木簡解読支援のための情報検索

高倉 純[†], Somayeh Sherini[†], 来代 誠仁[††], 石川 正敏[††], 中川 正樹[††], 馬場 基[†††], 渡辺晃宏[†††]
[†]東京農工大学 工学部 [††]東京農工大学 大学院工学府 [†††]奈良文化財研究所

本稿では，古代木簡解読支援のための2種類の情報検索について述べる．古代木簡の多くは汚損，破損による情報損失が著しい．そのため，既に解読された木簡の情報を用いた釈文の補完・検証が必要である．我々は文脈処理技術とパタンマッチング技術を用いることで，釈文の補完と検証を支援する情報検索を提案する．文脈処理技術は地名・人名・物産名・カバネなど，木簡に多く見られる記述の横断的検索を提供する．これは広範囲にわたる木簡の汚損・破損に対応する．また，パタンマッチング技術は，欠損を伴う文字パタンに類似した字体をWeb上のデータベース内で検索する．これは墨の欠損を伴う字体の解読に有用である．現在，専門家による評価が進行中である．

Preparing a camera-ready manuscript for the Jinmonkon 2008 conference

Jun Takakura[†], Somayeh Sherini[†], Akihito Kitadai[††], Masatoshi Ishikawa[††],
Masaki Nakagawa[††], Hajime Baba[†††] and Akihiro Watanabe[†††]
[†]Faculty of Engineering, Tokyo University of Agriculture and Technology
[††]Graduate School of Engineering, Tokyo University of Agriculture and Technology
[†††]Nara National Research Institute for Cultural Properties

This paper presents two methods for information retrieval to support reading historical mokkans. Most of the mokkans have defacement parts that lead to much information loss. Therefore, complementation and verification of the translations using the information of decodable mokkans are necessary. Therefore, we propose a context processing method that provides cross-over information retrieval of place-names, personal-names, product-names and dignities commonly written on the mokkans. Also, we propose a pattern matching method that searches character patterns in the historical mokkan database on the web to obtain the patterns similar to the unreadable character pattern. The evaluation by historians and archaeologists is in progress.

1. まえがき

本稿では，古代木簡解読支援のための 2 種類の情報検索について述べる．

古文書の一種である古代木簡の解読は，それが作成・利用された時代を知るための重要な作業である．しかし，長い年月の中で多くの木簡は汚損・破損しており，多くの文字が解読不可能または困難な状態になっている．また，文字の誤用・誤記，および語順の変化も少なからず見られる．このため，木簡解読は考古学・史学の専門家にとっても容易ではない．

一般的に，文書の解読支援には文脈処理と呼ばれる技術が用いられる．文脈処理は，文書中の解読可能な文字・文字列を入力として，解読不可能または困難な文字の推読を支援し，また文字の誤用・誤記の修正に有用な情報を利用者に提供する．古文書に対しては，山田らが n-gram を用いた文脈処理を提案し，江戸時代の文書を用いて有効性を示している[1]．しかし，木簡の場合は解読不可能な文字の割合が比較的高く，語順の変化も頻繁に発生することから，より強力な文脈処理が必要となる．また，部分的に残存した文字の解読を支援する情報検索技術を実現することで，未解読の文字を減らすことも重要である．

我々は，古代木簡解読のための文脈処理手法として Extended Aho-Corasick 法を提案した[2]．EAC 法は誤記または文字の部分的欠損によって未解読となっている釈文の補完を支援する文脈処理の一種で，釈文に含まれる誤字，語順の変化に頑健である．また，パタンマッチング技術を応用した字体解読支援のための情報検索手法（以下，グレーゾーン法）を提案した[3]．グレーゾーン法は，専門家とコンピュータの協調作業による欠損文字の解読を可能にする．

しかし，EAC 法については語彙データベースの充実，およびデータベースを効果的に利用するユーザインタフェースの実現が，またグレーゾーン法については精度向上，および専門家に提示する情報の整備が，それぞれ課題となっていた．

そこで，本研究では我々が EAC 法のために作成した地名・人名・物産名・カバネのデータベースと，それらを横断的に検索可能なユーザインタフェースについて，およびグレーゾーン法の精度向上と Web 上にあるデータベースとの連携について，それぞれ述べる．

2．古代木簡とその解読

木簡は，木片に文字が書かれた文書の総称である．耐候性に優れ，媒体となる木片が安価で入手性も高かったことから，奈良時代およびその前後には数～数十の文字を記録，伝達する手段として広く利用されたと考えられている．これまでに，奈良平城宮とその周辺から 170,000 点以上，全国の古代遺跡を合わせると 320,000 点余りの古代木簡が発見されている（図 1）．

これらの木簡には，荷札として利用されたと考えられるものが多数含まれている．荷札は当時の地名・各地の産業・地域間交流・人名・敬称などを表す貴重な文書である．この他にも，役所・寺院の間での連絡に用いられたと考えられる木簡も数多く見つかっており，当時の政を示す重要な資料となっている．このような理由から，木簡の解読結果は史学・考古学の分野において高い注目を集めている．

しかし，木簡の解読は容易ではない．現存する古代木簡の多くは遺跡の地下などから発掘されたものである．1,300 年近くも地中にあったことから，汚損，風化などによる損傷は著しい．また，古代木簡には人為的に破壊されたものも多く見られる．典型的なのは，木片の再利用のために鉋を掛けられた木簡である．この場合，鉋屑に残る文字を解読することになる．しかし，木簡一点分の鉋屑が完全に発見されることは希である．他にも，意図的に 2 つ以上に折られたもの，焼けた跡があるものなど，文書としての保存状態は良好とはいえない（図 2）．

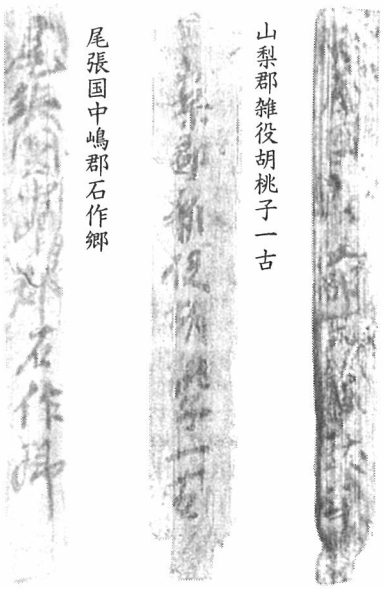

図 1　古代木簡
Figure 1　Historical Mokkans.

したがって，古代木簡の解読を進めるためには，損傷によって失われた情報の補完が重要となる．補完に用いる情報は，解読済の他の古文書（含む古代木簡），および古文書について書かれた解説書などから得ることができる．しかし，統一された索引のない大量の書物の中から適切な情報を得ることは史学・考古学の専門家にとっても困難を伴う．

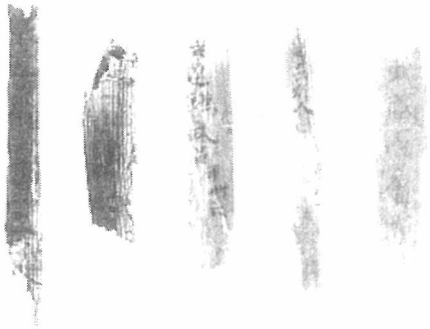

図 2　損傷した木簡
Figure 2　Broken Mokkans.

3．情報検索の基本方針

デジタル情報化された古文書は，現物の古文書，書物などに代わる補完のための情報源として有望な存在である．デジタル情報をネットワーク上に置くことで，地理的に分散していた古文書と書物は空間的制限から解放される．しかし，集約されただけの情報は索引のない分厚い辞書でしかない．デジタル情報を有効活用する上で，適切な情報検索手段の提供は不可欠といえる．

我々は，古代木簡の画像とその釈文を，形状，発見場所などのメタ情報と共にデジタル化し，デジタルアーカイブ「木簡字典」として Web 上で公開している．木簡字典では，釈文の一部，メタ情報などをキーとした検索機能を提供することで，デジタル情報の効果的な理由に配慮している．しかし，汚損，破損などによって形状，釈文が曖昧になった木簡の解読支援に対して，これらの検索機能だけでは十分とはいえない．

そこで，本研究では文脈処理技術，およびパタンマッチング技術を発展させた新しい情報検索の実現を考える．木簡解読支援のための文脈処理としては，我々が提案した EAC 法（Extended Aho-Corasick 法）がある．これまで，我々は EAC 法を，木簡に多く見られる地名に関する記述の補完と検証に限って適用してきた．本研究では，この適用範囲を人名・物産名・カバネ（敬称）にまで拡大すると共に，地名・人

名，物産名，カバネを対話的かつ横断的に検索するためのユーザインタフェースを実装することで，有効性の高いテキスト検索の実現を目指す．

パタンマッチング技術についても，部分的に欠損した文字パタンを認識するために我々が提案したグレーゾーン法がある．グレーゾーン法を用いることで，デジタルアーカイブに登録された字体の中から解読困難な文字に類似したものを検索することができれば，未解読の文字を減らし解読作業を進める有効な支援になる．しかし，グレーゾーン法は対話性に課題があり，そのままでは情報検索技術として不十分である．本研究では，グレーゾーン法を拡張し，木簡字典と連携した字体検索の実現を目指す．

4. 文脈処理技術を用いたテキスト検索

EAC 法は，不完全な釈文をキーとして，それ自身を含む完全な文字列を検索する文脈処理手法である．この際，キーに含まれる誤字および文字順序の変化に頑健でありながら，同時に文字順序の評価を行うことができることが EAC 法の特徴である[2].

EAC 法では，最初にキーからマシン EAC と呼ばれる状態遷移モデルを作成する．図 3 に，不完全な釈文“蔵父郡”が得られた場合のマシン EAC を示す．

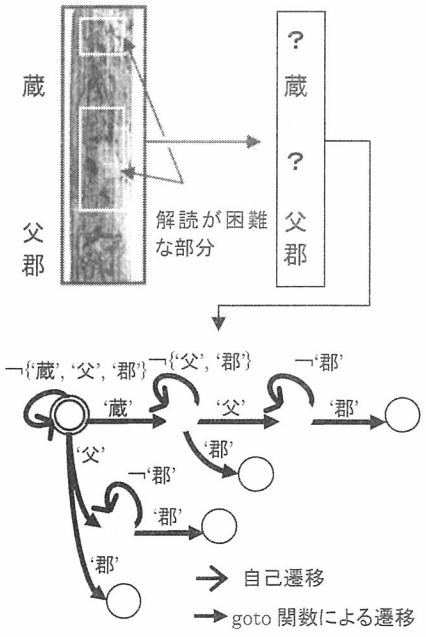

図 3　“蔵父郡”に対するマシン EAC
Figure 3　Machine EAC for “蔵父郡”.

マシン EAC では，遷移の発生条件としてそれぞれキーに含まれる文字が与えられる．また，自己遷移および始状態への遷移にはスコア 0 が，それ以外の遷移には始状態からの最小ホップ数に応じたスコアがそれぞれ与えられ，文字順序の評価に利用される．

検索処理時には，検索対象となる文字列（テンプレート文字列）を用いてマシン EAC 上で状態遷移を行い，その経路上に存在するスコアを合計する．こうして求まったスコアの合計値によって，文字列とキーとの類似度を評価することができる．

EAC 法では，テンプレート文字列をデータベースとして予め用意しておく必要がある．我々は，奈良時代とその前後に存在したと考えられる 4,041 の地名についてデータベース化を行い，これを用いた文脈処理を提供してきた．本研究において文脈処理の対象を人名・物産名・カバネを合わせた 4 カテゴリに拡張するにあたって，我々はこれまでに確認された 20,000 以上の人名（ただし，同姓同名などによる重複を含む），3,000 以上の物産名（ただし，産地・使途が異なる場合の重複を含む），および約 50 のカバネ（読み方，異表記による重複を含む）についてそれぞれデータベースを作成した．また，カテゴリ間で関連の強い項目をリンクし，カテゴリを跨いだ横断的なテキスト検索を可能にした．

図 4 に，テキスト検索のためのユーザインタフェースを示す．

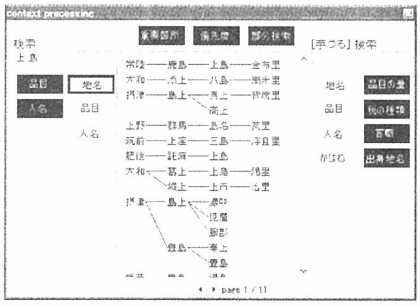

図 4　テキスト検索のためのユーザインタフェース
Figure 4　User Interface for Text Retrieval.

木簡解読支援システムでは，専門家の高度な思考を支援するため，電子ペンを用いた直感的な操作に対応している．文脈処理のためのユーザインタフェースについてもこの点を考慮し，ボタンの大きさ，位置などを電子ペンによる操作に合わせて設計・実装した．

図 5 に，地名・物産名・人名の横断的検索の例を示す．

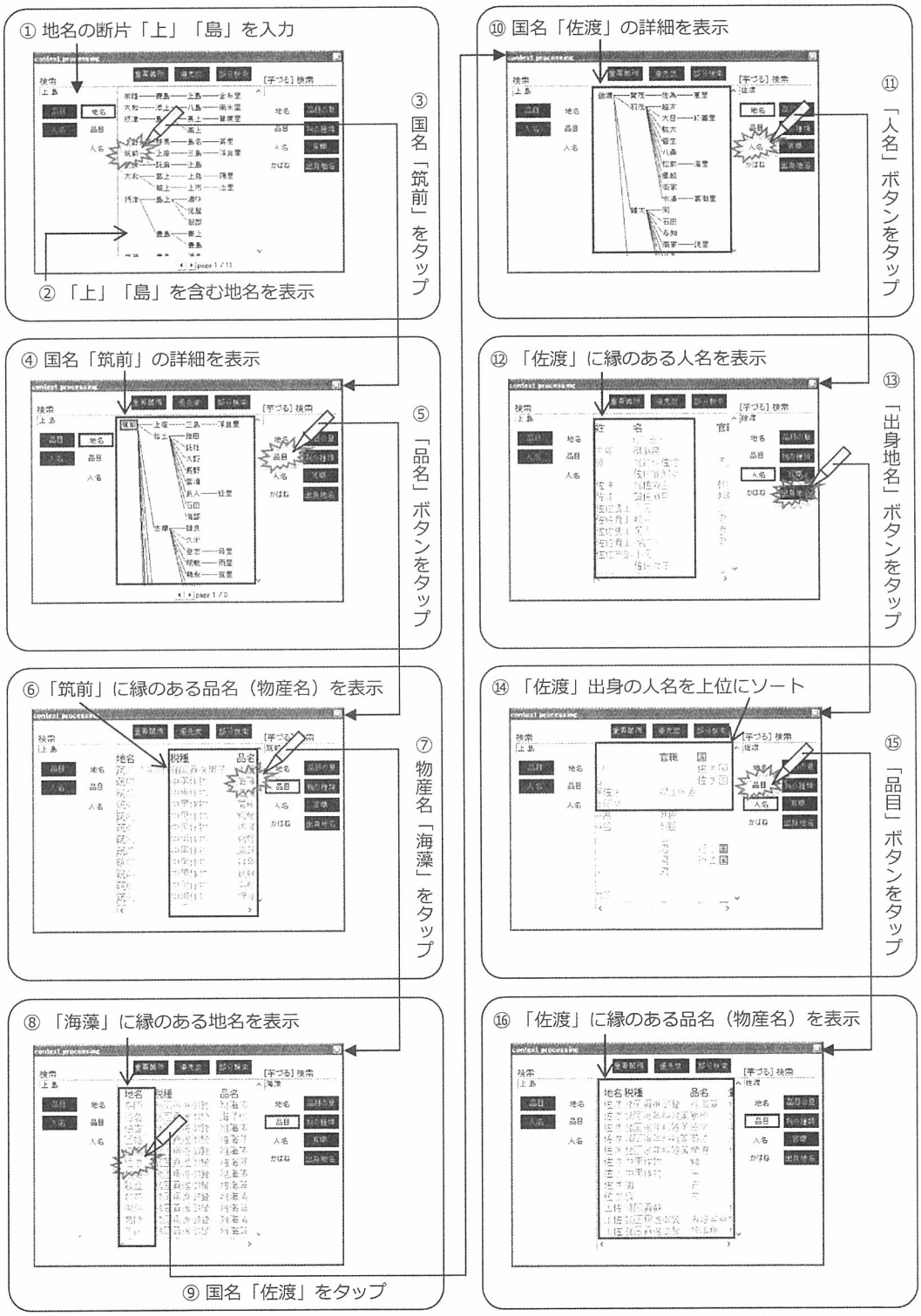

図 5　横断的なテキスト検索（文脈処理）の例
Figure 5　Example of Crossover Text Retrieval (Context Processing).

5．パタンマッチング技術を用いた字体検索

グレーゾーン法は，木簡の損傷によって欠損を伴った文字パタンの認識を目的として我々が研究を進めてきた手法である[3].

文字を表す墨が失われた場合，従来のパタンマッチング技術では認識対象とテンプレートの類似度を適切に評価することが難しかった．これは失われた墨の存在を，文字パタンの非線形化処理に反映できないことに起因する．しかし，木簡の損傷の状態から墨が失われた大まかな位置を推定することは専門家にとって不可能ではない．このような失われた墨に関する曖昧な情報をグレーゾーンとして専門家に入力してもらい，それをパタンマッチングの精度向上に利用するのがグレーゾーン法である．図 6 に，右側が失われた木簡を用いた例を示す.

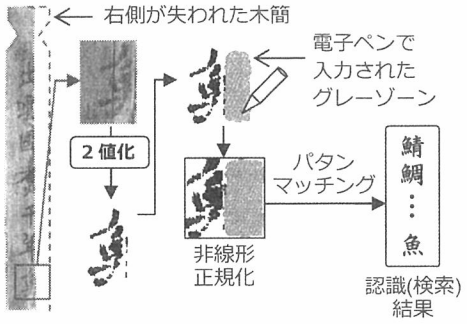

図 6　グレーゾーン法による処理例
Figure 6　Example Using Gray-zone Method.

我々は，線密度を用いた非線形正規化を用いることで，木簡に記された文字を高精度に認識できることを示した．ただし，グレーゾーンは筆画（ストローク）の形状を表すものではないため，線密度を直接算出することができない．そこで，これまではグレーゾーンを黒画素化，および白画素化した 2 つのパタンに対して線密度 D_{black} と D_{white} を算出し，それらの単純平均 D_{ave} を用いて非線形正規化を実現してきた（図 7）.

しかし，自動化を目標とする文字認識とは異なり，字体検索では検索パラメータの変更による対話的検索の実現が重要である．特に，失われたと考えられる墨の推定量（密度）を変更することで検索結果の絞り込みを行うことができれば，欠損文字パタンに対する有効な字体検索が実現できると考える.

そこで，本研究では D_{ave} に代わって数式 1 で表される加重平均 $D_{w\text{-}ave}$ を採用し，係数λを利用者が変更できるようにした.

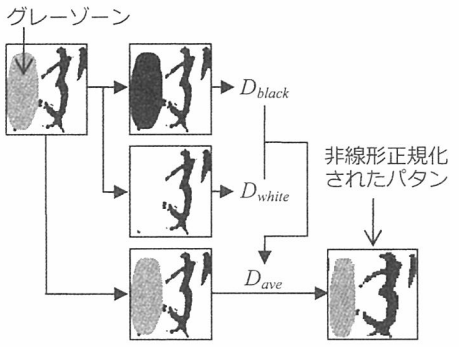

図 7　単純平均による非線形正規化
Figure 7　Non-linear Normalization Using Simple Average.

$$D_{w\text{-}ave} = \lambda \cdot D_{black} + (1-\lambda) \cdot D_{white} \quad (0 \leqq \lambda \leqq 1) \qquad (1)$$

図 8 に，λの変化と非線形正規化の結果の関係を示す．式(1)を用いる場合，グレーゾーン内で字形を表す多くの情報が失われたと考える場合はλに大きな値を，逆にグレーゾーン内で失われた字形の情報は少ないと考える場合はλに小さな値をそれぞれ設定することで，利用者の考えを反映したパタンマッチング処理，およびそれを用いた字体検索を実現できることになる.

式(1)のλには 0 から 1 の任意の値を指定することができる．しかし，実際に数値を直接入力する操作は煩雑であり，高度な思考を必要とする木簡解読には適さない．そこで，字体検索のためのユーザインタフェースでは 0/0.25/0.5/0.75/1 の 5 段階（表記は%単位）から利用者が選択する方法を採用した．このユーザ

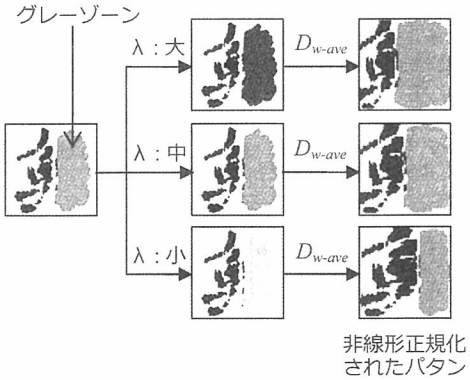

図 8　加重平均による非線形正規化
Figure 8　Non-linear Normalization Using Weighted Average.

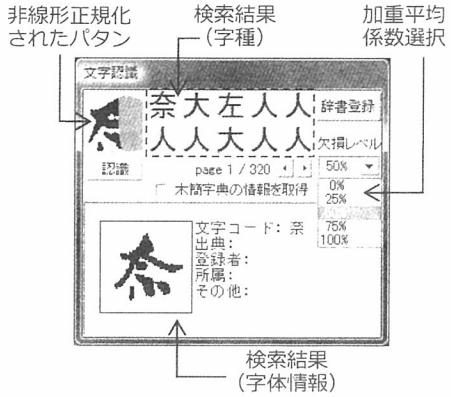

図 9　字体検索のためのユーザインタフェース
Figure 9　User Interface for Character Pattern Retrieval.

インタフェースはテキスト検索と同様に木簡解読支援システムの一部として動作し，電子ペンによる操作を考慮したものとなっている．

字体検索の結果となる字種を電子ペンなどで選択することで，その字種に関する字体情報を表示することが可能である．字体情報は我々の所有する古代木簡データベース「木簡字典」とリンクしており，ネットワーク接続された環境であれば該当する字体が記載された木簡の画像（カラー，モノクロ，赤外など）を参照することも可能である[4].

表 1 に，グレーゾーン法を用いた字体検索，および 5 段階のパラメータ選択の有効性を示す実験とその結果を示す．この実験では，実在の古代木簡から抽出した 2,108 の文字パタンを用いた．これらは出典が明らかで，字体の損傷も少ない．これに 10 種類の疑似的な欠損マスクを付加し，各 21,080 の欠損文字パタンおよびグレーゾーン付文字パタンをテストパタンとして生成した（図 10）．初めから欠損した文字パタンを使用しないのは，解読結果（字種）の信頼性を重んじてのことである．

検索率としては，leave-one-out 法（試行あたりのテストパタンを 1 とした cross-validation 法）による 10 位候補含有率を採用した．また，テンプレートには欠損マスクを適用しない．したがって，試行ごとにテストパタン自身を除く 2,107 の文字パタンがテンプレートとなる．

λ の値については，試行ごとに 5 段階から最適なものを選択する場合，および 0.5 に固定する場合のそれぞれについて実験を行った．前者は，専門家が最適な λ を選択した場合のシミュレーションである．一方，後者は専門家が λ の値を指定しなかった場合の初期設定値となる．λ の値については，常に最適なものを選べるとは限らないため，後者の実験を併記することは有意であると考える．

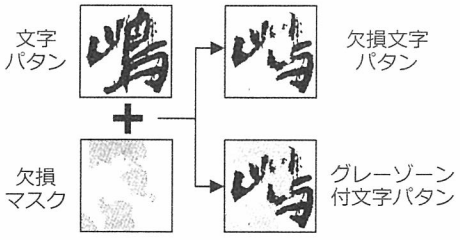

図 10　テストパタンの生成
Figure 10　Example Using Gray-zone Method.

表 1　検索率
Table 1　Retrieval Rates.

キー	検索率
欠損なし（理想値）	76.0% (1,601/2,108)
欠損文字パタン	43.4% (9,139/21,080)
グレーゾーン付文字パタン (λ=最適値)	69.6% (14,668/21,080)
グレーゾーン付文字パタン (λ=0.5)	58.6% (12,359/21,080)

表 1 の結果から，文字パタンの検索にグレーゾーン法が有効であることが示された．

6. おわりに

本稿では，木簡解読を支援する 2 種類の情報検索について述べた．提案手法については，専門家による評価実験が進行中である．今後の課題としては，評価実験の結果を反映した手法の改善があげられる．

謝辞

本研究は科研費基盤 S-20222002 および若手 B-19720202 の助成を受けたものである

参考文献

[1] 山田奨治, 柴山守: n-gram による古文書証文類翻刻支援の検討, 人文科学とコンピュータシンポジウム論文集, Vol.2000, No.17, pp.185-192,(2000).

[2] 耒代誠仁, 西嶋佳津, 斎藤恵, 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, 日本情報考古学会誌, Vol.13, No.1, pp.7-21 (2007).

[3] M. Nakagawa, K. Saito, A. Kitadai, J. Tokuno, H. Baba and A. Watanabe "Damaged character pattern recognition on wooden tablets excavated from the Heijyo palace site," Proc. 10th IWFHR, La Baule, France, vol. 1, pp.533-538 (2006).

[4] http://jiten.nabunken.go.jp/