平成１５年度採択分
平成２０年３月３１日現在

# 推論機能を有する木簡など出土文字資料の文字自動認識システムの開発

The development of the automatical system with a reasoning function which recognizes the character of the excavated written sources such as Wooden Tablets

渡辺　晃宏（WATANABE　AKIHIRO）
国立文化財機構 奈良文化財研究所・都城発掘調査部・史料研究室長

研究の概要
木簡の釈読を支援するシステムとして、①木簡情報デジタル化簡易システム、②木簡の文字画像データベース「木簡字典」、③木簡釈読支援データベース群、④劣化や欠損により不完全な文字を類推して釈読する木簡の文字自動認識（ＯＣＲ）システム「Mokkan　Shop」を研究・開発・公開し、木簡など出土文字資料釈読技術の汎用化・効率化を図った。

研　究　分　野：人文学
科研費の分科・細目：史学・日本史
キ　ー　ワ　ー　ド：木簡、出土文字資料、データベース、ＯＣＲ、漢字、文脈処理

## １．研究開始当初の背景・動機

全国で 30 万点を超えた木簡の半数以上を発掘調査し整理・保管・研究する機関として、従来熟練した研究員の長年の経験と知識、及び勘に頼らざるを得なかった木簡釈読の効率化を図り、またそのノウハウを客観的かつ汎用性のあるシステムとして学界共有のものとし、木簡など出土文字資料の研究環境を一新する必要があった。

## ２．研究の目的

近年めざましい進展を遂げた古文書における手書き文字ＯＣＲ技術による文字自動認識システムを応用して、木簡など出土文字資料の文字の読み取り、釈文作成作業を自動的に行う方法を開発して釈読技術の汎用化と効率化を図り、奈良文化財研究所のもつ厖大な文字資料を画像とともに検索に耐え得るデータとして蓄積し、広く研究者や一般の便に供することを目的とした。

## ３．研究の方法

(1)木簡情報デジタル化簡易システム―木簡の文字情報をできるだけ鮮明にデジタル化して蓄積するシステム
(2)木簡の「文字」の事典―個々の文字について、サンプルとなる画像をセットにしたデータベース
(3)木簡釈読支援システム―横断検索によって木簡の釈読を支援するデータベース群
(4)木簡用ＯＣＲ―劣化や欠損で不完全な状態の文字を類推して釈読するシステム。

上記システムの統合により、客観的かつ汎用性のある、推論機能をもつ木簡釈読システムを呈示する。

## ４．研究の主な成果

### ａ文字画像データベース「木簡字典」の開発・公開

「木簡字典」は、木簡に書かれた文字ごとの画像データベースで、文字種ごとに実際に書かれた字体の事例を、モノクロ・カラー・赤外線写真、私たちが解読した記録（記帳ノート）も含め、複数の画像で紹介する画期的なシステムである。また、従来から奈良文化財研究所が公開している木簡データベースのデータを用いて、その画像の文字がどのような文脈で用いられたかなど、その文字が書かれた木簡そのものの基礎データが全てわかるようになっている。

2005 年２月８日にまず単数文字の検索システムとして奈良文化財研究所のホームページ上でＷＥＢ公開し、その後複数文字検索システムへのヴァージョンアップを図り、2007 年２月 20 日に新システムへの切り替えを行った。収録木簡点数はカラー約 1,000 点･モノクロ約 800 点･赤外約 200 点・記帳約 50 0 点、文字種類 1,200 種、文字数約 20,000 文字である。約 1,500 種とい


検索機能を有する木簡データベース
奈良文化財研究所Ｗebページで
公開中

われる木簡に使われる文字のうち、主要な文字をほぼカヴァーしている。

奈良文化財研究所以外の機関が調査した木簡についても順次許可を得ながら掲載を進めており、時間的にも空間的にも広がりをもつデータベースとし、今後さらにデータの拡充に努めていきたい。アクセス件数は月 1000 件程度を維持し、これまでの総アクセス件数は約 25000 件に及んでいる。海外からのアクセスも多い。なお、ＷＥＢ公開とは別に、印刷版木簡字典を刊行した。対象は『平城宮木簡』１～６所収木簡で、親字約 940 文字、延べ 5000 文字を掲載。

**ｂ文字自動認識システム（ＯＣＲ）の開発**

「Mokkan　Shop」はオフラインの文字認識処理システムで、文字画像の切り出し、墨部の抽出、文字認識、認識結果の検証という手順で、可能性の高いものから順に認識候補を表示して木簡解読を支援するものである。これまでに認識対象とできた文字パターンは約 500 字種、約 4,000 パターンである。墨の部分を抽出するための画像処理手法、及び欠損文字の認識についても有効な文字認識システムの開発に成功した。全体が残るとは限らない、また劣化の著しい、いわば不完全な状態にあるのを特徴とする木簡の文字認識は、文字の自動認識として画期的なシステムといえる。

「Mokkan　Shop」には、日本古代の人名・地名・物品名について今回新たに作成した釈読支援データベースに基づく文脈処理モジュールを開発し、搭載した。これにより釈読の有効性を高めるのに成功し、文脈による文字列の釈読が容易に行えるようになった。なお、この研究を発表した『情報考古学』Vol,13No.1 掲載の論文「木簡解読支援のための文脈処理」は、2007 年度の日本情報考古学会論文賞を受賞した。

「Mokkan　Shop」により、これまで調査者の経験と勘に頼らざるを得なかった木簡の釈読作業の効率化を図ることができ、そうした経験と勘を汎用化することが可能になった。

今後、出土文字資料釈読支援システムとしての統合を図るとともに、「木簡字典」と「Mokkan　Shop」を基本的なツールとして活用することによって、木簡など出土文字資料に関する綜合的な研究拠点機能の構築をめざしたい。


**５．得られた成果の世界・日本における位置づけとインパクト**

「木簡字典」と「Mokkan　Shop」は、日本の木簡以外の出土文字資料のほか、東アジアの漢字文化圏の文字資料研究にとっても重要なツールとなると考える。また、今回の研究は、歴史学と情報工学の分野の共同研究としても画期的な研究開発であり、コンピュータと人間との関わりの新しい形を切り開いたものと考える。

６．主な発表論文

（研究代表者は太字、研究分担者には下線）


未代誠仁、西嶋佳津、齋藤恵、石川正敏、中川正樹、 馬場基、**渡辺晃宏**、木簡解読支援のための文脈処理、『日本情報考古学会論文誌』、13(1)、p.p.7-21、2007 年

未代誠仁、戸根康隆、石川正敏、中川正樹、馬場基、**渡辺晃宏**、木簡解読支援システムの改善に向けた取組み、『日本情報考古学会第 23 回大会』、p.p.33-38、2007 年

A. Kitadai, K. Nishijima, K. Saito, M. Nakagawa, H. Baba and **A. Watanabe**, Context Processing to Read Text on Damaged Wooden Tablets, 『Proc. 10th International Workshop on Frontiers in Handwriting Recognition』 Vol. 1, p.p.581-586, 2006 年

齋藤恵、未代誠仁、西嶋佳津、中川正樹、馬場基、**渡辺晃宏**、「平城京跡出土木簡上の欠損文字パターン認識」、『電子情報通信学会技術報告（信学技報）, PRMU』、p.p.85-90、2006 年

Akihito Kitadai, Kei Saito, Daisuke Hachiya, Masaki Nakagawa, Hajime Baba and **Akihiro Watanabe**、Design and Prototype of a Support System for Archeologists to Decode Scripts on Mokkan, Proc. 13th Conference of the International Graphonomics Society (IGS), p.p.54-58、2005 年

齋藤恵、蜂谷大翼、未代誠仁、中川正樹、馬場基、**渡辺晃宏**、木簡画像から墨の部分を抽出するための画像処理手法、『電子情報通信学会技術報告』、PRMU2005-03-18、p.p.163-168、2005 年

蜂谷大翼、齋藤恵、未代誠仁、中川正樹、馬場基、**渡辺晃宏**、欠損を含む文字パタンを対象とした文字認識システムの試作、『電子情報通信学会技術報告』、PRMU2005-03-19、p.p.169-174、2005 年

未代誠仁、齋藤恵、蜂谷大翼、中川正樹、馬場基、**渡辺晃宏**、木簡解読支援システムの基本設計と試作、『人文科学とコンピュータシンポジウム論文集』５-Ａ-１、p.p.215-220、2004 年


ホームページ等

http://hiroba.nabunken.go.jp/