from the Russian
List of Khazar/Bulgarian word candidates

Fig. 1. Screening for Khazar/Bulgar candidate words

known to be descending from an ancient language and try to reconstruct the ancestor language seeking for language anomalies and comparing the languages to discover common lexicon.

Thus, there are many works on reconstructing the proto-Indoeuropean language known to be the ancestor of all Indoeuropean languages and the works on reconstructing the proto-Turkic language [4]. In this article, we attempted to reconstruct some Khazar words which once were used in the Khazar khaganate.

The country spread from the Aral Sea in the East to the Crimean peninsula in the West in early Medival times. Scientists know very little about these mysterious people; still, their language has left no written evidence other than some personal names and toponyms we can find from the Arabian and Byzantium historians' works [5].

There are a lot of linguists and scientists who tried to unravel this language. However, they could know only a bit.

This example shows that an extinct language is a key to understanding the natural history of a particular nation. Reconstructing extinct languages is a challenging problem of an interdisciplinary nature, touching such areas of research as history, geography, linguistics, Computer Science (Artificial Intelligence, Computational Linguistics, Natural Language Processing), and others.

Our approach employed a comparative method of language reconstruction using Chuvash, Tatar, and Kazakh languages. It consisted of (1) normalizing the words by eliminating the vowel characters and mapping consonant characters of the compared languages to a standard alphabet and (2) finding matches between normalized Chuvash and Tatar words, which additionally share the same semantic concept to tackle the homonymy problem, (3) filter out the common Turkic words by eliminating the matches between Chuvash and Kazakh languages (as the Kazakh language is known to have no Khazar/Bulgar background), (4) filter out the words borrowed from the Russian language; see Figure 3.

The contribution of this work to the scientific knowledge is in (1) the approach and algorithm for discovering the Khazar/Bulgar word candidates in modern Chuvash and Tatar languages and (2) dataset with normalized words for Chuvash, Tatar, Kazakh and Russian languages[1].

[1] The code and data are available at github.com/iskander-akhmetov/Khazar-language-resurrection

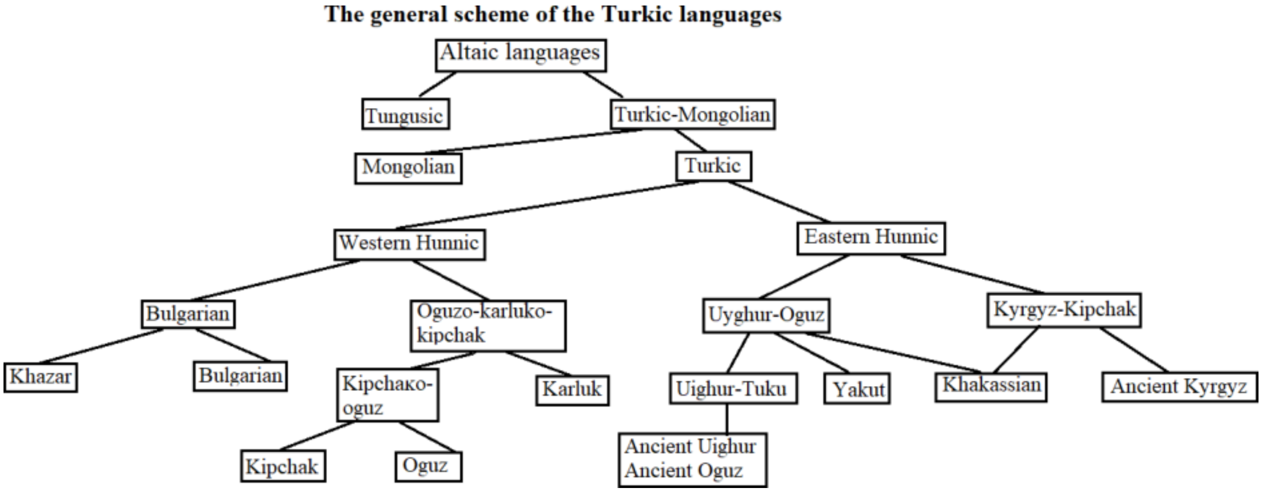

Fig. 2. The general scheme of the Turkic languages

In the following sections of the article, we will give an overview of the Khazar history and language and talk about the descendants of the Khazar people living nowadays and their ethnic groups. Then we will talk about the use of modern technologies in language reconstruction, data, methodology, experiments, and results.

2 Khazar Language History

Before observing the Khazar language, we must see how this ancient nation lived. At the very beginning, we have to observe their culture. We do it for a purpose because language cannot exist without history. We must mention that the Khazar language and culture are similar to the Tatar, Bulgarian, and Chuvash ones.

That is why it is essential to analyze the history of its neighbors and languages too. Based on the information about Khazars and their neighbors, we can find the common features between them.

At the very beginning need to start with the history of the Khazar nation, mainly how it was founded. There are many issues about this exciting nation like Khazar.

Some scientists consider that their language belonged to the Semitic language family; others attribute it to the Bulgarian branch of the Turkic language family. Still, there are plenty of questions about their history and culture.

2.1 Bulgars

History. The first step of the beginning of Great Bulgaria was not an easy job. The Bulgars nation decided to create their own country when they tried to escape from the powerful Khazar Khaganate. During some time, When the Bulgars finally created their own “Empire”, the ruling elite formed a unique ethno-political identity and culture.

However, their country did not exist for an extended period of time. Bulgars became the dominant tribe and formed the military service elite of society [10].

Language. The Bulgarian language is a part of the Turkic languages. Today this language does not exist anymore. The Bulgarian language was widespread in the 13th-14th centuries in the Volga region. Arabo-graphic epitaphs were found first on the territory of Volga-Kama Bulgaria.

The Bulgar language and the modern Chuvash language make up the Bulgar group of Turkic languages. Their main regularity lies in the transition from *r’>r, as well as *-d->-r-, by the transition *-l’>-l at the end of the syllable [12].

Bulgar, like all ancient languages, used the runic alphabet. Moreover, the Bulgar language has two main dialects: the Bulgar language and the

Table 1. NorthEuraLex corpora word content by the language used in this study

Language	Number of words
Chuvash	1,210
Tatar	1,149
Kazakh	1,312
Russian	1,037

Suvar language. The second one is nowadays the Chuvash language.

Additionally, scientists consider the Khazar language similar to the Bulgar language. People from Southern Bulgaria could understand people from the Khazaria. However, nowadays, we can see their footprints only in the Chuvash language [20].

2.2 Khazars

History. First, we must mention that different linguists have different points of view about the Khazar language. Some scientists and linguists consider that the Khazar Khaganate has the same roots as the Uighur Khaganate.

“Based on the fact that the Chinese name of the Khazars = k’o-sa closely resembles the name of six of the nine Uighur tribes of Kesa, some researchers classify the Khazars as Uighurs and believe that they appeared in Europe together with the Huns or after them in the VI century” [2, 15].

However, the author refutes this version. The language of Khazar khaganate is similar to the Bulgarian language, and it is close to the Turkic languages [2, 7]. Later there was a battle between the Armenian ruler and the Khazar nation. Whereas as a result, Armenian won [2].

Language. In the previous section, we talked about the history of the Khazar nation, and here we will see what the Khazar language looked like. Khazar language does not have many texts, so we must reconstruct it. Moreover, scientists still cannot understand what kind of language it is. That is still a question. The only source of Khazar words are names of kings and toponyms of Khazaria from non-Khazar historical manuscripts available to researchers [16].

Ibn Hordabeh states that the Khazar language is identical to the Bulgar language but different from the Burtas, Persian, and Russ (people of Scandinavian origin, known as Vikings or “varyags” languages [8]. From this information, we can conclude that it belongs to the family of Turkic languages; see Fig. 2, or, more specifically, to its oldest branch, which separated from the general Turkic unity first of all [16].

Nevertheless, the Khazar language presumably belongs to the Bulgarian group of languages. However, unfortunately, we have only one alive language from the Bulgarian language family. This alive language is the Chuvash language, which is commonly spoken in the Chuvash Republic in Russia.

“That is why the data of the Chuvash language is essential for studying the question of the Khazar language. In addition, the analysis of the early Turkisms in the Hungarian language, many of which are borrowings from the Khazar language, testifies in favor of the version about the Turkic affiliation of the Khazar language” [16].

Moreover, two types of alphabets were used by Khazars. The Don letter, represented by the inscriptions of the Mayak settlement, and the Kuban letter, are the only monuments that are inscriptions found during archaeological research of the Humarin fortress [16].

Furthermore, some linguists consider that this language can be similar to the Ossetian language. “The text written by this hand should be read in a language close to the Digor dialect of the modern Ossetian language. Thus, the language of the texts written in Runic script, distributed on the territory of the Khazar Khaganate, is not Turkic but Iranian in origin. That is, it is not a proper Khazar language.” [16, 13].

2.3 Contemporary Descendants

2.3.1 Chuvash

History. The nation that can attract people’s attention is Chuvash. Today we can observe the territory of the Chuvash in the Middle Volga region. The Chuvash speak the Turkic language, a linguistic relic of the Western ancient Turkic language also called “Bulgar” or “Ogur”.

Table 2. NorthEuraLex semantic concepts

id	Name	English	German	Russian
1	EYE	eye [[anatomy]]	Auge [[Anatomie]]	глаз [[анатомия]]
2	EAR	ear [[anatomy]]	Ohr [[Anatomie]]	ухо [[анатомия]]
3	NOSE	nose [[anatomy]]	Nase [[Anatomie]]	нос [[анатомия]]
4	MOUTH	mouth [[anatomy]]	Mund [[Anatomie]]	рот [[анатомия]]
5	TOOTH	tooth [EX:human incisor]	Zahn [BSP: Schneidezahn]	зуб [НАПР:человека]

Their neighbors are speakers of Eastern Turkic, Finn-Ugric, and Slavic languages, and historically in contact as in the Iranian world, the Chuvash, in many respects, is an excellent, illustrative example of the complexity of ethnogenesis, the mixing of ethnic groups, languages, and cultures that make up the people.

In the past, the Chuvash led a fairly diverse lifestyle, following various economic pursuits (sedentary agrarian lifestyle, pastoral nomadic lifestyle, hunting, and gathering) in the steppe, forest-steppe, and forest zones into clans, tribes, tribal unions, states, and sometimes empires. The Chuvash rarely engaged in any business alone often, they joined groups for this [1]. Some scientists firmly believe that the ancestors of the Chuvash were known as Savirs/Suvars [19].

Language. Scientists say that the ancestors of the Chuvash were Turkish nomads, and they immigrated from the West to middle Asia and moved off to Eastern Europe. This country's language is unique because it resembles the Mongolian and Finno-Ugric languages. However, scientists still argue that the Chuvash language belongs to the Turkic languages. Bulgarian Turks, the ancestors of Chuvash people, were the first Turkic clan that immigrated to the West and separated from the Central Asia Turkic community.

This immigration is thought to have happened at the beginning of the first centuries AD. For this reason, the Chuvash language, among the Turkic languages, is the oldest and represents Turkic all by itself. Because this language has Mongolian and Finno-Ugric characteristics, some scientists consider that this language was connected with the Mongolian and had similar culture and language in the past.

However, after some time, this language started to develop itself due to historical events [23]. In the past times, people used the same alphabet (runic alphabet as Bulgars did), and here there is a modern Chuvash alphabet: Аа, Ăă, Бб, Вв, Гг, Дд, Ее, Ёё, Ĕĕ, Жж, Зз, Ии, Йй, Кк, Лл, Мм, Нн, Оо, Пп, Рр, Сс, Çç, Тт, Уу, Ӳӱ, Фф, Хх, Цц, Чч, Шш, Щщ, ъ, Ыы, ь, Ээ, Юю, Яя .

2.3.2 Tatars

History. After the breaking of the Eastern Turkic khaganate, Kimaks and Kipchaks created their khaganate and called it "Kimak khaganate". At the same time, their powerful neighbor Bulgars created their own country and called it "Great Bulgaria".

After some time, when Great Bulgaria was broken and divided into two parts, "Danube Bulgaria and Volga-Kama Bulgaria", Danube Bulgaria combined with Slavic nations and accepted Orthodox religion meanwhile another part Volga-Kama Bulgaria combined with Turkic and Ugric tribes and accepted Islam religion.

After it, Volga-Kama Bulgaria, was conquered by the Mongols and used to be a part of the Golden Horde. When the Golden Horde was separated into several independent states such as Astrakhan, Crimea, and Kazan khanates, all of these gradually became the part of Russian Empire on its rise, and contemporary Tatar ethnic groups formed within it in the 19th century as local Muslim and Turkic communities [11].

Language. The Tatar language is widely spoken in the Tatarstan Republic. This language has several dialects, and all of these dialects are different. At the beginning of the 20th century, Tatar nations were combined. Additionally, this language

Table 3. Results sample of list of possible Khazar/Bulgar words

Norm.	Tatar	Chuvash	Concept
бс [bs]	бәс [bæs]	пас [pas]	HOARFROST
сдсб [sdsb]	савыт-саба [savɨt-saba]	савăт-сапа [savət-sapa]	DISHWARE
бсбк [bsbk]	башмак [başmak]	пушмак [puşmak]	SHOE
бндр [bndr]	мендәр [mendær]	минтер [minter]	PILLOW
ср [sr]	чир [tʃir]	чир [tʃir]	DISEASE
сд [sd]	оста [osta]	ăста [əsta]	MASTER
сл [sl]	усал [usal]	усал [usal]	EVIL
кскр [kskr]	кычкыру [kɨʧkɨru]	кăшкăр [kəʃkər]	SHOUT
сл [sl]	сулау [sulau]	сывла [sɨvla]	BREATHE
рд [rd]	ярату [jaratu]	юрат [jurat]	LOVE
сдр [sdr]	өстерәү [østeræw]	сĕтĕр [sətər]	DRAG

is a part of the Turkic languages and its Kipchak branch; they have three dialects of their language (Western, Eastern, and Middle).

In the middle is Zakamsky, Paranginsky, Nagorny, Menzelinsky, Birsky, Perm, Nokratsky, Kasimov; In the west people speak Sergachsky, Drozhzhanovsky, Chistopolsky, Melekessky, Temnikovsky, Kuznetsky; Finally, in the east there are Tobolo-Irtysh, Tyumen, Barabinsky, and Tomsk. During the creation of the Tatar Republic, their language was mixed and interacted with other languages.

Its neighbors are Bashkirs, Finno-Ugric, Mordovian, Mari, Udmurt, and Slavic languages [18]. Tatars traditionally adopted the Arabic alphabet, which was replaced for a short time by the Latin alphabet used by all Turkic people, and finally converted to a Cyrillic alphabet adaptation [22]. That is a modern version of the Tatar alphabet:

А а, Ә ә, Б б, В в, Г г, Д д, Е е, Ё ё, Ж ж, Җ җ, З з, И и, Й й, К к, Л л, М м, Н н, Ң ң, О о, Ө ө, П п, Р р, С с, Т т, У у, Ү ү, Ф ф, Х х, Һ һ, Ц ц, Ч ч, Ш ш, Щ щ, Ъ ъ, Ы ы, Ь ь, Э э, Ю ю, Я я.

3 Computer Science and Extinct Languages

Computers and different technologies can help us to solve many problems. One of them is the decryption of extinct languages. It can be complex and lengthy work if done manually; meanwhile, the technologies can solve it faster.

Let us see how it works. Instead of spending half of their life trying to get something from an extinct language, for instance, as people did with the Egyptian language, computers can take just several hours for this work. For example, utilizing computer technologies, it was possible to decrypt the Ugaritic language for several hours [6].

First, if we want to decrypt the target language, we need to know which languages can be similar to the target language. In the case of the Ugaritic language, scientists discovered that the most similar language is Hebrew. Without this comparison, it would be hard for the computer to find common features.

Computers can also help a lot with the computation of statistical features of a language, such as character or word distributions, word co-occurrences, and many others. The main thing scientists can do for the extinct language is to find out the “possible” language family of the target non-decrypted language [6].

Table 4. Examples of Out of Vocabulary (OV) words in Kazakh language

Norm.	Tatar	Chuvash	Kazakh OV	Concept
жр [ʐr]	җир [ʒir]	çĕр [stʲer]	жер [ʐer]	SOIL
д [d]	ут [ut]	вут [vut]	от [ot]	FIRE
сск [ssk]	чәчәк [tʃætʃæk]	чечек [tʃetʃek]	шешек [ʂeʂek]	FLOWER
срк [srk]	сарык [sarɨk]	сурăх [surəx]	сарық [sarɨq]	SHEEP
кккк [kkkk]	кәккүк [kækkyk]	куккук [kukkuk]	көкек [køkek]	CUCKOO

Additionally, nowadays, people speak around 6,000 languages; meanwhile, in the past, people spoke approximately 31,000 languages. As a result, people started to lose history. Via the languages, people can know the history.

In 2010 people had to know the relationships between languages to decrypt extinct languages with the help of AI. Today, machines can decrypt it without any comparison, in other words, no need to know the language family of the extinct language if we want to decrypt it in AI. To sum up, technology can help in linguistic research affairs in many ways.

Even linguists use it to know the history of the past. Machines, without any doubt, are developing year by year. However, it will take more time. With the help of machines, people can decrypt or reconstruct extinct languages faster. That is why we need to incorporate computer technologies in our research [9].

4 Data

NorthEuraLex 0.9 corpora[2] amongst 107 languages of Northern Eurasia contains datasets for Tatar, Chuvash, Kazakh and Russian languages (Table 1, and includes orthographic form of words with International Phonetic Alphabet (IPA) transcription and the semantic concept labels (Table 2).

Corpora contains 1,016 semantic concept tags explained in English, German, and Russian languages [3].

5 Methodology

The linguistic reconstruction task is to recover the lexicon, grammar, and syntax of an extinct language with no written text artifacts (unattested language) but known to be the ancestor of one or more live languages. A word rooting down to a proto-language is called reflex, and reflexes from the same root are cognate. The task can be approached in two major ways:

1. **Internal Reconstruction** exploits single language anomalies and irregularities to infer about earlier stages of language development, collecting the facts within the language studied.

   In internal reconstruction, the language is compared with itself, as it has changed over time, and we are looking for anomalies in morphology and grammar that may indicate linguistic features of the proto-language.

2. **Comparative Reconstruction** is finding a common ancestor for two or more languages from the same language group using the comparative method. The ancestor language is referred to as the proto-language of a given language family.

   The most famous examples of Proto-languages are Proto-Indo-European, Proto-Semitic, Proto-Turkic, and Proto-Dravidian because they are the most popular and common proto-languages that are being constantly researched by the scientific community.

   Languages, that are thought to have a common proto-language, are grouped together according to following criteria [17]:

[2] www.northeuralex.org/

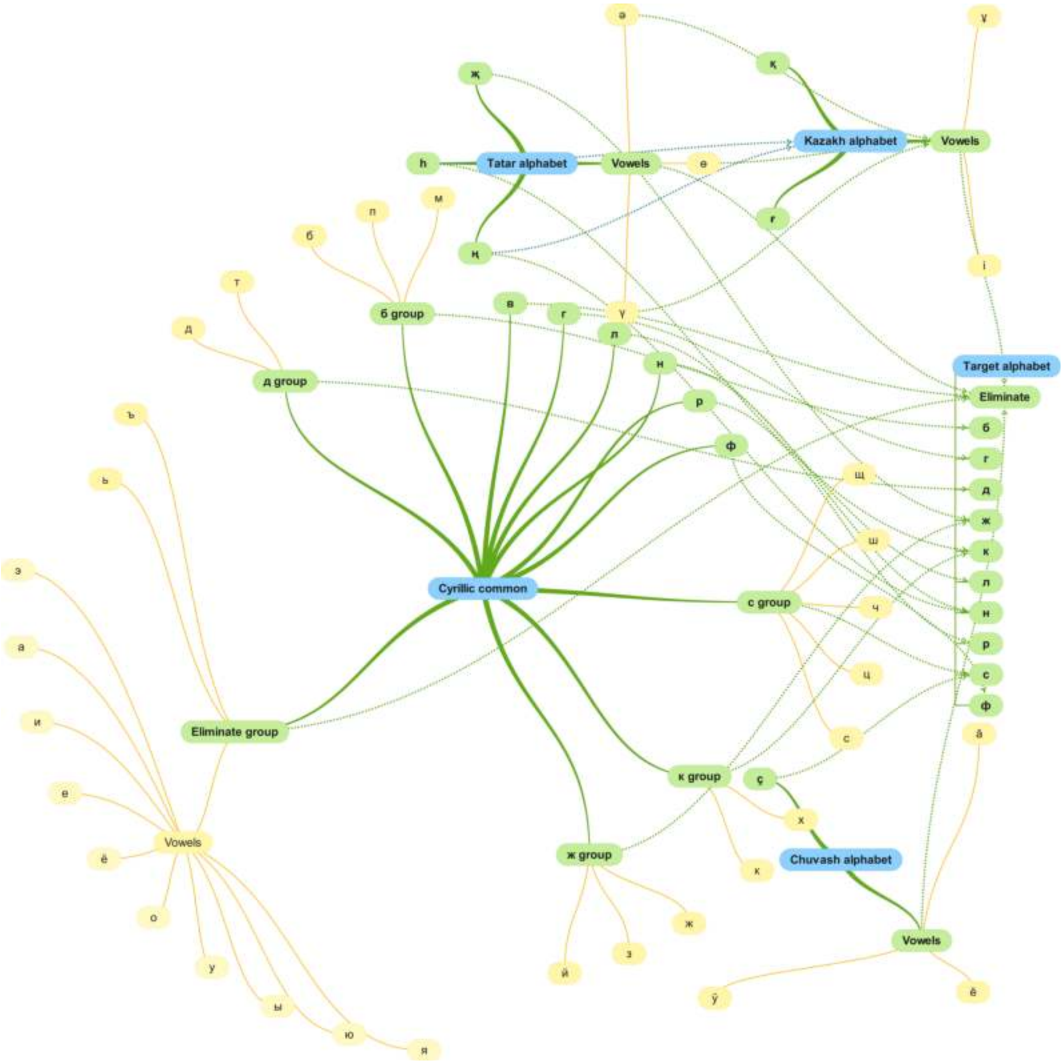

Fig. 3. Character mapping rules

- **Shared Innovation** meaning that the languages show common changes throughout time.
- **Shared Retention** which is opposite to the first criterion, meaning that languages preserve common features.

Comparative reconstruction exploits two major principles [24]:

- **The Majority Principle**, which observes that if cognates display a pattern, similar to repeating letter appearing in certain position within a word, then it is possible that the pattern was retained from the proto-language.
- **Most Natural Development Principle** proposes commonly appearing changes in languages throughout the time:
  - Omitting of final vowel in a word.
  - Consonants at the end of words become voiceless.

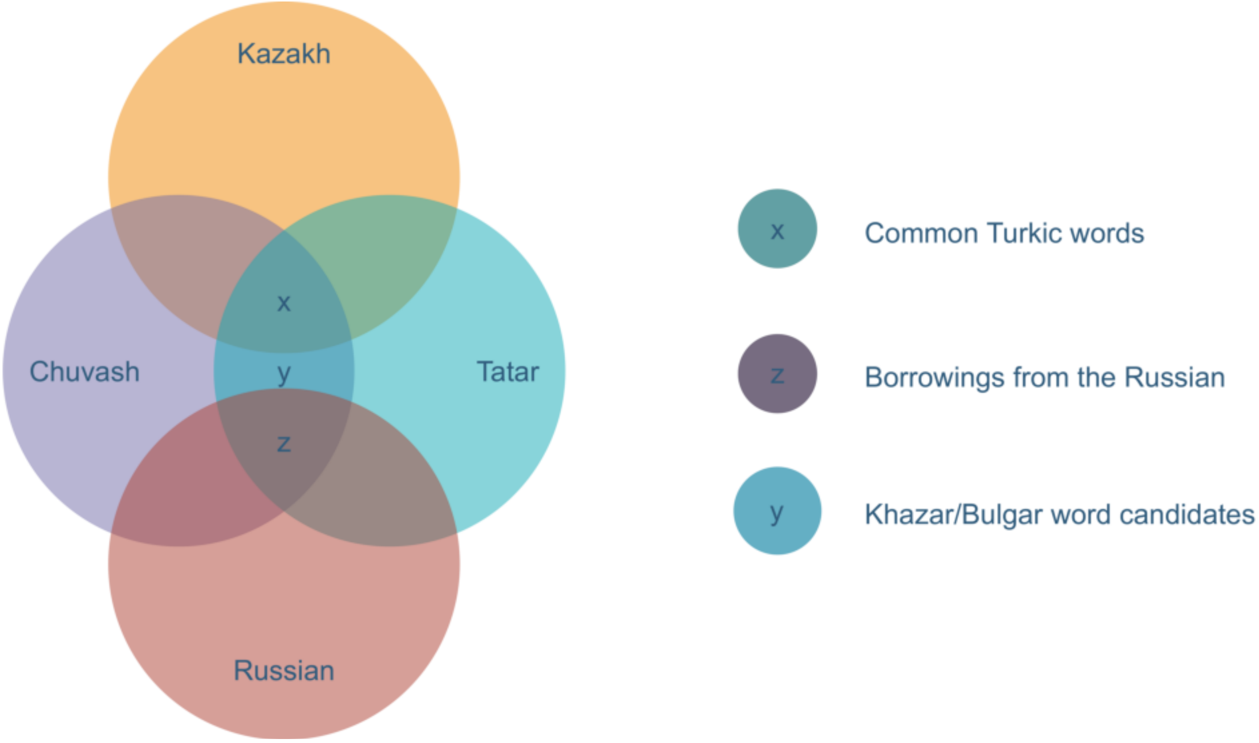
Kazakh
Chuvash
Tatar
Russian
x
y
z
x Common Turkic words
z Borrowings from the Russian
y Khazar/Bulgar word candidates

Fig. 4. Khazar/Bulgar word candidates selection in Venn diagram view

- Voiceless sounds appearing between vowels become voiced.
- Phonetic termination becomes fricative.

6 Experiment

Using the comparative method of language reconstruction, we have compared Chuvash and Tatar languages to find common words of possible Khazar/Bulgar origin.

1. Normalizing the words:
   - Remove vowels.
   - Character mapping rules; see Figure 3.
2. Find matching of normalized words in Tatar and Chuvash languages, with matching concept.
3. Exclude common Turkic words which match with the Kazakh language.
4. Exclude borrowed words from the Russian language.
5. Obtain the list of Khazar / Bulgar word candidates.

The overall process of obtaining the Khazar/Bulgar word candidates can be expressed by the Venn diagram shown in Figure 4.

7 Results

Some 185 normalized word and concept matches between Tatar and Chuvash languages were found (Figure 4 X, Y, and Z combined). Furthermore, 64 matches were left after filtering out common Turkic words (matches with the Kazakh language) and borrowings from the Russian language ((Figure 4 Y only); see Table 3 for a sample of 10 words of possible Khazar/Bulgar origin.

8 Discussion

8.1 Validity of Filtering Common Turkic Words

Briefly, the experiment included the stage where we filtered out presumably common Turkic words, which were indicated by the matching between normalized words in Chuvash and Kazakh language datasets. We assumed that the Kazakh language has no traces of Khazar/Bulgarian origin.

However, there might be some interactions as the Khazar khanate included some parts of modern Kazakhstan territory and bordered Khorezm in the past, which imposes 2 crucial questions:

– How different was the Khazar/Bulgar language from all the other Turkic languages back then and from the contemporary Turkic languages?

– How to differentiate words of Khazar/Bulgar origin in contemporary Turkic languages?

We also noticed that among those 64 words we obtained, there are still common Turkic words for which we have analogs in Kazakh, but they were not in the Kazakh language dataset we used; see Table 4. Therefore, we must repeat our experiments for all four languages on much larger corpora.

8.2 Finn-Ugric Components in Chuvash and Tatar Languages

Chuvash and Tatar languages might also share a lexicon borrowed from their Finn-Ugric neighbors: Mari, Udmurt, and Mordva people. Therefore to better distill the results, we need to account for the possible admixture from their languages and filter them out. Moreover, the neighbors could also borrow these words from ancient Bulgars or Khazars. We will need to compare their languages with their language family members who have no known contact with Khazars fixed in the history.

On the other hand, we might get better results by adding Karaim, Kumyk, and Balkar languages to the comparison, benefiting from the fact that these ethnic groups are also closely related to Khazars and Bulgars have no or little contact with Finn-Ugric people. However, they might have words from the Arabic, Persian, and neighboring Caucasian languages.

9 Conclusion

In conclusion, we want to emphasize the importance of the research in the direction of reconstruction and decrypting of extinct languages. Because it allows us to understand ancient scripts and, at the same time, makes it possible to look at the world with the eyes of our ancestors through the prism of their language. For future works, we plan:

1. Perform the experiments on significantly larger corpora.
2. Include the Karaim, Kumyk, and Balkar languages in the analysis.
3. Search for Khazar/Bulgar words in non-Turkic languages, such as Hungarian, Russian, Ukrainian, Bulgarian, and Chechen.
4. Use Bulgar vocabulary to find analog words in other Turkic and non-Turkic languages and then train a classifier model to find other possibly Khazar/Bulgar words.
5. Perform etymological analysis of the candidate Khazar/Bulgar words.

Acknowledgments

This work was supported by the Ministry of Education and Sciences of the Republic of Kazakhstan under the grant #AP14871214: “Development of machine learning methods to increase the coherence of text in summaries produced by the Extractive Summarization Methods”.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. **Anton, S. (2014).** Savirs - Bulgars - Chuvash. LAP Lambert Academic publishing.
2. **Artamonov, M. (1962).** Khazars’ history. Hermitage.

3. **Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J., Jäger, G. (2019).** NorthEuraLex: A wide-coverage lexical database of northern Eurasia. Language Resources and Evaluation, Vol. 54, No. 1, pp. 273–301. DOI: 10.1007/s10579-019-09480-6.

4. **Doerfer, G. (1976).** Proto-turkic: Reconstruction problems. Belleten-Türk Dili Araştırmaları Yıllığı, Vol. 23-24, No. 1975-1976, pp. 1–59.

5. **Golden, P. B. (2007).** Khazar studies: Achievements and perspectives. New Perspectives. Selected Papers from the Jerusalem 1999 International Khazar Colloquium, Brill, pp. 7–57. DOI: 10.1163/ej.9789004160422.i-460.7.

6. **Hardesty, L. (2010).** Computer automatically deciphers ancient language.

7. **Khamidullin, S. I. (2021).** Relations between the Bashkirs and the Volga Bulgars in the 10th–13th centuries. Ural Historical Journal, Vol. 71, No. 2, pp. 137–145. DOI: 10.30759/1728-9718-2021-2(71)-137-145.

8. **Khordadbeh, I. (1889).** The Book of Ways and Countries. Palmarium.

9. **Kumar, V. (2020).** Deciphering extinct ancient languages with machine learning. Analytics Insight.

10. **Leon, I. (2012).** The formation of the Volga Bulgaria: From tribe to state. St. Petersburg Slavic and Balkan Studies.

11. **Marjani Institute (2013).** Tatars' history.

12. **Mudrak, O. (2004).** Bulgar language.

13. **Mudrak, O. (2016).** Notes on the foreign language vocabulary of Khazar-Jewish documents. Space-2000 Moscow, Vol. 14, pp. 349–379.

14. **Papia, S. (2009).** Endangered languages: Some concerns. Economic and Political Weekly, Vol. 44, No. 32, pp. 17–19.

15. **Parker, E. (1895).** A thousand years of the tartars. Wentworth Press, 1st edition.

16. **Rashkovsky, B. (2014).** Khazars and judaism in the biblical commentaries of Yefet ben Ali. A New Medieval Jewish Source for Eastern European History, Vol. 1, No. 3, pp. 210–230. DOI: 10.14653/ju.2014.13.

17. **Reiss, C., Fox, A. (1996).** Linguistic reconstruction: An introduction to theory and method. Language, Vol. 72, No. 2, pp. 387. DOI: 10.2307/416657.

18. **Safarov, A., Gabrakhmanov, G., Galimova, E., Zagidullina, D., Izamaylov, I., Salikhova, A., Sitdikov, A., Shklyaeva, L. (2019).** Tatar language and written culture: From word to book. Tatar World, pp. 1–392.

19. **Salmin, A. (2013).** Ethnographical sources about the origin of the Chuvash. Japanese Slavic and East European Studies, Vol. 34, pp. 95–104. DOI: 10.5823/jsees.34.0_95.

20. **Salmin, A. (2015).** The Bulgarian language in the context of the history of the Chuvash. Bulletin of the Chuvash University.

21. **Sapargali, E., Akhmetov, I., Pak, A., Gelbukh, A. (2021).** Determining the relationship between the letters in the Voynich manuscript splitting the text into parts. Proceedings of the Mexican International Conference on Artificial Intelligence, Advances in Soft Computing, pp. 163–170. DOI: 10.1007/978-3-030-89820-5_13.

22. **Tatar, M. (2021).** Tatar alphabet: From ancient runes to modern Cyrillic. Billion Tatars.

23. **Yilmaz, E. (2002).** Chuvash and chuvash language. The Turks.

24. **Yule, G. (2010).** The study of language. Cambridge University Press.

Article received on 20/07/2023; accepted on 02/12/2023.
** Corresponding author is Alexander Gelbukh.*