overseas. For examples, Huang Lunsu graduated from Fuzhou Shipyard, Zhong Jincheng studied in the United States, Pan Yingqi graduated from Shixue College (實學館). A third group was trained in professional Western-style surveying and mapping techniques. Gu Chao, Zou Jin, and Lao Yingan had been involved in the process of settling a boundary dispute between China and Vietnam. From this we see the gradual introduction on modern cartography from the Daoguang period to the end of Qing dynasty.

Access to Logart

Although LoGaRT was developed as software to work with any collection of digital local gazetteers, it currently links to only the *Zhongguo fangzhi ku* from Erudition (*Airusheng* 愛如生) that is housed at the Berlin State Library. This dataset currently includes 3,999 local gazetteers, with some four million scanned pages, dating from the Tang dynasty through the early twentieth century. This accounts for roughly half of known extant gazetteers. Because this requires a special license, currently only MPIWG affiliates are entitled to use LoGaRT during their affiliation periods. We have been exploring ways to extend the access to LoGaRT, and we are planning on releasing a LoGaRT instance with open access contents in the spring of 2020.[12] In the future we will release LoGaRT as a stand-alone software package, allowing any institution to install it and link it with collections of digital local gazetteers for which they have rights.

Using Philologic For Digital Textual and Intertextual Analyses of the *Twenty-Four Chinese Histories* 二十四史

Jeffrey Tharsen and Clovis Gladstone

doi:10.1017/jch.2020.27

Abstract
What does it mean to be able to study Chinese history at scale? What methods, tools, and approaches will allow us to understand Chinese history and historiography from a larger perspective over the longue durée, including linguistic, philosophical, ethnographic, and literary concerns? In this article we present what we feel is one potential key to answering these questions and provide an overview of the utility and value of harnessing this framework for text-based historical research as a means to expand one's scholarship to virtually limitless scales.

[12] This Open LoGaRT will contain around 400 titles of local gazetteers from the Harvard Yenching Library rare book collection. We thank Chiang Ching-kuo Foundation and the Max Planck Society for their funding to digitize this set of local gazetteers as searchable full texts.

Jeffrey Tharsen, University of Chicago, email: jcarlsen@uchicago.edu and Clovis Gladstone, University of Chicago, email: clovisgladstone@uchicago.edu

https://doi.org/10.1017/jch.2020.27 Published online by Cambridge University Press

A Brief History of Philologic

Concordances are the oldest, simplest, and in many ways the most powerful tool for navigating and exploring texts.[1] Concordances provided the opportunity to compare the different uses of the same word in the context of surrounding words within a coherent set of texts. Centuries later, this same tool was at the origin of the digital humanities: in 1951, a Jesuit researcher, Roberto Busa, undertook to create a complete concordance of the entire work of Thomas Aquinas with the help of IBM, leading, decades later, to the publication of the *Thomistica Index* in fifty-six volumes.[2] Based on the recognized power of this approach to text analysis, the ARTFL Project began in the late 1980s and early 1990s to work on PhiloLogic, an open source full-text search, retrieval, and analysis system powered by concordances. Originally developed to serve the needs of the ARTFL Project[3] and its large user base, in particular the exploration of the "ARTFL-Frantext"[4]and "ARTFL-Encyclopédie"[5] collections, PhiloLogic has been extended to handle an ever increasing array of collections, text encodings, and languages.[6] The ARTFL Project and the Textual Optics Lab[7] relies on this software not only to deliver many different types of text collections and dictionaries to scholars, but also to host databases for collaborative projects such as the "Opera del Vocabolario Italiano,"[8] the "History of Black Writing,"[9] or the "Shanghai Library Republican Journal corpus."[10] As mass digitization efforts such as HathiTrust, Google Books, the Internet Archive, and the Digital Library of America Project transform the conditions (and stakes) of humanities text research, PhiloLogic is designed to provide rigorous analysis and discovery tools on these larger and more complex collections while accommodating the growing demand for new kinds of text analysis and visualization of results.

A considerable amount of work has been devoted to designing a user-friendly interface, most notably by making all of the functionality of the software easily accessible. In many ways, PhiloLogic's user experience is meant to provide a workspace for the researcher,

[1] We can trace concordances back to the Dominicans, who in the thirteenth century had the idea of providing a new way to study the Bible.

[2] See Thomas Nelson Winter, "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance," *The Classical Bulletin* 75.1 (1999), 3–20

[3] Founded in 1982, the ARTFL Project is one of the oldest Digital Humanities Lab in North America. See http://artfl-project.uchicago.edu/ for more details.

[4] See http://artfl-project.uchicago.edu/content/artfl-frantext.

[5] See http://encyclopedie.uchicago.edu/.

[6] PhiloLogic is completely open-source; installation instructions for Linux and Mac OS are available here: https://artfl-project.github.io/PhiloLogic4/installation.html. For those who prefer GitHub, the current PhiloLogic repository is https://github.com/ARTFL-Project/PhiloLogic4.

[7] The Textual Optics Lab is a Digital Humanities Lab at the University of Chicago which comprises scholars from the ARTFL Project and the Chicago Text Lab. See https://textual-optics-lab.uchicago.edu/ for more details.

[8] A collaboration between the ARTFL Project, the Centro di studi Opera del Vocabolario Italiano (Florence, Italy), the William and Katherine Devers Program in Dante Studies (University of Notre Dame), and the Department of Italian Studies at the University of Reading. See http://artfl-project.uchicago.edu/content/ovi for more details.

[9] A collaboration between the University of Chicago and The Project on the History of Black Writing (HBW) at the University of Kansas. See https://textual-optics-lab.uchicago.edu/black_writing_corpus for more details.

[10] A collaboration between the University of Chicago and the Chinese Periodical Database 全国报刊数据库 at the Shanghai Library. See https://textual-optics-lab.uchicago.edu/shanghai-library-republican-journal-corpus and http://www.cnbksy.cn/ for more details.

https://doi.org/10.1017/jch.2020.27 Published online by Cambridge University Press

where hypotheses can be formulated and verified with just a few clicks, guiding scholars deeper and deeper into the collections as they leverage the frequency distributions of results displayed by the faceted browser, to finally arrive at the actual text. The seamless navigation between queries at the corpus level and the text is an essential aspect of PhiloLogic: it gives the researcher the opportunity to change reading scales, from a more distant perspective appropriate for finding general patterns, to the close reading experience, which has been further enhanced by the information gathered at the corpus level queries.

Additionally, PhiloLogic also provides a number of APIs and data outputs for further text-mining experiments. In particular, the TextPAIR sequence alignment tool can leverage the PhiloLogic index, text structure, and metadata gathered during the parsing stage to conduct a document to document comparison to uncover text-reuses within and between PhiloLogic databases. Future work on PhiloLogic will involve a closer integration between the results of such data-mining tasks into the main search and navigation interface, therefore allowing researchers to gain new insights into an ever-growing array of text collections.

Philologic Concordance Search, Kwic, Collocations and Time Series

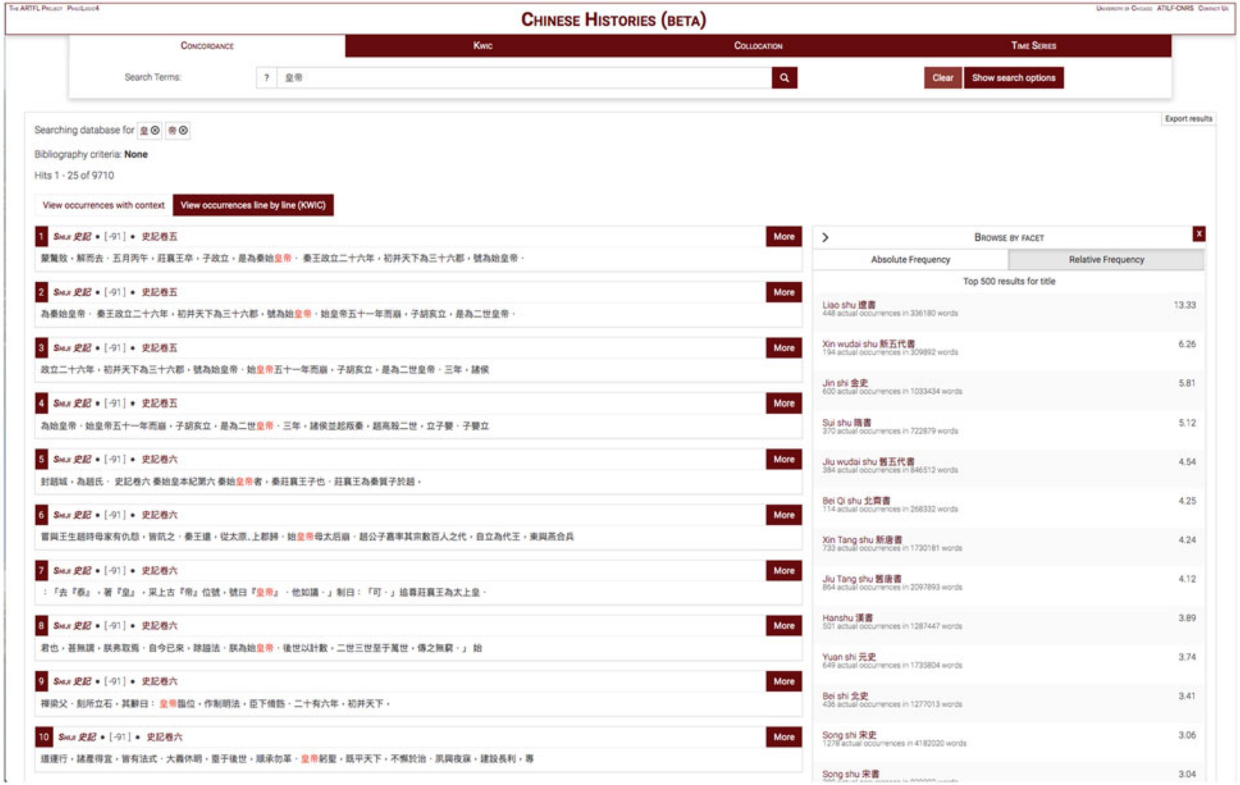
The ability to perform targeted searches is a key component of any digital textual data repository. The PhiloLogic Concordance function[11] is designed to search the corpus by harnessing the power of regular expression syntax, including full support for wildcards and complex sequences of search terms. Beyond the concordance search function, the use of facets, drawn from the metadata that accompanies each TEI-encoded document in the source corpus, provides an on-the-fly means to further constrain and explore the output; this becomes particularly valuable when thousands or millions of results are returned. Metadata categories within the facet browser are sorted by default, with the largest frequencies (both absolute and relative to the total size of each document) presented at the top of the list. Common facets include title, date, author, and section type, but any metadata tags can be employed as facets and thereby provide users with a highly individualized and configurable experience while browsing through search results.

By default, each search term returned is highlighted within its specific textual context of twenty-five words or graphs on each side; clicking on "View occurrences by line (KWIC, Key Word In Context)" changes the output to a vertical display with the primary search term in the center and its context on either side. This display gives the user the ability to quickly scan through and sort by the words directly preceding and following the highlighted search term, providing penetrating insights into which sequences of words and phrases containing the search terms occur most frequently. Clicking on "Export results" in the upper right produces a JSON file object in Unicode encoding in a new window, containing the complete data for the concordance produced by the search.

Beyond the concordance functions, PhiloLogic provides two additional algorithmically generated types of results: collocations and time series. The Collocations function generates a word cloud visualization of the top 100 most frequent words (or in the Chinese case, characters) that co-occur in the sentences retrieved; the terms in the word cloud

[11] See https://artflsrv03.uchicago.edu/philologic4/histories_5_7 for the Concordance UI for the *Twenty-four Chinese Histories*. Figure 1 contains a screenshot of the UI.

https://doi.org/10.1017/jch.2020.27 Published online by Cambridge University Press



The ARTFL Project PhiloLogic4
University of Chicago ATILF-CNRS Contact Us

CHINESE HISTORIES (BETA)

Concordance
Kwic
Collocation
Time Series

Search Terms: ? 皇帝

Clear
Show search options
Export results

Searching database for 皇 ⊗ 帝 ⊗

Bibliography criteria: None

Hits 1 - 25 of 9710

View occurrences with context
View occurrences line by line (KWIC)

1 Shiji 史記 • [-91] • 史記卷五

蒙驁敗，解而去．五月丙午，莊襄王卒，子政立，是為秦始皇帝． 秦王政立二十六年，初并天下為三十六郡，號為始皇帝．

2 Shiji 史記 • [-91] • 史記卷五

為秦始皇帝． 秦王政立二十六年，初并天下為三十六郡，號為始皇帝．始皇帝五十一年而崩，子胡亥立，是為二世皇帝．

3 Shiji 史記 • [-91] • 史記卷五

政立二十六年，初并天下為三十六郡，號為始皇帝．始皇帝五十一年而崩，子胡亥立，是為二世皇帝．三年，諸侯

4 Shiji 史記 • [-91] • 史記卷五

為始皇帝．始皇帝五十一年而崩，子胡亥立，是為二世皇帝．三年，諸侯並起叛秦，趙高殺二世，立子嬰．子嬰立

5 Shiji 史記 • [-91] • 史記卷六

封趙城，為趙氏． 史記卷六 秦始皇本紀第六 秦始皇帝者，秦莊襄王子也．莊襄王為秦質子於趙，

6 Shiji 史記 • [-91] • 史記卷六

嘗與王生趙時母家有仇怨，皆阬之．秦王還，從太原、上郡歸．始皇帝母太后崩．趙公子嘉率其宗數百人之代，自立為代王，東與燕合兵

7 Shiji 史記 • [-91] • 史記卷六

：「去『泰』，著『皇』，采上古『帝』位號，號曰『皇帝』．他如議．」制曰：「可．」追尊莊襄王為太上皇．

8 Shiji 史記 • [-91] • 史記卷六

君也，甚無謂，朕弗取焉．自今已來，除謚法．朕為始皇帝．後世以計數，二世三世至于萬世，傳之無窮．」始

9 Shiji 史記 • [-91] • 史記卷六

禪梁父．刻所立石，其辭曰：皇帝臨位，作制明法，臣下修飭．二十有六年，初并天下，

10 Shiji 史記 • [-91] • 史記卷六

道運行，諸產得宜，皆有法式．大義休明，垂于後世，順承勿革．皇帝躬聖，既平天下，不懈於治．夙興夜寐，建設長利，專

BROWSE BY FACET

Top 500 results for title

Liao shu 遼書
448 actual occurrences in 336180 words 13.33

Xin wudai shu 新五代書
194 actual occurrences in 309892 words 6.26

Jin shi 金史
600 actual occurrences in 1033434 words 5.81

Sui shu 隋書
370 actual occurrences in 722879 words 5.12

Jiu wudai shu 舊五代書
384 actual occurrences in 846512 words 4.54

Bei Qi shu 北齊書
114 actual occurrences in 268332 words 4.25

Xin Tang shu 新唐書
733 actual occurrences in 1730181 words 4.24

Jiu Tang shu 舊唐書
864 actual occurrences in 2097893 words 4.12

Hanshu 漢書
501 actual occurrences in 1287447 words 3.89

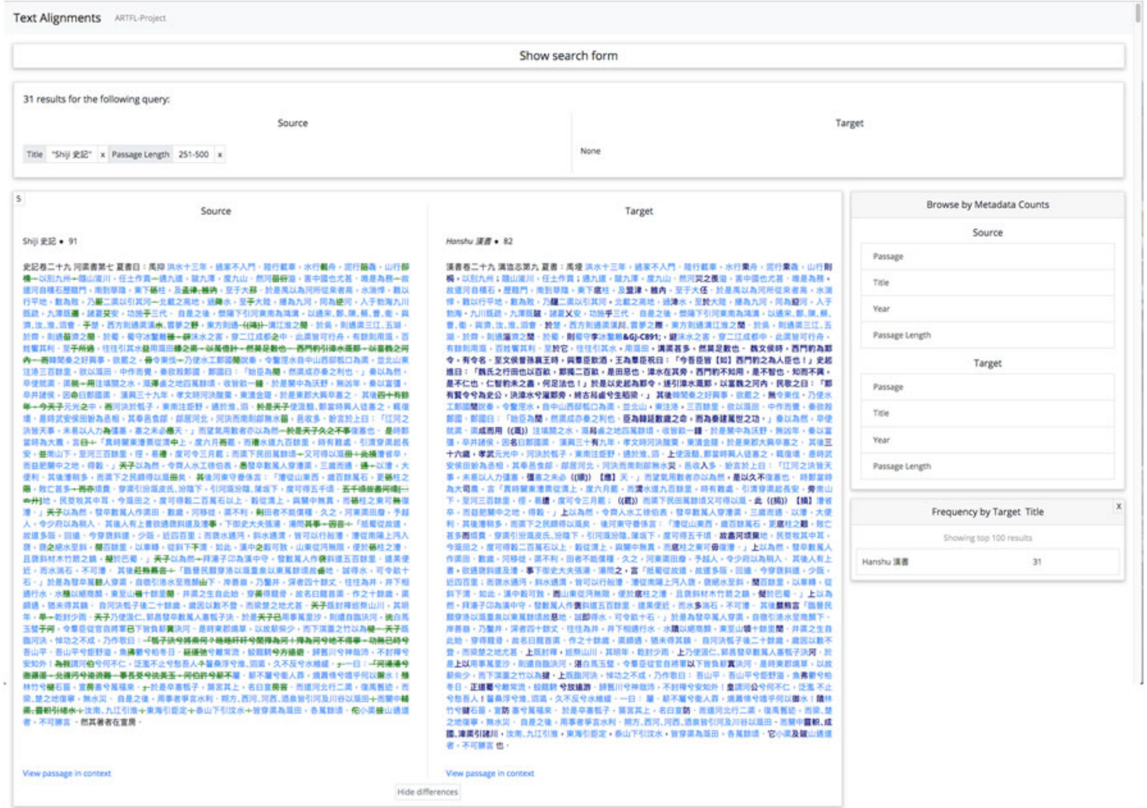
Yuan shi 元史
649 actual occurrences in 1735804 words 3.74

Bei shi 北史
436 actual occurrences in 1277013 words 3.41

Song shi 宋史
1278 actual occurrences in 4182020 words 3.06

Song shu 宋書 3.04

Figure 1. Screenshot of Concordance Search in PhiloLogic



Text Alignments ARTFL-Project

Show search form

31 results for the following query:

Source	Target
Title "Shiji 史記" × Passage Length 251-500 ×	None

5

Source	Target
Shiji 史記 • 91	*Hanshu 漢書* • 82
史記卷二十九 河渠書第七 夏書曰：禹抑洪水十三年，過家不入門。陸行載車，水行載舟，泥行蹈毳，山行即橋。[illegible] 不可勝言。然其著者在宣房。	漢書卷二十九 溝洫志第九 夏書：禹堙洪水十三年，過家不入門。陸行載車，水行乘舟，泥行乘毳，山行則梮。[illegible] 不可勝言也。
View passage in context	View passage in context

Hide differences

Browse by Metadata Counts

Source

Passage

Title

Year

Passage Length

Target

Passage

Title

Year

Passage Length

Frequency by Target Title ×

Showing top 100 results

Hanshu 漢書	31

Figure 2. Screenshot of the TextPAIR Sequence Alignment in PhiloLogic

are sized by their relative frequency so that the most frequent collocates are displayed largest. The time series visualization provides a simple horizontal timeline of the works returned from the concordance search, with vertical bars representing the absolute counts or relative frequency of the search terms in each document in the corpus.

Textpair : Large-Scale Intertextual Analysis (Sequence Alignment)

Algorithmically based intertextual analysis is an increasingly popular analytical method through which all exact or similar textual phrases or passages in a text or corpus can be detected using a simple matching algorithm that relies on an n-gram representation of the text with a sliding window for evaluation.[12] For a corpus like the *Twenty-four Chinese Histories*, algorithmic intertextual analysis allows the user to efficiently assemble and review graph-by-graph similarities and differences (changes, interpolations and deletions) between passages of any length greater than a sentence in each source.[13] As with the concordance functions, full corpus-level intertextual results can also be further constrained by facets, allowing a user to quickly assemble and review all passages shared between two or more sources, authors, or any other criteria included in the metadata.[14] These results can also be constrained by passage, title, passage length, or time period.

In PhiloLogic's TextPAIR framework, under each pair of similar passages, clicking on "Show differences" provides the at-a-glance results of the similarity matrix mapped onto each source, with identical sections in light blue type, interpolations in dark blue boldface type, and deletions marked in boldface green type with strikethrough. Figure 2 shows a passage from the twenty-ninth chapter of the *Records of the Historian* (*Shi ji* 史記) compared with its counterpart from the twenty-ninth chapter of the *Book of Han* (*Han shu* 漢書).[15] It bears noting that this result does not necessarily indicate a direct linear connection between the two passages being compared (à la the *stemma codicum* model in redaction criticism), all that is necessary is that it meet the mathematical threshold of a minimum shared number of lexical units (words or graphs) in sequence, and thus this type of intertextual analysis equally indicates citations, allusions, and paraphrased portions in any combination of sources from any time period, genre, or style.

[12] For a description of the algorithm, see Mark Olsen, Russell Horton, and Glenn Roe, "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections," *Digital Studies / Le Champ numérique* 2.1 (2010), DOI: http://doi.org/10.16995/dscn.258. A similar method is used in D. A. Smith, R. Cordell and E. M. Dillon, "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers," *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, 2013, doi: 10.1109/BigData.2013.6691675.

[13] See http://anomander.uchicago.edu/text-pair/histories.

[14] For a similar example focused upon Ming literature, see Paul Vierthaler and Mees Gelein, "A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora," *Journal of Cultural Analytics*, March 2019, http://doi.org/10.22148/16.034.

[15] For example, for the *Records of the Historian* (*Shi ji* 史記) compared with the *Book of Han* (*Han shu* 漢書), TextPAIR identified 3,933 shared passages.

https://doi.org/10.1017/jch.2020.27 Published online by Cambridge University Press