Chapter 8

# Cross-Lingual and Cross-Chronological Information Access to Multilingual Historical Documents

Biligsaikhan Batjargal

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72421

**Abstract**

In this chapter, we present our work in realizing information access across different languages and periods. Nowadays, digital collections of historical documents have to handle materials written in many different languages in different time periods. Even in a particular language, there are significant differences over time in terms of grammar, vocabulary and script. Our goal is to develop a method to access digital collections in a wide range of periods from ancient to modern. We introduce an information extraction method for digitized ancient Mongolian historical manuscripts for reducing labour-intensive analysis. The proposed method performs computerized analysis on Mongolian historical documents. Named entities such as personal names and place names are extracted by employing support vector machine. The extracted named entities are utilized to create a digital edition that reflects an ancient Mongolian historical manuscript written in traditional Mongolian script. The Text Encoding Initiative guidelines are adopted to encode the named entities, transcriptions and interpretations of ancient words. A web-based prototype system is developed for utilizing digital editions of ancient Mongolian historical manuscripts as scholarly tools. The proposed prototype has the capability to display and search traditional Mongolian text and its transliteration in Latin letters along with the highlighted named entities and the scanned images of the source manuscript.

**Keywords:** historical documents, multilingual databases, information access, information retrieval, digital edition

## 1. Introduction

As historical materials are increasingly being digitally preserved, multilingual materials concerning a diversity of languages and historical periods have been made available to the public

on the Internet. Recently, a number of large-scale digital library projects have been launched, e.g., Europeana, World Digital Library, HathiTrust and Google Book Search. These websites make multilingual materials covering various languages and historical periods available to the public.

There are various technical challenges, however, in implementing universal integrated access to these digital collections due to this great diversity, and difficulties occur in accessing these information sources, mainly due to the diversity of languages. Even within the same language, considerable differences exist in grammar, vocabulary and script depending on the historical period, and this is the primary cause of the difficulties in implementing universal information access. Thus, this chapter presents our approach to providing cross-lingual and cross-chronological access to historical documents that account for evolution of languages over periods ranging from ancient to modern. Particularly, in this chapter, we introduce our approach in providing cross-lingual and cross-chronological information access to historical materials in a less-researched language such as ancient Mongolian.

In Section 2, we discuss the current situation of digitized ancient historical materials written in ancient Mongolian and the challenges in providing universal information access to them in the digital era. Then, our proposed method for cross-lingual and cross-chronological information access to ancient Mongolian historical materials is discussed in Section 3. Finally, in Section 4, we discuss the future prospects of this research.

## 2. Ancient Mongolian manuscripts

This section briefly explains certain characteristics of Mongolian manuscripts and current situation of digitized ancient historical materials written in ancient Mongolian and challenges they present in the digital era.

### 2.1. A brief introduction of Mongolian manuscripts

Mongolian historical documents have been written in numerous scripts, i.e., the traditional Mongolian script, Square or Phags-pa script, Soyombo script and Horizontal square script [1]. Among them, the traditional Mongolian script is the most popular and longest-surviving script for over 800 years and has better supports with the computer systems recently since its integration to the Unicode Standard [2] in September 1999. On the 20th of June, 2017, the Soyombo and Horizontal square scripts (a.k.a. Zanabazar scripts) were standardized in the most recent version of the Unicode Standard [3]. However, this research focuses on the traditional Mongolian script because of its popularity, availability of digital texts and improved supports at the computers.

In 1946, Mongolia has made language reforms to eliminate a difference between written and spoken Mongolian language, and the Cyrillic script was adapted to Mongolian. The spelling of modern Mongolian in the Cyrillic alphabet was based on the pronunciations in the Khalkha dialect, the largest Mongol ethnic group [4, 5]. Such a radical change separated the Mongolian people from their historical archives written in traditional Mongolian script. Manuscripts in traditional Mongolian script preserve the ancient writing, while modern Mongolian reflects

http://dx.doi.org/10.5772/intechopen.72421

the unique pronunciations in modern dialects. Understanding historical documents in traditional Mongolian script is becoming as equally important a consideration for Mongolians as modern Mongolian in Cyrillic script. However, reading traditional Mongolian documents by using literacy in modern Mongolian is not a simple task. Traditional Mongolian is a distinct dialect with grammar different from that of modern Mongolian. The traditional Mongolian script is written vertically, from top to bottom, in columns advancing from left to right. This script has four derivative scripts: Todo or Clear, Manchu, Vaghintara and Sibe (Xibe) script. The Todo script was used by the Oirats and Kalmyks, and the Manchu script was a writing system in the Qing dynasty. The Sibe script is used in Xinjiang, in the northwest of China. The Vaghintara script was used by the Buryats.

Moreover, the circumstances that the manuscript passed through a process of copying or reprinting with possible alterations, corrections and unintended errors makes researchers wonder which ancient spelling is correct or what the ancient word originally meant. Scholars had been pointing out from time to time that copies could not meet the requirements of scholars who want to study them as a source material [6]. Moreover, various different commentaries, transcriptions, annotations and interpretations have been suggested by humanities researchers. Besides, manuscripts are vulnerable to degradations and might have lacunas, physical damages or missing parts, which require costly reconstructions of the original text.

In general, there are two main demands from both users and researchers for making ancient Mongolian manuscripts usable in this digital era. Firstly, a digital representation that explains a given manuscript in a modern language is helpful for users who want to read, search and browse ancient Mongolian manuscripts. Secondly, in the field of humanities, getting knowledge by analysing various historical documents is an important task. There are increasing demands from Mongolian humanities researchers to perform text analysis at massive scale with prompt and accurate results. Having a digital representation that fully reflects a given manuscript is an awaited demand for researchers who want to study it as a scholarly source using a computer.

Nevertheless, computerized text analysis of Mongolian historical documents has not been done due to the lack of natural language processing (NLP) tools that can handle ancient Mongolian. Such demands have encouraged us to introduce our approaches in providing universal information access to ancient Mongolian historical documents.

### 2.2. Ancient Mongolian manuscripts in the digital age

To the best of our knowledge, there are a small number of digital texts of ancient Mongolian manuscripts. A few ancient Mongolian historical manuscripts including (1) 'Qad-un ündüsün-ü quriyangγui altan tobči neretü sudur' (the Altan Tobchi or the Golden Summary: Short history of the Origins of the Khans) (written in 1604) a.k.a. 'Little' Altan Tobchi and (2) the 'Asaraγči neretü-yin teüke' or 'Asragch nėrtĭĭn tüükh' (the Story of Asragch) (written in 1677), which were written in traditional Mongolian script, have been converted to digital texts and made publicly available through the traditional Mongolian script digital library (TMSDL) [7]. **Figure 1** shows a folio of the 'Little' Altan Tobchi in the TMSDL with keywords' highlights.

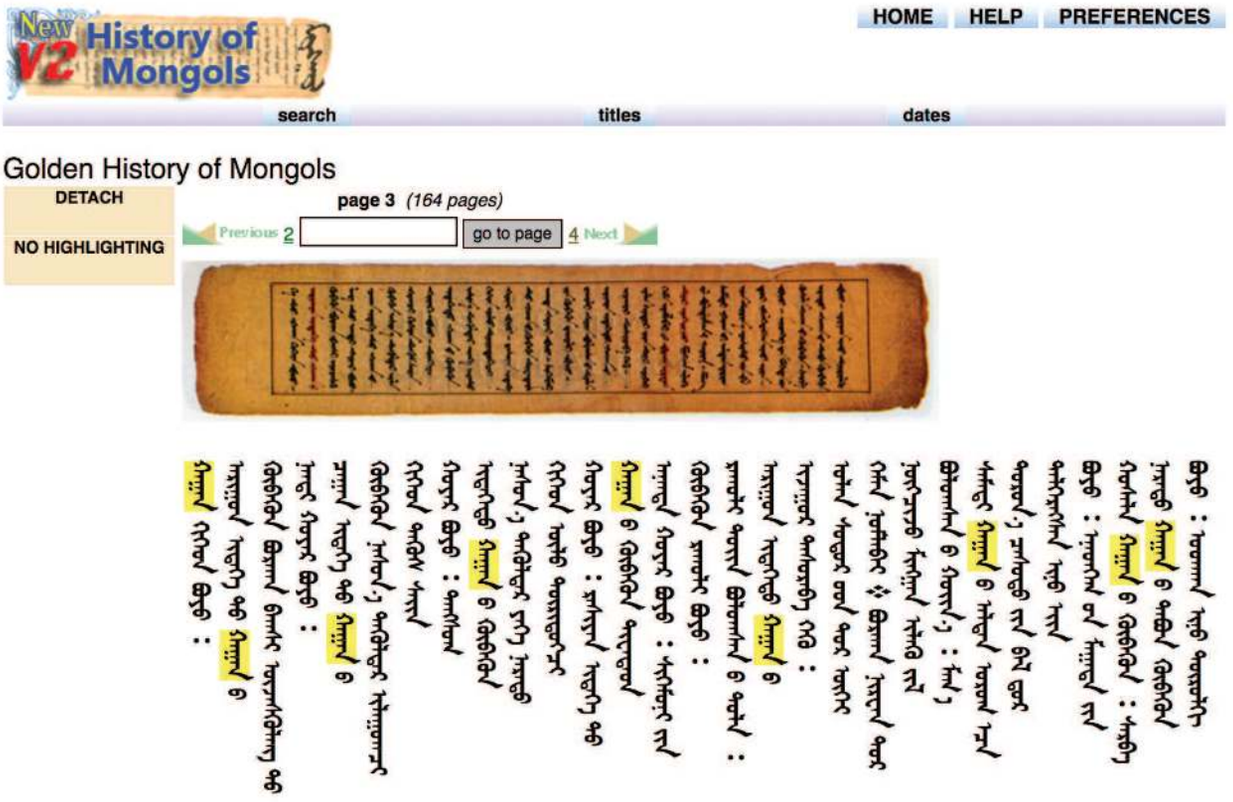


**Figure 1.** A folio of the 'little' Altan Tobchi in the TMSDL with keywords' highlights.

TMSDL can be used to access and retrieve the historical manuscripts written in traditional Mongolian script using a query in modern Mongolian (Cyrillic). The research achievements, as well as the experiences obtained from the development of the TMSDL, have motivated us to share further research results in developing methods to providing cross-lingual and cross-chronological information access to ancient Mongolian historical documents.

Certainly, there has been a little research on text mining for Mongolian language, and none of the research has considered text mining on ancient Mongolian historical documents due to the lack of research in those areas. Because of the notable difference between mediaeval Mongolian and modern Mongolian, the existing NLP tools, which were designed on modern Mongolian, do not perform well on ancient Mongolian texts. Therefore, further computerized analyses of ancient Mongolian historical documents are necessary.

## 3. Information access to Mongolian historical documents

In the recent years, the needs for utilizing digital representations and proving access to historical documents encouraged the development of various tools for transcribing, annotating and publishing of historical manuscripts. In order to provide computer technology-driven solutions to

solve the facing challenges of Mongolian humanities scholarship as well as to benefit the recent achievements in the digital humanities worldwide, it is necessary to analyse the requirements of Mongolian historical documents for digital tools.

In this section, we describe our methods for implementing integrated access to historical documents that are capable of coping with linguistic transformations from ancient times to the present. First, we propose an information extraction method for digitized ancient Mongolian historical documents. The proposed method extracts named entities from historical manuscripts by utilizing machine learning techniques. Results will be utilized for building digital text representations that encode named entities, the possible alterations, corrections, errors and interpretations of ancient Mongolian words in a modern language. In the later sections, we discuss how to develop a digital edition of Mongolian historical documents by considering various features and requirements of Mongolian manuscripts.

### 3.1. Information extraction from ancient Mongolian documents

This section discusses an information extraction method for digitized ancient Mongolian documents by using the features of traditional Mongolian script. Named entities such as personal names and place names are extracted automatically from digitized text of ancient Mongolian documents by employing support vector machine (SVM) for aiming to reduce the labour-intensive analysis on historical text. Information extraction, named entity extraction (NEE) and tagging or annotations are able to turn plain text into structured data for analysis or effective use, via NLP applications and analytical methods. State-of-the-art NEE systems for English produce near-human performance to extract named entities [8]. However, there has been little research on text mining or NEE for Mongolian language, and none of the research has considered text mining on ancient Mongolian historical documents due to the lack of research in those areas. Therefore, proposing an information extraction method for ancient historical documents in traditional Mongolian script is crucial.

#### *3.1.1. The proposed approach*

The flowchart in **Figure 2** shows an overview of the main steps and components of the proposed approach. The proposed approach starts with preprocessing tasks where an ancient Mongolian corpus gets tokenized, each token gets annotated and gold standard annotations are prepared for inputting into SVM for learning. The proposed method learns the extraction rules of personal names from annotated training corpora and then extracts personal names from ancient Mongolian texts by using SVM. The following sections explain the main three components: (1) pre-processing, (2) annotating and (3) named entity extraction.

##### *3.1.1.1. Preprocessing step*

The first step is to divide digitized ancient Mongolian plain text of into tokens. This is necessary because we want to mark up each token in the next tasks. A token is quite often a word delimited by space, but there exist some unique features for traditional Mongolian script. For

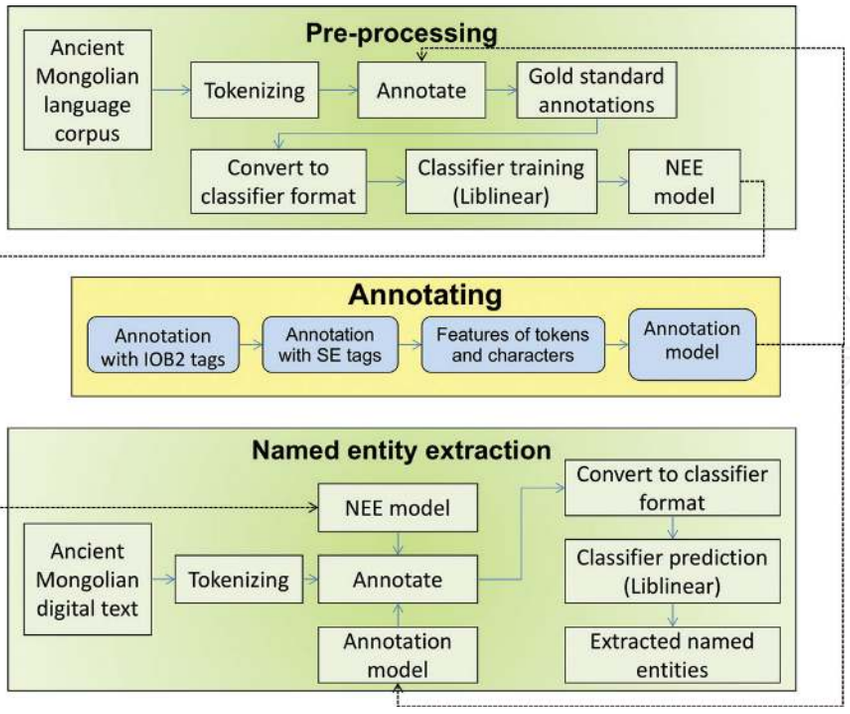


**Figure 2.** Overview of the main steps and components of the proposed approach.

instance, in traditional Mongolian script, certain words with a final vowel letter 'a' or 'e' are separated visually from the preceding consonant by a narrow gap. Moreover, some suffixes are visually separated from the stem of a word or from other suffixes. However, the 'a' or 'e' is an integral part of the word stem, as well as any attached suffixes are considered to be an integral part of the word as a whole. In Unicode, control characters Mongolian Vowel Separator (MVS) and narrow no-break space (NNBSP) handle the behaviour of Mongolian suffixes and vowels 'a'/'e' in the end of a word [2]. This information can be used as a feature in SVM. Other features are discussed in Section 3.1.1.3.

The next step is to annotate tokens and prepare gold standard annotations. Because of the lack of NLP tools and part of speech data for ancient Mongolian manuscripts, we first annotate all the personal names in the 'Little' Altan Tobchi using the manually compiled personal names' indices (lists of personal names) obtained from the 'Qad-un ündüsün quriyangγui altan tobči-Textological Study' [9]. After converting to a format that is suitable for a linear classifier, we input that data into the classifier for training, which returns a probability matrix (i.e., a model). The classifier is trained with gold standard annotations of tokens with known classes (i.e., personal names). The classifier calculates weights for each feature in correlation to each class. This can be seen as a probability of an object belonging to a certain class (i.e., personal names) when having those specific characteristics. These weights are saved in a probability matrix (i.e., NEE model), which will be used for classifying unseen named entities in the next steps.

*3.1.1.2. Annotating step*

In this step, each token of digitized ancient Mongolian manuscript will be annotated with the correct tag. We use the IOB2 [10] format for tagging tokens. 'B' tag indicates the beginning of a

http://dx.doi.org/10.5772/intechopen.72421

personal name, and 'I' tag indicates the tokens inside a personal name. 'O' tag indicates other tokens not belong to personal names. An example of the IOB2 annotation of the text in traditional Mongolian script can be seen in **Table 1**.

Because of some unique features of traditional Mongolian script, we also use 'Start/End' (SE) chunk tag set [11], which represents the character position in a word, along with the IOB2

| Token | Transliteration | IOB2 tag |
|---|---|---|
|  | ejen | O |
|  | naiiman | O |
|  | sirɣ-a_yi | O |
|  | ögelen | B |
|  | eke_dür_iyen | I |
|  | abču | O |
|  | irebei | O |
|  | iregsen_ü | O |
|  | qoyin-a | O |
|  | qan | O |
|  | yeke | O |
|  | oru | O |
|  | saɣubai. | O |
|  | tengri_eče | O |
|  | jayaɣ-a_bar | O |

| Token | Transliteration | IOB2 tag |
|---|---|---|
|  | törügsen | O |
|  | temüjin | B |
|  | činggis | B |
|  | qaɣan | I |
|  | buyu | O |

**Table 1.** An example of the IOB2 annotation of personal names in traditional Mongolian script text.

tags. 'S' tag is attached to the first character of each word including the personal names and 'E' tag to the last character. Therefore, each token will include the (1) IOB2 tag and (2) SE tag. SE tags are useful when there is a difference in word boundary between the test data and trained data [11, 12]. Particularly, an approach based on SE tags could improve the SVM prediction when there is no stemmer for traditional Mongolian. After attaching the IOB2 and SE tags to each token, we extract the features for chunking that will be used to learn the rules of personal name extraction. The features, i.e., characteristics of a token are explained in the next section.

*3.1.1.3. Named entity extraction step*

In this step, the proposed approach had to find the personal names in ancient Mongolian digitized texts. This method conducts the classification and grouping of tokens by SVM. The classifier in the SVM calculates a probability of a token belonging to personal names by inputting the extracted features to SVM. The features of a token might be possible clues to the proposed approach of whether or not this token is a named entity. In other words, we need some features to distinguish personal names.

We consider the following features of traditional Mongolian script for distinguishing personal names.

- **Preceding information of the current token**: If the preceding token is generational or dynastic information, an inherited or lifetime title of nobility, or a traditional descriptive phrase, it could indicate that current token is a personal name.
- **Beginning of a sentence**: For example, subjects or personal names are often at the beginning of a sentence.

http://dx.doi.org/10.5772/intechopen.72421

- **Suffix**: In traditional Mongolian script, many living being and humankind proper names take only certain plural suffixes such as ᠨᠠᠷ nar or ner and possessive suffixes [13].
- **Special non-word boundaries:** In traditional Mongolian script, some suffixes are visually separated from the stem of a word or from other suffixes, although they are an integral part of the word. Moreover, in some words with a final vowel letter 'a' or 'e', final vowel letters 'a' and 'e' are separated visually from the preceding consonant by a narrow gap although they are an integral part of the word stem.
- **End of token or special word delimiters**: A token is usually a word delimited by space, but there exist some unique features in traditional Mongolian script.
- **Information of the preceding and following tokens**: We also extract a feature by looking at the context of the current, preceding and succeeding IOB2 annotations (currently, the window stretches from $C_{n-2}$ *to* $C_{n+2}$) as visualized in **Table 2**. Such a feature could correct mislabelled IOB2 annotations.

The final task in this step is to extract the personal names, which have the proper names' markups, from the ancient Mongolian digital text.

### *3.1.2. Performance of extracting named entities from Mongolian historical documents*

The proposed method [14] is capable of extracting proper nouns from digitized text of ancient Mongolian manuscripts with 0.6993, 0.5679 and 0.6268 of precision, recall and F-measure, respectively, when utilizing a SVM tool LIBLINEAR with the L2-regularized L2-loss support vector classification (dual) solver [15].

When conducting experiments in extracting personal names from traditional Mongolian historical documents, we utilized digitized text of a chronological book of ancient Mongolian kings and the Mongol Empire—'Little' Altan Tobchi—which was made using bamboo pen xylograph technique as the experimental corpus. The 'Little' Altan Tobchi consists of 164 pages that contain approximately 16,200 words. The average number of words is 100 per page, with the longest one having 115 words and the shortest one 75 words. Precision, recall and F-measure were calculated by the fivefold cross-validation for extracting personal names.

Manually annotated named entities, extracted named entities [14], manually compiled scholar's commentaries and interpretations [9], as well as digital texts of ancient Mongolian manuscripts [7], will be utilized for building a digital edition of ancient Mongolian manuscripts. The next sections discuss how to develop a digital edition of Mongolian historical documents by describing some features and requirements of Mongolian manuscripts.

| Tokens | $W_{n-3}$ | $W_{n-2}$ | $W_{n-1}$ | $W_n$ | $W_{n+1}$ | $W_{n+2}$ | $W_{n+3}$ |
|---|---|---|---|---|---|---|---|
| IOB2 tags | $C_{n-3}$ | $C_{n-2}$ | $C_{n-1}$ | $C_n$ | $C_{n+1}$ | $C_{n+2}$ | $C_{n+3}$ |

**Table 2.** A feature of the preceding and following two tokens.

### 3.2. Making a web-based system by utilizing research outcomes

The past achievements in developing the TMSDL and the research outcomes of extracting named entities from Mongolian historical text allow us to create a digital representation that reflects ancient Mongolian historical manuscripts. This section covers our development in creating a web-based prototype system, which browses ancient Mongolian historical manuscripts.

#### *3.2.1. A digital edition of Mongolian manuscripts*

We utilized Edition Visualization Technology (EVT) for creating and browsing a digital edition of Mongolian manuscripts, which is encoded according to the Text Encoding Initiative (TEI) XML schemas and guidelines [16]. The named entities including the historical figures and place names are explicitly encoded using the TEI guidelines along with the additional data such as editorial markup, various commentaries, transcriptions and interpretations that have been suggested by researchers [9], etc., [17]. Well-known historical figures including generational or dynastic information, an inherited or lifetime title of nobility, or a traditional descriptive phrase or nickname are also marked. In the proposed digital edition, Unicode is chosen at the character level, and TEI P5 is applied on higher levels. As shown in **Figures 3** and **4**, all the personal names and place names in the 'Little' Altan Tobchi are visualized and highlighted in both transliteration and traditional Mongolian text. Image-to-text feature can

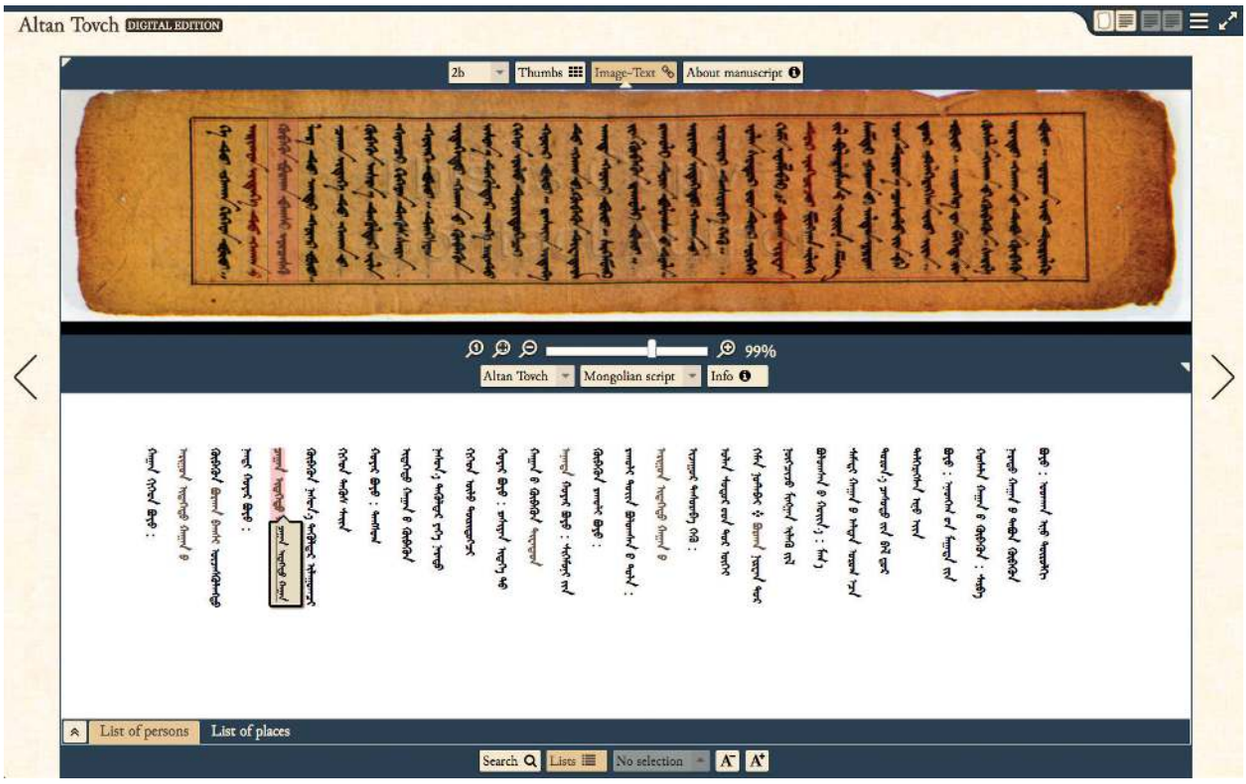


**Figure 3.** A digital edition with image-to-text link and personal names' highlights.

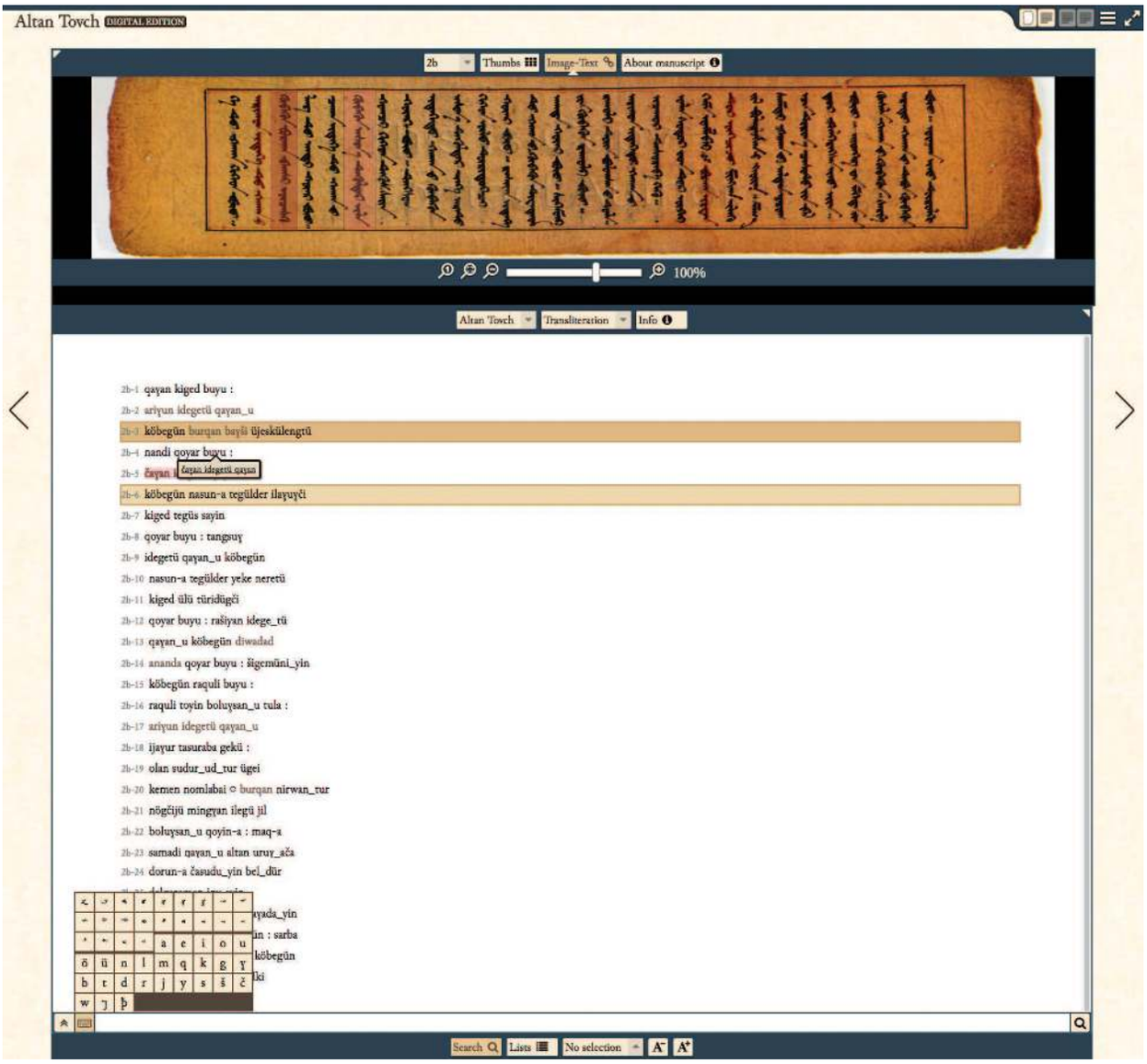


**Figure 4.** A digital edition with image-to-text link, a virtual keyboard and personal names' highlights in transliteration.

link a column in a manuscript folio image to the corresponding text and highlight them in all edition levels. As shown in **Figure 5**, all the named entities are listed as a full list with hyperlinks to the folios that appear certain named entity.

In addition, we made the following customizations in EVT to make it suitable for Mongolian manuscripts in traditional Mongolian script.

*3.2.1.1. Parallel-text editions with transliteration*

The proposed prototype can present scanned image-based editions with two edition levels: (1) diplomatic interpretative and (2) transliteration. Transliteration is helpful for those who are not familiar with a script of a certain language but understands that language. Transliteration in Latin letters of Mongolian historical documents is popular among scholars.

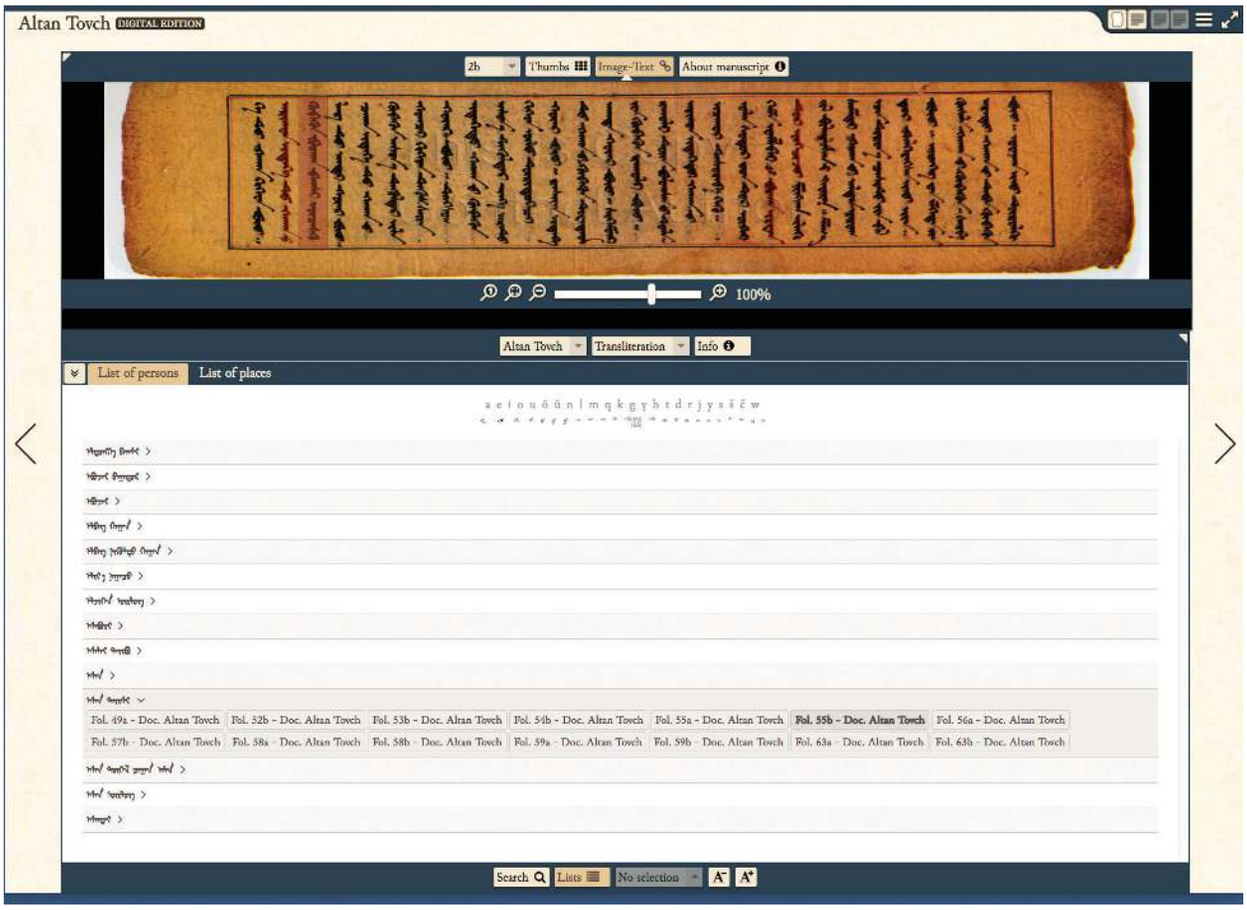


**Figure 5.** A list of named entities with hyperlinks to the folios of a Mongolian manuscript.

There is a limited recommendation to encode transliterations in TEI. Soualah and Hassoun [18] proposed to implement transliteration by using a specific model, which uses the [18] element with the *@xml:lang*, *@target* and *@type* attributes. However, we consider transliteration as a separate edition and use it as parallel-text editions as shown in **Figure 6**.

### *3.2.1.2. Supporting the traditional Mongolian script*

A unique feature of traditional Mongolian script is displaying vertically, from top to bottom, in columns advancing from left to right. Due to poor support for traditional Mongolian script at the EVT, we customized it to display the scanned images at the top and the corresponding text in traditional Mongolian script below with the direction top to bottom and left to right. We also set to display text in traditional Mongolian script on the left, and the corresponding transliteration in Latin letters on the right that can be used to compare them. Additionally, as shown in **Figures 4** and **6**, we added a simple virtual keyboard composed of 22 traditional Mongolian letters and their corresponding Latin letters to help users to input a Mongolian keyword to benefit free-text search and keyword highlighting.

http://dx.doi.org/10.5772/intechopen.72421

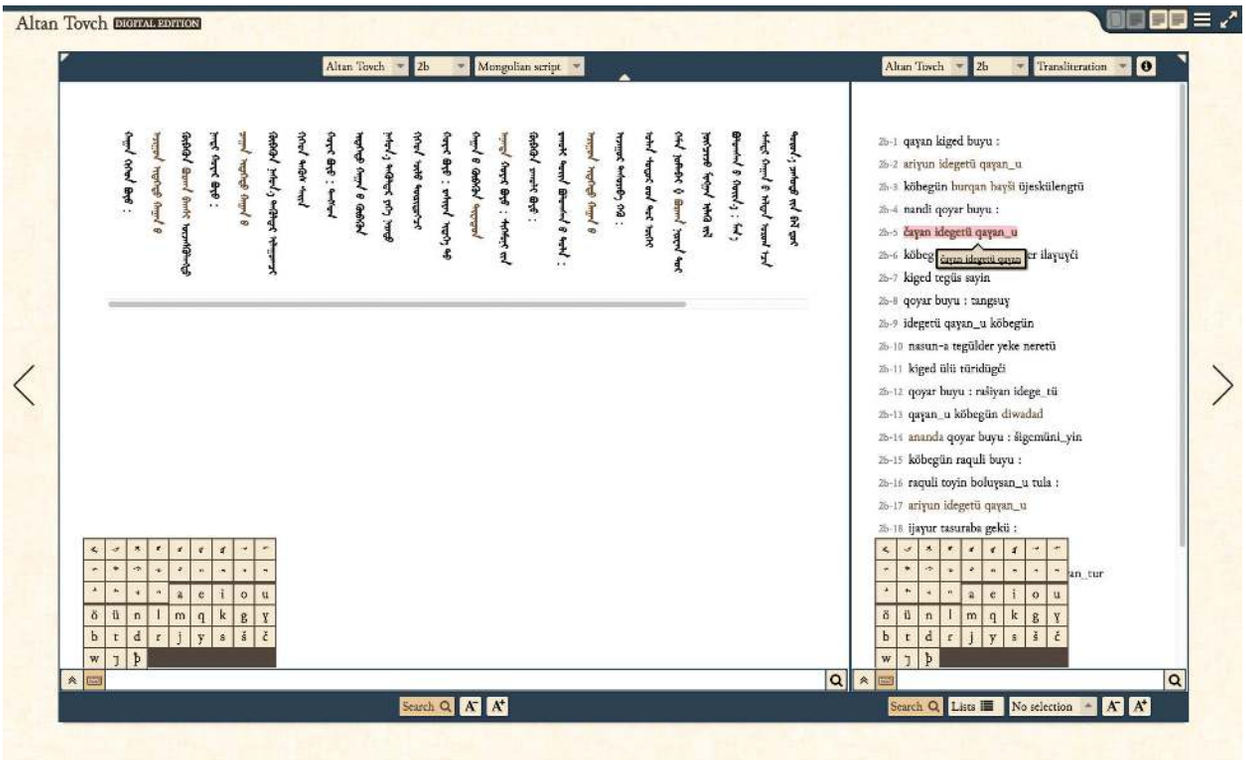


**Figure 6.** Parallel-text editions with personal names' highlights and virtual keyboards.

### 3.3. Applying and extending the proposed method to across languages

This section discusses (1) how the existing cross-language information retrieval techniques can be utilized in the proposed prototype system and (2) how the proposed approach can be applied to other languages in order to provide cross-lingual and cross-chronological information access to multilingual historical documents.

#### *3.3.1. Adopting cross-language and cross-chronological information retrieval techniques in historical documents*

There has been little research in information retrieval techniques for historical documents, and almost none of the breakthroughs in research in information retrieval and information access have aimed at retrieving information in the native language from ancient, cross-chronological and/or cross-script foreign language documents. Few approaches that could be considered a cross-chronological information retrieval have been proposed, and there has been little research in information retrieval techniques for historical documents. Ernst-Gerlach and Fuhr focused on modern and archaic German and developed a retrieval method that considers the spelling differences and variations over time [19]. Koolen et al. considered the spelling and pronunciation differences between ancient and modern Dutch [20], while Gotscharek et al. [21] and Hauser et al. [22] considered the spelling differences and variations between

modern and archaic German. Pilz et al. considered spelling variations of English and German historical texts [23]. In general, the main challenge for historical European languages like Dutch, English and German is the spelling variants.

Furthermore, Kimura and Maeda proposed a retrieval method that considers not only language differences over time but also cultural and time differences in modern and archaic Japanese [24]. Tripathi developed a retrieval system that considers the differences in various scripts and writing systems of Brahmic (Indic) and proposed a method to retrieve Sanskrit documents written in Sanskrit script or Brahmic families' scripts, using scripts such as Devanagari, Kannada, Telugu and Bengali [25]. To cope with cross-chronological and cross-script Mongolian documents, Khaltarkhuu and Maeda proposed a retrieval technique that is capable of searching traditional Mongolian script documents using modern Mongolian query [26–28].

We improved Khaltarkhuu and Maeda's grammatical-rule-based approach [26–28] and proposed an 'ancient-to-modern information retrieval' method [7, 29] by adding a dictionary-based query translation technique in order to consider cross-chronological differences in the writing systems of the ancient and modern Mongolian languages for accessing cross-chronological and cross-script ancient Mongolian documents by using a query in modern Mongolian in Cyrillic. To boost the quality of the translation, the 'ancient-to-modern information retrieval' approach [7, 29] matches query terms to words in a dictionary. If no exact match is found, the grammatical-rule-based approach [26–28] is used. In other words, the grammatical-rule-based query translation approach is used for inflected words, words with ancient spellings or grammar or the words missing from the dictionary. For the word sense disambiguation, in case if there are words which have multiple candidates, we choose the most frequent words. In our approach, we merge spelling variants of ancient Mongolian words.

We have already integrated the 'ancient-to-modern information retrieval' method in the TMSDL, and it can be easily applied to our digital edition for accessing ancient Mongolian historical collections written in traditional Mongolian script.

*3.3.2. Applying the proposed approach to other languages*

We have been demonstrating a facility for cross-language searching between English and Japanese for enabling English-speaking users to search Ukiyo-e databases available in Japanese by using English queries [30–32]. Such a feature is very useful for users, since the Ukiyo-e databases in Japanese institutions are mostly available in Japanese, so that users who do not understand Japanese may not find the desired information. Ukiyo-e, a Japanese traditional woodblock printing, is known worldwide as one of the fine arts of the Edo period (1603–1868). The texts of Ukiyo-e databases contain archaic Japanese words which reflect the Japanese language of the Edo period.

Like the 'ancient-to-modern information retrieval', a dictionary-based query translation approach is adopted by utilizing a domain-specific dictionary, which contains the terms related to Japanese arts and cultures. The proposed feature works well with a variety of keywords (i.e., no full sentences) that may include the personal names, specific terms such as 'Geisha', traditional Japanese female entertainers; 'Fuji', Mount Fuji, the highest mountain in

http://dx.doi.org/10.5772/intechopen.72421

Japan; and 'Sumo', Japanese traditional wrestling. For instance, if the search query submitted by the user is a name of the Ukiyo-e artist, i.e., 'Utagawa Hiroshige', then the query 'Utagawa Hiroshige' is translated into Japanese as '歌川広重' and sent to Japanese databases.

We are conducting further research to generalize the proposed method to other historical documents in various languages. We also believe that the proposed prototype could be applied to other historical documents in Todo, Manchu and Sibe, which are the derivative scripts of traditional Mongolian.

## 4. Summary and future directions

In this chapter, we have described our research to achieve cross-lingual and cross-chronological information access to ancient Mongolian historical materials. More specifically, we have introduced methods for providing information access that cuts across different historical periods and dialects.

We introduced an information extraction method for digitized ancient Mongolian historical manuscripts of the 13–16th century in Sections 3. The proposed information extraction method for ancient Mongolian historical documents performs computerized massive analysis on Mongolian historical documents. It can reduce traditional labour-intensive manual analysis on Mongolian historical text significantly. Named entities such as historical figures and places of ancient Mongolia that are difficult for manual examination are recognized from historical manuscripts.

The extracted results are utilized for building a digital edition of an ancient Mongolian historical document and made available through a web-based system.[1] We also believe the TEI-encoded digital edition that reflects the ancient Mongolian manuscripts would help scholars conducting research in the ancient history for digging hidden knowledge of the Middle Ages of Mongolia in ancient Mongolian historical documents that is not available in modern-language documents. Furthermore, explicitly encoded digital text enables users to search and browse ancient Mongolian manuscript using the named entities' visualization, i.e., it allows not only retrieving information but also analysing and visualizing the contents of the information. We also hope digital editions along with the scanned images would recreate the experience of encountering the original manuscripts. Its information visualization feature of ancient Mongolian texts and a TMSDL's feature that can retrieve ancient manuscripts written in traditional Mongolian script using a query in modern Mongolian (Cyrillic) would help researchers who are interested in using digital representations of ancient historical manuscripts as scholarly tools by using a modern language. Such a feature is very useful, since the needs of humanities researchers are diverse and might require access to information in ancient languages, rather than searching and browsing limited collections in modern languages. Indeed Mongolian ancient documents are mostly available in ancient scripts and dialects, so users who do not understand ancient Mongolian may not find the desired information.

[1] http://www.dl.is.ritsumei.ac.jp/AltanTovch/

Finally, the proposed prototype could be applied to other documents in Todo, Manchu and Sibe, which are the derivative scripts of traditional Mongolian. The systems introduced in this chapter are targeted primarily at researchers in the humanities field. Nevertheless, these systems are expected to be useful to users other than researchers, in the sense that they open up new possibilities for acquiring the kinds of information that cannot be found solely in modern documents available on the web.

## Author details

Biligsaikhan Batjargal

Address all correspondence to: biligsaikhan@gmail.com

Research Organization of Science and Technology, Ritsumeikan University, Japan

## References

[1] Shagdarsüren Ts. Study of Mongolian Scripts (Graphic Study or Grammatology). 2nd ed. Ulaanbaatar: Urlakh Erdem Khevleliin Gazar: Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia; 2001. 299 pp

[2] The Unicode Consortium. Chapter 13: South and Central Asia-II, Other Modern Scripts. The Unicode Standard, Version 10.0.0 [Internet]. 2017 June 20 [Updated: 2017 June 20]. Available from: http://www.unicode.org/versions/Unicode10.0.0/ch13.pdf

[3] The Unicode Consortium. Chapter 14: South and Central Asia-III, Ancient Scripts. The Unicode Standard, Version 10.0.0 [Internet]. 2017 June 20 [Updated: 2017 June 20]. Available from: http://www.unicode.org/versions/Unicode10.0.0/ch14.pdf

[4] Sečenbaγatur, Qasgerel, Tuyaγ-a, ǰirannige B, Ying ǰe U. Mongγul kelen-ü nutuγ-un ayalγun-u sinǰilel-ün uduridqal. Kökeqota: Öbür mongγul-un arad-un keblel-ün qoriy-a; 2005

[5] Svantesson J, Tsendina A, Karlsson A, Franzén V. The Phonology of Mongolian. New York: Oxford University Press; 2005. p. 336

[6] Bira Sh. The Golden Summary which Relates Briefly the Deeds of Civil Governing Established by Ancient emperors (The Mongol Chronicle of the 17th Century). Ulaanbaatar, Mongolia: Mongolia: Ulsiin hevleliin gazar; 1990. 222 pp

[7] Batjargal B, Khaltarkhuu G, Kimura F, Maeda A. Developing a digital library of historical records in traditional Mongolian script. International Journal of Digital Library Systems. 2012;**3**(1):33-53. DOI: 10.4018/jdls.2012010103

[8] Finkel JR, Grenager T, Manning CD. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting

on Association for Computational Linguistics (ACL '05); June 25-30, 2005; Ann Arbor, Michigan. Stroudsburg, PA, USA: Association for Computational Linguistics; 2005. pp. 363-370. DOI: 10.3115/1219840.1219885

[9] Choimaa Sh. Qad-un ündüsün quriyangγui altan tobči (Textological Study). vol. 1 ed. Ulaanbaatar: Urlakh Erdem Khevleliin Gazar: Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia; 2002. 276 pp

[10] Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, editors. Natural Language Processing Using Very Large Corpora. Dordrecht: Springer Netherlands; 1999. pp. 157-176. DOI: 10.1007/978-94-017-2390-9_10

[11] Asahara M, Matsumoto Y. Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1; May 27-June 01, 2003; Edmonton, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics; 2003. pp. 8-15. DOI: 10.3115/1073445.1073447

[12] Yoshimura M, Kimura F, Maeda A. Personal name extraction from ancient Japanese texts. In: Proceedings of the Exploration, Navigation and Retrieval of Information in Cultural Heritage ENRICH 2013 Workshop; 1 August 2013; Dublin, Ireland. New York, NY, USA: ACM; 2013. pp. 31-34

[13] Chinggaltai, editor. A grammar of the Mongol language. Revised ed. New York: Frederick Ungar Publishing Co; 1963. 173 pp

[14] Batjargal B, Khaltarkhuu G, Kimura F, Maeda A. An approach to named entity extraction from Mongolian Historical Documents. In: Proceedings of the 2015 International Conference on Culture and Computing; 17-19 October; Kyoto, Japan. IEEE Computer Society; 2015. pp. 205-206. DOI: 10.1109/Culture.and.Computing.2015.41

[15] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research. 2008;**9**:1871-1874

[16] Del Turco RR, Buomprisco G, Pietro CD, Kenny J, Masotti R, Pugliese J. Edition visualization technology: A simple tool to visualize TEI-based digital editions. Journal of the Text Encoding Initiative [Online]. 2014;**8**. DOI: 10.4000/jtei.1077

[17] Batjargal B, Khaltarkhuu G, Kimura F, Maeda A. Applying text encoding initiative guidelines to a historical record in traditional Mongolian script. In: Proceedings of the 2013 International Conference on Culture and Computing; 16-18 September; Kyoto, Japan. IEEE Computer Society; 2013. p. 141-142. DOI: 10.1109/CultureComputing.2013.36

[18] Soualah MO, Hassoun MA. TEI P5 manuscript description adaptation for cataloguing digitized Arabic manuscripts. Journal of the Text Encoding Initiative [Online]. 2012;**2**:DOI: 10.4000/jtei.398

[19] Ernst-Gerlach A, Fuhr N. Retrieval in text collections with historic spelling using linguistic and spelling variants. In: Rasmussen E, editor. Proceedings of the 7th ACM/IEEE-CS

Joint Conference on Digital Libraries; June 18-23; Vancouver, BC, Canada. USA/New York City: ACM; 2007. p. 333-341. DOI: 10.1145/1255175.1255242

[20] Koolen M, Adriaans F, Kamps J, Rijke M. A cross-language approach to Historic Document retrieval. In: Lalmas M, MacFarlane A, Rüger S. M, Tombros A,Tsikrika T, Yavlinsky A, editors. Proceedings of the 28th European conference on Advances in Information Retrieval ECIR. Heidelberg: Springer-Verlag Berlin; 2006. pp. 407-419. DOI: 10.1007/11735106_36

[21] Gotscharek A, Reffle U, Ringlstetter C, Schulz K, Neumann A. Towards information retrieval on historical document collections: The role of matching procedures and special lexica. International Journal on Document Analysis and Recognition (IJDAR). 2011;**14**(2):159-171. DOI: 10.1007/s10032-010-0132-6

[22] Hauser A, Heller M, Leiss E, Schulz K, Wanzeck C. Information access to Historical Documents from the early new high German Period. In: Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data; January 8; Hyderabad, India. 2007. p. 147-154

[23] Pilz T, Ernst-Gerlach A, Kempken S, Rayson P, Archer D. The identification of spelling variants in English and German historical texts: Manual or automatic? Literary and Linguistic Computing. 2008;**23**(1):65-72. DOI: 10.1093/llc/fqm044

[24] Kimura F, Maeda A. An approach to information access and knowledge discovery from Historical Documents. In: Conference Abstracts of the Digital Humanities 2009; June 22-25; College Park, Maryland. 2009. pp. 359-361

[25] Tripathi A. Saraswati: Cross-lingual Sanskrit digital library. Library Hi Tech News. 2009;**26**(10):1-5. DOI: 10.1108/07419050911022252

[26] Khaltarkhuu G, Maeda A. Retrieval technique with the modern Mongolian query on traditional Mongolian text. In: Sugimoto Sh, Hunter J, Rauber A, Morishima A, editors. Digital Libraries: Achievements, Challenges and Opportunities, 9th International Conference on Asian Digital Libraries, ICADL 2006; November 27-30; Kyoto, Japan. Heidelberg: Springer, Berlin; 2006. p. 478-481. DOI: 10.1007/11931584_5

[27] Khaltarkhuu G, Maeda A. Building a digital library of traditional Mongolian Historical Documents. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries; June 18-23; Vancouver, BC, Canada. NY, USA: ACM New York; 2007. p. 483. DOI: 10.1145/1255175.1255282

[28] Khaltarkhuu G, Maeda A. Developing a Traditional Mongolian Script Digital Library. In: Buchanan G, Masoodian M, Cunningham SJ, editors. Digital Libraries: Universal and Ubiquitous Access to Information, 11th International Conference on Asian Digital Libraries, ICADL 2008; 2-5 December; Bali, Indonesia. Heidelberg: Springer, Berlin; 2008. pp. 41-50. DOI: 10.1007/978-3-540-89533-6_5

[29] Batjargal B, Khaltarkhuu G, Kimura F, Maeda A. Ancient-to-modern information retrieval for digital collections of Traditional Mongolian Script. In: Chowdhury GG, Koo C, Hunter J, editors. The Role of Digital Libraries in a Time of Global Change,

12th International Conference on Asia-Pacific Digital Libraries, ICADL 2010; June 21-25; Gold Coast, Australia. Heidelberg: Springer, Berlin; 2010. p. 25-28. DOI: 10.1007/978-3-642-13654-2_4

[30] Batjargal B, Kimura F, Maeda A. Approach to Cross-language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database. In: Lalmas A, Jose J, Rauber A, Sebastiani F, Frommholz I, editors. Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2010); September 6-10; Glasgow, UK. Heidelberg: Springer, Berlin; 2010. p. 518-521. DOI: 10.1007/978-3-642-15464-5_71

[31] Batjargal B, Kimura F, Maeda A. Realizing bilingual and parallel access to Ukiyo-e databases in the World. In: Proceedings of the International Conference on Culture and Computing (Culture and Computing 2011); October 20-22; Kyoto, Japan. IEEE; 2011. pp. 165-166. DOI: 10.1109/Culture-Computing.2011.48

[32] Batjargal B, Maeda A, Akama R. Providing bilingual access to multiple Japanese humanities databases: Text retrieval using English and Japanese queries. In: Hsiang J, editor. Digital Humanities: Between Past, Present, and Future. Taipei, Taiwan: National Taiwan University Press; 2016. pp. 351-367

IntechOpen