2016 届研究生硕士学位论文

分类号：　　学校代码：10269

密　级：　　学　　号：51132901114



華東師範大學

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

论文题目：漢簡語料庫的設計與應用

院　　系：中國語言文學系

专　　业：漢語言文字學

研究方向：古文字信息化處理

指导教师：張再興 教授

学位申请人：溫博雅

2016 年 3 月

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2016 届研究生硕士学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51132901114



華東師範大學

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

论文题目：漢簡語料庫的設計與應用

院　　系：中國語言文學系

专　　业：漢語言文字學

研究方向：古文字信息化處理

指导教师：張再興 教授

学位申请人：溫博雅

2016 年 3 月

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2016 Master Graduates University Code:10269

Student Number: 51132901114

East China Normal University

The Design and Application of the Han Dynasty Bamboo Slips Corpus

College: Chinese Language and Literature

Major: Chinese Language and Graphology

Specialty: Ancient Text Information Processing

Supervisor: Prof.Zhang Zaixing

Author: Wen Boya

March, 2016

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《汉简语料库的设计与应用》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：温博雅　　日期：2016 年 5 月 31 日

华东师范大学学位论文著作权使用声明

《汉简语料库的设计与应用》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（　）1.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，于　　年　　月　　日解密，解密后适用上述授权。

（√）2.不保密，适用上述授权。

导师签名　　本人签名　温博雅

2016 年 5 月 31 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

温博雅 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
王元鹿	教授	华东师范大学	主席
刘志基	教授	华东师范大学	
潘玉坤	教授	华东师范大学	
董莲池	教授	华东师范大学	
吕志峰	副教授	华东师范大学	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

摘要

計算機科學的發展衍生出了信息處理技術、信息存儲技術等高科技手段，為傳統的漢語言文字學帶來了新的研究手段、研究方法、研究理念的創新。漢代簡牘作為我國漢代歷史的珍貴實物資料和載體，具有極高的研究價值。因為數字化技術獨有的便捷、快速、信息處理能力強等優點，因此在傳統的研究之上，現在主流的研究機構都開始進行著古籍數字化的工作。通過搭建漢代簡牘語料庫，可以從數據收集、自動標註、網絡平台搭建幾個方面，逐步完成漢簡數字化的工作。

本文在計算機數據庫知識、網絡編程知識、簡牘研究知識以及中文信息自動化處理知識的指導之下，進行了漢簡語料庫搭建的探究與嘗試。在此期間完成了漢簡數據庫的建設工作、漢簡語料庫網絡平台設計工作、漢簡自動分詞、詞性自動標註工作等。解決了在數據庫建設、網絡平台建設中遇到的一些實際應用問題，在對簡牘自動分詞和詞性標註的領域進行了探索性的方法嘗試，并分析了西北漢簡中的專有名詞和古書類材料的虛實詞的標註結果，提出了進一步改進方法。最後，粗略統計了西北漢簡中的人名、地名等專有類名詞并形成了詞表。

論文主體包括五部分：

1) 緒論。分別介紹了選題依據、研究意義、漢簡語料庫建設與應用現狀、現狀分析與改善目標、 研究方法和創新點。
2) 第二章介紹了漢簡語料庫設計的過程。從數據庫和網絡平台兩個方面介紹了本漢簡語料庫的功能和特點以及搭建步驟。
3) 第三章是以漢簡語料庫為基礎，進行了漢簡文本自動分詞與詞性標註研究。重點選取了西北漢簡的人名、地名等專有名詞標註，以及古書類簡帛的虛詞標註作為探究的對象。
4) 第四章是結語。對本文用到的技術、研究過程、研究成果等作了總結。
5) 參考文獻、附錄和後記部分。

關鍵詞：漢簡，語料庫，數據庫，網絡平台，自動分詞，詞性標註

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ABSTRACT

With the development of computer science, the technology of information processing and information storage technology has been developed. It has brought new research methods, research methods and research ideas to the traditional Chinese language and characters. The Han Dynasty bamboo slips as precious materials and carrier of China's history of Han Dynasty, have high research value. Because of its unique advantages such as convenient, fast and strong information processing ability, the digital technology has made the mainstream research institutions begin to work on the digital work of ancient books. By building a corpus of bamboo slips of Han Dynasty, from data collection, automatic annotation, network platform to build several aspects, the gradual completion of the bamboo slips of digital work.

In this paper, under the computer database knowledge, knowledge of network programming, bamboo and wooden slips of knowledge and the automation of Chinese information processing knowledge guidance, build a corpus of bamboo slips and try to explore. During this period the bamboo slips of the database construction, Han Jian automatic word segmentation, part of speech automatic annotation work, Han corpus network platform design etc.. Some of the practical problems encountered in the construction of database, network platform construction is solved, and gives an exploratory attempt in the field of bamboo slips automatic word segmentation and part of speech tagging. Automatic identification technology and the use of corpus, rough statistics of bamboo slips of the Han Dynasty unearthed in northwest of names, in the name, and Zhangjiashan Bamboo Slips of Han Dynasty, Mawangdui Han Dynasty, Wuwei Han Dynasty, and other ancient bamboo slips of function words and content words.

The thesis consists of five parts:

1) introduction. Respectively introduce the background, research significance,

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

construction and application status of bamboo slips of the Han Dynasty corpus, the present situation analysis and improvement goals, research methods and innovations.

2) the second chapter introduces the process design of the corpus of Han dynasty. From two aspects of database and network platform, introduces the function and characteristics of the Han Dynasty and the corpus building steps.

3) the third chapter is a corpus based study of Han Dynasty, the Han Dynasty text automatic word segmentation and POS tagging do. The key of the research in the northwest Han Dynasty name, name and topic, gerunds, words of ancient silk books inquiry class.

4) the fourth chapter is the conclusion. In this paper, the technology used in the research process, research results have been summarized.

5) references,appendix and postscript.

Key words: Bamboo Slips, Corpus,Database, Network Platform, Automatic Word Segmentation, POS Tagging

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目錄

摘要 I

ABSTRACT II

第一章　緒論 1

第一節　選題依據和研究意義 1

一、　選題依據 1

二、　研究意義 3

第二節　相關研究現狀綜述 5

一、　漢簡語料庫建設與應用現狀 5

二、　現狀分析與改善目標 13

第三節　研究方法和創新點 13

一、　研究方法 13

二、　創新點 14

第二章　漢簡語料庫系統設計 15

第一節　漢簡語料庫的數據庫設計與建立 15

一、　漢簡語料庫收錄材料範圍 16

二、　數據庫中字、詞、簡、篇的關聯方法設計 25

三、　自動分詞方法探析 28

四、　各家考釋的數據庫表示方法 30

五、　數據庫打印功能 31

第二節　漢簡語料庫網絡平台 33

一、　平台簡介 33

二、　全文檢索功能 34

三、　考釋檢索 39

四、　詞性檢索功能 41

五、　著錄顯示功能 42

六、　簡牘拓片顯示功能 43

七、　編連綴合考釋信息顯示功能 43

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

第三章　基於漢簡語料庫的詞彙自動化標註研究.....45
第一節　研究材料概說.....46
一、　西北漢簡.....46
二、　古書類漢簡.....48
三、　材料版本選擇.....49
第二節　語料預處理與詞典製作.....50
一、　語料預處理.....50
二、　詞典製作.....53
第三節　自動化分詞與詞性標註.....55
一、　分詞算法介紹.....56
二、　詞性自動化標註方法.....60
三、　實驗過程.....61
第四節　標註結果分析及改進方法.....64
一、　標註結果分析.....64
二、　改進方法.....75
第四章　結語.....76
參考文獻.....78
附錄.....82
附表一　西北漢簡人名表.....82
附表二　西北漢簡姓氏頻次表.....85
附表三　西北漢簡地名表.....87
附表四　西北漢簡職官表.....89
後記.....90

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

第一章　緒論

第一節　選題依據和研究意義

一、　選題依據

中國語言文字歷史悠久，博大精深，記錄承載著中華文明的歷史、文化、政治、科學等豐富的信息。漢簡，即漢代簡牘，則是研究我國漢代歷史的珍貴實物資料和載體，一代又一代的古文字學家們，踏實治學，精益求精，為漢簡研究做出了不可磨滅的貢獻，為後人提供了可靠的研究方法和知識基礎。在過去的短短幾十年間，大量的出土文獻陸續問世，尤其是秦漢時期的古文字文獻的出土，更刺激著我們不斷去研究其中新的問題與發現。如何將這些新的材料與已有的古文字材料融合，如何將文煙浩淼的材料準確的辨識、理解，都是新時代的學者們要面臨的艱巨考驗。

然而，傳統的研究方法有其侷限性。首先，傳統的材料研究方法依賴於個人對材料的熟悉程度和知識積累，是經驗積累型研究，加之研究對象的規模巨大，這就造成了研究時容易遺漏、看錯、記錯等問題時有發生。其次，由於個人時間精力有限，傳統的研究往往集中在單類材料的字形分析、釋文校補或某類專題研究上，人力的局限性導致難以匯聚所有資源，綜合全面的統計分析，因此對漢簡整體面貌分析、詞彙使用統計等規律性問題的研究依然欠缺，導致一些潛在的、深入的資訊無法被統計、挖掘。再次，傳統材料由於檢索困難，人們不得不通過編纂利於治學的工具書，如字書、辭書、目錄等方式來幫助學術研究。這些紙質材料檢索麻煩，檢索耗時長，占空間大，不利於跨地區、跨單位的高效學術交流。此外，對於沒有缺乏經驗積累的文字學初學者來說，由於基本功有限，不善於發散聯想舉一反三，面對單一材料難以發現研究點。最後，古文字學研究經常需要花大量的時間在人工處理材料這件事上面，缺乏自動化分析手段，效率較為低下。

在計算機技術迅速發展的今天，資訊技術已經廣泛運用於各個行業，給學者們提供了新的研究方法和研究思路。時至今日，絕大多數語言文字工作者已經加入了運用電子資料的行列。相比傳統的紙上作業，資訊技術具有容量大、體積小、

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

檢索快、正確率高、善於處理大量數據等優點。因此，充分發揮計算機技術的優勢，利用信息技術擅長處理“大數據”的優勢，通過合理搭建數字平臺，在對漢簡文字信息處理的基礎上，通過科學的數據分析，數據處理，更加深入系統的挖掘出漢簡中蘊含的文化歷史信息，使它能夠為文字研究工作提供便利，是極有價值的。

然而，和其他學科相比，漢語言文字學對信息技術的運用是開始的比較晚，發展水平也相對落後的。首先，現有的出土文獻數字化研究，還停留在數據的收集錄入、關鍵字檢索、拓片顯示等初步階段。對於如何發揮計算機“智能分析”的功能，利用“大數據”統計優勢，從詞彙意義、詞性分析等方面對出土文獻進行研究，還是近乎空白的狀態。其次，現有的秦漢簡語料庫存在重複建設，材料不全或材料互異的問題，始終缺乏一個內容全面，體例統一的語料庫。並且在線語料庫的功能十分簡單，傳播範圍很小，並不能真正起到幫助文字研究者更快更好的查閱到需要的資料的功能，反而因為建設不到位，現實中的一些技術問題，出現檢索內容不全，檢索內容有誤等問題。

因此，在計算機技術高速發展的大數據時代和出土材料日益豐富、文字研究日以精細化這兩大趨勢的交集下，漢字數字化，或者說語料庫的建設，就是最急迫也是最需要加大力度去完善的一項工作。如何突破技術瓶頸，通過技術創新，將兩者更好地融合發展、相輔相成，使得出土文獻更便利、更有效、更全面的為我們所利用，從而為漢字研究工作煥發出新鮮的活力，刺激產生新的研究方法或學術議題，是擺在當前每一位文字研究工作者面前的問題。

本文的選題依據主要有以下幾個方面：

1. 作者在信息技術方面的優勢。

作為一名讀過四年計算機科學技術專業的研究生，來到漢語言文字學專業，由衷希望能夠揚長避短，利用計算機技術的優勢彌補自己其他方面的不足。作者希望能夠利用所學知識來為傳統文字學研究帶來一些新的研究理念、研究方法和研究手段。通過自身對數據庫、網頁編程、自然語言科學、數據挖掘、機器學習等方面的知識，來開啟將信息技術與傳統漢簡研究相結合的大門。希望在此過程中能夠切實解決一些實際的問題，并得到一些有價值的研究結果，即使只能解決

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一小部份問題，對作者來說也是非常大的成長與收穫。

2. 漢簡數字化資源的完善與豐富。

得益於文字中心很早就開展的古籍數字化項目，目前我們已經擁有了完備的漢簡數字資源，其中囊括了絕大部份的出土漢簡，這樣的數據基礎，為下一步研究提供了極大的便利，有了這樣扎實牢靠的數據基礎，就為漢簡內容的深入挖掘及漢簡語料庫的建設提供了可操作性。

3. 漢簡內容本身具有重要的研究價值。

“兩漢時代，簡仍是主要的書寫材料，牘也用的相當普遍。……一直等到南北朝時代，紙才比較徹底的把簡牘排擠掉。在晉代，雖然紙已經用的很普遍，政府的文書簿籍卻往往仍然使用簡牘。”[1]可見簡牘，或者漢簡本身，是研究漢代邊塞制度、歷史、文化、政治、軍事、經濟、法律、民俗、文學、天文、地理、曆法等方方面面的重要的資料。而通過“簡史互證”，漢簡的內容又大大補充了史籍的不足，以此來補史之闕、證史之謬，有著很高的研究價值。同時，漢代簡牘數量龐大，已經出土的漢簡已有百萬字，這其中蘊含的大量“數據”亟待我們去“挖掘”，大量的語料也為計算機語義自動化研究提供了很好的數據支撐。

二、 研究意義

研究意義體現在以下幾個方面：

1. 解決漢簡語料庫建設中的實際問題

目前古籍數字化工作還主要集中在整理、錄入釋文、簡單檢索為主，對於古籍本身內容的分析以及深度挖掘還需要進一步發展。相比現代漢語中對信息技術的應用深度，古代漢語的信息處理之路才剛開始。目前已有的漢簡數據庫有許多功能可以補充完善，還有一些實際的問題需要解決。例如本文後面將要介紹到的：料庫檢索功能、語料庫字形顯示功能、語料庫統計功能、數據庫優化設計與實現、數據庫中字詞簡篇的關聯方法設計、詞彙屬性的自動標註方法、同義詞集合與反義詞集合的數據庫表示方法、各家研究成果的數據庫表示方法、數據庫打印功能等等。雖然只是解決了漢簡語料庫的部份問題，但本文研究過程中遇到的實際技

[1] 裘錫圭.《文字學概要》.北京.商務印書館，1988年

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

術問題、處理方法等，可以作為今後語料庫研究的借鑑和參考。

2. 提高漢簡語料庫的應用價值

漢簡語料庫的建設，或者說古籍數字化，不是目的，而是途徑。只有能為文字工作者提供便利，幫助文字工作者更好、更快的解決問題，才是有價值有意義的。否則就是捨本逐末，顛倒主次。有的文字專家不重視古籍數字化工作，認為其價值不大，這並不奇怪。以漢簡為例，目前的漢簡語料庫集中在釋文搜集，檢索功能上，幾乎無法為研究者提供深入研究的功能。從目前的研究現狀來看，缺乏一個全面綜合的漢簡語料庫系統，既能涵蓋大部分的漢代出土文獻，提供準確的檢索、字形顯示等功能，又能夠深入挖掘材料中的字詞含義，將簡單的文本檢索提升到智能分析詞性詞義的高度。因此，作者希望通過一系列技術手段與方法，從“詞性”、“詞義”、“詞類”等幾個方面推進語料庫的建設與完善，提高語料庫應用價值，使其能夠更好地為文字工作者所使用，也引起更多人對古籍數字化工作的的重視和認可，形成良性循環。

事實證明，漢簡語料庫的建設擁有很大的進步空間，其價值也是值得肯定的。舉例而言，傳統引得的製作方法，是通過卡片的方式，工作量浩大，耗時很長又容易出錯，通過漢簡語料庫，只需點擊按鈕就可以自動打印出全部引得，工作效率提高了不止百倍。又如，傳統的搜集各家考釋的方法，主要通過人力查找，既容易遺漏，又費時費力，而在漢簡語料庫中，各家考釋可以集中反映在相應的簡之後，對於使用者來說方便又快捷。

漢簡語料庫的建設有很多基礎性的工作，工作量大、難度高，需要投入很多人力物力。設計完成的漢簡語料庫開發了網絡平臺，這樣不同的研究者可以在該平臺上進行工作，避免了重複勞動，將時間精力節省到更有價值的研究中去。

在紮實積累，踏實治學的基礎之上，合理利用信息技術的便利性、有效性和全面性，使其為文字研究提供便利和新的研究思路，讓兩者能夠相輔相成互相促進，是作者設計研究該語料庫的最終目的。

3. 基於漢簡語料庫的應用：自動分詞與詞性自動化標註研究

利用設計完成的漢簡語料庫數據全面，數據量龐大，檢索快速、自動化分析等優點，可以進行進一步的材料處理，即自動分詞與詞性標註。通過結合古漢語

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

詞彙特點、標註語料自身特點，設計出合理的試驗方法，并在實驗過程中對詞彙自動標註結果進行分析，得到改進方法。同時補充完善了西北漢簡專有名詞詞表，有利於推動對漢代文化研究的進一步深入。

第二節　相關研究現狀綜述

一、　漢簡語料庫建設與應用現狀

對於出土文獻的電子資料庫或稱語料庫的建設，早在上世紀九十年代就已展開。目前，以漢代簡牘為主體內容的網絡在線語料庫已經存在一些，內容較為完整且可訪問的主要有以下幾個：

華東師範大學中國文字研究與應用中心開發的“秦漢簡牘帛書語料庫”[2]、香港中文大學中國文化研究所中國古籍研究中心開發製作的“漢達文庫”[3]、中央研究院歷史語言研究所文物圖像研究室製作的“簡帛金石資料庫”[4]、中央研究院歷史語言研究所製作的“漢代簡牘數位典藏”[5]、武漢大學簡帛研究中心製作的“中國古代簡帛字形、詞例資料庫”[6]。

其中除作者正在參與建設的“秦漢簡牘帛書語料庫”之外，另外四個語料庫將從收錄材料範圍和功能及使用說明兩方面來介紹各個語料庫。

漢達文庫

(一)　材料範圍

馬王堆、武威、睡虎地、銀雀山、居延等

(二)　功能及使用說明

查閱資料

[2] 网址：http://www.wenzi.info/confs_web/NewWeb/ExpertsLogin.aspx

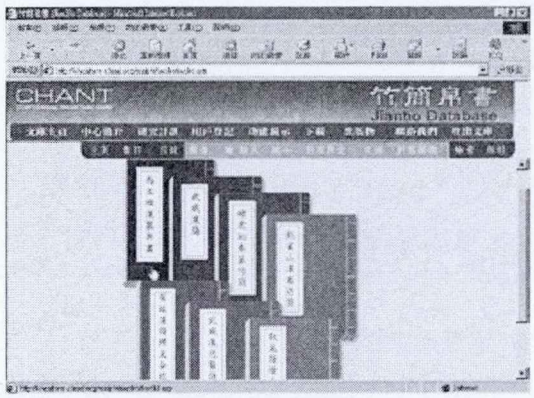
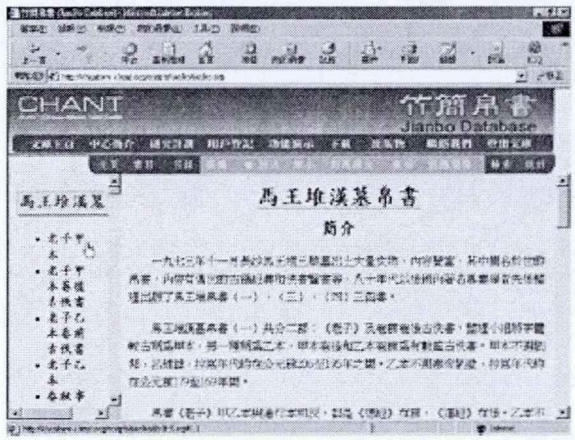
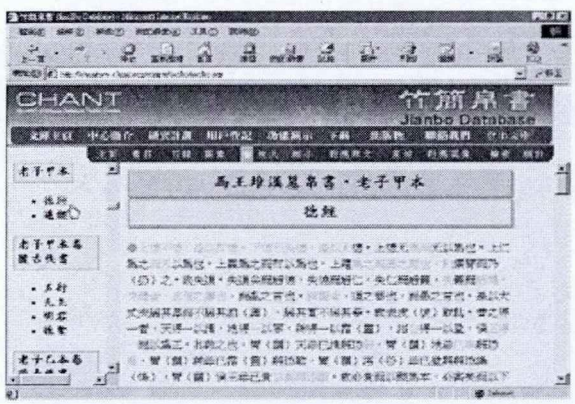
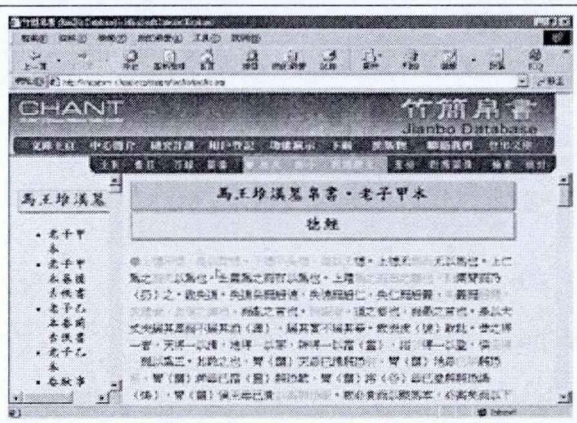
[3] 网址：http://www.chant.org/info/trial_sub.asp

[4] 网址：http://saturn.ihp.sinica.edu.tw/~wenwu/search.htm

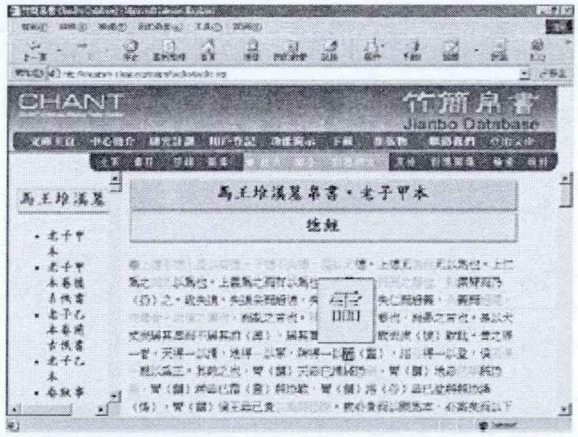
[5] 网址：http://rub.ihp.sinica.edu.tw/~woodslip/index.htm

[6] 网址：http://www.bsm-whu.org/zxcl/index.php

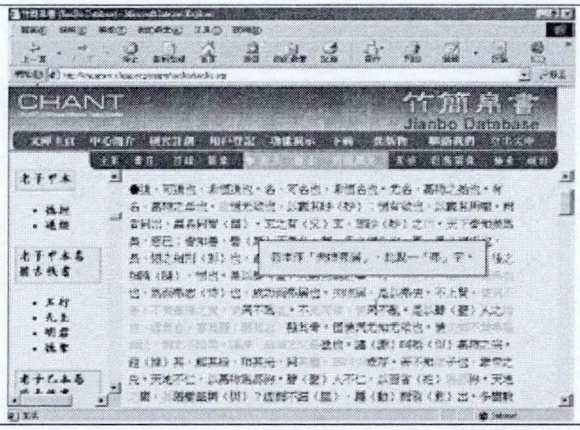
(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

	可選取需要查看的種類，進行查閱。
	選取需要查看的書目。
	選取需要查看的篇目。
	灰色字體表示原資料殘脫，有關字詞據傳世資料增補。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

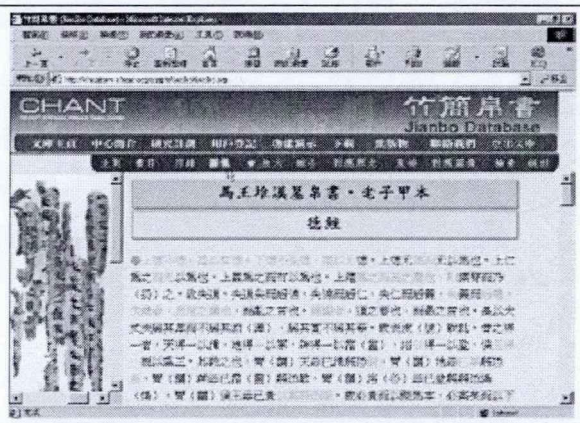


在字體上連續點擊可放大該字。

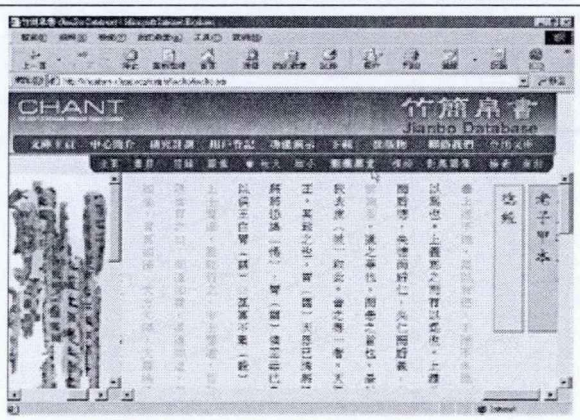


所有校改記錄均以內嵌式視窗顯示，只要把滑鼠座標移至已標示的字詞，有關資料便會自動顯示。

顯示圖像



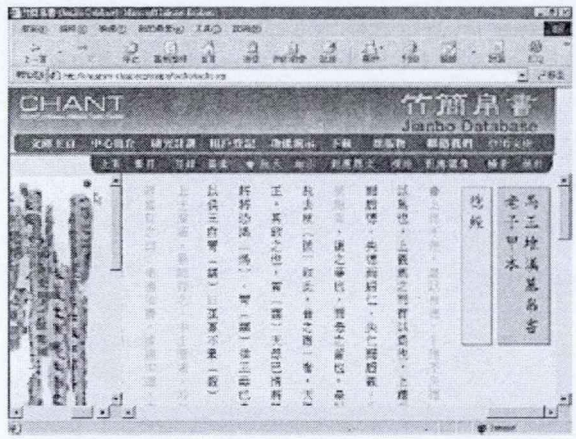
按「圖像」可顯示原竹簡圖像。



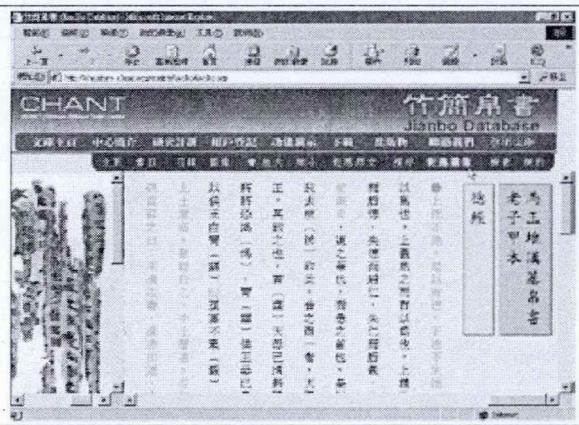
對應原文

按「對應原文」後，在竹簡上端點擊，便可對應該竹簡的原文。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

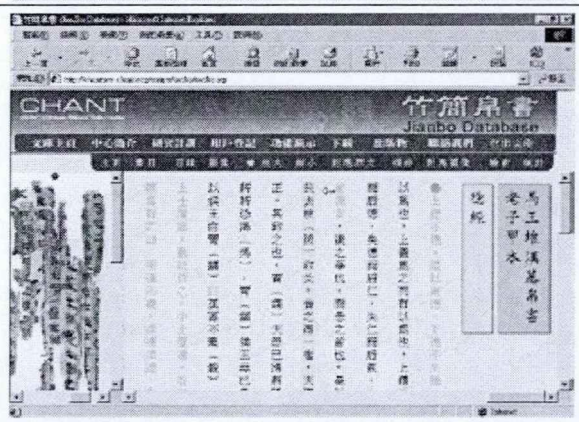


竹簡上端的紅點，表示希望對應的部份，而相關的原文則會以黃色標示。

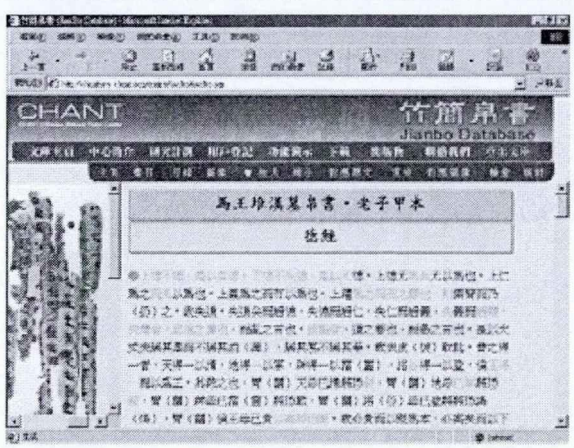


對應圖像

按「對應圖像」後，在原文上端點擊，便可對應該原文的竹簡。

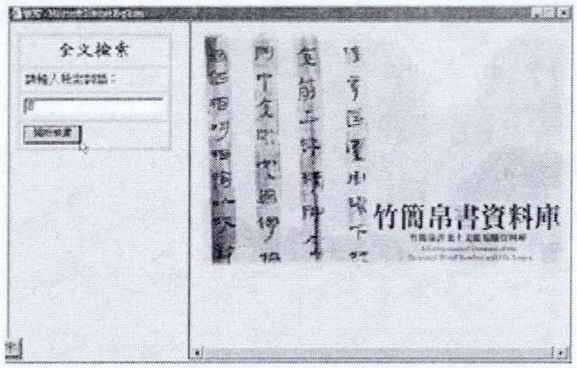
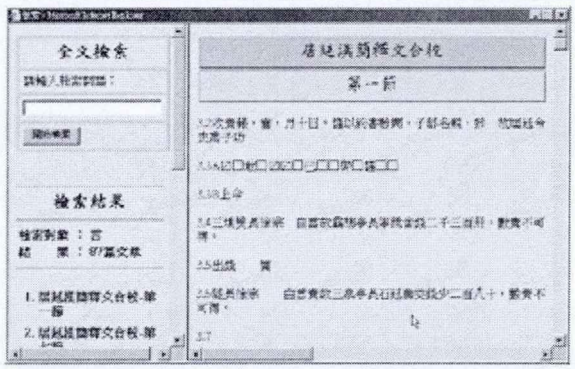
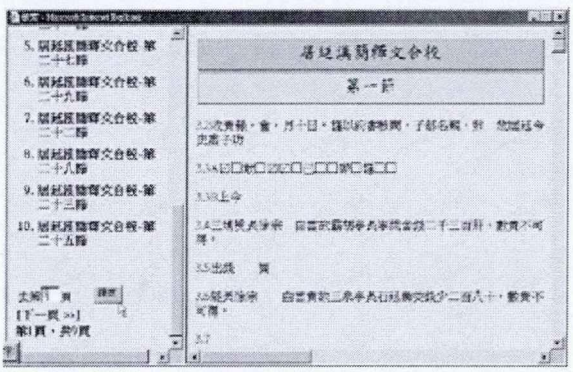


表示希望對應的原文會以黃色標示，竹簡上端的紅點顯示對應的部份。

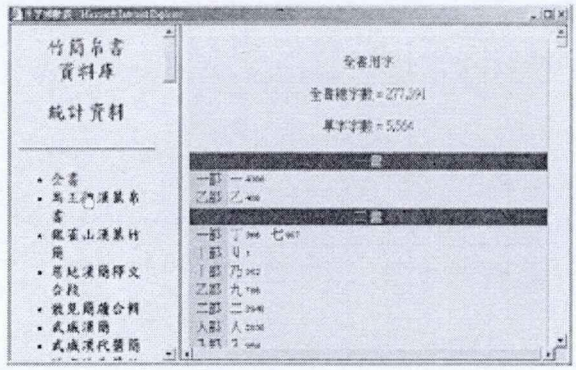


「對應原文」和「對應圖像」均可以在橫排的情況下使用。

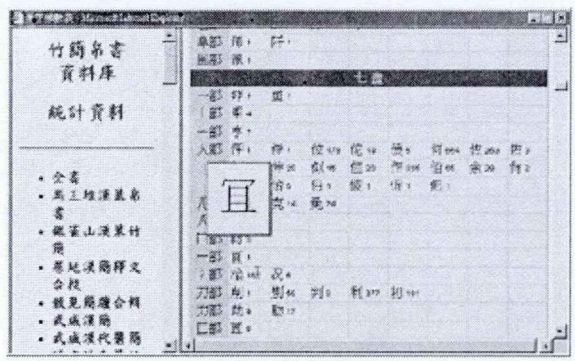
(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

檢索功能	
	在新增視窗輸入所需的字詞，然後按「開始檢索」。
	顯示全文檢索結果，有關的字詞以黃色標示。
	跳至 因檢索結果繁多，可到頁底輸入頁碼，以便跳查頁數。
統計功能	
	按「統計」後，「竹簡帛書資料庫統計資料」顯示整個資料庫的總字數、單字字數及圖片數目的資料。

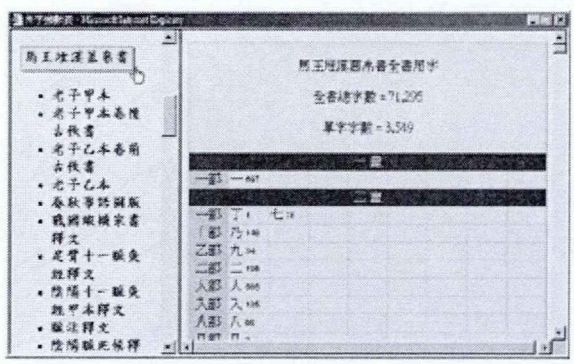
(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



按「全書」可得整個資料庫的用字資料。

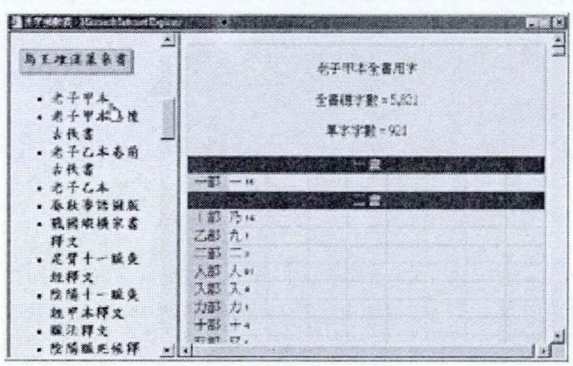


各單字按筆劃、部首排列，把滑鼠座標移至各單字便會自動放大。



個別種類

點擊個別種類便會顯示該書的統計資料。

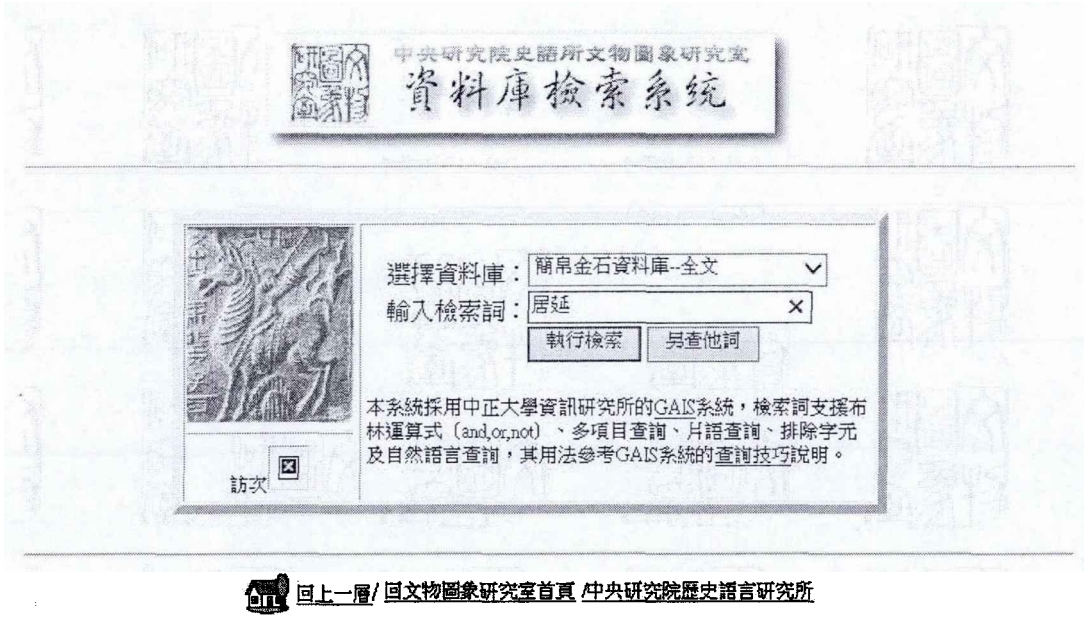


個別書目

點擊個別書目便會顯示該篇的統計資料。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

簡帛金石資料庫



(一) 材料範圍

天水放馬灘秦簡甲種《日書》釋文、睡虎地秦墓竹簡、居延新簡、居延漢簡甲乙篇、疏勒河流域出土漢簡、敦煌漢簡、散見簡牘合輯、馬王堆 、銀雀山漢墓竹簡(壹)孫子兵法等、武威漢代醫簡、 張家山漢簡《引書》釋文 、 脈書

(二) 功能及使用說明

該語料庫提供釋文檢索和論文檢索。

由截圖可見，可通過輸入關鍵詞的方式檢索含有關鍵詞的釋文及論文題目

漢代簡牘數位典藏

(一) 材料範圍

[居延漢簡]為主，約 11000 餘枚，另外還包括西元 1930、1934 年黃文弼在新疆鹽澤發現的「羅布淖爾漢簡」(58 枚)，以及西元 1944、1945 年夏鼐、閻文儒勘查玉門關、陽關及漢代邊防烽燧遺址路線，所發現的「敦煌小方盤城漢簡」(76 枚)、「武威剌麻灣漢簡」(7 枚)。

(二) 功能及使用說明

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.簡易檢索：

簡易檢索適用於有特定檢索目標之使用者，可藉由鍵入簡號、品名(該枚簡牘的名稱)、部份釋文（如檢索簡牘釋文中含有「肩水候官」此一字串的的簡）、遺址代碼（如檢索遺址代碼為「A8」或遺址名稱為「破城子」出土的漢簡）、遺址名稱等查詢資料庫。

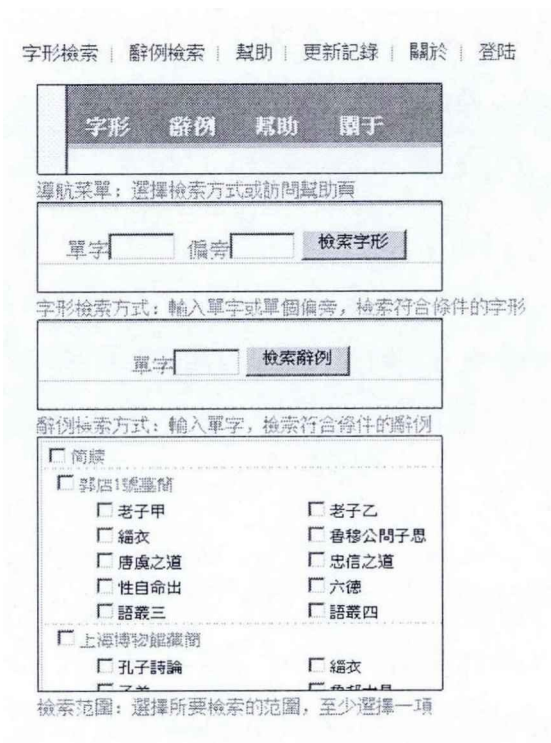
2.進階檢索：

進階檢索適用于對簡牘資料已有基本概念之使用者，透過給定兩項以上的查詢值，進行交叉查詢（此處使用布林邏輯"and"），如查詢出土遺址為「A35 大灣」且釋文含有「肩水都尉」字串的簡。

3.流覽檢索：

流覽檢索適用於沒有特定檢索目標 或對簡牘資料不熟悉者可就某一類別或範圍資料進行流覽。該資料庫提供簡號、品名、年代、遺址、形制等資料流覽。

中國古代簡帛字形、詞例資料庫



(一) 材料範圍

裡耶秦簡、張家山漢簡（曆譜、二年律令、奏讞書、脈書、算數書、蓋廬、引書、遣策）

(二) 功能及使用說明

如截圖所示。可檢索字形及辭例。

字形：輸入單字或偏旁進行字形檢索。

辭例：輸入單字進行辭例檢索。

檢索範圍：需要檢索的材料範圍。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、 現狀分析與改善目標

依上所見，目前已有的漢簡語料庫具有以下特點：

1. 已有語料庫以顯示釋文、圖片為主體。其中漢達文庫可做到釋文原文與原簡互相對應，但還停留在整條簡的對應層面，無法做到單個字與字形的對應。

2. 已有語料庫檢索類型較為豐富。可檢索的內容包括著錄、文例、字形、參考文獻、考釋資訊、遺址、簡牘資訊等。

3. 各個語料庫收錄材料的範圍不同，其圖像、釋文也往往互異，需要互相參照相輔相成。

4. 缺少深層次的內容分析，導致缺乏例如分詞、詞性、同義詞反義詞檢索等功能。

改善目標：

1. 提高漢簡語料庫覆蓋材料的全面性。建設好的語料庫應該覆蓋了所有已經發表的出土漢簡內容，并不斷保持更新。

2. 豐富語料庫檢索功能。不侷限於關鍵詞與釋文匹配的簡單查詢方式，還應能夠從內容和語義方面進行檢索。如本文研究的分詞與詞性標註，就可以實現更深層次的檢索功能。例如可檢索人名、地名、動詞、虛詞、職官、數量詞等等。

3. 豐富語料庫統計功能。通過對語料庫數據的分析和詳細標註，應該能夠提供更多維的統計功能。例如根據前後搭配統計詞彙的使用情況等。

4. 增加單個字與字形圖片對應顯示的功能，方便查看字形。

第三節 研究方法和創新點

一、 研究方法

本文的研究方法主要分為以下幾個步驟：

1. 設計數據庫字段，使得数据库结构中字、词、簡、篇能夠合理的關聯起來。

2. 漢簡釋文及相關信息錄入，即基礎材料準備部份。漢簡的數字資源是漢簡語料庫的構成基礎，也是最為耗時耗力的工作。這一步由課題組成員共同努力完成，已經錄入了 2014 年以前發表的大部分漢代簡牘材料，該語料庫收錄的材

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

料還將不斷更新擴展。

3. 進一步完善數據庫功能。解決各家考釋表示問題、數據庫打印功能問題等等。主要的方法是數據庫編程。

4. 漢簡語料庫網絡平臺功能豐富。增加字形圖片與釋文對應顯示功能、統計功能等。主要用到的方法是 ASP 網頁編程。

5. 對語料進行分詞以及詞性標註。分詞主要用到的方法是基於詞典、基於規則和基於統計的分詞。詞性標註用到的方法是基於詞典和利用模型自動化分析。

6. 結果分析。對於分詞以及詞性標註的結果進行結果分析，通過分析結果得到進一步改進的方法。

7. 根據分析結果，進行西北漢簡人名、地名、職官等專有名詞統計。主要運用的方法是數據庫操作以及總結分析。

二、 創新點

作為交叉學科選題，本文的創新點有以下幾個方面：

1. 利用自身學科背景優勢，將漢語言文字研究與計算機技術相結合。通過深度開發數據庫功能與網絡平臺功能，構建了功能豐富、內容全面的漢代簡牘文獻的數字化平台。

2. 解決了數據庫使用過程中的一些實際應用問題。例如不同層次簡牘信息如何關聯的問題、各家考釋的不重複表示問題等。合理的數據庫設計使得漢簡語料庫能夠從單純的數據存儲進入到對簡牘內容、語義等方面深層次研究的領域。

3. 搭建了漢簡語料庫網絡平臺。該網絡平台增加字形圖片與釋文相對應顯示功能，以及增加不同維度統計功能。在同類型網絡平臺中具有更好的實用性，通過開放性平台實現了研究成果共享，方便其他研究學者使用。

4. 利用漢簡語料庫進行詞彙自動化標註探究，豐富了利用大規模語料和計算機技術，對漢簡自動化分詞及詞性標註等方面研究的經驗，并統計出了人名、地名、職官等專有名詞詞表。大大提高了材料加工的效率，并為相關研究領域提供了更全面、更準確的研究資料。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二章 漢簡語料庫系統設計

本文研究的漢簡語料庫系統，是基於中國文字研究與應用中心張再興教授領導開發的“秦漢簡牘帛書語料庫”的基礎之上，繼續優化開發得到的。主要包含兩個模塊：漢簡語料庫的數據庫系統和漢簡語料庫網絡平台。其中數據庫系統是整個語料庫的數據基礎，語料庫的所有文本信息、詞彙標註信息、詞性詞義標註信息、圖片信息等都來源於數據庫。而漢簡語料庫網絡平台提供一個在線網站形式的檢索平台，能夠實現各種檢索功能。下面就將從這兩部分出發分別介紹漢簡語料庫系統的設計思路與實現功能。

第一節 漢簡語料庫的數據庫設計與建立

竹簡是木製載體，在被發掘出土之後，隨著時間推移，其本身容易出現氧化、斷裂、霉變等諸多問題，對簡牘內容造成不可挽回的損毀。為了長久保存簡牘記載的珍貴內容，伴隨著每一批新的簡牘出土，往往緊接著的就是拍照、整理、校對等工作。通過學者的努力，一批批新出土的或完好或殘缺的簡牘被編序、翻譯釋文、增補校對后進行印刷發表，以供廣大文字專家進行後續的研究。新世紀計算機科技的進步，使得簡牘內容數字化成為了最新的發展趨勢。簡牘數字化不僅有利於數據的長時間保存，更大的優勢在於，和傳統的大部頭印刷書籍相比，其特有的便攜性、共享性和數據分析的靈活性。

簡牘的數字化有兩種主要的處理方式。一種是將竹簡掃描或拍照，取得數字圖像，經過計算機的調整，比如增加圖像的對比度，去除圖像中的噪聲等方式，呈現出較為清晰的簡牘原始圖版，俗稱“切條”。此外可以在此基礎上，將簡牘上的內容以字為單位進行“切字”，得到單個字的圖像。這種處理方式能夠最大程度保存簡牘的原貌，用最直觀的圖像還原簡牘的色彩、圖案、文字內容、排版、殘缺情況等。另外一種是只提取簡牘中的文本信息，即“釋文”。由於技術水平的限制，目前的計算機文字識別技術還不成熟，加上簡牘文字的手寫特點、大量異體字通假字、以及內容殘缺損毀等客觀因素，即便是主流的 OCR（Optical

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Character Recognition，光學字符識別）技術的正確率也非常之低。因此釋文的錄入目前還是依靠人工錄入進行。

對於漢簡數字化後的文本存儲，最好的工具就是數據庫。相比其他的文字處理軟件，例如 Microsoft Word 或者 Microsoft Excel，數據庫具有更加靈活，更加方便，功能豐富等優點。因此漢簡語料庫的建設採用數據庫作為基礎支撐，在此基礎上開發了漢簡語料庫的網絡平台。本文研究的漢簡語料庫用到的數據庫工具是 Microsoft Access 2010。經過長時間的建設，目前在課題組全體成員的努力下，已經完成了對大部分 2014 年之前已經發表的漢代簡牘的釋文錄入工作，由於數據量龐大，且不斷有新的材料被整理公佈，因此這是一項需要長期堅持的工作量很大的任務。在今後該語料庫還會不斷更新維護，將及時納入新發表的研究成果。經過不斷設計和建設，該數據庫已經成為了一個集簡牘釋文、簡牘圖版、簡牘字形、各家考釋信息、著錄信息、打印功能為一體的綜合數據庫系統。

一、 漢簡語料庫收錄材料範圍

(一)漢簡語料庫所收錄漢簡

漢簡，即漢代簡牘，是兩漢時期人們主要的書寫材料。“從清末（二十世紀初年）到解放後，在西北地區兩漢和魏晉時代的邊塞遺址里，陸續發現了很多漢簡和一些魏晉簡。解放後，還在湖南、湖北、安徽、山東、河北、甘肅、青海等地的漢墓里發現了很多簡冊。”[7] 20 世紀以來的簡牘出土，帶動了簡牘學的興起，掀起了研究簡牘的學術潮流。

本文介紹的漢簡語料庫，收錄了大部份在 2014 年之前已經發表的各個時期出土漢簡，基礎材料十分全面，這也是本語料庫的一大突出特點。下面各表從收錄漢簡名稱、出土時間、出土數量、已發表數量、總字數、簡介和發表情況幾方面，簡要介紹了本語料庫收錄的各個時期的漢簡情況。表格中有一些尚未統計出的數據暫時為空。

[7]裘錫圭.《文字學概要》.北京.商務印書館，1988 年

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格 1 漢簡語料庫簡況表

（1）兩漢時期（九種）

文献名称	出土时间	出土数量	已发表数量	总字数	简介	发表情况
漢代散簡		200	200			《散見簡牘合輯》
敦煌漢簡		2878	2878	42660	敦煌漢簡是二十世紀初以來在河西疏勒河流域漢塞烽燧遺址中陸續出土的竹、木簡牘。本數據庫以甘肅省文物考古研究所編《敦煌漢簡》（中華書局 1991 年 6 月第 1 版）釋文爲底本，共編號 2878。	甘肅省文物考古研究所編《敦煌漢簡》（中華書局 1991 年 6 月第 1 版）
居延漢簡		20000	20000	100000		居延漢簡甲乙編
居延新簡		8400	8400	104765	居延甲渠候官破城子、甲渠塞第四隧，1972-1974 居延漢簡的時代跨度較大，早的在西漢武帝時代，晚的已屬東漢	
悬泉漢簡	1990-1992	23000	272		悬泉汉简於 1990—1992 年间在敦煌悬泉置汉晋遗址出土。全部出土简牍 23000 多枚，其中已整理编号者 17900 多枚，内容包括诏书、律令、科品、檄记、爰书、薄籍、符传、历谱、术数、字书、医方以及一些古籍残篇，完整或基本完整的册子 40 多个。	《文物》2000 年第 5 期刊发了悬泉置的发掘简报和汉简的内容概述。另外，《敦煌悬泉汉简释文选》一文还发表了部分简的照片和释文。2001 年上海古籍出版社出版了胡平生、张德芳编纂的《敦煌悬泉汉简释粹》，共發表 357 件，包括《敦煌悬泉月令诏条》。中華書局也在 2001 年出版了《敦煌悬泉月令诏条》，该书释文、注释部分也由胡平生执笔。
额济纳漢簡	2005	500	500	8000	2005 年，广西师范大学出版社出版了魏坚主编的《额济纳汉简》。收录 1999、2000、2002 年内蒙古额济纳地区出土的	本数据库释文依据孙家洲《额济纳汉简释文校本》（文物出版社 2007 年）。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

					汉简 500 余枚。	
肩水金关汉簡		2351	2351	26229	甘肃金塔县肩水金关。与居延新简共 19000 余	
武都漢簡	2000	12	12		2000 年陕西历史博物馆征集，12 枚。据说是 2000 年 5 月出土於甘肃省武都琵琶乡赵坪村。均木质，两面书写。	《文物》200 3 年 4 期发表王子今、申秦雁《陕西历史博物馆藏武都汉简》，对简文进行了考释。
香港中文大学藏漢簡		200	200	0	漢代簡牘包括日書（109 枚，漢初惠帝以後）、遣策（11 枚）、奴婢廩食粟出入簿（69 枚，西漢中期）、河隄簡（26 枚）、序寧簡（14 枚，東漢章帝）。	《香港中文大學文物館藏簡牘》，香港中文大學文物館 2001 年

（2）西漢早期簡（八種）

文献名称	出土时间	出土数量	已发表数量	总字数	简介	发表情况
孔家坡漢簡	2000	500	500		2000 年出土於土於湖北省隨州市城關東北孔家坡 M8。内容包括日書和曆日，日書共 478 枚，另有殘片 48 片。曆日簡 60 枚。木牘四方，只有一方墨書，爲告地書。景帝后元 2 年（公元前 142 年）	湖北省文物考古研究所、隨州市考古隊《隨州孔家坡漢墓簡牘》，文物出版社 2006 年 6 月。
馬王堆漢墓簡帛	1972-1973	1500	1500	100000	一號墓，1972 年發掘，出土竹簡 312 枚，2 063 字，内容爲遣策。汉文帝 15 年（公元前 165 年）三號墓， 1973 年發掘，出土簡牘包括：遣策：牘 6 件；簡 402 支（含三支残简）。医书： 200 支。甲卷包括十问、合阴阳方；乙卷包括杂禁方、天下至道谈。還有大量帛書，約 11.5 万字。	长沙马王堆一号汉墓（上下册，文物出版社 1973 年）發表遣册釋文和圖版 长沙马王堆二三号汉墓 第一卷 田野考古发掘报告（文物出版社 2004 年）發表釋文和圖版
張家山漢	1983	1200	1200	36048	张家山二四七号汉墓位	张家山二四七号汉墓竹简整

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

簡					于湖北省江陵县（今荆州市荆州區）城外西南一点五公里处的江陵磚瓦厂内，因取土而发现，一九八三年十二月由荆州博物馆配合准行发掘。它是一座土坑木椁墓，墓坑上部已遭破壞。	理小组《张家山汉墓竹简（247号墓）》，文物出版社2006年。
銀雀山漢簡	1972	7500	2211	37085	1972年發掘。一號墓出土竹簡4942？7500？，二號墓出土32枚。內容包括孙子兵法、孙膑兵法、尉繚子、晏子、六韜、守法守令等十三篇 等。時代爲文景至武帝初。	银雀山汉墓竹简（壹）文物出版社1985-09 银雀山汉墓竹简整理小组《银雀山汉墓竹简（贰）》文物出版社2010-01
阜陽漢簡	1977	6000	1000	5000	1977年，安徽省阜阳市博物馆在阜阳双古堆1号汉墓中发掘出大批竹简，墓主为第二带汝阴侯夏侯竈。因此，漢簡大抵為漢初遺物。《詩經》殘存170餘條，有國風、小雅兩種。《周易》752片，計3119字。其中經文1110字，卜辭2009字。《阜阳汉简》为西汉初期简册。1977年在阜阳县双古堆1号汉墓中发掘出土。阜阳汉简经整理编纂分为10多种古籍，因原有书名标题无一残存，今名或用传本原名，或据内容另拟。其中有《苍颉篇》、《诗经》、《周易》、《万物》、《吕氏春秋》等。	（1）《文物》1983年第2期发表了《阜阳汉简简介》，对出土竹简的内容作也全面介绍。（2）同期发表《阜阳汉简释文》及《仓颉篇及其研究》。（3）《文物》1988年4期发表《阜阳汉简释文》及《略说》。（4）《文物》1984年8期发表《阜阳汉简释文》及《阜阳汉简简论》。1988年上海古籍出版社出版了胡平生、韩自强的《阜阳汉简研究》，有圖版。（5）1998年，《出土文献研究》第四辑发表胡平生的《阜阳双古堆汉简数术书简论》，介绍和考释了《日书》等。（6）2004年上海古籍出版社出版社韩自强的《阜阳汉简研究》，有圖版，附《儒家者言》、《春秋事語》。安徽省文物工作队、阜阳地区博物馆、阜阳县文化馆：《阜阳双古堆汝阴侯墓发掘简报》，刊于《文物》1978年8期。文章介绍了墓葬发掘及出土简牍概况。文物局古文献研究室、

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

						安徽省阜阳地区博物馆阜阳汉简整理组：《阜阳汉简简介》，刊于《文物》 1983 年 2 期
萧家草场汉简	1992	35	35	164	1992 年湖北省荆州市沙市區關沮鄉岳橋村太湖港東岸蕭家草場 26 號墓出土竹簡 35 枚。內容為遣冊。	圖版著錄於《關沮秦漢墓簡牘》，中華書局 2001 年出版。
定县汉簡	1973	900		10000	1973 年河北省定县 40 号墓（西漢中山懷王刘脩墓）出土。竹簡出土時已經散亂殘斷，碳化後的簡文墨字已多不清晰。竹簡內容多爲先秦文獻。計有：（1）《論語》 620 多枚，多殘簡。共 7576 字。（2）《儒家者言》，存 104 枚，出土時已殘斷。（3）《哀公問五義》。（4）《保傅傳》。（5）《六韜》出土時已嚴重殘損，現存 144 枚，約 1402 字。（6）《文子》。（7）《六安王朝五鳳二年正月起居記》。（8）《日書》、占卜等殘簡。	（1）《河北定縣 40 漢墓發掘簡報》，《文物》 198 1 年 8 期，發表摹本 96 個編號。（2）《定縣 40 號漢墓出土竹簡簡介》，《文物》 1981 年 8 期，刊布了《儒家者言》釋文。（3）《定州西漢中山懷王墓竹簡釋文》、《定州西漢中山懷王墓竹簡校勘記》、《》，《文物》1995 年 12 期。（4）李學勤《試論八角廊簡》，《文物》1996 年 1 期。（5）《文物》1997 年 5 期發表《論語》釋文選、校注等三文。（6） 1997 年文物出版社出版定州漢墓竹簡《論語》一書。（7） 2001 年《文物》 5 期發表了《六韜》釋文及校注。
江陵凤凰山漢簡	1973-1976	600	600		1973-1976 年，湖北省春秋战国时期楚都纪南城遗址的考古勘探调查中，从其东南隅的凤凰山（今隶湖北省荆州市荆州区纪南乡）八、九、十、一六七、一六八、一六九等六座西汉墓葬发现了一批竹简。本数据库释文依据湖北省文物考古研究所编《江陵凤凰山西汉简牍》（中华书局 2012 年）。各篇释文作者：八号墓：彭浩；九号墓：李家浩；十号墓：裘锡圭；	（1）《江陵凤凰山八号汉墓竹简试释》，《文物》 1976 年 6 期。（2）《江陵凤凰山十号汉墓出土简牍考释》，《文物》 1974 年 7 期。（ 3）《江陵凤凰山一六七号汉墓遣策考释》，《文物》 1976 年 10 期。（4）《江陵凤凰山一六八号汉墓》，《考古学报》 1993 年 4 期。（5）湖北省文物考古研究所编《江陵凤凰山西汉简牍》，中华书局 2012 年，收八、九、十、一六七、一六八、一六九號墓簡牘。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

					一六七号墓：姚孝遂、陈雍等；一六八号墓：陈振裕；一六九号墓：李天虹。	

（3）西漢中晚期簡（五種）

文献名称	出土时间	出土数量	已发表数量	总字数	简介	发表情况
北京大学藏西漢竹书		221	221	5200	現存完整竹簡 176 枚，殘斷竹簡 105 枚，共有完整及接近完整的竹簡 211 枚，殘簡 10 枚，另有兩枚完整竹簡遺失。正文現存 5200 字，另有重文 110 字，計字尾題 13 字，簡背篇題 8 字。	已出版《北京大學藏西漢竹書 貳》，内容爲《老子》。
尹湾漢簡	1993	157	157	40000	1993 年出土於江蘇省連雲港市東海選溫泉鎮尹灣村。 M2 出土木牘一方，其餘 23 方木牘、133 枝竹簡均出土於 M6。M6 不晚於西漢晚期成帝末年。M2 屬新莽時期。内容包括（一）簿籍類：集簿、吏員簿、長吏名籍、兵車器集簿、贈錢名籍。（二）數術類：神龜占、六甲占雨、博局占、刑德行事、行道吉凶。（三）曆譜。（四）衣物疏、名謁、日記。（五）神烏傅（賦）。《尹湾汉墓简牍》收录了尹湾汉墓出土的全部木牍和竹简。	連雲港市博物館、東海縣博物館、中國社會科學院簡帛研究中心、中國文物研究所編《尹灣漢墓簡牘》，中華書局 1997 年 9 月。
扬州仪征胥浦汉墓简牍		19	19		江蘇儀徵胥浦 101 號西漢墓，出土"先令券書"竹簡 16 枚，媵贈木牘、木觚各一方，木牘衣物券 1 塊。"先令券書"是墓主朱姣臨終前所立的遺囑文書。是我國考古發現中最早的遺囑。	扬州市博物馆：《江苏仪征胥浦 101 号西汉墓》（《文物》1987 年 1 期） 陈平等：《仪征胥浦 101 号西汉墓初考》（《文物 1987 年 1 期》
武威漢簡	1959	469	469	27332	1959 年 7 月，甘肅省武	《武威漢簡》，文物出版社

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

					威县磨咀子六號汉墓出土。此墓屬於王莽時期。墓主經師的活動時代爲宣帝以後的西漢晚期。內容主要是比较完整的九篇仪礼，一篇写於竹简，9篇写於木札。共 469 簡，273 32字。另有 7 枚日忌木簡、雜占木簡 4 枚。《武威漢簡》一書另收有 18 號墓出土的王杖十簡及 4 號、15 號、22 號、23 號墓出土的柩銘四條。一九五七年七月至一九五九年十一月间，甘肃省博物馆在武威县先后清理了三十七座汉代古墓，在磨咀子六号墓发现了四百六十九仪礼木简以及七枚日忌木简。	1964年出版。甘肃省博物館、中國科學院考古研究所編《武威漢簡》，中華書局 2005年9月。甘肃省博物馆：《甘肃武威磨嘴子6号汉墓》，《考古》1960—5期。甘肃省博物馆：《武威汉简在学术上的贡献》，《考古》1960—8期：《甘肃武威磨嘴子汉墓发掘》，中国科学院考古研究所编辑室：《武威磨嘴子出土王杖十简释文》，《考古》1960—9期。武威"王杖诏书令"首见于武威县博物馆：《武威新出土王杖诏令册》，刊于甘肃省文物工作队、甘肃省博物馆合编的《古代简牍研究》文物出版社，2003年 04月第 1版 .
上孫家寨漢簡	1978	240	240		1978年出土於青海省大通县上孙家寨村西北 115 汉墓。出土木简计 240 枚。简长 25 厘米，宽 1 厘米，厚 0.2 厘米。文字为墨书隶体。出土时尚有残简 400 篇。简文可分作（1）兵法类（2）军法、法令、军爵类（3）篇题目录。	《大通上孙家寨汉简释文》，《文物》1981 年 2 期。《上孙家寨木简初探》，《文物》1981 年 2 期。李零《青海大通上孙家寨汉简性质小议》，《考古》 1983 年 6 期。陈公柔、徐元邦、曹延尊、格桑木《青海大通馬良墓出土漢簡的整理與研究》，《考古學集刊》第 5 期(1987 年)。青海省文物考古研究所《上孙家寨汉晋墓》，文物出版社 1993 年，公布了全部照片和释文。

（4）西漢簡（三種）

文献名称	出土时间	出土数量	已发表数量	总字数	简介	发表情况
纪庄木牍	2004	34	2500		2004 年 11 月，安徽省天长市考古工作者在该市安乐镇纪庄（汉晋东阳古	[1]天长市文管所、天长市博物馆《安徽天长西汉墓发掘简报》，《文物》2006， 11

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

					城遗址附近）发掘了一处西汉墓葬（19号墓），出土文物119件，其中有木牍34件(编为1—34号)，削2件，砚盒1件。木牍尺寸长22.2—23.2cm、宽3.6—6.9cm，共书有2500字左右，整理者公布了十件（14面）木牍及文字内容，内容有户口簿、算簿、书信、名谒、药方、礼单等。根据牍文及其他出土文物得知，19号墓主人为"谢孟"，是当时东阳县一位官吏。[1]除1号牍为官文外，其余木牍均为私人文书（包括信牍、名谒、药方、礼单等）。1号牍正背两面分别题有"户口簿"和"算簿"，表明其为上计文书，属汉代"户口簿"实物形式及汉代县级上计文本的首次发现。[2]对墓葬时间有两种观点：西汉早期、西汉中晚期，根据牍文、墨迹风格及其他情况看在武帝中期至宣帝之间的可能性较大。[3]木牍字体为隶书，少量字迹不太清晰，分正、草两种字体，为多人所书，从字迹看，书写者都有较好的书写功底且书风各不相同。1、14号牍写得最工整；2、5、12、13、15、18、20号牍于当时的正体隶书中加入"行意"，显得灵活随意、不拘格套；10号牍写得更草率些。	（pdf文件976頁）發表了彩版圖及釋文。[2]卜宪群、蔡万进《天长纪庄木牍及其价值》,《光明日报》2007，6，29 天长纪庄木牍《算簿》与汉代算赋问题
光化五座坟西					光化五座坟西汉墓简牍	湖北省博物馆:《光化五座坟西汉墓》,《考古学报》 1976

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

汉墓简牍						年 2 期
未央宫簡	1980	115	115	515	前殿 A 區建築遺址，時代上限為西漢初年，下線不會晚於西漢晚期。	《漢長安城未央宮》238-248 頁有摹本和釋文。

（5）東漢簡（三種）

文献名称	出土时间	出土数量	已发表数量	总字数	简介	发表情况
武威漢代醫簡	1972	92	92		1972 年 11 月，甘肃武威旱滩坡发掘到东汉早期墓葬。内有木制医简 92 枚。体例为先记述病症，然后列出治疗该病的处方。除药物疗法外，还涉及针灸疗法。武威医简对于研究中国传统医学及医家哲学思想，具有一定价值	甘肃省博物馆、武威县文化馆《武威旱滩坡汉墓发掘简报——出土大批医药简牍》，刊于《文物》1973 年 12 期。1975 年 10 月，文物出版社出版甘肃省博物馆、武威县文化馆合编《武威汉代医简》。
長沙東牌樓漢簡	2004	200	200		2004 在长沙东牌楼古井中发现了 426 枚东汉时期的简牍。其中有 216 余枚写有文字，其内容大致是官府文书、邮驿签牌和私人信件，第一次发现了书写函件的大型封泥匣。	長沙市文物考古研究所、中國文物研究所《長沙東牌樓東漢簡牘》，文物出版社 2006 年 4 月。
张家界漢簡	1987	90			这批竹简发现于 1987 年，出自张家界（原大庸市）古人堤东汉时期水塘中，共计 90 片，其中有东汉永元、永初等年号。一般长 22.2-23 厘米，宽 2.6-3 厘米，厚 0.3 厘米，文字的书写可分为单面书写、双面书写、双面倒错书写等形式。根据简牍内容可分 6 类：汉律、医方、官府文书、书信及礼物揭、历日表、九九乘法表。	（1）张春龙、胡平生、李均明《湖南张家界古人堤遗址与出土简牍概述》，《中国历史文物》2003 年 2 期。（2）湖南省文物考古研究所《湖南张家界古人堤简释文与简注》，《中国历史文物》2003 年 2 期。（3）湖南省文物考古研究所、中国文物研究所《湖南张家界古人堤遗址与出土简牍概述》，2001 年 8 月"长沙三国吴简暨百年来简帛发现与研究国际学术研讨会"。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、 數據庫中字、詞、簡、篇的關聯方法設計

數據庫系統是語料庫建設的基礎和核心。是發揮計算機在存儲容量、計算能力等方面優勢的集中體現，如何設計出科學合理，適合古文字研究內在邏輯的數據庫，對語料庫的建設至關重要。好的數據庫設計，不僅僅局限於將簡牘內容堆積錄入，還應該通過合理的數據表設計、數據關聯、有效標註，使得數據內涵得到充分的表示，從而更好地從整體上考察漢代簡牘的文獻內容、文字特點、語言特點、歷史面貌等等。

1. 字、詞、簡、篇的邏輯關係

從實際情況來看，出土文獻的基本單位是“簡”。由數條簡經過編連構成的一段完整的內容，成為“篇”。例如敦煌漢簡中的《倉頡篇》、《急就篇》等。簡牘上書寫的單個“字”，又稱為“字頭”。由字組成了“單音詞”或“多音節詞”等。因此，總的來說字、詞、簡、篇是依次從屬的關係，在設計數據庫的時候，要依據他們的層次關係，進行合理的數據結構設計，將各自的層次關係表述清楚。從實際研究的需要來看，數據庫應當解決以下幾個問題：

(1) “字”。字形的釋讀應分層次表示。由於實際研究的時候，對字形的釋讀包含了隸定、考釋、破讀等三個層次。數據庫應當能夠反映這三個層次的信息，對於不同寬嚴標準和各家考釋造成的隸定、考釋、破讀方法不一樣的各家意見分歧，也應當在數據庫中能夠體現。此外為了便於研究，文字拓片的原始圖版是必不可少的。因此數據庫應該能夠將釋文與文字原型拓片相關聯。對於古文字手寫特征，現代輸入法無法準確表達字形的問題，需要對古文字字形進行嚴格意義上的隸古定，并重新造字。為了檢索方便，還需要能夠將文字進行偏旁部首分析后進行拆分，從而方便用戶通過拆分檢索找到需要檢索的字形。

(2) “詞”。為了方便詞彙研究的需要，對出土文獻進行詞彙方面的標註十分必要，這也是本語料庫的一大特色。在進行詞彙標註之後，就可以從內容、搭配、語義等更深層次的方面對數據進行統計，體現出計算機技術對大數據信息處理的優勢。上古漢語以單音詞為主，因此對於單音詞，直接標註其詞義（義項）、詞性等信息。不可忽視的是，上古時期的多音節詞已經大量出現，數據庫應當能夠解決多音節詞的標註問題。同時，在詞義、詞性的標註基礎之上，可以進一步將

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同義詞、反義詞、類義詞等進行匯集。這個過程如果通過人工標註將會是巨大的工作量，本文重點研究的就是如何通過機器學習的方法，自動識別、標註詞彙，減少人工勞動量。當然，目前的自動識別準確率不高，還需要後期的人工校對，但是能夠在一定程度上減輕人工操作難度。

(3) “簡”。數據庫應當能夠將簡號與簡牘釋文相對應、將簡牘整體與其中的單個字相對應、簡牘與所在篇章相對應。還應當能夠展示簡牘原版圖像。

(4) “篇”。數據庫中應當能將篇章名和其包含的簡號、釋文、字形相關聯。以上的字、詞、簡、篇類型都應當標示出其所在的文獻種類，即其所屬的是哪一種簡。

2. 數據庫中字、詞、簡、篇的關聯方法設計

基於以上的內在邏輯，為適應語料庫實際使用的需求，數據庫設計包含了以下四張較為核心的數據表。

表格 2 字頭表

dbo.qczitou	
列	
id (PK, int, not null)	ID
sid (int, not null)	字型大小
zihao (nvarchar(255), null)	文獻種類號
wxzlhao (nvarchar(255), null)	資料庫簡號
sjkujhao (nvarchar(350), null)	簡內字序
jlzixu (nvarchar(255), null)	句子序號
juzixuhao (nvarchar(255), null)	行號
hanghao (nvarchar(255), null)	釋文字頭
swzitou (nvarchar(255), null)	統計字頭
tjzitou (nvarchar(255), null)	隸古定
lgd (nvarchar(650), null)	讀為字
duweizi (nvarchar(255), null)	文字構成
wzgc (nvarchar(450), null)	來源
laiyuan (nvarchar(350), null)	詞義
ciyi (nvarchar(350), null)	詞性
cixing (nvarchar(3500), null)	接前簡
jqjian (nvarchar(350), null)	接後簡
jhjian (nvarchar(350), null)	組詞
zuci (nvarchar(350), null)	詞長
cichang (nvarchar(350), null)	字位置
ziweizhi (nvarchar(255), null)	字性質
zixingzhi (nvarchar(255), null)	字詞關係
ziciguanxi (nvarchar(350), null)	字形註腳
zixingjz (nvarchar(450), null)	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格 3 新造字表

dbo.qcxinzaozi	
列	
id (PK, int, not null)	
zixing (nvarchar(255), null)	字形
xzzlx (nvarchar(255), null)	新造字類型
xzzlx_b (nvarchar(255), null)	新造字類型_B
bihuashu (nvarchar(255), null)	筆劃數
diyibi (nvarchar(255), null)	第一筆
dierbi (nvarchar(255), null)	第二筆
disanbi (nvarchar(255), null)	第三筆
pinyinjs (nvarchar(355), null)	拼音檢索
jgoujs (nvarchar(255), null)	結構檢索
bushoujs (nvarchar(255), null)	部首檢索
tupianhao (nvarchar(255), null)	圖片號

表格 4 釋文表

dbo.New_QJ_ShiWen	
列	
id (int, not null)	
sid (nvarchar(50), null)	ID
wxzlhao (nvarchar(50), null)	文獻種類號
jianxu (nvarchar(50), null)	簡序
pianhao (nvarchar(350), null)	篇號
pianmin (nvarchar(550), null)	篇名
duanhao (nvarchar(150), null)	段號
zfmian (nvarchar(350), null)	正反面
sjjianhao (nvarchar(200), null)	資料庫簡號
ctzluhao (nvarchar(150), null)	出土著錄號
zlizlihao (nvarchar(150), null)	整理著錄號
ltjhao (nvarchar(150), null)	連圖簡號
shiwen (nvarchar(3500), null)	釋文
jianjiaozhu (nvarchar(750), null)	簡註腳
swjiaozhu (nvarchar(750), null)	釋文校注
plxuhao (nvarchar(150), null)	編連序號
tbly (nvarchar(350), null)	圖版來源

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格 5 文献目录表

dbo.qhwxcategory	
列	
id (PK, int, not null)	
wxbhao (nvarchar(255), null)	文獻編號
wxmc (nvarchar(255), null)	文獻名稱
shidai (nvarchar(255), null)	時代
chutusj (nvarchar(255), null)	出土時間
chutusf (nvarchar(255), null)	出土省份
chutudd (nvarchar(255), null)	出土地點
chutusl (nvarchar(255), null)	出土數量
jianjie (text, null)	簡介

這樣的表結構設計不僅很好地解決了上面提到的將字、詞、簡、篇相關聯的要求，並且在實際使用中也證明了表結構的合理性和易操作性。通過各個表之間的關鍵欄位鏈接，能夠提供不同需要的查詢、檢索功能。

三、 自動分詞方法探析

1. 自動分詞的重要性

“詞是語言中最小的能夠獨立活動的有意義的語言成分"[8],是信息處理的基本單位。在簡牘研究領域,分詞是詞彙標註、句子含義分析的第一步,是後續語法和語義分析的基礎。語義是語言中的概念與概念之間的關係,而詞是表達概念的,沒有詞,就無所謂概念,因而也就無所謂語義研究。詞頻統計、詞性自動標注、自動分類、詞類檢索等方面的研究,也必須首先分詞,在這些應用和研究領域,沒有準確高效的分詞策略,漢語的進一步分析必將受到嚴重影響。

而自動分詞，目的是為了通過計算機技術的優勢，做到自動化材料加工，從而大大減輕人力工作，提高處理效率。漢簡出土量巨大，材料規模達到數十萬字，單純依靠人力加工，切詞、標註詞性，將會耗費大量時間經歷，效率很低。而計算機自動分詞的處理速度大概是每分鐘五萬字，處理一種材料只需要不到一分鐘時間，將會大大提高材料處理的效率，減輕人力負擔。

2. 古漢語自動分詞的難點

[8]朱德熙. 語法講義[M]. 北京:商务印书馆，1982.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

古漢語自動分詞過程中的主要困難有以下四個方面:

首先,“詞“這個概念一直是語言學界糾纏不清而又無法回避的問題。在漢語中什麼是詞,到現在並無公認的定義。在古漢語中，語法句法靈活多變，沒有統一的句法規則，存在大量的倒置、缺省、活用等現象。詞與詞之間的關係情況複雜，因此想做到準確分詞十分困難。主要困難在兩方面,一方面是古漢語以單音詞居多，如何做到單音詞與多音節詞之間的劃界;另一方面是多音節詞與短語(片語)的劃界。到目前為止沒有公認的具有權威的詞表。其次,對漢語切分會產生切分歧義。歧義切分是影響分詞系統切分正確率的重要因素,也是分詞階段最困難的問題之一。再次,古漢語中未錄入詞的識別問題。漢語分詞問題中的歧義切分固然困難,但是未錄入詞問題可能比上述分詞的問題更嚴重。因為缺少權威的全面的古漢語詞典，許多詞語由於未被錄入而導致無法識別。最後，漢代出土簡牘中有大量的殘簡、斷簡、模糊不清的簡，導致釋文內容的破碎、不連貫，甚至無法呈現出一句完整的句子。這給通過上下文搭配進行分詞來了很大的難度。綜上所述，雖然自動分詞技術在現代漢語領域已經有了非常廣泛的應用，相關的研究論文也很多，但是古代漢語領域，特別是出土文獻的釋文分詞尚處在探索階段。在這方面的理論研究並不多見的情況下，要實現漢簡釋文的自動分詞，是一件難上加難的技術難題。

3. 分詞方法

基於規則的方法一般都需要事先有人工建立好的分詞詞典和分詞規則庫。主要是基於字串匹配的原理進行分詞,往往以足夠大的詞表為依據,採用一定的處理策略將漢語文本的字串與詞表中的詞逐一匹配,如若成功,就認為該子串為詞。主要有正向最大匹配法、逆向最大匹配法、雙向匹配法、逐詞遍歷匹配法、設立切分標誌法、正向最佳匹配法和逆向最佳匹配法等。如果分詞詞典規模小,覆蓋程度有限,則會影響分詞的正確率。[9]

基於統計的方法首先為要解決的語言處理問題建立統計模型,並且訓練語料庫來估計統計模型中的參數,然後把參數值應用到模型中處理分詞問題。訓練語料庫的方法是事先對有一定規模的語料進行標註，通過機器學習的方式對模型進

[9]熊冬明. 汉语自动分词和中文人名识别技术研究[D].浙江大学,2006.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

行“訓練”，從而得到相關的參數。一旦語料庫被訓練完成，即模型參數得到，就可以用它來進行分詞。常用的數學模型有隱馬爾可夫模型、條件隨機場模型和N-gram 模型等。其中最為流行的是隱馬爾可夫模型。隱馬爾科夫模型(Hiddne Mkarov Model,簡稱 HMM)是現代語音辨識系統中構建統計模型的重要手段。現在 HMM 模型已經廣泛的應用到自然語言處理領域。HMM 模型的優點主要有:研究已經非常透徹,演算法成熟,效率高,效果好,易於訓練。

基於人工智能的分詞方法分為专家系统方法和神经元网络方法。北京师范大学研制的书面汉语自动分词专家系统就首次将专家系统方法引入到分词系统中。

在這些主流的分詞方法中，處理現代漢語材料都能夠取得不錯的效果，但是在古代漢語分詞領域，針對性的研究並不多，一來是上文提到的古代漢語語法複雜、材料不全等客觀因素造成分詞的種種難度，二來是古代漢語進入數字化階段較晚，目前還集中在數字材料的收集和整理階段。因此用目前的分詞系統對古代漢語文本進行分詞，正確率很低，還是需要大量的人工勞動。但是通過比對分析，在切分人名、地名、職官、數量詞、干支、年號、連綿詞等具有典型特征的詞彙時，自動分詞系統的準確率較高，是可以進行自動分詞嘗試的領域。總之，自動分詞一直以來都是中文信息處理的重難點課題，如果能夠實現，將會對整個古漢語研究帶來極大的便利和好處。

本語料庫所採用的是基於規則的方法、基於詞典的方法和基於統計的方法相結合的方式。具體操作辦法將在第三章詞彙自動化標注研究中說明。

四、 各家考釋的數據庫表示方法

為了清晰地反應各家考釋信息，我們設計了各家考釋表以及考釋字表。通過這兩張表，能夠清楚的將被考釋字與考釋信息相關聯。解決了各家考釋的信息存儲問題。表結構如下所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格 6 考釋目錄表

dbo.qckaoshimu	
列	
id (PK, int, not null)	
bianhao (nvarchar(50), null)	文獻編號
author (nvarchar(50), null)	作者
tmpm (nvarchar(50), null)	題目
smkm (nvarchar(50), null)	出處
cbshe (nvarchar(50), null)	出版社
cbsjqh (nvarchar(50), null)	出版時間
ywlj (nvarchar(max), null)	原文連結

表格 7 考釋字表

dbo.qckaoshi	
列	
id (PK, int, not null)	
wid (int, null)	原庫的唯一編號
zihao (nvarchar(50), null)	ID_序號
wxid (nvarchar(50), null)	簡牘文獻種類編號
xuhao (int, null)	簡號
swztou (nvarchar(50), null)	釋文字頭
yxtp (nvarchar(255), null)	原形拓片：字形路徑及檔案名
lgd (nvarchar(50), null)	隸古定
kaoshi (nvarchar(max), null)	考釋提要
kscc (nvarchar(50), null)	考釋出處，對應文獻目錄
zlum (nvarchar(200), null)	著錄名
zluhao (nvarchar(50), null)	著錄號
pianming (nvarchar(255), null)	著錄篇名
sortzlm (int, null)	

在這樣的表設計支撐之下，關於漢簡的一切所需信息、數據，都能夠得到合理的存儲、關聯，便於提取信息和檢索等操作的實現。在設計完成表結構之後，就是基礎的數據錄入工作，例如簡牘釋文、字形圖片、考釋信息等。這一步由課題組成員共同分工參與錄入，在錄入過程中，還需要解決如何做到各個成員之間的數據同步、修改及時覆蓋的問題，我們採用單一時間段內只有一人能操作數據庫的運作模式，保證數據庫信息不會被錯誤覆蓋。

五、 數據庫打印功能

傳統的引得、字典的製作需要耗費大量的人力，里用數據庫技術，可以大大簡化這一過程。通過窗體編程，可以自動打印出引得、字典等工具書，並且自己

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

編輯打印格式。目前漢簡數據庫系統支持詞典打印、構件打印、釋文打印、引得打印、字形打印等多種功能。部分打印實例截圖如下圖所示。

序號	構件	組字次數	所構字
1	水	101	暴 冰 波 漕 池 滑 涑 瀆 犢 讀 法 繁 氾 泛 浮 溝 沽 海 漠 浩 河 涸 滑 淮 澮 汲 犰 濟 減 漸 澗 濺 津 洒 沮 涓 決 塡 渴 梁 涼 泠 凌 溜 流 洛 沒 淖 泥 溺 凝 汸 淖 溥 漆 淺 清 渠 溗 濡 洳 瀝 沙 涉 深 濕 泰 水 泰 湯 涕 涂 沚 渭 溫 汙 污 溪 洗 消 泄 洫 泫 沇 衍 洋 沂 益 淫 渤 [illegible] 漁 浴 淵 源 澤 湛 汁 治 注 瀆
2	人	99	佰 北 倍 備 俾 佖 便 伯 參 倀 償 偖 侈 飭 傳 倅 代 但 伐 仮 伏 付 負 傅 佝 光 何 佫 侯 厚 候 化 偭 及 亟 夾 監 件 介 俓 竟 覓 競 咎 俱 蚀 俊 僂 僇 倫 侵 伓 耆 侵 寢 囚 備 儴 人 仁 任 色 傷 伸 什

圖 1 打印構件

一部	字頻	字形
一一弌	2328	T01:045 T01:103 T01:123 T01:170 T03:071 T03:094 T03:102 T04:064 T04:098B T05:026 T06:019 T06:041A T06:041A T06:054 T09:055 T09:057 T09:082 T09:082 T09:091

圖 2 打印字形編

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二節　漢簡語料庫網絡平台

一、　平台簡介

本文研究的漢簡語料庫網絡平台，如圖 3 所示，是基於中國文字研究與應用中心，由作者導師負責的“秦漢簡牘帛書語料庫”的基礎之上，繼續優化開發得到的。該語料庫網絡平臺中除了收錄第二章第一節第一部分中所提到的十餘種漢簡之外，還收錄了八種秦代簡牘，主要包括睡虎地秦簡木牘、天水放馬灘秦簡、龍崗秦簡木牘、楊家山秦簡、王家台秦簡竹牘、關沮周家台秦簡木牘、裡耶秦簡、岳麓秦簡等。

平台所用數據庫為 Microsoft SQL Server 2005，其中的數據是由上一節介紹的的漢簡語料庫數據庫系統中的數據作為支撐的，由 Microsoft Access 2010 數據庫導入。平台的開發語言為 PHP+ASP。



圖 3

由於該語料庫具有開放性，除了會不斷補充簡牘材料之外，語料庫功能目前也在不斷開發優化之中，因此本文將從平台已有功能和正在開發功能兩方面來介紹。

漢簡語料庫網絡平臺已有的兩個主要功能包括了全文檢索功能和考釋檢索功能。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

二、 全文檢索功能

漢簡語料庫全文檢索功能包括按文獻出處檢索、字詞檢索、字形檢索、字際關係檢索四類。

1. 按文獻出處檢索

著錄檢索提供以簡牘為單位的各種簡的流覽、查找。左側列出資料庫中的簡的類型（見圖 4），點開後可以顯示每種簡的各篇及各簡。右側上部可以提供著錄書名、篇名、簡號的檢索，下部顯示該簡的所有原始字形圖片及逐字對應的釋文。

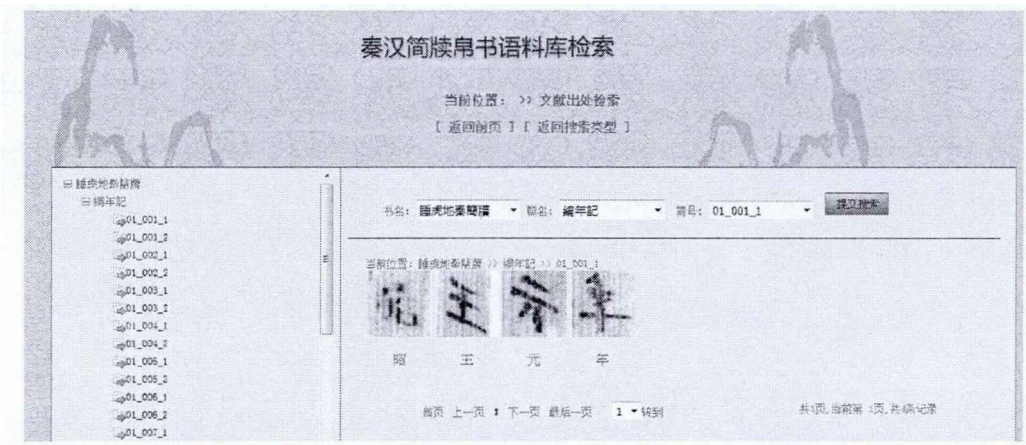


圖 4

各簡字形及釋文分別鏈結字形表和考釋表。圖 5 顯示的是從圖 4 的第一個字“昭”鏈結的有關該字的考釋內容。



考释字形	考释提要	著录名	著录号	篇名	考释作者
	昭王省稱	睡虎地秦墓竹簡	01.003.2	編年記	睡虎地秦墓竹簡整理小組
	秦昭王	睡虎地秦墓竹簡	01.001.1	編年記	睡虎地秦墓竹簡整理小組
首页 上一页 1 下一页 最后一页 1 转到			共1页,当前第 1页,共2条记录		

圖 5

2. 字詞檢索

字詞檢索提供釋字字頭、隸古定字頭或詞義的檢索。檢索範圍可以進行多重選擇。見圖 6。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

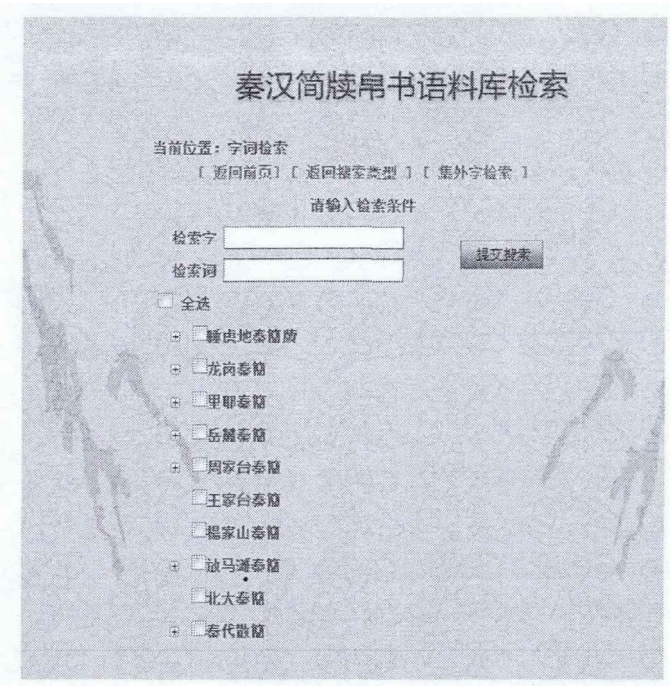


圖 6

字詞檢索結果顯示所有文例，見圖 7、圖 8。按簡的出處分列文例。如果語料庫已經經過詞義分析標注，例句將按義項排列。尚未經過詞義分析者，則如普通引得直接排列例句。例句後標記篇名和簡號。單擊簡號，可以直接鏈結到圖 4 所示的相應簡的頁面。

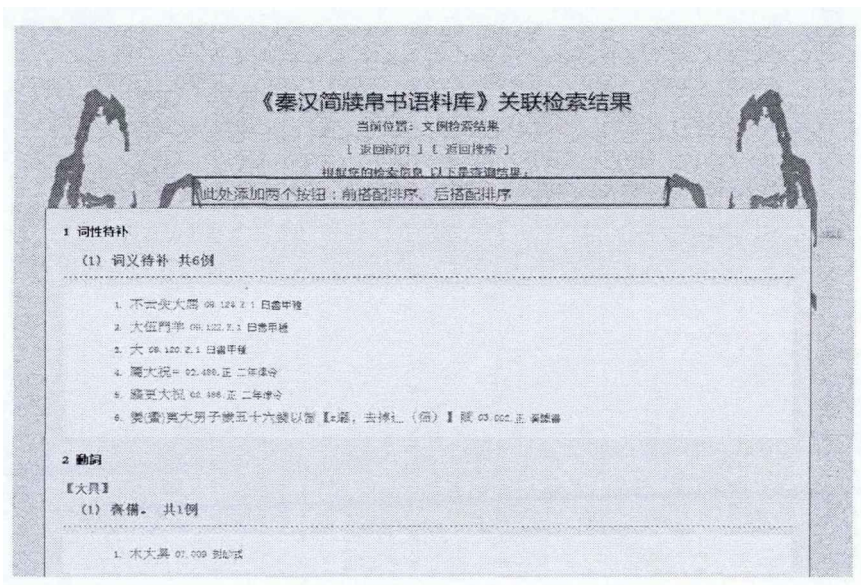


圖 7

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

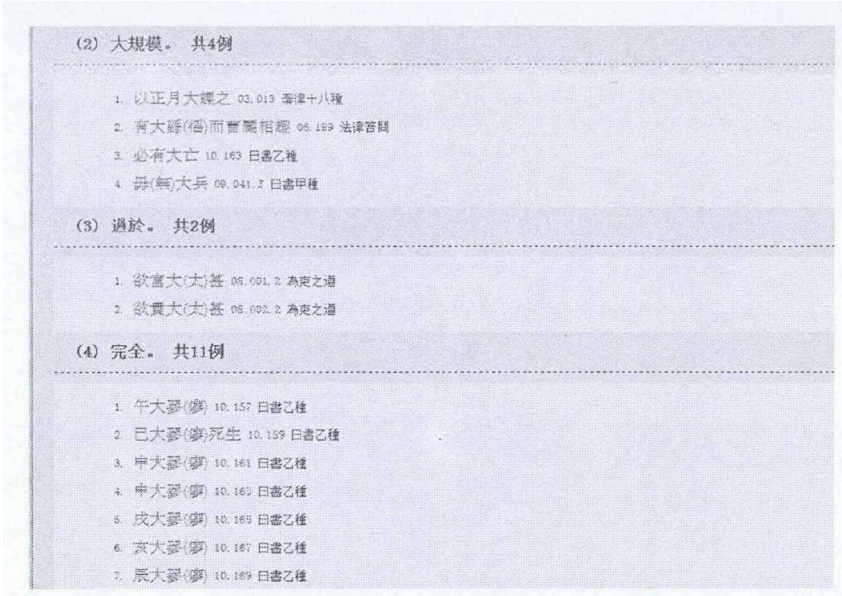


圖 8

此外字詞檢索中添加了集外字檢索，如圖 9 所示。可通過筆畫數和偏旁拆分，搜索集外字符。

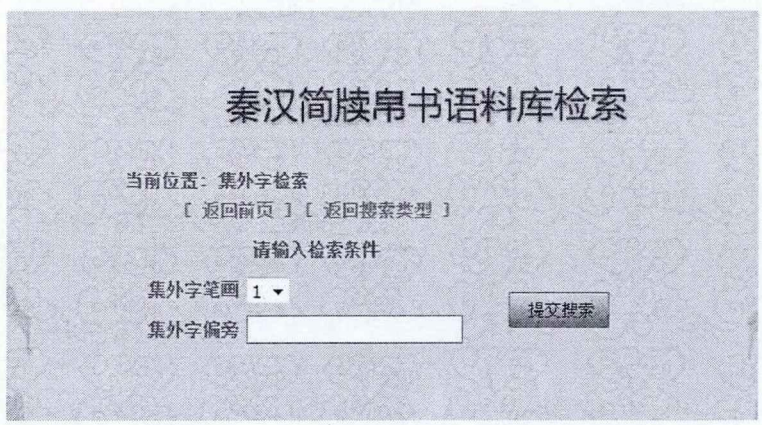


圖 9

3. 字形檢索

字形檢索提供秦漢簡全部字形的檢索。檢索條件可以通過字形構件、釋字字頭兩者進行。檢索範圍也可以進行選擇。見圖 10。檢索結果顯示符合條件的全部秦漢簡字形。圖 11 顯示里耶秦簡、睡虎地秦簡這兩種材料中的“昭”字檢索結果第 1 頁的字形（由於網絡問題，字形圖片顯示錯誤，在網絡條件好的情況下可正常顯示）。通過簡號，也可以鏈結查看圖 4 所示的該簡全部字形情況。字形有考釋的圖片，用紅色外框突出顯示，點擊字形圖片，可以鏈結該字形的各家考

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

釋情況，如圖 5 所示。

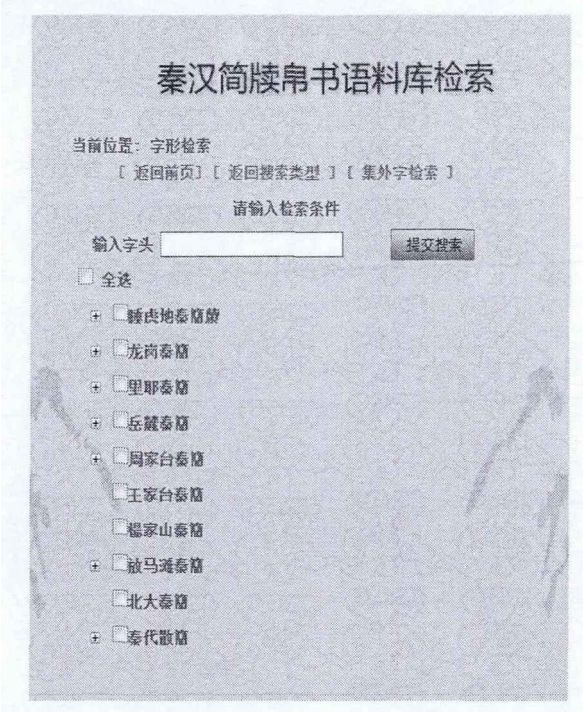


圖 10

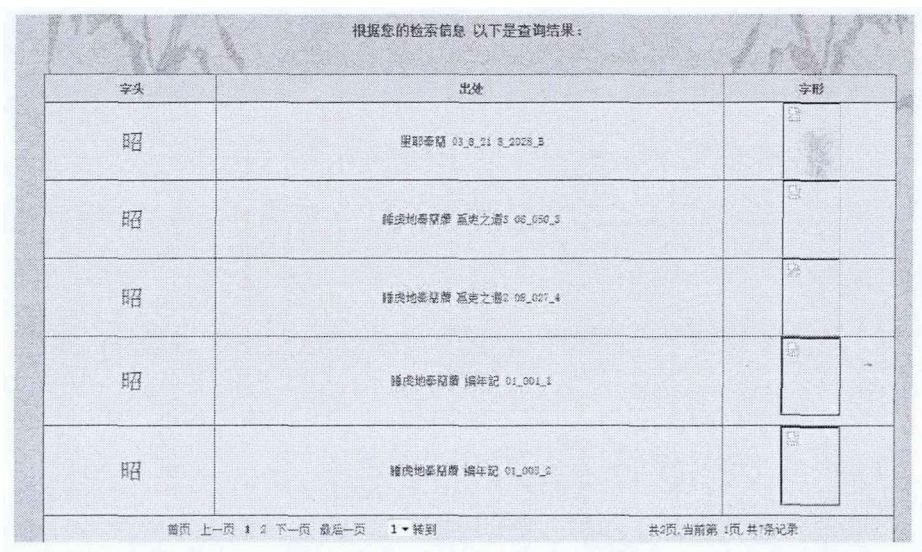


圖 11

4. 字際關係檢索

異體字、通假字在古漢語中十分常見，也是重要的研究方向。因此我們提供字際關係檢索，包含異體檢索和通假檢索兩種類型。如圖 12 所示。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

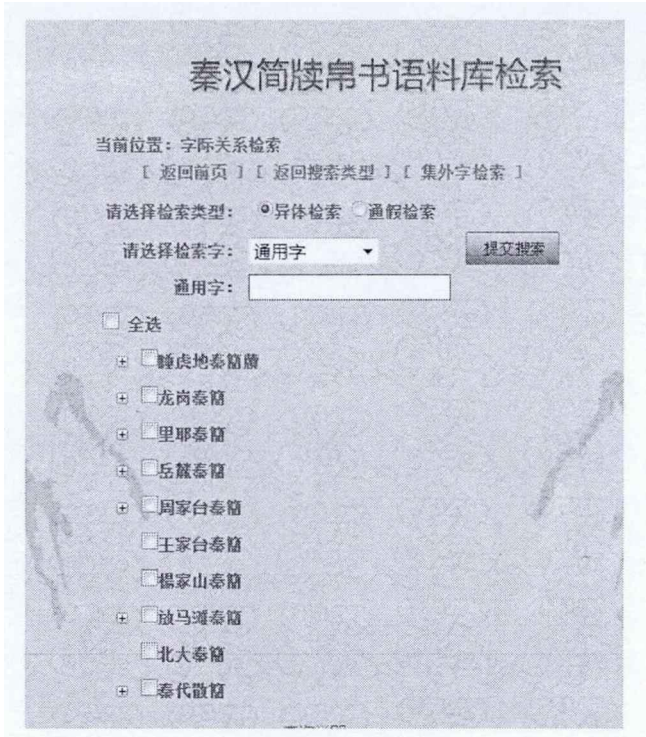


圖 12

a) 異體檢索

異體檢索可在異體字和通用字之間互相檢索。

當檢索字為通用字時，顯示對應的異體字。例如檢索通用字“來”，則顯示異體字“逨、漓、澾”等。檢索“明”，則顯示異體字“明、朙”。

當檢索字為異體字，則顯示對應的通用字。如檢索異體字“逨”，則顯示通用字“來”。

以上兩種檢索結果中，點擊通用字，可鏈接至字詞檢索條目，如圖 7 所示。

b) 通假檢索

通假檢索可在假借字和本字之間互相檢索。

當檢索字為本字時，顯示對應的假借字。例如檢索本字“卑”，則顯示假借字“俾、脾”等。檢索“安”，則顯示異體字“洝、焉、晏”。

當檢索字為假借字，則顯示對應的本字。如檢索異體字“白”，則顯示通用字“柏、日”。

以上兩種檢索結果中，點擊本字，可鏈接至字詞檢索條目，如圖 7 所示。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

三、 考釋檢索

1. 被釋字檢索

提供被釋字的考釋信息檢索，如圖 13 所示。輸入被釋字后，檢索出相應的考釋提要、考釋出處等信息，如圖 14 所示。點擊詳細信息，會彈框出現考释者、篇目、书名、出版社、出版日期等詳細的考釋信息。如圖 15 所示。

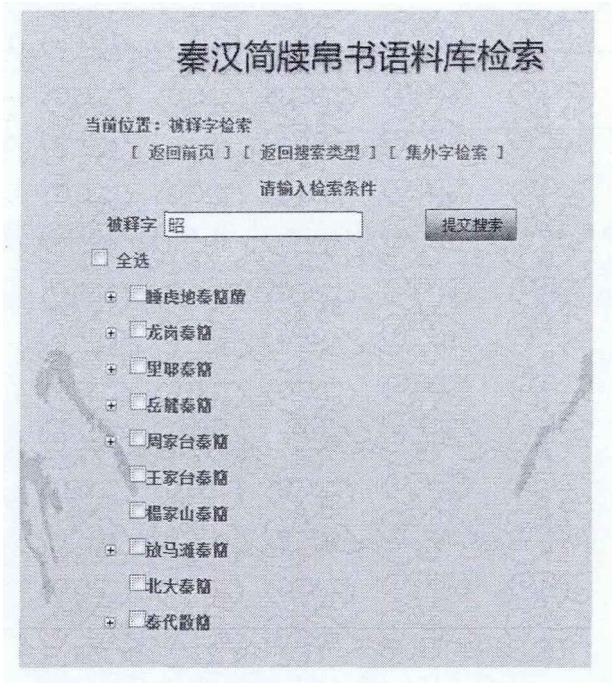


圖 13

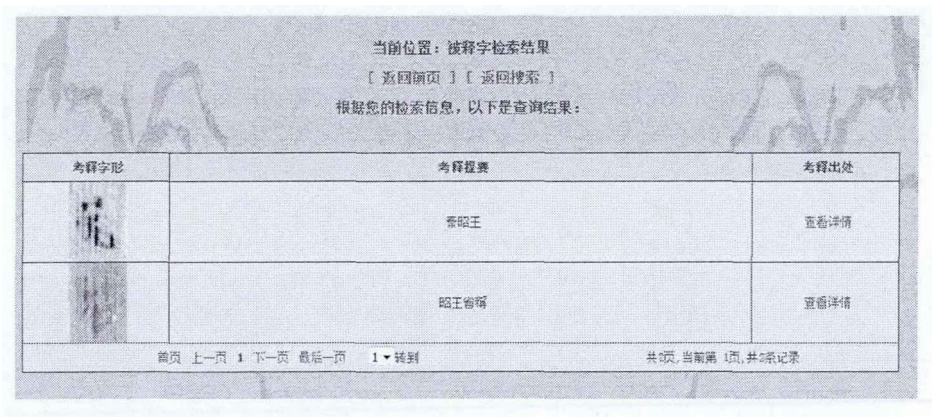


圖 14

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

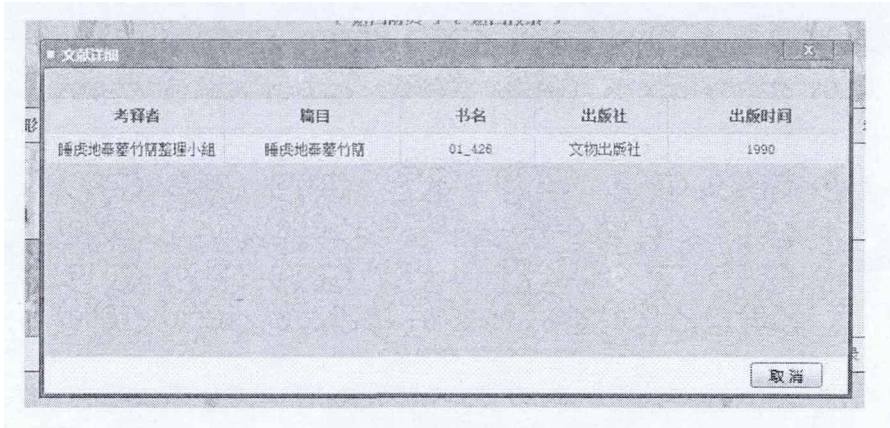


圖 15

2. 考釋論著檢索

提供有關秦漢簡牘考釋論著的檢索。檢索途徑見圖 16。檢索結果顯示符合條件的所有考釋論著目錄。見圖 17。通過“查看詳細”鏈結，可以鏈結該論著的具體釋字情況，如圖 5 所示。

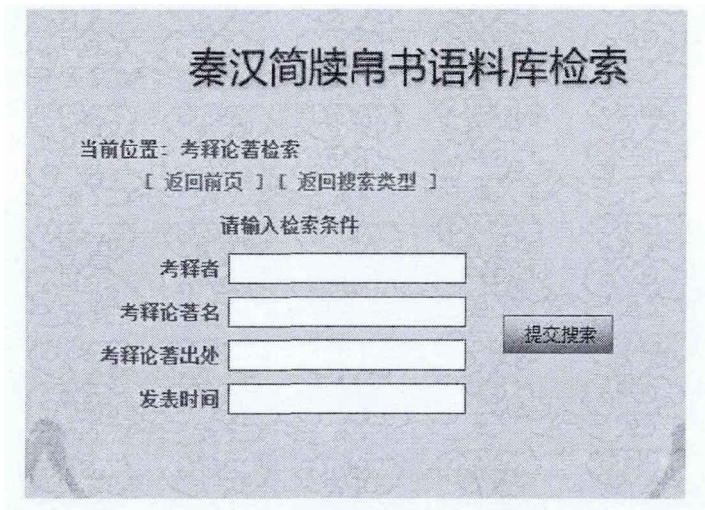


圖 16

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



作者	題目篇名	书名刊名	出版社	出版时间	查看详细
李學勤	論包山楚簡魯陽公城郟	清華大學學報(哲學社會科學版)		2004年第3期	查看详细
李學勤	上博楚簡《魯邦大旱》解義	《上博館藏戰國楚竹書研究續編》(上海大學古代文明研究中心 清華大學思想文化研究所編)	上海書店出版社	2004年	查看详细
李學勤	楚簡《子羔》研究	《上博館藏戰國楚竹書研究續編》(上海大學古代文明研究中心 清華大學思想文化研究所編)	上海書店出版社	2004年	查看详细
李學勤	釋《詩論》簡"兔"及從"兔"之字	《北方論叢》		2003年第1期	查看详细
李學勤	《詩論》說《周頌》等七篇釋義	《齊魯學刊》		2002年第2期	查看详细

首页 上一页 1 2 下一页 最后一页 1 转到 共2页,当前第 1页,共7条记录

圖 17

以上為漢簡語料庫網絡平台已經開發的功能，除此之外此平台還有一些已經設計好的功能正在開發之中，分別是詞性檢索功能、著錄顯示功能、簡牘拓片顯示功能和編連綴合考釋信息顯示功能。

四、 詞性檢索功能

為了更好的利用數據庫中已經標註好的詞性、詞義等信息，也適應平台使用者的研究需要，我們設計開發詞性檢索功能。

1. 界面設計：

提供詞性搜索框。可選的限制條件為著錄名、篇名等。

输入词性

2. 結果顯示

例如：輸入的是“名詞”，顯示結果為：

名詞。

【姑】

1. 名詞。“古先”通“姑洗”，即姑洗。音韻十二律之一。

句子 ：

1. 夷則黃鐘姑先之卦曰此是可亡不復可求弗得中聞不樂又若虎□上下行往莫中吾步 放馬灘秦簡：一號秦墓竹簡乙種《日書》釋文 6 ： 02_246_1

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

五、 著錄顯示功能

著錄，可以理解為相關文獻的意思。秦漢簡中的某一條簡，可能被收錄在不同的著錄中，因此應在該簡所在頁面顯示出此簡還收錄在哪些額外的著錄之中。

1. 界面設計

著錄內容顯示在“文獻出處檢索”模組，點開某條簡，右側頁面中。具體位置圖 18 所示：

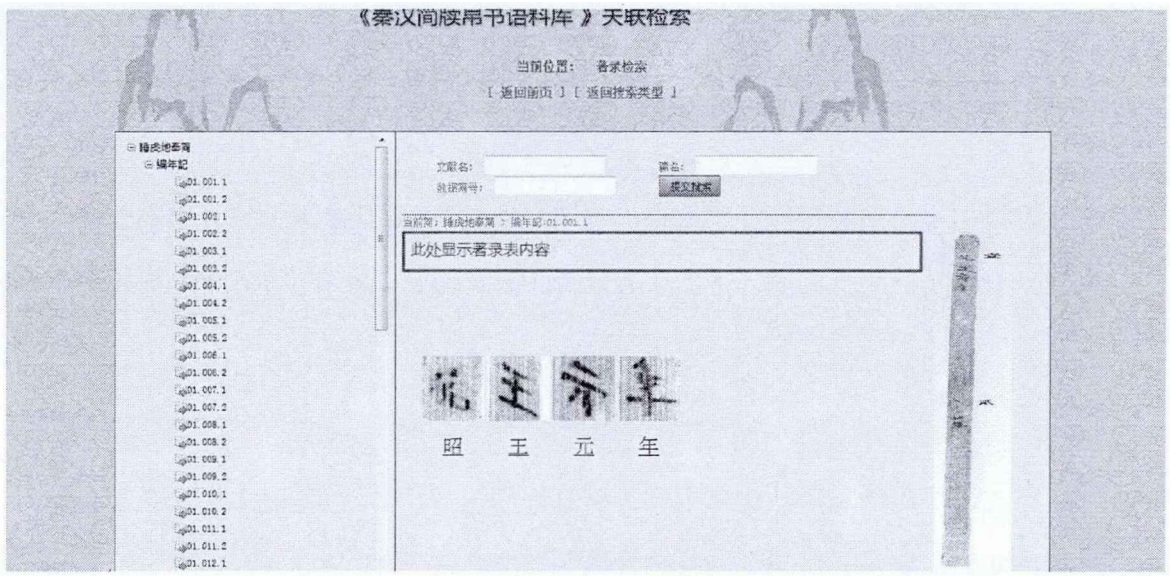


圖 18

2. 數據庫中增加著錄表

應在數據庫中增加一張新的表：著錄表（zhulu）。表的設計如下：

表格 8 著錄表

欄位名	類型	能否為空	備註
ID	自動增加	Notnull	主鍵，昇冪序號
Jianhao	Nvarchar(350)	Notnull	簡號。用於關聯著錄表和樹狀圖中的簡號。如 01.001.1
wenxian	Nvarchar(255)	Null	所在著錄的文獻名
page	Nvarchar(255)	Null	所在著錄的頁碼
currentjhao	Nvarchar(255)	null	當前文獻使用的簡號

3. 顯示格式

文獻名，頁碼，當前簡號

Eg. 蔡偉《讀竹簡劄記四則》，P30，1.3.23

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

六、 簡牘拓片顯示功能

1. 在“文獻出處檢索”模組，右側添加簡號對應的簡牘拓片。如圖 19 紅色方框所示。

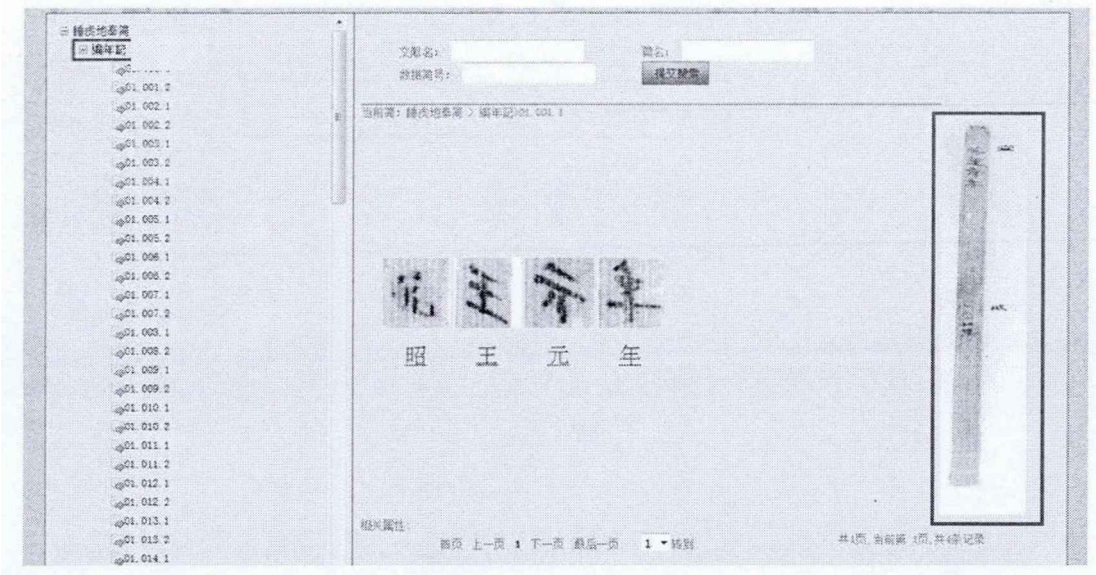


圖 19

2. 增加圖像放大功能。

當鼠標置於簡牘拓片之上時，實時放大鼠標所在區域，從而便於清楚查看簡牘上的文字。

七、 編連綴合考釋信息顯示功能

1. 目的

簡的銜接順序，即編連綴合順序不同，應將關於編連綴合的不同各家考釋顯示在相對應的簡頁面。便於研究者全面了解簡牘編連的信息。

2. 顯示位置

編連綴合考釋信息同樣顯示在“文獻出處檢索”模組。顯示位置在著錄信息下方。如圖 20 所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



圖 20

以上這些功能大部分已經完成，部分還在不斷的開發優化當中，綜合來看，將建成的漢簡語料庫網絡平台將具有材料全、檢索方便、可實現深層次檢索、字形拓片顯示方便等優點。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第三章 基於漢簡語料庫的詞彙自動化標註研究

在第二章第一節中已經提到，詞彙的自動標註分析研究，是在語料庫豐富的語料資源基礎之上，通過計算機智能分析、自動分詞、詞性標註等技術的運用，從而實現對語料庫龐大的數據進行自動化、智能化的分詞、標註、統計的研究。通過理論分析與實際操作相結合的研究方式，探索出合理的自動分詞、自動化標註詞性的操作辦法，是對已建成的漢簡語料庫最難也最有實用價值的應用。

之所以有價值，是因為首先，目前的材料加工主要依靠人力，對於幾十萬字規模的材料，人工分詞、標註詞性工作量巨大，效率十分低下。而由計算機進行自動分詞、詞性標註等材料處理工作，充分發揮了計算機處理速度快運算量大的優勢，例如對于整部敦煌漢簡的詞彙自動化分析只需要不到一分鐘的時間即可完成，因此，在提高正確率的前提之下，進行漢簡詞彙自動化標註將大大減輕人力工作量，提高工作效率。其次，計算機的數據處理功能還能夠在標註結果的基礎之上進行更深層次和多維度的統計，得到一些規律性的結論。經過標註後的簡牘釋文，能夠以其數據量大，規律性強等優勢更好地為我們提供研究數據，通過對數據的分析發現一些規律性問題，從而開闢出新的研究課題。所以探索研究出正確率較高的自動化標註方法，對漢簡研究是非常有必要和有價值的。

之所以最難，是因為古漢語文獻的詞彙自動化標註相關研究較少，古漢語語法較為複雜。所以本文的許多方法都參考了現代漢語分詞、詞性標註的方法。從結果來看，由於是機器處理，標註的結果的準確度依然低於人工標註，這也是我們需要不斷探索完善古漢語詞彙自動化標註技術的動力所在。

本章選用了機種西北漢簡、張家山漢簡、馬王堆漢墓帛書、武威漢簡等幾種材料做為詞彙自動化標註研究的對象，因為以上材料分別代表了文書類和古書類兩種用詞風格的材料，具有一定的典型性。本章將從材料介紹、材料準備、實驗原理、實驗過程、結果分析與改進方法等幾方面，來介紹詞彙自動化標註的研究過程。根據文獻的自身特點不同，研究過程中有針對性的設計了不同的標註方法和步驟，以期能夠達到科學準確的識別結果。具體來講，對於文書類的西北漢簡，

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

將重點進行人名、地名、職官、年號等專有名詞類命名實體的識別研究及結果分析，并統計出相關的專有名詞詞表；而對與古書類的張家山漢簡、馬王堆漢墓帛書、武威漢簡等，則重點分析虛詞這類詞彙的識別結果。

第一節　研究材料概說

一、　西北漢簡

西北漢簡指的是在西北地區發現的幾種典型文書類簡牘，本文選擇的材料包括：敦煌漢簡、居延漢簡（舊、新）、肩水金關漢簡。

敦煌漢簡

敦煌漢簡是指20世紀初至80年代在甘肅西部疏勒河流域漢代關塞遺址陸續出土的漢代簡牘，其在漢代敦煌郡範圍內發現的世紀最早、數量最多，共計25000餘枚。這些漢簡的成書年代大概在西漢中後期至東漢中後期，也就是西元一世紀到西元二世紀中葉。敦煌漢簡的內容主要有詔書、律令、官府文書、簿籍等多方面，是漢代在西北屯軍的有關屯戍活動檔案。可以說，敦煌漢簡不僅代表了漢代邊郡防禦制度的情況，同時也揭示了漢代時期與西域各個國家的邦交和文化交流。為我們研究漢王朝時期的西北史地提供了絕好的原貌素材，具有極高的研究價值。

敦煌漢簡的發現：二十世紀初，匈牙利考古學家斯坦因在中亞考古活動中考察了新疆和甘肅等我國西北疆域。1907 年在敦煌的數個漢代烽燧中，發現了許多漢文簡牘，由此開啟了敦煌漢簡的發掘。斯坦因、周炳南、西北科學考察團、嘉峪關市文物保管所和甘肅省文物工作隊等考古科學家和團隊先後在敦煌及其附近地區發掘獲得數量不等的漢代簡牘。

敦煌漢簡的整理：甘肅省文物研究所將各批出土的簡牘進行了細緻的整理，基本解決了釋讀研究等問題。

居延漢簡

居延漢簡是指中外學者在居延烽燧遺址當中發掘出的漢代簡牘，居延漢簡因其在今天的內蒙古自治區西部額濟納旗的居延地區被發現而得名。居延在西漢時

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

屬於張掖郡居延都尉，是我國西北地方的軍事重地和漢長城要塞。居延漢簡包括了居延舊簡和居延新簡。其中，居延舊簡的最早紀年簡為西元前 102 年，最晚紀年簡為西元 30 年，也就是從武帝太初三年到東漢建武六年這段時間。居延漢簡是我國漢代居延和肩水都尉府的行政文書檔案，內容非常豐富，記載了居延地區的屯戍活動、歷史興衰，極具科學研究、歷史文物價值，也是簡牘學研究的重要根基。居延漢簡被譽為 20 世紀中國檔案界的“四大發現”之一。

居延漢簡的發現：1930 年至 1931 年期間，以貝格曼為首的瑞士科學家和中國科學家組成的西北科學考察團在居延地區的漢代邊塞遺址中發現了漢代的八個塞牆，並在其中發掘出 10200 枚漢代簡書，這批出土的簡牘被稱為《居延漢簡》，也就是現在的“居延舊簡”。1972 年至 1976 年，甘肅居延考古隊在額濟納旗流域的進行文物發掘工作，新獲得近兩萬枚漢代簡牘，這批出土的漢簡被稱為“居延新簡”。

肩水金關漢簡

肩水金關漢簡是指出土于肩水金關遺址的漢代簡牘。同居延漢簡一樣，肩水金關漢簡的出土時期也是分為 20 世紀 30 年代和 20 世紀 70 年代兩個階段，並且居延漢簡和肩水金關漢簡之間的關係非常緊密。20 世紀 30 年代出土的肩水金關漢簡被包含在《居延漢簡》的“居延舊簡”部分，此次出土的肩水金關漢簡簡牘數量約 850 枚，其中大部分簡牘紀年集中在西元前 82 年至西元前 4 年，也就是漢昭帝到漢哀帝的西漢中晚期時期；而 20 世紀 70 年代出土的居延地區的漢簡，則分為兩部分，其中一部分即為《肩水金關漢簡》，此次發掘出來的肩水金關漢代簡牘數量約 11000 枚。肩水金關漢簡為研究漢代的政治、經濟、歷史和當時的社會狀況等提供了寶貴的原始資料。

西北漢簡內容屬於文書類，記錄了大量西漢中期到東漢初期西北邊境戍卒的政治、經濟、軍事、文化以及社會生活等現象，其中有大量的涉及到人名、地名、職官、機構等的內容，此外還有專門用於記錄人名的名籍簡等等，因此用來對漢代專有名詞類詞彙進行分析非常適合。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、 古書類漢簡

馬王堆漢墓帛書

馬王堆漢墓帛書出土於20世紀70年代初發現的位於中國湖南省長沙市馬王堆三號漢墓。馬王堆三號漢墓的葬墓時期可推斷為西元前168年，也就是漢文帝前元十二年。帛書內容涉及戰國時期至西漢初這段時間的政治、軍事、思想和文化科學等各個方面，具有非常重要的學術研究價值。

其中：一號墓，1972年發掘，出土竹簡 312枚，2 063字，內容爲遣策。漢文帝 15年（西元前 165年）三號墓， 1973年發掘，出土簡牘包括：遣策：牘6件；簡 402支（含三支殘簡）。醫書： 200支。甲卷包括十問、合陰陽方；乙卷包括雜禁方、天下至道談。還有大量帛書，約 11.5萬字。

長沙馬王堆三號漢墓出土的帛書，其字體有篆書和隸書之分，大部分用墨書寫，帛書作為戰國時期的文字記錄載體，承前甲骨金文，啟後魏晉及以後的紙張書寫，對古漢字的演變和發展都有十分重要的意義。出土帛書共有28種，計十二萬餘字，均破損嚴重。依《漢書・藝文志》分類 ，六藝類的有《周易》、《喪服圖》、《春秋事語》和《戰國縱橫家書》；諸子類的有《老子》甲本（後附佚書4種）、《九主圖》、《皇帝書》和《老子》乙本（前附佚書 4種）；兵書類有《刑德》甲、乙、丙3種；數術類有《 篆書陰陽五行》、《隸書陰陽五行》、《五星占》、《天文氣象雜占》、《彗星圖》、《出行占》、《木人占》、《符籙》、《神圖》、《築城圖》、《園寢圖》和《相馬經》，其中《五星占》是中國現存最早的天文書；方術類有《五十二病方》 (附佚書4篇)、《胎產圖》、《養生圖》、《雜療方》、《導引圖》(附佚書2篇)，其中 《五十二病方》是中國已發現的最古老醫書。另外，還有《長沙國南部地形圖》、《駐軍圖》、《城邑圖》3幅地圖。

本文將選取其中的《老子》甲本（後附佚書4種）、《老子》乙本（前附佚書4種）進行詞彙分析研究。

張家山漢簡

張家山漢簡出土于張家山二四七號漢墓，位於湖北省江陵縣（今荊州市荊州區）城外西南的江陵磚瓦廠內。1983年12月至1984年1月，荊州地區博物館

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在湖北江陵張家山發掘了二四七、二四八和二四九三座漢墓，發現了大量的竹簡，其中以二四七號漢墓出土的竹簡數量最多，保存得也較好，字跡清晰可辨。二四七號漢墓中竹簡共計 1236 枚，出土竹簡內容為漢代典籍，有《曆譜》、《二年律令》、《奏讞書》、《脈書》、《算數書》、《蓋廬》、《引書》和遣策共八種，涉及漢代時期的法律、軍事、曆法、醫藥和科技等諸多方面，具有重要的學術研究價值。張家山漢墓竹簡中的《漢墓書》是比《九章算術》還要早的數學佚籍。

本文件選擇《二年律令》、《奏讞書》、《蓋廬》幾篇作為詞彙研究對象。

武威漢簡

武威漢簡是指出土于中國甘肅省武威市漢墓中的簡牘，其中主要有《禮儀》簡、王杖詔令簡和醫藥簡牘三大國寶級文物。武威漢簡具有數量多、保存好、內容豐富等特點，是我國簡牘學研究的重要組成部分。今天甘肅的武威縣，在漢代時期為武威郡治所在的姑臧縣，漢武帝置郡以後，興築長城，設立烽燧，屯戍兵卒，改變了河西走廊的政治、經濟和文化的面貌。

1959 年 7 月，甘肅省博物館的文物工作隊在武威縣發掘六號漢墓得到一片竹、木簡牘。六號漢墓屬於王莽時期，共出土 469 簡，之後甘肅省博物館在武威縣先後清理了三十七座漢代古墓。武威醫藥簡牘對於研究中國傳統醫學及醫學哲學思想具有一定價值。

內容主要是比較完整的九篇儀禮，一篇寫於竹簡，9 篇寫於木劄。共 469 簡，273 32 字。另有 7 枚日忌木簡、雜占木簡 4 枚。《武威漢簡》一書另收有 18 號墓出土的王杖十簡及 4 號、15 號、22 號、23 號墓出土的柩銘四條。

選擇以上幾種古書類簡牘作為研究材料，著重進行虛詞類詞彙標註進行探究，是由於古書類文獻內容相對完整、規範、用詞類型豐富，所以是進行以上類型詞彙分析的很好的選擇。

三、 材料版本選擇

基於實際操作的考慮，漢簡語料庫中對每種材料選取一種版本進行錄入。版本選擇的依據主要是以較新版本、較權威的整理結果、電子材料較為方便取得等

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

為主。具體的錄入方法是：由某一種材料作為基礎材料錄入，構成基本的語料庫，然後在此基礎上，再根據其他已發表的整理版本進行釋文校對，通過和原簡拓片對比、綜合各家意見等方法，選取較為合理的結果進行備註、說明等。另外語料庫還有專門的各家考釋表、著錄表等，力求能夠最大程度反應不同專家的研究成果。

本文選取的材料為單一版本的基礎材料，因為在分詞和詞性標註中，個別字詞的不同考釋對結果正確率影響不大，因此只選取一個版本進行詞彙標註研究。

敦煌漢簡：選取版本為《中國簡牘集成·甘肅省卷》，2004 年由中國簡牘集成編輯委員會主編，敦煌文藝出版社出版。

居延漢簡：選取版本為《居延漢簡甲乙編》，1980 年由中國科學院考古研究所出版。

肩水金關漢簡：選取版本為《肩水金關漢簡》1-3 卷，2011 至 2013 年由甘肅省簡牘保護中心、甘肅省文物考古研究所、甘肅省博物館、中國社會科學院簡帛研究中心等主編，中西書局出版。

馬王堆漢墓帛書：選取版本為《長沙馬王堆漢墓簡帛集成》，2014 年由復旦大學出土文獻與古文字研究中心、湖南省博物館編纂，中華書局出版。

張家山漢簡：選取版本為《張家山漢墓竹簡（二四七號墓）》，2006 年由張家山二四七號墓竹簡整理小組主編，文物出版社出版。

武威漢簡：選取版本為《武威漢簡》，1964 年由甘肅省博物館、中國科學院考古研究所及陳夢家先生整理主編，文物出版社出版。

第二節　語料預處理與詞典製作

待分析的語料以數據庫的形式存儲在建設好的漢簡語料庫中，因此在進行分詞及詞性標註之前，首先要對被分析的語料庫進行預處理，使其內容和形式符合分析模型的要求。同時，由於自動分詞和詞性標註採用了基於規則的方法、基於詞典的方法和基於統計的方法相結合的方式，應當預先製作好用于匹配的詞典。

一、　語料預處理

語料預處理主要目的是將需要分析的語料從漢簡語料庫的表中提取出來，并

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

對文本進行處理，使得後面的自動化處理結果更為準確、方便。由於出土簡牘的不完整性和古文獻中大量存在的通假字、異體字、殘缺字、待造字等實際情況，漢簡語料庫中存儲的原始語料藉助了大量的特殊符號，在預處理中需要將其剔除。下面將具體介紹語料預處理的方法。

1. 特殊符號處理

對表示簡首尾的⚌、◇、●、■、☑等，由於其不具有詞彙屬性，進行刪除處理。以及其他一些簡牘中的特殊符號，例如▩、◎、△、〇等，還有表示空缺或者無法辨認的字的符號□，凡是不具備詞彙含義的，都進行刪除處理。而對於表示段落、句子等的標點符號例如“，”、“。”、“？”、“：”、“、”等符號進行保留處理，有助於分詞的準確性，在詞性標註時統一標為“標點符號”。

2. 通假字、異體字、古今字處理

例如：

天下无（無）道，式〈戎〉馬生於郊。34_01_01_019 《馬王堆 老子甲本・德經》

“（）”中的為古今字，“<>”中的為通假字。在分詞時，如果按照原文分詞，則分詞結果是：

天下/无/（/無/）/道/，/式/〈/戎/〉/馬/生/於/郊/。

這樣的結果顯然是不準確的。因此對於異體字、通假字、古今字等特殊文字的處理，處於以詞彙意義完整性為主的原則，將只保留括號內的字，以保證詞彙的連貫性。例如上文在經過預處理后，將變為：

天下無道，戎馬生於郊。

這樣處理後的文本雖然無法準確揭示原簡字形，但是對於詞彙意義的研究是較為科學合理的。

3. 新造字替換處理

漢簡語料庫的字形顯示是一大難點，許多古漢字在現代字符集中並不存在，因此我們採取了重新造字的方法，對需要重新造的字採用 Unicode 國際編碼體系進行自主編碼，建立了各種新造字體用於字型顯示。但在進行自動分詞時，由於需要在 TXT 文檔中進行匹配，新造字的編碼在字符集之外無法進行匹配，因此就會出現新造字無法識別的問題。為解決這一問題，對語料庫語料進行了新造字

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

替換處理，對於原始字形，通過在《說文解字》或後世字書中找到對應的字，確定該字形是後世的何字，即“考釋字”。對於在原文中意義不通的“考釋字”，還需要進行破讀，替換為“破讀字”。這樣處理后的文本材料，能夠最大程度還原語義，使得基於語義的分詞結果更加的精準。

4. 語料庫合併處理

在語料庫中，不同的簡牘材料分別存儲在不同的表中，這樣在進行分析時比較麻煩，因此對語料庫進行了合並處理。將敦煌漢簡、居延漢簡、居延新簡、肩水金關漢簡幾種合並為西北漢簡材料庫，將馬王堆漢簡（老子部分）、張家山漢簡、武威漢簡合並為古書類漢簡材料庫。

5. 語料庫導出

本文採用的分詞系統無法讀取數據庫中的文獻材料，需將待分析材料從語料庫導出為 TXT 格式文檔，導出過程中注意編碼選擇為 UTF-8 編碼，不然會出現編碼錯誤問題。導出的文本由 ID 和釋文兩部分構成，ID 與原語料庫 ID 相對應。導出前後對比如下圖所示：

表_釋文

ID	釋文
1	制詔御史：秦年七十以上，比吏六百石，出入官府不趨，毋二尺告劾（劾），吏擅徵召◇
3	變事吏上，毆擊之。召，爰書。變事痛所毆以不能言變事，皆大逆不道□□◇
4	民無爵裏名姓，吏擅事使，有行事穎川東鄉佐坐論◇
5	諸自非九月，吏不得發民車馬牛給縣官事。非九月時，吏擅發民車□◇
6	民占數以男為女，辟（避）更徭，論為司寇
7	吏金二兩，在田律。民作原蠶，罰金二兩，令在乙第廿三◇
8	坐臧（贓）為盜，在公令第十九丞相常用第三◇
9	不道，在御史挈令第廿三◇
10	赦，不得下蠶室，在蘭台挈令第□◇

圖 21 武威漢簡語料庫

武威.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
ID      释文
1       制詔御史：秦年七十以上，比吏六百石，出入官府不趨，毋二尺告劾，吏擅徵召
3       變事吏上，毆擊之。召，爰書。變事痛所毆以不能言變事，皆大逆不道
4       民無爵裏名姓，吏擅事使，有行事穎川東鄉佐坐論
5       諸自非九月，吏不得發民車馬牛給縣官事。非九月時，吏擅發民車
6       民占數以男為女，避更徭，論為司寇
7       吏金二兩，在田律。民作原蠶，罰金二兩，令在乙第廿三
8       坐贓為盜，在公令第十九丞相常用第三
9       不道，在御史挈令第廿三
10      赦，不得下蠶室，在蘭台挈令第
```

圖 22 武威漢簡導出文本

導出的預處理后的語料庫文本，將作為分詞系統的分析對象，進行下一步的分詞操作。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、 詞典製作

本文採用的自動分詞工具為基於 Python 的“結巴中文分詞”系統，本系統支持基於詞典的分詞方式，因此在進行分詞前重要的一步就是進行詞典製作。

基於詞典的分詞方式擁有成熟的算法，並且很易於實現，是最為常用的分詞方法之一，通常配合基於規則和基於統計的方法一起進行分詞。而分詞正確率的關鍵就在於完善詳盡的詞典製作。不論什麼樣的分詞方法，優秀的詞典必不可少，因此在進行詞典製作時，作者盡可能多的利用了線上、線下能夠找到的古漢語文獻、詞典等材料，完成了基礎詞典製作。基礎詞典包含了單音詞、多音節詞、專有名詞等，并標註了每種詞的詞性以及每種詞性的概率。

1. 統一詞性標註集

不同的字典中對詞性的表示方法不一樣，此外，為了便於進一步的研究，特地在詞典中加入了人名、地名等專有名詞詞性。因此，為了便於標註，首先需要進行詞性集的統一。經分析后統一確定詞性為以下幾種：形容詞、副詞、名詞、連詞、嘆詞、方位詞、數詞、量詞、介詞、代詞、虛詞、時間詞、助詞、動詞、語氣詞，以及人名、地名、職官、機構、干支等專有名詞類，共計 20 種詞性。另外增加一項標註集“標點符號”，用於對標點符號進行標註。

2. 詞典處理與製作

(1) 《古漢語常用字字典》

1979 年商务印书馆出版，作者王力、岑麒祥、林焘、戴澧、唐作藩、蒋绍愚、张万起、徐敏霞等。其中包含了字頭 4564 個，詞條 3629 個。根據其體例形式，需要對字頭的詞性進行自動化提取，并統計每種詞性概率。例如詞條“阿”：

> 阿 ē ①〈名〉山陵；大丘。王勃《滕王阁序》：“访风景于崇～。”②〈名〉山湾。屈原《山鬼》：“若有人兮山之～。”③〈名〉屋角翘起来檐。《古诗十九首·西北有高楼》：“～阁三重阶。”④〈动〉曲从；迎合。《韩非子·有度》：“法不～贵，绳不挠曲。”（贵，地位高的人。）⑤〈动〉偏私；袒护。屈原《离骚》：“皇天无～私兮。”〖引〗亲近。《后汉书·文苑传下》：“苟失其道，则兄弟不～。”⑥通“婀”。柔软而美丽的样子。《诗经·小雅·隰桑》：“隰桑有～，其叶有难。”（难，茂盛的样子。）

經過程序自動化提取后的信息如下表所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表格 9

字頭	詞性	概率
阿	名詞	0.5
阿	動詞	0.33
阿	形容詞	0.17

概率統計的方法為：概率=每種詞性個數/總的詞性個數

統計概率的目的在於計算出一個多詞性的詞在每種詞性的可能性有多大。例如“阿”字出現時，則更有可能是作為名詞使用，較不可能是形容詞。這樣的概率有助於後面的詞性標註工作的順利進行。

(2) 《漢語大詞典第一版》

由於《漢語大詞典》2.0 版的資料無法提取的條件限制，本文採用的是《漢語大詞典》第一版，羅竹風主編。中國漢語大詞典編輯委員會、漢語大詞典編纂處編纂，漢語大詞典出版社出版。全書正文 12 卷，共收單字 2.27 萬，多音節詞 37.5 萬，全書約 5000 萬字。本文中主要提取了漢語大詞典中的多音節詞加入基礎詞典。

(3) 《漢語大字典》八卷本

《漢語大字典》全書約 2000 萬字，共收楷書單字 56000 多個，凡古今文獻、圖書資料中出現的漢字，幾乎都可以從中查出，是當今世界上規模最大、收集漢字單字最多、釋義最全的一部漢語字典。由四川、湖北兩省 300 多名專家、學者和教師經過 10 年努力編纂而成的，以解釋漢字的形、音、義為目的的大型漢語專用工具書。

由於《漢語大字典》體例過於複雜，提取難度較大，所以僅僅提取字頭約六萬個作為單音詞參考。

(4) 專有名詞

對於各類專有名詞，也進行了收集整理。

地名類：參考提取了《漢書地理志 上下卷》以及相關論文期刊中的里名、縣名等。

人名類：參考《居延漢簡人名編年》李振宏，孫英民著，中國社會科學出版

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

社，《百家姓》作為姓氏參考，以及相關簡牘的人名研究類文獻期刊中統計的人名、姓氏等。

職官與機構類：參考網絡資源中的漢代職官表、機構表。

干支與年號類：參考網絡資源中的漢代年號表。

(5) 其他

由於資源條件限制，還有一些材料，在整理提取其信息后可用於詞典製作的，尚未被收錄入基礎詞典中。這部分資料在今後還會不斷被補充收錄進入基礎詞典，用於完善詞典豐富性，提高分詞正確率。待整理收錄的材料如下表所示：

表格 10

文献名称	文献类型
說文解字	古代字書
玉篇	古代字書
廣韻	古代字書
爾雅	古代雅書
康熙字典	古代字書
故訓匯纂	古代訓詁
釋名	古代字書
篆隸萬象名義	古代字書
金文形義通解	出土文獻詞典
睡虎地秦簡語義詞典	出土文獻詞典
包山楚簡語義詞典	出土文獻詞典
郭店楚簡語義詞典	出土文獻詞典
呂氏春秋詞典	出土文獻詞典
史記辭典	出土文獻詞典
經籍纂詁	古代訓詁
說文段註	古代字書
小爾雅	古代雅書
廣雅	古代雅書
集韻	古代字書

3. 字典合併與導出

將各類詞典進行數據合併后，導出為 TXT 格式的基礎詞典，基礎詞典包含詞頭、詞性、詞性概率三個屬性

第三節　自動化分詞與詞性標註

通過比較，本文的自動分詞採用了基於 Python 的“結巴”中文自動分詞系

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

統[10]。選用該系統主要基於以下幾點考慮：1.該分詞系統支持用戶自定義詞典。2.該係統採用了基於規則、基於詞典、基於統計等多種方法相結合的分詞方式，正確率較高。3.該係統操作接口友好、系統完全免費、支持繁體分詞等。系統支持三種分詞模式：

精確模式：試圖將句子最精確地切開，適合文本分析。

全模式：把句子中所有的可以成詞的詞語都掃描出來，速度非常快，但是不能解決歧義

搜尋引擎模式：在精確模式的基礎上，對長詞再次切分，提高召回率，適合用於搜尋引擎分詞。

以現代漢語語境下舉例說明，例如對於“我来到北京清华大学”這句話，經測試，三種分詞模式返回的結果如下：

精確模式：我/來到/北京/清華大學

全模式：我/來到/北京/清華/清華大學/華大/大學

搜尋引擎模式：我/來到/北京/清華/華大/大學/清華大學

根據實際需要，本文中選擇精確模式進行分詞。當然該分詞系統也有其不足之處，因為其算法是針對現代漢語設計的，因此在分析漢簡語料的時候，正確率不如分析現代漢語那麼高，有一些分詞規則也不盡完善。這也正是本文需要著重解決的問題，即如何在現有的現代漢語分詞系統下，通過改進算法，自定義規則，使其對簡帛類語料的分詞結果更加精確。

一、 分詞算法介紹

1. 基於詞典的分詞算法

這是一類較為機械的分詞方法，基本原理就是將製作好的詞典中的字串，即已登錄詞，放在語料中進行掃描匹配，如果發現字串的子串和詞相同，就算匹配。不難理解，這種演算法的分詞精確度很大程度上取決於詞典的完善程度，如果能得到完善精確的詞典，這類算法就將具有較高的精確度。在對古代漢語運用詞典分詞算法時，主要是用於匹配其中的多音節詞，因為古代漢語以單音節詞為主，

[10] 結巴分詞系統網站：https://github.com/fxsjy/jieba

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

所以除字典匹配出的多音節詞外，基本可以判斷為單音節詞。

按照不同的對文本的掃描方向和詞條匹配的精度，基於詞典的分詞算法又可以劃分為正向掃描、逆向掃描、最大匹配和最小匹配。舉例來說，本文中採取的"正向最大匹配規則"的簡單過程是：在系統中設定好已經製作好的用戶詞典，從被切分的語料中，按正方向順序截取一個定長的字串，這個字串的長度，叫做最大詞長。把這個具有最大詞長的字串與詞表中的詞相匹配，若匹配成功，則可確定這個字串為詞，電腦程序的指針向後移動匹配成功串長度個字元，繼續進行匹配；否則，則把該字串長度減一，再與詞表中的詞進行匹配，直至成功為止。[11]

基於詞典的分詞方法優點是：

1) 不需要大量的訓練語料庫，代碼編寫容易，開發週期短。

2) 比較符合自然語言的構成方式，能夠概括的統計出基本詞彙，統計結果符合語言習慣。

3)對於某些特殊的歧義組合，可以通過對規則設置、對搭配的細緻區分、對詞語所處語境的特征等獲得很好的排歧效果。

缺點是：

1)對於未登錄詞，即詞典中沒有出現的詞，當然也就處理不好。

2) 處理歧義的能力差。歧義的例子很簡單，在現代漢語中："長春市/長春/藥店"和 "長春/市長/春藥/店"是對同一句話不同的的分詞方法，這樣就產生了歧義。對於古代漢語來說，同樣存在歧義問題，例如對於多音節詞的不同切割劃分導致的歧義問題。此外，對於詞彙之間的邊界如何確定，詞彙分割的精度如何統一等問題，也無法很好地解決。

因此，基於詞典的分詞方法往往還要結合規則、統計等方法來使用。

本文採用的“結巴”中文分詞系統，支持用戶自定義詞典，並且採用了正向最大匹配規則，能夠基於詞典實現高效的詞條掃描，生成句子中漢字所有可能成詞情況所構成的有向無環圖 (DAG)，並且採用了動態規劃查找最大概率路徑，找出基於詞頻的最大切分組合，在同類分詞系統中具有相對較高的正確率。所選用的詞典即為本章第二節介紹的基礎詞典。

[11] 代建英. 汉语自动分词系统的研究与实现[D].重庆大学,2005.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2. 基於規則的分詞算法

基於規則的分詞算法是基於字典算法的輔助，是根據語料特點，通過設定的一系列識別規則，使分詞結果更加準確可靠的方法。

在本文中，具體採用了以下規則：

(1) 對未識別詞以單音詞處理

古漢語與現代漢語最大的不同，在於古漢語中的單音詞佔了絕大多數，只有少量的詞彙是以多音節詞形式出現，且大多為專有名詞。基於古代漢語語法特點中單音詞占絕大多數這一事實，本文作出假設，即被處理的語料中除去可匹配出的多音節詞外，其他的未識別詞統一以單音詞處理。因此設置規則：除分詞系統匹配出的多音節詞外，其餘未識別詞均單字切分，成為單音節詞。這樣的假設雖然會存在誤差，但是基本符合古代漢語的實際情況，也便於實際操作。

(2) 人名的識別規則

以西北漢簡作為主要的人名識別材料，通過分析發現，在西北漢簡中，人名的出現具有一定的規律。除了專門的名籍簡可以析出人名外，作者發現人名往往出現在一些特定的詞彙之後，例如“隧長”、“亭長”、“候長”、“卒”等 職官后，其次還出現在“年”字的前面。

例如：

元康𤼵令史昌敢言之。遣亭長宋建□□□□大守府與從者富里王 《居延漢簡》37.56

□□市陽里上遁王福年六十長七尺二寸黑色 《居延漢簡》14.13

根據這一發現，設定規則：凡是出現在職官詞之後，或“年”字之前的詞彙，優先預判斷為是人名，然後進行比對基礎詞典中的人名常用字詞典、百家姓詞典等，確定是否為人名類詞彙。

(3) 地名識別規則

同樣在西北漢簡中，里名、縣名等地名類專有名詞也有一定的規律，即通常這類地名都以“里”、“縣”、“郡”等結尾。例如：

□□市陽里上遁王福年六十長七尺二寸黑色 《居延漢簡》14.13

因此設立規則，以“里”、“縣”、“郡”、“隧”等行政單位結尾的詞彙，優先預判斷為地名。在完善詞典地名類詞彙的基礎之上，通過這一規律可以更好地識

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

別出地名類詞彙。

3. 基於統計的分詞算法

詞典的詞畢竟是有限的，如果只用機械的字符匹配的方法，即簡單的“是”與“非”來進行分詞，會發現有大量的未登錄詞無法識別，也無法很好的消除匹配過程中的歧義問題。因此，還需要藉助基於統計的分詞算法，也就是人們常說的機器學習，或模式識別算法。簡單來說，基於統計的分詞方法就是根據人工標注好的詞性和統計特徵，對中文進行建模，即根據標注好的語料對模型參數進行估計，即訓練模型。在分詞階段再通過模型計算各種分詞出現的概率，將概率最大的分詞結果作為最終結果。常見的序列標注模型有 HMM（隱馬爾可夫模型）和 CRF 模型。

基於統計的分詞算法的優點：

1) 根據數學理論基礎，提供了消歧的方式。

2) 運用大規模語料庫更容易，大規模語料庫能提供足夠的實例模型化知識。

3) 如果訓練的語料足夠大，能更客觀反映語言學中的規律。

缺點：

1) 對古漢語語言的處理和表示比較不適用。

2) 需要大規模的標注語料庫，人工工作量較大。

3) 分詞速度較慢。

“結巴”中文分詞系統，採用了已經被訓練好的 HMM（隱馬爾科夫） 模型，並使用了 Viterbi 演算法。“隱馬爾科夫模型是現代語音辨識系統中構建統計模型的重要手段，現在 HMM 模型已經廣泛的應用到自然語言處理領域”，Viterbi 演算法是其中的一種算法，由於該模型和算法的數學理論非常複雜，本文在這裡不做過多介紹。基於統計的分詞算法除了有助於自動分詞以外，對於詞性標註也是極為重要的，因為存在大量的一詞多義的情況，在這種情況下，需要藉助訓練好的模型，自動根據訓練時所用語料的搭配情況做出智能的判斷，依據搭配概率最大化原則進行詞性標註。

總的來說，本文所採用的“結巴”分詞系統這一工具，綜合運用了以上三種算法，通過基於詞典的分詞算法、基於規則的分詞算法和基於統計的分詞算法相

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

結合的方式，對漢簡語料庫的語料進行自動化分詞。

二、 詞性自動化標註方法

對古漢語詞性的自動化標註是本文研究的重點和難點。

詞性標註之所以重要，是因為只有對分好詞的語料進行進一步的詞性標註，才使得語料有了語義層面的現實意義，才有可能實現下一步的語義分析、詞彙類型研究等工作，只有進行詞性標註，才能為我們進一步提供由大數據支撐的各類統計結果，否則對材料的自動化處理就是沒有實際意義的。

詞性自動標註之所以是難點，是因為漢語詞彙，尤其是古漢語詞彙的特殊性。不同語境下的一詞多義、一詞多詞性、詞類活用、詞類兼用等複雜的情況，人工標註尚且會有不同意見，對於計算機自動化標註工作更是極大的挑戰，因此這也是中文詞性標註研究領域一直以來的難題。本文設計的詞性標註方法，是對古漢語詞彙詞性標註的一次探究和實驗，還存在一些不科學不合理的地方，針對出現的問題，作者會在下一節對標註結果進行具體的分析以及提出相應的改進意見。

詞性標註的核心，是依據已經建好的用戶詞典中的詞性進行匹配。在上一節中已經提到，製作完成的基礎詞典包含了詞條、詞性、詞性概率等幾項基本信息，因此在分詞完成後，就將按照一定的先後順序，對所分詞彙依據詞典中的詞性進行詞性標註。

1. 專有名詞詞性標註。

專有名詞由於義項較為特殊，因此分詞結果相對來說比較明確清晰，詞的詞性大多數情況下比較單一，很少出現一詞多詞性的現象。因此，凡是從詞典中匹配出的專有名詞，直接標註詞典中的詞性。對於未從詞典中識別出，而是根據規則識別出的專有名詞，則標註相應規則下的詞性（如人名、地名等）。

專有名詞的標註相對較為簡單，準確率也比較高。

2. 其他類詞彙詞性標註。

除專有名詞以外，其他的諸如名詞、動詞、虛詞、形容詞、副詞、代詞、助詞等等詞彙，標註起來較為複雜和困難，準確率相對來說也較低。首先，古漢語材料中詞彙一詞多義、一詞多用現象十分常見，因此在標註詞性時還需要考慮上下文搭配等因素才能獲得較為準確的結果。其次，已經建成的基礎詞典中，有部

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分詞彙屬於單一詞性詞彙，但是更多的詞條屬於多詞性詞彙，一個詞後面可能標註著三到五種不同的詞性，在選擇詞性的時候同樣要依據詞彙所處的上下文環境。因此，語料和詞典中都存在的一詞多義現象是造成詞性標註正確率偏低的主要問題。

針對這一問題，在具體實驗中，對於詞典中為單一詞性的詞，就直接標註該詞為詞典中的詞性。對于有多種詞性的詞彙，詞性標註系統一方面會參考人名、地名等專有名詞標註規則的算法，另一方面利用訓練好的 HMM 模型對詞性進行自動化詞性標註。除此之外，我們會參考詞典中已標註好的詞性出現的概率，對於模型無法判斷的詞彙優先選擇其出現概率較大的詞性進行標註。

對於在以上步驟后依然無法標註的，則詞性標註為“未識別”。

三、 實驗過程

1. 語料預處理與詞典製作。該過程在前文已有詳細描述。

2. 安裝“結巴”分詞系統

在計算機中配置好 Python 語言環境后，對分詞系統進行安裝。該係統提供三種安裝方式：

全自動安裝：調用 easy_install jieba 或者 pip install jieba / pip3 install jieba

半自動安裝：先下載 http://pypi.python.org/pypi/jieba/ ，解壓後運行 python setup.py install

手動安裝：將 jieba 目錄放置於目前的目錄或者 site-packages 目錄通過 import jieba 來引用

本文採用半自動安裝模式。

3. 添加用戶詞典

用法： 系統中 jieba.load_userdict(file_name) # file_name 語句為加載自定義詞典的文件路徑。詞典的格式為，一個詞占一行；每一行分三部分：詞語、詞性、詞頻（即概率），用空格隔開。file_name 必須為 UTF-8 編碼。同時，可自定義詞典匹配的順序。在本次試驗中，詞典匹配的順序為：先匹配專有名詞詞典，再匹配專有名詞外的多音詞詞典。

4. 調整詞典

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

本系統支持用戶通過自定義詞典來增強歧義糾錯能力。

使用 add_word(word, freq, tag) 和 del_word(word) 兩個語句，可在程式中動態修改詞典。使用 suggest_freq(segment, tune=True) 語句可自行調節單個詞語的詞頻，使其能（或不能）被分出來。

5. 訓練 HMM 模型

利用約 300 條已經分詞、標註好詞性的熟語料（手動標註的《敦煌漢簡》語料），對分詞系統中的 HMM 隱馬爾科夫模型進行訓練。

6. 編寫規則代碼

為分詞系統系統設定分詞規則，例如未識別詞以單音詞處理規則、人名識別規則、地名識別規則等。

7. 進行分詞

調用結巴分詞系統進行自動分詞。主要用到了 jieba.cut 語句。

jieba.cut 方法接受三個輸入參數: 需要分詞的字串、cut_all 參數用來控制是否採用全模式、HMM 參數用來控制是否使用 HMM 模型。在本次試驗中 cut_all 參數為否，HMM 參數為是。

待分詞的字串即待分詞文本可以是 unicode 或 UTF-8 字串、GBK 字串。具體在本次試驗中採用的是 UTF-8 字符串。

分詞算法同時綜合了基於詞典、基於規則和基於統計的算法，即首先進行詞典匹配分詞，然後依據人名、地名規則和 HMM 模型繼續分詞，最後對於依舊無法匹配識別出的詞語，按單音詞進行切割。經過這樣的步驟后，基本能將詞典中的已登錄詞和部分人名地名較為精準的分出來，剩下的多為單音節詞。

jieba.cut 返回的數據結構是一個可反覆運算的生成機制，可以使用 for 循環來獲得分詞後得到的每一個詞語，或者用 jieba.lcut 直接返回 list。本次試驗即採取直接返回 list 的模式，直接生成一個分好詞的 TXT 文檔。

8. 詞性標註

在進行分詞的同時，進行詞性標註工作，通過編寫代碼完成。具體步驟是:

1) 對於能夠在詞典中匹配出的詞語

無論是單音詞還是多音節詞，如果詞典中該詞條只有唯一詞性，則標註該詞

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性。如果是多義詞，則利用訓練好的 HMM 模型進行自動選擇。如模型仍舊無法確定詞性的，則按照詞典中詞頻概率分佈，預測該詞詞性。例如對於“阿”這個詞的標註，依據上文提到的詞頻概率統計結果，詞性為名詞、動詞、形容詞的概率分別為 0.5、0.33、0.17。因此，例如有 10 次無法確定詞性的標註，通過程序控制，使其被標註為名詞的頻率為 5 次，標註為動詞的為 3 次，標註為形容詞為 2 次。選擇這樣的標註方法主要有兩點考慮：1.HMM 模型無法標記的詞語並不多，因此只有少量詞彙會出現需要依據概率標記詞性的情況，對整體正確率影響較小。2.這個依據概率預測詞性的方法對於某一次標註來說并不準確，但是對於整篇文章來說，某個詞在大多數情況下被標記為常用詞性，少部分情況被標記成不常用詞性，可以預見仍然有較大的可能性大部分詞語標註結果是正確的。當然這個方法嚴格來說不是十分的科學，從結果來看，對標註結果的正確率也有一定的影響，在後文的實驗結果分析中將提出具體的改進方法。

2) 對於未被詞典匹配出的詞語（數量不多）

在完成分詞后，嘗試依據訓練好的 HMM 模型根據前後詞語搭配進行詞性預測，此種預測會比較容易出錯。

3) 以上兩步仍未確定詞性的，標記詞性為“未識別”。

經過以上步驟，就完成了對語料的分詞以及詞性標註工作。整個實驗的前期準備時間較長，但利用系統進行分詞和詞性標註的過程大概只持續了二十分鐘左右。返回的結果以 TXT 文檔形式存儲，形式如下圖 23 所示：

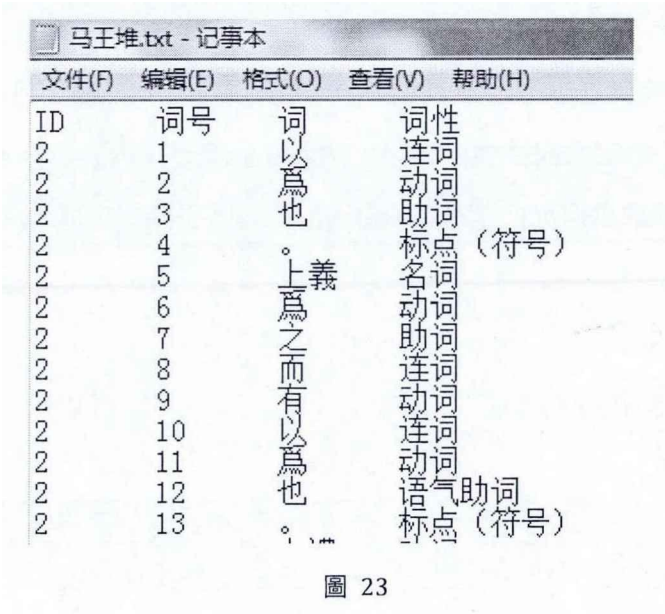


圖 23

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中，ID 為該詞在原簡中的唯一標記簡號 ID，詞號是該詞在該條簡中的詞序，後面是分好的詞以及標註的詞性。

在得到標註結果后，需要把 TXT 文本導入數據庫中，進行下一步的結果分析、結果統計等研究。

第四節　標註結果分析及改進方法

一、　標註結果分析

上述實驗步驟完成後，得到了六張分詞、詞性標註完成的數據表。下面將分為西北漢簡和古書類漢簡兩類材料，對標註結果進行分析。其中對西北漢簡，將重點分析其中的人名、地名、職官這幾類專有名詞的識別情況，以及所出現的問題，并將統計結果做成詞表。對於古書類漢簡，根據其特點將重點分析虛詞這類詞彙的識別情況，并分析存在的問題。

由於詞彙數量龐大，因此對於正確率較低的人名類詞彙，採用抽樣統計的方法對標註結果進行分析，即在西北漢簡材料中隨機選取十至二十條符合要求的簡牘，考察分詞及詞性標註的實際情況。簡牘選取需滿足一定的要求，標準為：首先，選取字數較多、內容較為完整的簡牘，防止因為內容殘缺、字數較少導致的結果誤差。其次，選取內容較有代表性的、符合一些判定規則的簡牘，以更好的考察對於分詞規則的執行情況。

對於識別正確率較高的地名類、職官類專有名詞詞彙，和古書中的虛詞類詞彙，則不再一一舉例分析。

說明：詞彙標註結果及分析以表格形式呈現，每一條例子包含了簡牘編號、原簡釋文、部分分詞及詞性標註結果、結果分析這幾項內容。為了節省篇幅，部分較長釋文只選取需研究的部分的分詞結果，同時去除了分詞結果中所有的標點符號。

1. 西北漢簡

1) 人名類詞彙識別情況

《敦煌漢簡》

簡号	原文	词	词性	结果分析
0046	多問陳司馬、航司馬，願數數	多	副词	1.“陳司馬”一詞分

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

	相聞，為檄欲移鄯善，毋使行也。	問	动词	詞錯誤，應為一個詞2. 此處司馬應為 職官，不是人名
		陳司	人名	
		馬	名词	
		航	名词	
		司馬	人名	
0243	原匡叩頭白：謹使卒張常奉記，叩頭再拜白，……因卒張記	原匡	名词	“卒”后人名識別正確
		叩頭	动词	
		白	形容词	
		謹使	名词	
		卒	副词	
		張常	人名	
		奉記	名词	
		叩頭	动词	
		再拜	动词	
		白	形容词	
		因卒	名词	
		張記	人名	
0250	厭胡卒，龍勒萬年里，孫良，年卄七。	厭	名词	“年”字前面人名識別正確
		胡卒	名词	
		龍勒萬年	名词	
		里	名词	
		孫良	人名	
		年	名词	
		卄	名词	
		七	数词	
0252	戍卒，效穀常利里，張詡，年三十。	戍卒	名词	“年”字前面人名識別正確
		效	动词	
		穀	名词	
		常利	地名	
		里	名词	
		張詡	人名	
		年	量词	
		三	数词	
		十	名词	
0848	東聞喜利里，魏延年	東聞	名词	根據詞典匹配出正確人名
		喜利里	名词	
		魏延年	人名	
1164	粟望之長，趣李順依將來。	粟望	名词	“長”字后人名識別正確
		之長	代词	
		趣	名词	
		李順	人名	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

		依將	名词	
		來	动词	
1403	青堆隧長，效穀宜禾里，李廣利。　　始元七年五月戊戌，除。	青堆	名词	“隧長”及職官詞后的人名，能夠正確判定。
		隧長	职官	
		效	动词	
		穀	形容词	
		宜禾	名词	
		里	名词	
		李廣利	人名	
		始元	年号	
		七年	时间词	
		五月	时间词	
		戊戌	干支词	
		除	动词	
2394	受王猛堅書一封。孫小六日付鄭。	受	动词	動詞后接人名，識別正確
		王猛	人名	
		堅書	名词	
		一封	数词	
		孫小	名词	
		六日	时间词	
		付鄭	名词	

《居延漢簡》

数据库简号	释文	词	词性	結果分析
29.1	永光四年正月己酉 橐佗延壽郃長孫時符妻大女昭武萬歲里孫第卿，年廿一。 子小女王女年，三歲。 弟小女耳，年九歲。 皆黑色	永光	年号	職官詞后搭配人名
		四年	数词	
		正月	时间词	
		己酉	干支	
		橐	名词	
		佗	代词	
		延壽	人名	
		郃	名词	
		長	形容词	
		孫時	人名	
		符	名词	
33.3	入粟大石廿五石　十二月乙亥令史□受陽	入粟	名词	地名后搭配人名
		大石	名词	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

	里王宣	廿五	数词	
		石	名词	
		十二月	数词	
		乙亥	干支	
		令史	职官	
		受陽里	名词	
		王宣	人名	
37.56	元康令史昌敢言之。遣亭長宋建□□□□□大守府與從者富里王	元康令	名词	職官詞后搭配人名
		史昌	名词	
		敢言	名词	
		之	代词	
		遣亭長	名词	
		宋建	人名	
		大守府	名词	
		與	介词	
		從者	名词	
		富里	名词	
		王	名词	
40.29	□□問諸大夫曰□□諸大夫之論莫及寡人也，居有閒而三稱之。吳起進對曰：不審亦	問諸	名词	依據詞典匹配人名正確
		大夫	名词	
		曰	动词	
		諸	副词	
		大夫	名词	
		之論	连词	
		莫及	地名	
		寡人	名词	
		也	语气词	
		居有	名词	
		閒	名词	
		而	名词	
		三稱	名词	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

		之	助词	
		吳起	人名	
		進對	名词	
		曰	助词	
		不審	名词	
		亦	副词	
55.8	執胡郃卒張平　十二月食	執胡	名词	“卒”字后搭配人名
		郃	名词	
		卒	动词	
		張平	人名	
		十二月	数词	
		食	动词	
57.15	次吞郃長徐當時，卒四人　卒犢利親　卒祝自為　卒吳道　用粟十石	次	量词	“卒”字后搭配人名，有兩個人名未被正確分詞，“犢利親”和“祝自為”
		吞	动词	
		郃	名词	
		長	名词	
		徐當時	人名	
		卒	副词	
		四人	名词	
		卒	名词	
		犢	名词	
		利親	名词	
		卒	名词	
		崔利	人名	
		卒	动词	
		祝	动词	
		自為	名词	
		卒	动词	
		吳道	人名	
		用粟十石	名词	
58.2	居延甲渠第廿八郃長	居延甲渠	名词	職官詞后搭配人名

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

	居延始至里大夫孟憲，年廿六	第	副词	
		廿八	数词	
		郤	名词	
		長	名词	
		居延始	名词	
		至	动词	
		里	量词	
		大夫	名词	
		孟憲	人名	
		年	名词	
		廿六	数词	
142.35	□石　五月辛亥甲渠令史誼受當遂里張容	石	量词	“誼”作為人名未被正確識別；“張容”識別正確
		五月	时间词	
		辛亥	干支	
		甲渠	名词	
		令史誼	名词	
		受當	名词	
		遂	名词	
		里	名词	
		張容	人名	
122.17	給尉史郭常	給尉史	名词	職官詞后搭配人名
		郭常	人名	
127.35	以牒驗問官久，故吏令史范弘	以牒	名词	
		驗問	名词	職官詞后搭配人名
		官久	名词	
		故吏	名词	
		令史	职官	
		范弘	人名	
131.1	日　水都尉李由勞二歲五月二日	日	名词	職官詞后搭配人名
		水	名词	
		都尉	职官	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

		李由	人名	
		勞二歲	名词	
		五月	时间词	
		二日	时间词	
132.24	史鄭忠南界士吏張放守令史張武第一守右尉	史	名词	職官詞后搭配人名
		鄭忠	人名	
		南界	名词	
		士吏	名词	
		張放	人名	
		守令史	名词	
		張武	人名	
		第一	数词	
		守右尉	名词	
511.28	隧長李廣 元鳳二年十月辛亥除	隧長	职官	“隧長”后搭配人名
		李廣	人名	
		元鳳	年号	
		二年	数词	
		十月	时间词	
		辛亥	数词	
		除	名词	

《肩水金關漢簡》

新简号	釋文	词	词性	結果分析
38_T01_042_0	日勒丞王勝， 馬二匹軺車二乘， 小史脩里王奉光	日勒丞	名词	依據詞典匹配，人名匹配正確
		王勝	名词	
		馬	名词	
		二匹	数词	
		軺	名词	
		車二乘	名词	
		小史	名词	
		脩	动词	
		里	名词	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

		王奉光	人名	
38_T01_078_0	日勒騎士延壽里張定	日勒	地名	里名后搭配人名，識別結果正確
		騎士	名词	
		延壽里	名词	
		張定	人名	
38_T02_048_0	沙上卒武強	沙上	名词	“卒”字后搭配人名
		卒	动词	
		武強	人名	
38_T05_018_0	戍卒魏郡鄴呂廣里士五馮長卿年	戍卒	名词	正確分詞應為“馮長卿”，分詞結果錯誤
		魏郡鄴	名词	
		呂廣里士	名词	
		五	数词	
		馮長	人名	
		卿	名词	
		年	名词	
38_T05_051_0	戍卒鉅鹿郡南蕝里張定年卅三	戍卒	职官	地名后搭配人名，識別正確。“南蕝里”作為一個完整的地名，分詞錯誤
		鉅鹿郡	地名	
		南	名词	
		蕝	名词	
		里	名词	
		張定	人名	
		年	名词	
		卅三	数词	
38_T10_246_B	白奉業卿， 賜明教弟李譚	白奉業	名词	依據詞典匹配，分詞標註正確
		卿	名词	
		賜	动词	
		明教	名词	
		弟	名词	
		李譚	人名	
38_T21_086_B	張忠字子仲 王充字少	張忠字子	名词	1.“張忠”未能正確識別 2.“王充“識別正確
		仲	形容词	
		王充	人名	
		字少	名词	
38_T23_017_A	居延都尉守屬郭良	居延	地名	依據詞典搭配正確

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

		都尉	职官	
		守屬	名词	
		郭良	人名	
38_T23_697_0	錢六百，給執適隧長王猛六月奉	錢	名词	“隧長”后搭配人名，結果正確
		六百	数词	
		給執適	名词	
		隧長	职官	
		王猛	人名	
		六月	时间词	
		奉	动词	
38_T23_944_0	受張忠四百七十	受	动词	“受”動詞后搭配人名，標註結果正確
		張忠	人名	
		四百	数词	
		七十	数词	
38_T06_028_0	東郡博平都鄉佐麥里公乘李安世年廿四長七尺四寸黑	東郡	地名	依據詞典搭配，結果正確
		博平	地名	
		都	动词	
		鄉佐麥里	名词	
		公乘	名词	
		李安世	人名	
		年	名词	
		廿四	数词	
		長	名词	
		七尺	数词	
		四寸	数词	
		黑	形容词	
38_T06_052_0	萬歲里公乘藉忠年卌八， 為姑臧尉徐嚴葆與嚴俱之官， 正月庚午入， 丿	萬歲	数词	職官詞后搭配人名，分詞正確
		里	名词	
		公乘	名词	
		藉	动词	
		忠	形容词	
		年	名词	
		卌	数词	
		八	名词	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

		為姑	名词	
		臧尉	名词	
		徐嚴	人名	
		葆	动词	
		與	介词	
		嚴俱之官	名词	
		正月	时间词	
		庚午	干支	
		入	名词	
		丿	名词	

通過以上抽樣統計結果來看，人名類詞彙的分詞和詞性標註確實較為複雜。由於人名用字豐富，用法不固定等自身特點，導致人名識別的過程中會出現分詞不準確、單字名字無法有效識別、姓氏和名字切割錯誤、名字和職官無法有效區分等問題。

從結果來看，對於職官類詞彙或地名類詞彙后搭配人名這種情況，該係統具有較高的正確率。而對於其他情況下，該係統仍存在較多的錯誤匹配。

根據詞彙標註結果，作者整理了西北漢簡人名錶，包含了識別出的超過 1200 個人名，詳細的人名錶參見附錄一。同時通過去重、分析、統計這些姓名，得出以下簡略結論：

(1) 姓名字數方面。姓名字數為兩個字的名字最多，約 685 個。其次是姓名為一個字的名字，約 290 個。姓名為三個字的有 140 個。姓名為四個字的有 2 個。

(2) 在姓氏方面。排名前五的姓氏為王、張、徐、李、孫，分別出現的頻次為 74 次、58 次、38 次、37 次和 27 次。具體的姓氏頻次表參見附錄二。

2) 地名類詞彙、職官類詞彙識別情況

由於地名、職官名屬於兩類用詞比較固定，用法上不易產生大的變化的專有名詞，通過對詞典中地名類詞彙和職官類詞彙的匹配，能夠準確識別出大部分的已登錄詞。例如：

《敦煌漢簡》.0025

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

隧長,	效	穀	閒田	常利	里,	上	牛康,	年	三十五
职官	名词	形容词	名词	地名	名词	动词	名词	名词	数词

又如：

《居延漢簡》.3.6

隧長	徐宗	自言責	故	三泉	亭長	石延壽	茭錢少	二百八十,	數責	不可。
职官	人名	名词	副词	地名	职官	人名	名词	数词	名词	动词

在對標註結果進行統計分析后,發現這兩類專有名詞具有較高的匹配正確率，經過抽樣統計，對這兩類詞彙的識別正確率達到了 85%以上。根據統計結果，整理出了“西北漢簡地名錄”和“西北漢簡職官表”，參見附錄三、附錄四。

2. 古書類漢簡

虛詞是古漢語的重要特征，古虛詞是在漢語中沒有實際意義的字。虛詞不能獨立成句，只有配合實詞來完成語法結構。虛詞對實詞有協助作用，這類詞包含介、連、助、歎、副、象聲六大類，雖然這類詞數量較少，但是卻在古漢語中有著舉足輕重的地位。常見的古漢語虛詞為以下 18 個：

而、 何、乎、乃、其、且、若 、所、為、 焉、也、以、因、 於、與、則、者、之

古書類漢簡中有大量的虛詞用法，因此對於虛詞的準確匹配是非常有必要的。實驗中採取的虛詞分詞方法主要是基於詞典的分詞方法，即以上 18 個詞基本上都判定為虛詞詞性（具體可判定為介詞、連詞、助詞等詞性）。因此分詞的結果是較為準確的，如下例所示：

《馬王堆漢簡》老子甲本・德經 002

以	爲	也	上義	爲	之	而	有	以	爲	也	上禮	爲	之	而	莫之應	也	則	攘臂	而	扔	之	故失	道	矣
虛词	动词	语气助词	名词	动词	代词	虛词	动词	虛词	动词	助词	动词	动词	代词	连词	未知	语气助词	连词	动词	连词	动词	代词	名词	名词	语气词

由標註統計結果來看，對於古書類的虛詞詞彙，本文所採用的分詞及詞性標註方法是有較高正確率的，正確率達到了 80%以上。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、 改進方法

根據對標註結果的分析，除了地名類、職官類、虛詞類詞彙的正確率較高外，其他類型詞彙由於用法非常複雜，例如動詞、名詞、形容詞、副詞、人名類詞彙等，正確率還有待提高。基於以上情況，提出以下幾點改進方法。

1. 進一步完善用戶詞典。將為完成錄入的詞典材料盡快納入到基礎詞典中。從而提高在基於詞典分詞時的分詞正確率。
2. 對自動分詞和詞性標注的算法進行優化。具體來說就是設定更多的外部規則，例如人名、地名、職官等詞彙的分詞規則，使得分詞的結果更加精準可靠。
3. 解決 HMM 模型的訓練集過小問題。應該使用大量的熟語料對 HMM 模型進行訓練，訓練集越大，模型的精確度就越高，相應的標注結果就更加準確。
4. 詞性標註中採取的基於概率的標註方法錯誤率較高，應該嘗試通過詞彙前後搭配詞的詞性、詞義等線索預測詞性，結果會更加準確。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第四章 結語

近年來許多高校機構都開始展開了漢語古籍的電子化、數字化工作。目的就是利用不斷發展的計算機技術，為古漢語研究提供便利，從而激發出新的活力和創造力。漢代簡牘作為漢語古籍中非常重要的一類，在整個古文字研究中有著十分重要的地位，其大量的數據基礎和較為完整的語料材料，都非常適合進行數字化處理。本文作為應用型研究，本著解決實際問題的原則出發，選取漢代簡牘這類材料，探索實現了漢簡語料庫的建設和在自動分詞、詞性標註這方面的應用，並將設計思路和設計過程進行記錄。

第一，本文首先對現有的電子語料庫進行了梳理，并總結出其各自的優點與不足之處，在借鑒已有的研究成果基礎之上，結合實際學習中遇到的需求，對語料庫的功能提出了新的改進方案。

第二，通過運用 Access 數據庫技術和 ASP 網頁編程技術，從語料庫的數據庫平台和網絡平台兩方面對其進行了具體的需求梳理、邏輯設計、問題解決和功能實現的過程。經過不斷努力，目前已經基本建成了一個數據基礎龐大，功能完善，操作簡便，方便實用的漢簡語料庫。可以為更多的專家學者提供研究上的便利。目前的語料庫已經具備了各家考釋檢索，引得、字形編打印，詞性檢索，簡牘拓片顯示，著錄信息顯示等一系列方便快捷的功能。

第三，應用語料庫進行材料深層加工，運用現代漢語分詞、詞性標註的技術支持和理論模型，結合古漢語語言特點，探索了自動化分詞和詞性標註的方法。經過設計和實驗操作，完成了西北漢簡和幾種古書類材料的自動分詞和詞性標注。並通過分析實驗結果，對試驗方法提出了改進意見。同時對標注出的人名、地名、職官等專有名詞進行了統計分析。為漢簡材料自動化分詞研究提供了更充分的理論、資料支援，也為解決人工材料加工存在的效率低、工作量大等實際問題找到了解決辦法。

當然，語料庫還存在許多的問題和改進空間。本應實現更多的功能，但由於時間和作者本人的能力有限，還有更多的工作需要繼續完成。論文的不足之处主

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

要在以下幾點：1.由於時間原因，語料庫收錄的材料截止至 2014 年及以前發表的，2014 年後的材料還有待補充。2.漢簡語料庫的數據庫設計部分還有待進一步完善，例如同義詞、反義詞的表示方法，詞彙搭配的表示方法等。3.漢簡語料庫網路平臺的一部分功能只完成了設計，還沒有完全實現。4.在自動化分詞和詞性標注研究中，由於現有資源限制，用戶詞典的建設還很不完善，此外在基於統計的分詞方法中，對 HMM 模型未進行充分訓練，導致分詞結果正確率並不是非常高。在詞彙自動化標註領域還有非常多的問題等待作者去探索、解決。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

參考文獻

著作類

[1] 裘錫圭. 文字學概要[M]. 北京：商務印書館，1988 年

[2] 張再興. Access 數據庫在語言文字研究與教學中的應用[M]. 江西:江西高校出版社. 2003 年

[3] 朱德熙. 語法講義[M]. 北京:商務印書館，1982 年

[4] 張三夕 毛建軍. 漢語古籍電子文獻知見錄[M]. 廣州：世界圖書出版社. 2015 年

期刊類

[5] 張再興. 互聯網時代出土文獻數據庫建設的思考與實踐 [J]. 中國文哲研究通訊，2011 年, 21(2)

[6] 洪一梅. 甲骨學研究學術性系統發展芻議：理論與實踐[J]. 中國文哲研究通訊，2011 年, 21(2)

[7] 何志華. 香港中文大學中國古籍研究中心出土文獻數據化研究之回顧與前瞻[J]. 中國文哲研究通訊，2011 年, 21(2)

[8] 程少軒. [上博簡字詞全編資料庫]的設計和製作[J]. 中國文哲研究通訊，2011 年, 21(2)

[9] 魯家亮、李靜. 數位時代的簡帛研究[J]. 中國文哲研究通訊，2011 年, 21(2)

[10] 單育辰. 淺談數位化時代的出土文獻及相關資源資料庫[J]. 中國文哲研究通訊，2011 年, 21(2)

[11] 張再興. 秦漢簡牘語料庫建設中基礎語料的處理標準統一問題[J]. 中國文字研究, 2014, 02:53-56.

[12] 王輝. 漢簡人名“延年”身份考[J]. 南都學壇, 2013, 05:14-21.

[13] 馬孟龍. 談肩水金關漢簡中的幾個地名(二)[J]. 中國歷史地理論叢, 2014, 02:88-92.

[14] 張俊民. 敦煌懸泉出土漢簡所見人名綜述(二)——以少數民族人名為中心的考察[J]. 西域研究, 2006, 04:1-11+117.

[15] 趙久湘. 張家山漢簡《奏讞書》人名研究[J]. 樂山師範學院學

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

報, 2010, 04:58-61.

[16] 趙久湘. 《奏讞書》人名冷僻字淺析[J]. 蘭州教育學院學報, 2010, 01:7-10+138.

[17] 張俊民. 懸泉漢簡所見漢代複姓資料輯考——敦煌懸泉置出土漢簡所見人名綜述(三)[J]. 秦漢研究, 2008, 00:194-204.

[18] 馬孟龍. 談肩水金關漢簡中的幾個地名[J]. 中國歷史地理論叢, 2012, 03:57-63.

[19] 馬智全. 《肩水金關漢簡(壹)》校讀記[J]. 考古與文物, 2012, 06:107-109.

[20] 孫兆華. 《肩水金關漢簡(貳)》所見裡名及相關問題[J]. 魯東大學學報(哲學社會科學版), 2014, 02:72-80.

[21] 羅見今, 關守義. 《肩水金關漢簡(貳)》曆簡年代考釋[J]. 敦煌研究, 2014, 02:109-115.

[22] 黃豔萍. 《肩水金關漢簡(壹)》紀年簡校考[J]. 敦煌研究, 2014, 02:116-121.

[23] 黃豔萍. 《肩水金關漢簡(三)》紀年簡校考[J]. 敦煌研究, 2015, 02:111-118.

[24] 黃豔萍. 《肩水金關漢簡》所見“燧”及其命名探析[J]. 敦煌研究, 2016, 01:116-122.

[25] 劉金嶺. 基於《現代漢語語義分類詞典》的文本聚類方法[J]. 情報雜誌, 2010, 11:170-173+199.

[26] 鐘茂生, 胡熠, 劉磊. 基於詞典詞語量化關係的中文文本分割方法[J]. 電腦工程與應用, 2008, 21:25-29+88.

[27] 晏昌貴. 增補漢簡所見縣名與裡名[J]. 歷史地理, 2012, 00:249-255.

[28] 胥桂仙, 蘇筱蔚, 陳淑豔. 中文文本挖掘中的無詞典分詞的演算法及其應用[J]. 吉林工學院學報(自然科學版), 2002, 01:16-18.

[29] 留金騰, 宋彥, 夏飛. 上古漢語分詞及詞性標注語料庫的構建——以《淮南子》為範例[J]. 中文資訊學報, 2013, 06:6-15+81.

[30] 化振紅. 深加工中古漢語語料庫建設的若干問題[J]. 西南大學學報(社會科學版), 2014, 03:136-142+184.

[31] 石民, 陳小荷, 於麗麗, 李斌. 基於CRF的古漢語分詞標注一體化研究

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[A]. 中國中文資訊學會. 中國電腦語言學研究前沿進展（2007-2009）[C]. 中國中文資訊學會:, 2009:6.

[32] 石民, 李斌, 陳小荷. 基於 CRF 的先秦漢語分詞標注一體化研究[J]. 中文資訊學報, 2010, 02:39-45.

[33] 梁社會, 陳小荷. 先秦文獻《孟子》自動分詞方法研究[J]. 南京師範大學文學院學報, 2013, 03:175-182.

學位論文類

[34] 陳練軍. 居延漢簡量詞研究[D]. 西南師範大學, 2003.

[35] 任攀. 居延漢簡釋文校訂及相關問題研究（居延舊簡部分）[D]. 復旦大學, 2012.

[36] 白軍鵬. “敦煌漢簡”整理與研究[D]. 吉林大學, 2014.

[37] 李瑤. 居延舊簡文字編[D]. 吉林大學, 2014.

[38] 張國豔. 居延漢簡虛詞研究[D]. 華東師範大學, 2005.

[39] 葛紅麗. 《居延新簡》詞語通釋[D]. 華東師範大學, 2007.

[40] 趙小兵. 基於動態流通語料庫的現代漢語基本詞彙自動識別與提取方法研究[D]. 北京語言大學, 2007.

[41] 郝慧芳. 張家山漢簡語詞通釋[D]. 華東師範大學, 2008.

[42] 范董平. 《敦煌漢簡》詞語通釋[D]. 華東師範大學, 2008.

[43] 郭鵬. 漢語語法語料庫系統的基礎設計[D]. 天津師範大學, 2006.

[44] 楊豔輝. 《敦煌漢簡》整理研究[D]. 西南大學, 2007.

[45] 潘景劍. 具有中文人名識別功能的漢語自動分詞演算法研究[D]. 河北工業大學, 2006.

[46] 任達. 《肩水金關漢簡（壹）》文字編[D]. 吉林大學, 2014.

[47] 劉國慶. 《居延新簡》釋文合校[D]. 西南大學, 2008.

[48] 代建英. 漢語自動分詞系統的研究與實現[D]. 重慶大學, 2005.

[49] 劉禹孜. 漢語自動分詞中排除歧義欄位演算法的研究[D]. 重慶大學, 2005.

[50] 熊冬明. 漢語自動分詞和中文人名識別技術研究[D]. 浙江大學, 2006.

[51] 吳雲燕. 馬王堆漢墓帛書通用字研究[D]. 華東師範大學, 2006.

[52] 楊志勝. 基於字根法的古漢語分詞模型[D]. 天津大學, 2007.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[53] 王嘉靈. 以《漢書》為例的中古漢語自動分詞[D]. 南京師範大學, 2014.

[54] 朱曉. 古漢語編年體的人名實體識別與詞性標注[D]. 復旦大學, 2012.

[55] 馮秋香. 基於資料庫語義學的古漢語句法語義分析研究[D]. 大連理工大學, 2012.

網絡資源

[56] 香港中文大學中國文化研究所中國古籍研究中心 [漢達文庫]: http://www.chant.org/info/trial_sub.asp

[57] 中央研究院歷史語言研究所文物圖像研究室 [簡帛金石資料庫]: http://saturn.ihp.sinica.edu.tw/~wenwu/search.htm

[58] 中央研究院歷史語言研究所 [漢代簡牘數位典藏]: http://rub.ihp.sinica.edu.tw/~woodslip/index.htm

[59] 武漢大學簡帛研究中心 [中國古代簡帛字形、詞例資料庫]: http://www.bsm-whu.org/zxcl/index.php

[60] 中國文字研究與應用中心 [秦漢簡牘帛書語料庫]: http://www.wenzi.cn/objects/confs_web/NewWeb/ExpertsLogin.aspx

[61] 結巴分詞系統網站: https://github.com/fxsjy/jieba

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

附錄

附表一 西北漢簡人名表

□充	成則	范昌	郭崇	彊	藺禹
□定昌	成澤	范處	郭次都	焦永	劉武
□威	程宣	范當時	郭當	解□	劉歆
鈕昌	誠	范弘	郭謹	解宮	柳世
愛彭	池信	范護	郭良	解梁	六禹
安	充	范客子	郭市人	解清	隆
安昌	充國	范利中	郭事親	解事	龍世
安國	充輸	范臨	郭外人	解它	魯儅時
安世	崇	范勝	郭政	解萬年	魯侯
安主	儲壽	范同	邯	靳龜	路博
霸	淳于次孺	范翕	函何陽	景臨	路舒
包	淳于良	范翊	韓壽	竟	犖晏
褒	淳于樂	范尊	韓陽	九百翊	呂敞
薄進	淳于子遊	方子惠	闞褒	就	呂弘
豹	賜林	房	漢彊	居	呂護
暴便	崔廣	房岑	郝卿	駿	呂良
鮑小叔	崔黃	房日利	何望	孔光	呂普
鮑永	崔利	放	何武	孔勝之	呂孺
必良	崔自當	豐	何憲	寬	呂尚
便	戴通	馮白	何鎮	寬主	呂憲
丙吉	丹	馮長	和	匡	呂異眾
邴利	單昌	馮弘	賀	況	馬丙
并	單宮	馮彊	弘	曠	馬病已
並	單子巽	馮匡	弘羊	李安	馬定
博	黨	馮利	宏	李褒	馬谷
步遷	得	馮勉	侯倉	李敞	馬良
蔡褒	德	馮同	侯得	李城	萌
蔡崇	登	馮喜	侯賢	李充	孟倉
蔡喜	鄧祿	馮晏	侯禹	李崇	孟昌
參	狄望之	馮音	侯云	李丹	孟充
曹式	丁放	馮章	胡毋	李定國	孟賞
昌	丁宮	奉	胡霸	李奉	孟壽
昌林	丁竟	奉光	胡敝	李鳳	孟憲
長	定	奉親	胡敞	李宮	孟儀
長秋	定破胡	符陽己	胡超	李廣	孟幸
長生	定世	福	胡朝	李廣宗	孟延

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

長詡	定佐博	輔	胡地餘	李何傷	孟延壽
常	東郭昌	輔平	胡政	李鹹	苗訴
常池	董安國	傅勝	許昌	李赦之	眇
常調	董霸	傅育	許放	李生	閔立
常富	董放	富蓋邑	許鎮	李勝	沐得
常業	董輔	富邑	許柱	李識	沐惲
常有	董山	蓋	扈卿	李壽	聶伯
常樂	董習	泠臨	護	李壽王	寧
常卒	董誼	幹嘉	皇楚	李順	寧岑
敞	董雲	榦永	皇隆	李潭	寧常
陳安國	董惲	高何齊	黃禁	李譚	牛慶
陳褒	董忠	高建	黃利	李威	牛延壽
陳長子	董子方	高隆	黃良	李吳	彭
陳當	董子游	高明	黃山	李業	彭陽
陳豐	董子真	高青	黃子公	李欽	彭張成
陳光	董宗	高嚴	霍秀	李由	辟非
陳橫	竇何	革	獲	李有	平
陳尚	杜旦	耿世	畸	李稺君	平晏
陳通	杜福	公孫建	伋	李忠	破胡
陳宣	杜蓋之	公武	籍良	李子長	破奴
陳偃	杜雋	功孫	季宮	李子春	樸裏
陳陽	杜買	恭	季由	李宗	普
陳業	杜未央	躬	紀褒	李遵	齊
陳義	杜詡	宮	紀音	立	齊黨
陳友	段封	龔重	家橫	利親戀	齊憙
陳樂成	段尚	共龍	嘉	良	起
陳宗	段賢	關淳	嘉平	良林	遷
陳尊	段宗	關武	賈良	梁不示	前
諶	段尊	光	賈通	梁充	錢万年
成	兒政	廣	賈望	梁多	淺
成竟	樊奉口	郭并	堅	梁習	強
成赦之	樊立	郭昌	建	林	橋建
成遂	樊隆	郭常	建國	臨	秦恭
成宣	范安世	郭充	江勝客	臨平	勤光

慶	孫地餘	王慶	謝乙	晏望	張放	鄭吉
丘張	孫廣	王戎	辛崇	羊田	張奉世	鄭彊
萩嚴	孫宏	王赦之	辛戎	陽	張福	鄭柳
區	孫級	王審	信	楊霸	張廣德	鄭慶
仁	孫良	王勝之	信都	楊褒	張合眾	鄭少卿
任根	孫臨國	王士吏	刑博	楊充	張護	鄭赦

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

任林	孫猛	王舒	刑齊	楊黨	張獲	鄭湯
任喜	孫明	王彎年	邢	楊君	張嘉	鄭偉
戎	孫時	王望	邢延壽	楊山	張況	鄭習
榮福	孫受	王紋	邢忠	楊壽	張連	鄭孝
如昌	孫壽	王毋何	徐并	楊湯	張臨	鄭歆
山	孫壇雲	王武	徐充	楊武	張令史	鄭陽
商奉世	孫習	王翕	徐充國	楊音	張隆	鄭欽
賞	孫辛	王習	徐當時	楊禹	張曼	鄭忠
上官隆	孫翊	王憲	徐輔	楊元仲	張彭祖	鄭子孟
上官武	孫勛	王孝	徐貴	仰田殷	張平	鄭子真
尚	孫知	王歆	徐弘	羨	張容	直
尚林	孫宗	王翊	徐閎	野	張赦之	郅安世
赦之	覃嬴	王宣	徐惠倩	業	張譚	郅臨
審	譚	王尋	徐就	衣嚴	張桐	郅卿
盛	湯	王延	徐廉	宜王	張未央	郅嚴
勝胡	唐	王晏	徐隆	乙昌	張武	郅宗
勝有	桃勝之	王陽	徐猛	弋南	張翊	置佐博
勝之	騰譚	王殷昌	徐蒲	義	張宣	忠
師就	田林	王禹	徐遷	誼	張勛	忠藏
石伕	田隆	王則	徐卿	音	張循	忠臨
石彊	田彭	王子恩	徐勝客	欽	張偃	仲山
石匡	田惲	王尊	徐勝之	尹便	張業	眾
石隆	同	望	徐譚	尹就	張意	周并
石芹	屯仁	威	徐武	尹勝胡	張義	周駿
時報	脫毋紬	韋朋	徐相	尹野	張欽	周利
時尚	拓	韋受	徐歆	嬰齊	張永	周明
實	外人	隗忠	徐循	永	張禹	周年
史承祿	万	未央	徐延壽	于定國	張垣	周平
史充	万禹	魏便	徐嚴	禹	張惲	周仁
史弘	萬世	魏崇	徐揚	元	張札	周舒君
史監	王安世	魏相	徐業	垣并	張忠	周雅
史臨	王霸	魏延年	徐宜馬	爰應	張宗	周揚
始昌	王褒	魏羽	徐永	原憲	召浦	周育
始長	王并	溫常利	徐忠	願之	召仲	周惲
世	王博	溫申	徐子勝	樂博	趙倗	朱茂
壽王	王長	毋害	徐子禹	樂德	趙常	朱明
壽樂	王敞	毋憂	徐尊	樂護	趙賜	朱普
順	王承明	吾丘受	詡	樂奴	趙輔	朱千秋
司非子	王充	吳道	宣	樂勝之	趙輔進	朱卿
宋充貴	王當	吳建	宣博	云	趙蓋眾	朱武
宋黨	王豐	吳信	宣成	鄖宣	趙閎	騶毋害

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

宋建	王鳳	吳翊	宣建	允	趙回	祝自為
宋竟	王福	吳宗	玄	鄆翊	趙佳	莊宣
宋良	王輔	武	薛昌	臧	趙竟	莊禹
宋普	王恭	武光	薛崇	臧強	趙康	子淵
宋同	王宮	武賀	薛寄	臧禹	趙匡	子元
宋萬	王光	熹	薛君伯	則	趙良	子章
宋宗	王廣	習	薛隆	翟方進	趙臨	宗
蘇長	王弘	憙	薛少卿	翟遷	趙猛	鄒相
蘇敞	王宏	夏侯常	薛誼	詹奴	趙彭	鄒永
蘇蓋眾	王閎	夏侯譚	薛幼蘭	章	趙千	左長壽
蘇海	王護	夏侯掾	薛章	張	趙湯	左隆
蘇慶	王建國	夏良	循	張安	趙忘生	左朔
蘇循	王彊	夏奇	尋	張褒	趙翕	左忠
蘇晏	王禁	夏收	延	張并	趙宣	
蘇誼	王匡	賢	延年	張博	趙宜	
遂	王良	顯	延壽	張常	趙于	
遂昌	王隆	憲	閻少季	張常業	趙忠	
孫安世	王萌	相	閻樂成	張敞	趙壯	
孫昌	王猛	薤良	顏譚	張成	趙子回	
孫成	王明	肖并	嚴	張崇	政	
孫充國	王莫當	孝	嚴鳳	張德	政敞	
孫黨	王奴	謝鳳	嚴複	張定	鄭昌成	
孫地	王遷	謝橫	晏	張多	鄭馮	

附表二 西北漢簡姓氏頻次表

姓氏	頻次	姓氏	頻次	姓氏	頻次
王	74	韓	2	家	1
張	58	皇	2	建	1
徐	38	季	2	江	1
李	37	紀	2	焦	1
孫	27	嘉	2	靳	1
趙	26	九	2	景	1
陳	19	孔	2	寬	1
鄭	18	劉	2	利	1
董	15	魯	2	良	1
范	15	路	2	臨	1
楊	15	沐	2	蘭	1
馮	14	寧	2	柳	1

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

郭	14	牛	2	六	1
周	12	彭	2	龍	1
呂	10	破	2	犖	1
孟	10	齊	2	苗	1
宋	9	時	2	閔	1
薛	9	始	2	聶	1
安	8	壽	2	辟	1
常	8	韋	2	平	1
胡	8	未	2	樸	1
蘇	8	毋	2	錢	1
杜	7	武	2	橋	1
解	7	薤	2	秦	1
成	6	辛	2	勤	1
段	6	刑	2	丘	1
高	6	邢	2	萩	1
延	6	閻	2	榮	1
郅	6	嚴	2	如	1
朱	6	宜	2	商	1
侯	5	臧	2	尚	1
許	5	翟	2	赦	1
黃	5	忠	2	師	1
馬	5	莊	2	司	1
勝	5	鄒	2	遂	1
石	5	鈕	1	覃	1
史	5	愛	1	桃	1
魏	5	薄	1	騰	1
吳	5	暴	1	屯	1
樂	5	必	1	脫	1
左	5	丙	1	外	1
淳	4	邴	1	万	1
淳于	4	步	1	萬	1
崔	4	曹	1	隗	1
樊	4	昌	1	吾	1
何	4	程	1	肖	1
梁	4	池	1	信	1
田	4	儲	1	顏	1
尹	4	賜	1	晏	1
口	3	戴	1	羊	1
蔡	3	鄧	1	仰	1
長	3	狄	1	衣	1
單	3	東	1	乙	1

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

丁	3	竇	1	弋	1
定	3	兒	1	嬰	1
奉	3	方	1	于	1
賈	3	符	1	垣	1
任	3	輔	1	爰	1
上官	3	泠	1	原	1
溫	3	幹	1	願	1
夏	3	榦	1	鄖	1
夏侯	3	耿	1	鄆	1
謝	3	功	1	詹	1
宣	3	龔	1	政	1
召	3	共	1	置	1
子	3	函	1	仲	1
鮑	2	闞	1	騶	1
充	2	漢	1	祝	1
房	2	郝	1		
傅	2	弘	1		
富	2	扈	1		
公孙	2	霍	1		
關	2	籍	1		

附表三 西北漢簡地名表

阿昌	當利	海東	臨利	入時	西京	榆樹
阿南	當陽	邯鄲	臨泉	入市	西利	玉門
安昌	道人	邯鄲縣	臨邑	塞北	西門	玉門關
安成	德昌	漢昌	臨澤	塞蘭	西南	冤句
安定	德里	漢中	令居	三井	西平	元城
安東	登山	河北	六安	三中	西山	月氏
安國	定安	河池	六甲	沙裡	西首	樂安
安漢	定安裡	河東	六中	山東	西土	樂昌
安利	定國	河渡	龍勒	山林	西鄉	樂成
安陵	定陵	河津	龍馬	山西	西夷	樂浪
安南	定陶	河裡	隴西	山陽	西域	雲中
安平	定襄	河馬	盧舍	刪丹	西垣	鄆
安息	東阿	河南	魯國	陝	相國	張掖
安陽	東北	河南省	魯陽	陝縣	薤	昭武
安夷	東郡	河內	陸渾	鄯善	襄國	召陵
安邑	東門	河渠	祿福	上黨	襄平	趙國
安樂	東南	河水	潞縣	深山	襄垣	真定

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

安眾	東平	河西	洛裡	石安	蕭	正北
白馬	東武陽	河中	雒	石關	新安	中道
白石	東西	弘農	略陽	石南	新平	中東
拜拜	東鄉	侯官	茂陵	史國	新沙	中封
北倉	東陽	侯國	孟陽	始安	新野	中加
北地	東域	壺關	莫及	壽昌	信成	中居
北宮	都市	湖	墨城	壽光	信都	中門
北裡	渡河	許都	牡馬	蜀中	焉得	中省
北平	敦煌	華裡	南安	順陽	焉耆	中堂
北遷	多汗	淮	南北	四中	鄢陵	中田
北曲	法度	淮海	南昌	駟望	延安	中營
北鄉	繁陽	淮陽	南充	宋子	延昌	諸國
北郵	範陽	淮陽國	南川	睢陽	延陵	鄒
北至	房子	會稽	南渡	天河	延平	
貝丘	汾	會水	南海	天門	延中	
表是	封丘	琿	南劍	天水	兗	
博南	奉先	吉時	南郡	天陰	兗	
博平	奉元	起士	南口	通國	燕山	
博望	伏地	即丘	南裡	同縣	陽東	
不韋	伏首	棘陽	南平	屯留	陽關	
察北	扶風	稷米	南鄉	屯田	陽丘	
昌安	扶溝	冀州	南陽	宛陵	陽山	
昌裡	福大	濟陽	內黃	宛縣	陽武	
昌平	福利	濟陰	涅陽	萬國	陽夏	
昌邑	父城	嘉裡	蓬蓬	萬年	陽陽	
長廣	富平	夾河	平安	望城	陽翟	
常安	贛	郟	平安裡	望垣	揚威	
常利	高成	建成	平北	魏	楊氏	
朝裡	高平	建平	平川	巫山	鄴	
成昌	槁	江道	平東	屋蘭	宜昌	
成德	故道	江海	平陵	烏氏	宜春	
成都	官山	江河	平南	吳	宜都	
成皋	關津	金城	平丘	吳房	宜市	
成固	關南	京兆	平望	五陵	已氏	
城旦	關內	京兆尹	平西	五原	以博	
城父	關西	荊	平鄉	武安	以利	
城南	館陶	涇陽	平邑	武昌	易縣	
城西	廣阿	井水	平陰	武城	益昌	
城邑	廣昌	酒泉	平原	武當	益州	
乘氏	廣德	居延	濮陽	武都	意中	
茬平	廣穀	君臨	祁	武關	義成	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

充國	廣漢	君山	青黑	武口	陰安	
大伯	廣利	郡國	清河	武陵	陰陽	
大昌	廣陵	苦縣	清淵	武始	鄞	
大城	廣明	昆陽	曲河	武威	滎陽	
大東	廣年	黎陽	曲梁	武陽	潁川	
大都	廣田	涼州	曲周	舞陰	潁川	
大河	廣為	梁國	任城	西安	雍丘	
大威	廣武	梁期	日東	西北	雍州	
大夏	廣至	聊城	日勒	西海	永安	
大邑	龜茲	遼東	入春	西河	勇士	
大原	歸德	遼西	入閣	西華	右扶風	
儋	虢	臨汾	入門	西街	榆木	

附表四 西北漢簡職官表

兵曹	侍御史
博士	隧長
丞相	太尉
都尉	廷尉
隊長	亭長
國相	尉史
侯長	縣長
候長	縣令
祭酒	縣尉
金曹	御史大夫
郡守	員吏
令史	主簿
閭師	主客
三老	主事
少府	

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

後記

本論文的研究是在我的導師張再興教授的悉心指導下完成的。從論文的選題、研究工作的開展和論文撰寫等過程中，我都得到了張老師的耐心教導和無私幫助。在我讀研究生的三年時間里，張老師以其嚴謹求實的治學態度、平易近人的教學風格和高度的敬業精神，深深地感染、激勵著我，使我在學業上和生活中都受益匪淺。值此論文完成之際，謹向張老師表達我的崇高敬意和深深的感謝。

感謝實驗室的姜慧師姐、黃艷萍師姐以及师弟师妹们在學習、實踐以及生活上給予我的無私關懷和幫助。實驗室良好的共同學習與互幫互助的氛圍，讓我充分感受到了同學之間的美好友誼，這將成為我終生難忘的回憶。

感謝華東師範大學所有關心、幫助過我的各位老師。

衷心感謝我的父母這麼多年來的養育和鼓勵，你們是我永遠奮鬥的動力。

最後，感謝各位老師與專家百忙之中對我論文的審閱，感謝您提的每一點意見，感謝您的辛苦工作！

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net