Algorithmic Vowel Harmony Detection in the Khitan Small Script

Eric Le
Unversity of Nebraska-Lincoln
Lincoln, Nebraska, USA
eric.le@huskers.unl.edu

Abstract

Recent advances in computational analysis in linguistics have opened up the area for more study for the analysis of different scripts and the languages that they represent. Ancient scripts such as the Old Turkic Script are designed in a manner where the back vowels and front vowels are different symbols. Languages that have this vowel harmony characteristic usually employ a means of encoding it into their writing system if they have one. The Khitan Small Script is believe to encode the purpoted vowel harmony system of the Khitan language. In this paper, a vowel harmony algorithm will be tested on different corpus of the Khitan Small Script and will be used to analyze the feasibility of the algorithm to detect vowel harmony in the Khitan Small Script.

CCS Concepts: • **General and reference → General conference proceedings**.

Keywords: datasets, khitan, graphs, machine learning

ACM Reference Format:
Eric Le. 2022. Algorithmic Vowel Harmony Detection in the Khitan Small Script. In *International Database Engineered Applications Symposium (IDEAS'22), August 22–24, 2022, Budapest, Hungary.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3548785.3548787

1 Introduction

The main goal of this paper will be to employ a vowel harmony algorithm to see if the Khitan Small Script encodes vowel harmony or not. The paper will be divided into several sections. First, there will be an introduction of the Khitan people, the Khitan scripts, and vowel harmony in the Khitan language. Secondly, the paper will go into related works in the analysis of the Khitan Small Script and the linguistic analysis of vowel harmony. Thirdly, the paper will discuss the algorithm that is being used to detect vowel harmony in the Khitan Small Script. The paper will then discuss the dataset that is used to computationally analyze the Khitan Small Script. After the implementation of the algorithm and is application, the paper will then discuss the results from different experiments. Finally, future possible research will be proposed and then be finalized with a conclusion.

Figure 1. 五代胡瑰出猎图 The painting "Khitans using eagles to hunt" by Hugui (胡瑰)

1.1 Khitan People

The Khitan people (契丹人) were a nomadic people who lived in Northeast Asia. Their territory includes modern day Russia, China, Mongolia, and Korea. The had originally engaged in stockbreeding, fishing and hunting. But with contact with Chinese people, and after the establish of the Liao dynasty, that required a sedentary adminstration, the Khitans started to engage in farming and building of cities. They are believed to be a Para-Mongolic people people.

After the Kyrgyz takeover of teh Uyghur Khaganate and the collapse of the Tang Dynasty, the Khitan were able to establish their own state. In 907, Abaoji was able to unite the different Khitan tribes and founded the Liao dynasty. The dynasty existed from 916 AD to 1125 AD. Eventually the dynasty succumbed to the Jurchen who then established the Jin dynasty in the former terriroty of the Liao. The Liao

IDEAS'22, August 22–24, 2022, Budapest, Hungary

ACM ISBN 978-1-4503-9709-4/22/08…$15.00
https://doi.org/10.1145/3548785.3548787

Khitan eventually escaped westward where the established the Western Liao dynasty that lasted until from 1128 AD to 1218 AD. That dynasty eventually fell to the Mongols.

1.2 Khitan Scripts

In 920, Abaoji ordered the creation of a writing system for the Khitan language. This script was to be known as the Khitan Large Script (契丹大字) [2]. When an Uyghur delegation visited the Liao court in 924 or 925 CE, Abaoji ordered his younger brother to a new script. From this interaction, a new script was created. This script is called the Khitan Small Script (契丹小字).

1.2.1 Khitan Large Script. Khitan Large Script took its basis from Chinese characters [2]. The characters were written equally spaced and in vertical columns like Chinese characters. The script is thought to be for the most part logographic similar to its Chinese character counterparts, but there is reason to be believe that some of the characters used in the script are syllabograms that are used for grammatical meanings [3]. In total there are about 830 characters that can be identified as distinctly different. Figure 2 is an example of the Khitan Large Script called the Epitaph for Court Attendant Dorlipun (多罗里本郎君) and some analysis on possible meaning of some of the characters.

1.2.2 Khitan Small Script. The Khitan Small Script, however, developed in a different manner. Rather than having simple logographic characters laying in the usual square dimension of Chinese characters, it was formed by using logo-graphic, syllabic, and even maybe phonemic symbols and formed them into clusters for different words [3]. If we take a look at 3, the Khitan Small Script designed their blocks depending on how many symbols were used in a word. If the word contained a single symbol, that symbol would occur in the entire space. However, if the word needed two Khitan Small Script symbols in order to described the word, the first symbol would be to the left and then the second symbol would be to its immediate right. If there were three symbols needed, the third symbol would be placed in the middle underneath the first two symbols. This pattern than would go on until the words stop. Theoretically, the process could go to as large as seven symbols, but it seems that the largest Khitan word possible was that consisting of six symbols.

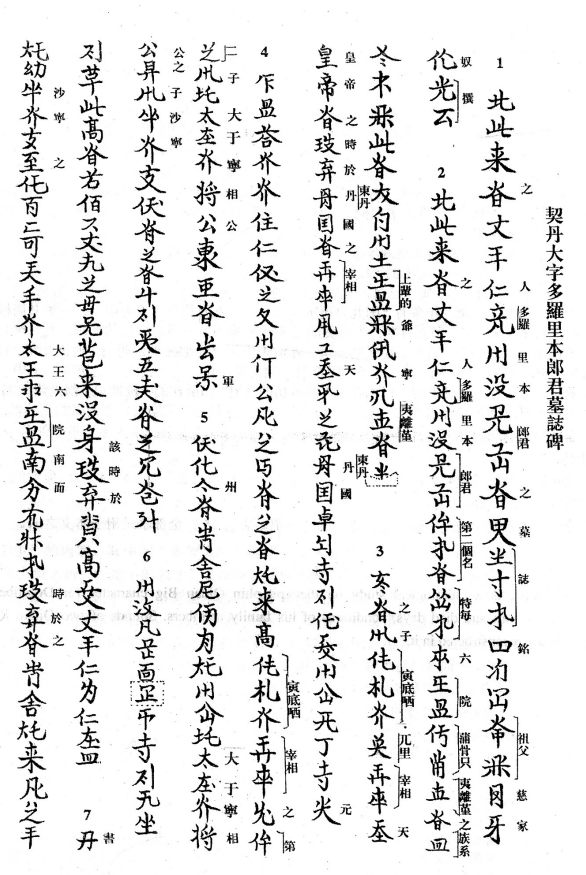


Figure 2. 多罗里本郎君 Epitaph for Court Attendant Dorlipun

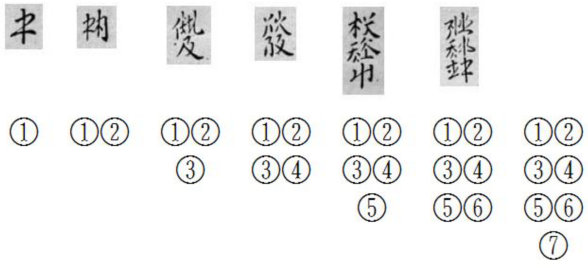


Figure 3. Khitan Small Script Construction [8]

1.3 Vowel Harmony

Languages in the Turkic, Mongolic, Koreanic and Tungusic language families show varying degrees of vowel harmony. Depending on the language, the characteristic of its vowel harmony can be different. There are four classes of vowel harmony that languages can take. These are Backness Harmony, Round Harmony, Height Harmony, and Tongue Root Harmony [6]. In other research these types of vowel harmony can also be called Palatal Harmony, Labial Harmony, Height Harmony, and Tongue Root Harmony [4]. These vowel harmony don't have to be mutually exclusive. The Azerbaijani system of vowel harmony for one contains both front/back and rounded/unrounded vowel harmony. In some languages, there is a class called the neutral vowel that allows itself to be with any class. This will be important for the analysis of the Khitan Small Script. Khitan has been shown to have a front-back vowel harmony [7].

2 Algorithm

The basis of this research is based on the Vowel Harmony Algorithm developed by Dr. Revesz of the University of Nebraska-Lincoln [5]. The basic assumption of the algorithm is that the script is a syllabary and that each symbol either represents a single vowel a consonant and vowel (CV) type syllable.

The vowel harmony algorithm has several different steps. These are the steps: Find frequencies of pairs and triples of symbols, find hypothetical root words, create an adjacency graph, test that the graph has at least two major connected components, and then return an answer.

Given that there was no publicly available code to use this algorithm, the code was rewritten. The code was written in Python. The program was run a simple Jupyter Notebook and given that the database was only text based and not large, could be ran a simple laptop.

3 Data sources

The Khitan Small Script was recently only digitally encoded into Unicode in 2020. Therefore, any research into the Khitan Small Script before was either done through a manual reading of the inscriptions or individually developed computational methods. In order to do a computational analysis of the script, computer texted base documents are needed. According to current knowledge, there are no know publicly available datasets that are written in the Khitan Small Script. Therefore, it was imperative to make one.

There are currently 33 known inscriptions that are written in the Khitan Small Script. Due the large amount of time that would be required to encode the inscription into digital format, the text that were read were limited to the inscriptions that were included in [3] and a few others. These would be the *Bronze Mirror*, the *Record of the Younger Brother of the Emperor of the Great Jin Dynasty (*大金皇弟都统经略郎君行记*)* now known for the rest of the paper as the Langjun inscriptions and the *Eptigraph for Yelu Dilie (*耶律迪烈*)*. The dataset is available at here for more computational analysis on the Khitan Small Script.

4 Experimental Results & Discussion

After the running the Vowel Harmony Algorithm on all three different inscriptions, three different types of results came out.

4.1 Bronze Mirror

As we can see from figure 7, there are two distinct groups pairings; the green and red. From the criteria of the algorithm, it would seem that according to the vowel harmony algorithm, Khitan Small Script encodes vowel harmony from the Khitan language.

(a) First Iteration

(b) Second Iteration

(c) Final Iteration

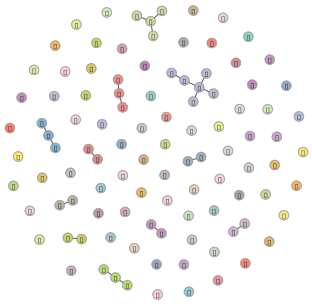
Figure 4. Inscriptions Used for the Database

4.2 Langjun Inscription

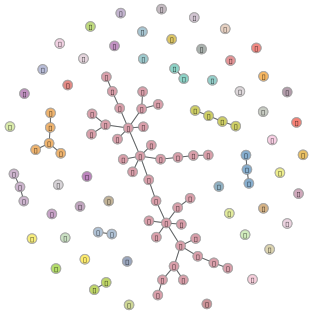
The results from the Langjun Inscription can be seen in figure 5. We can see from the network that at the beginning there were several groups. The ones that are linear are words. While there exist one grouping that seems to indicate at least some partial vowel group. However, when the next iteration of the algorithm was ran with a large corpus of words, the network created a large connected network. When all words were used in the creation of the network, most symbols were connected with each other.

4.3 Eptigraph of Yelu Dilie

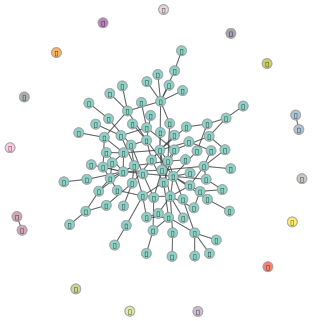
The results for the Eptigraph of Yelu Dilie can be seen in figure 6. Here at the beginning we can see some connected components. This can indicate to us that there exist some



(a) First Iteration

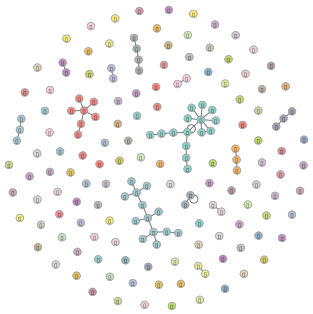


(b) Second Iteration

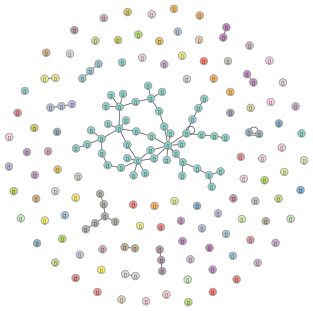


(c) Final Iteration

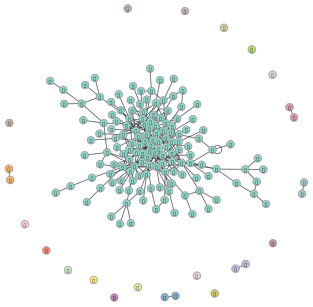
Figure 5. Vowel Harmony Network for the Langjun inscription



(a) First Iteration



(b) Second Iteration



(c) Final Iteration

Figure 6. Vowel Harmony Network in the Eptigraph of Yelu Dilie

groups in the text and shows us that there might be some vowel harmony. However, just like the Langjun inscription, as the corpus grew, the network started to connect the nodes of the graph more and more. The final version of the graph would indicate to us that there is no vowel harmony indicated in the Khitan Small Script.

4.4 Analysis

According to both [3] and [7], the current analysis of the Khitan script should include a front and back vowel harmony. So why did the algorithm fail to see that in the corpus of the text. This could be to several reasons. We do know that several of the symbols in the Khitan Small Script encode just a constant value in certain words. This could effect the algorithm if that symbol is being used for both front and back vowels. Another issue that could occur during the creation of the graph is neutral vowels. In languages such as Mongolian, there exist a category for an advance root tongue and retracted root tongue. But there also exist a neutral class. These are vowels that are allowed to go with both classes. Therefore, it is possible that the graph combined two connected components that were meant to be the vowel harmonic groups when it detected a neutral vowel.

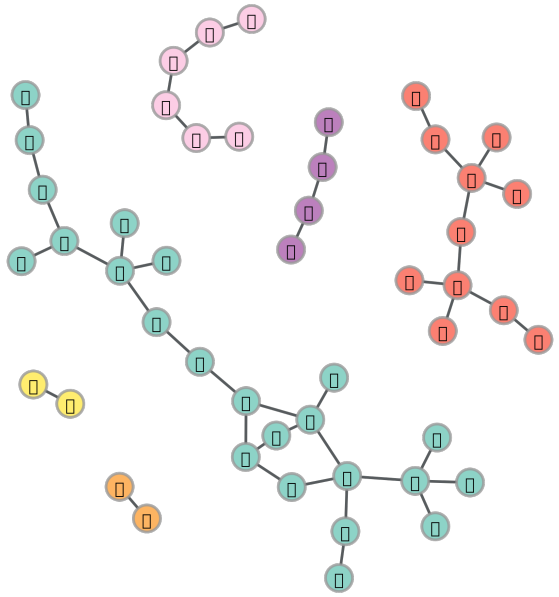


Figure 7. Bronze Mirror Vowel Harmony Graph

5 Further Research

From current understanding, this is the first attempt to try to computationally analyze any of the Khitan scripts. This current research has shown that there is potential use for the computational tools to analyze the Khitan scripts. However, there does need to be some improvements to the current algorithm to include some more nuances in different languages and scripts.

Given the current development of the vowel harmony algorithm, there are several further research areas that this can be taken into. Several future research areas are detecting the neutral vowels class, updating the the Khitan database, analysis of the Khitan Large Script.

5.1 Detecting the Neutral Class

As noted before, one major issue that arises when the vowel harmony algorithm is used, is that it is not able to detect a neutral class. Given that neutral classes are allowed to go with any of the two classes, it is highly likely that two connected components representing the vowel harmonic groups were connected by either a neutral vowel symbol or a constant based symbol. Therefore, it would most beneficial to include a method to detect this neutral vowel and not allow it to effect the creation of the graph.

5.2 Khitan Database

The current Khitan database consists of 3 different inscriptions; the Bronze Mirror, the Langjun, and the Yelu Dilie. If we look at all three texts, the number of unique symbols in total is about 200. As mentioned previously, the total number of the Khitan Small Script symbols is over 400. Therefore, in order to get a complete database, it would be required to input texts that would in total have all of the known Khitan Small Script symbols. An update to the database would allow for a better representation of the the script and a more accurate analysis of it.

5.3 Analysis of Khitan Large Script

Most research of the Khitan scripts has mostly been focused on the Khitan Small Script. This is due to there being one bilingual inscription that has allowed for us to learn about the pronunciation of the Khitan Small Script and the meaning behind the words. Given that Khitan Large Script is more similar in its structure to how Chinese is written and the lack of the bilingual gloss, it has been more difficult to understand Khitan Large Script. However, recent advances and applications of deep learning models might allow for some more rigorous analysis of the evolution of Khitan Large Script from Chinese characters. [1] has shown that through neural networks and SVN that you are able to see high relational probability that one script is derived from another script. Deriving Khitan Large Script from Chinese characters has actually been proposed before. created a list of his possible evolution of Khitan Large Script characters from Chinese characters. This, however, was done in, not utilizing the recent advances in computation methods. It would be worthwhile in future research to explore using neutral networks to derive a Khitan Large Script character from a Chinese character and possible retrieving the meaning from it.

6 Conclusion

This is the first known analysis of the Khitan Small Script using computational tools. Even though, the current implementation of the algorithm did not show strong signs of vowel harmony in the Khitan Small Script, it is a first step into using computation analysis tools in the research of the Khitan Small Script. The vowel harmony algorithm as strong potential in showing vowel harmony in the Khitan Small Script with some improvements. The future research areas might be even more interesting. This research has shown that the Khitan Small Script is available to analyze computationally and that there might be more areas to use such tools. Such research as deep learning in the Khitan Large Script. Given this start, it would be nice to see more progress in this area.

References

[1] Shruti Daggumati and Peter Z. Revesz. 2019. Data Mining Ancient Scripts to Investigate Their Relationships and Origins. In *Proceedings of the 23rd International Database Applications & Engineering Symposium* (Athens, Greece) *(IDEAS '19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 10 pages. https://doi.org/10.1145/3331076.3331116

[2] Jacques Gernet. 1996. *A History of Chinese Civilization.* Cambridge University Press.

[3] Daniel Kane. 2009. *The Khitan Language and Script*. Brill, Leiden Boston.
[4] Seongyeon Ko. 2018. Tongue Root Harmony and Vowel Contrast in Northeast Asian Languages. *Turcologica* 112 (2018).
[5] Peter Z. Revesz. 2020. A Vowel Harmony Testing Algorithm to Aid in Ancient Script Decipherment. *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)* (2020), 35–38.
[6] Sharon Rose. 2011. *Handbook of Phonological Theory*. Blackwell Publishing Ltd, Leiden Boston.
[7] András Róna Tas. 2017. Khitan Studies I. The Graphs of the Khitan Small Script. 2. The Vowels. *Acta Orientalia Academiae Scientiarum Hungaricae* 70, 2 (2017), 135–188.
[8] Andrew West. 2010. Babelstone blog. https://www.babelstone.co.uk/Blog/2010/12/mystery-of-two-khitan-scripts.html