

# 「縁」モデルにもとづく拓本文字データベース\*

安岡孝一†

## 1 はじめに

漢字は「形」「音」「義」の3要素から成り立っていると説かれてきた。現代では、それに加えて新たな要素「縁」があらわになってきている [1]。漢字の「縁」は、情報学でいうところの「リレーション」にあたり、漢字相互の関連性を抽象化した概念である。

筆者はこれまで、透明テキスト付き画像に関する研究をおこなってきた [2, 3]。本稿では、筆者があらたに構築した拓本文字データベースについて述べる。この拓本文字データベースは、京都大学人文科学研究所所蔵石刻拓本資料‡の透明テキスト付き画像データベースであると同時に、「縁」モデルにもとづいた文字データベースである。この拓本文字データベースに関して、2章では、透明テキスト付き画像データベースとしての側面から、3章では、「縁」モデルにもとづく文字データベースとしての側面から、それぞれ述べる。

## 2 拓本文字データベースの概要

この章では、拓本文字データベースの WWW インターフェースと、そのバックグラウンドとなるデータ構造について述べる。

### 2.1 WWW インターフェース

拓本文字データベースの WWW インターフェースは、検索画面、集字結果画面、拓本 DjVu 画面、の3種類の画面から構成される。検索画面 (図 1) は、収録されている拓本に対して釈文の全文検索をおこなうための入口であり、漢字一文字あるいは文字列を入力する。

検索の結果は、集字結果画面として表示される。集字結果画面では、検索にマッチした全拓本から、該当文字あるいは該当文字列を切り出して、拓本の年代順に表示する。例として「墓」を検索した場合の集字結果画面を図 2 に示す。各画像上にマウスを置くと、切り出し元となった拓本のタイトルがミニボックスに表示される。また、各画像をクリックすると、切り出し元の拓本 DjVu 画面へとジャンプする。

拓本 DjVu 画面では、拓本の透明テキスト付き画像を、テキストビハインド DjVu [2] で表示§する。この際に、集字結果画面で切り出された文字あるいは文字列は、赤

\*第 16 回「東洋学へのコンピュータ利用」研究セミナー (2005 年 3 月 25 日)

†京都大学人文科学研究所附属漢字情報研究センター

‡<http://kanji.zinbun.kyoto-u.ac.jp/db-machine/imgsrv/takuhon/>で公開中。

§表示には、<http://www.lizardtech.co.jp/download/djvu/>などで配布の DjVu プラグインが必要。

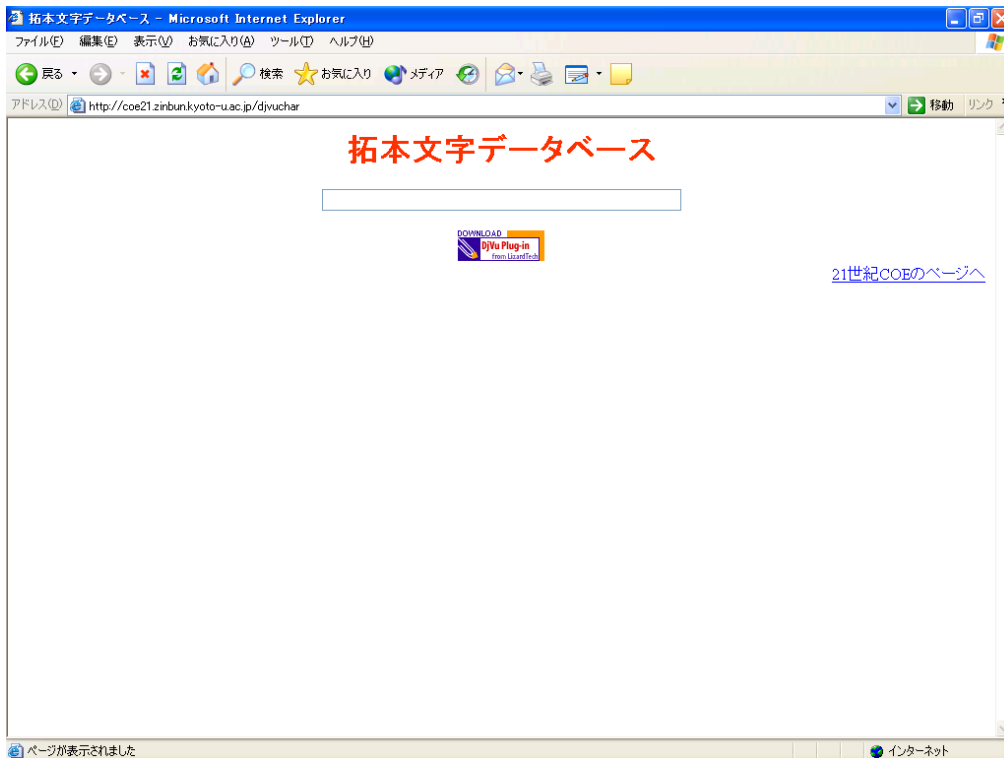


図 1: 拓本文字データベースの検索画面

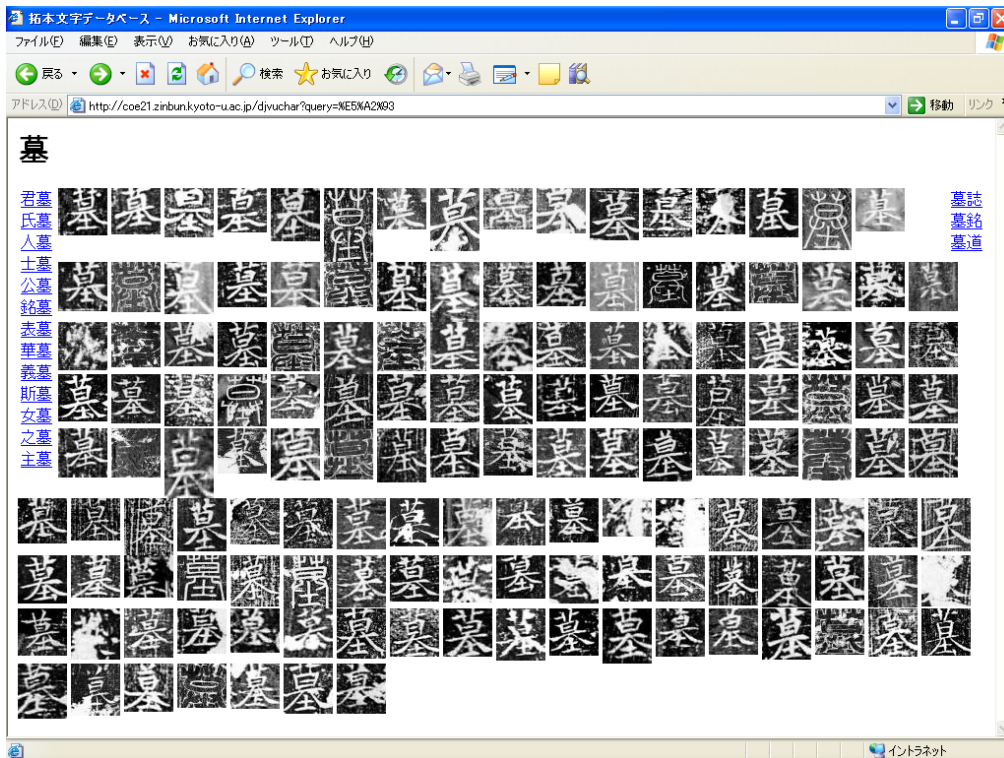


図 2: 「墓」の集字結果

色で反転してハイライト表示<sup>¶</sup>している。例として、図2の左上の「墓」画像をクリックした場合の拓本 DjVu 画面を図3に示す。拓本 DjVu 画面は、上部のナビゲーションペインの機能により、ズームイン、ズームアウト、文字列検索などがおこなえる。また、拓本の各文字上にマウスを置くと、その文字に対する釈文がミニボックスに表示<sup>||</sup>される。各文字をクリックすると、その文字に対する集字結果画面へとジャンプする。

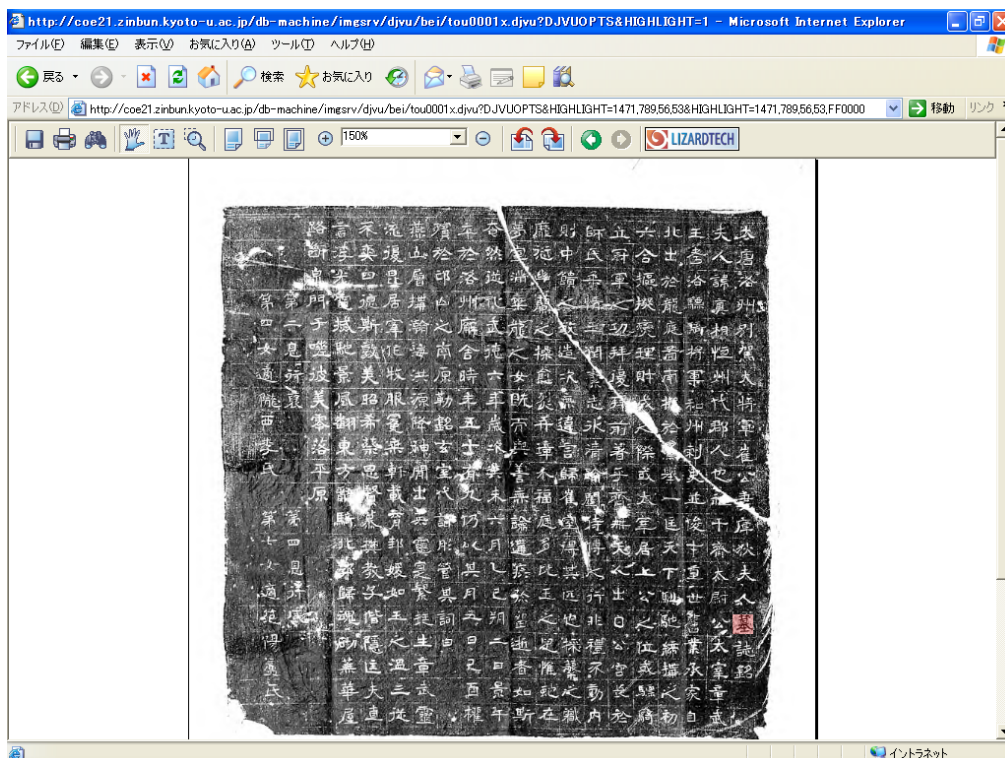


図 3: 「唐洛州別駕大將軍崔公妻庫狄真相墓誌銘」の DjVu 画面

## 2.2 データベースの構造

拓本文字データベースの中心にあるのは、各拓本画像に対する座標付釈文情報である。座標付釈文情報は、ttext-kanbunが出力する CSV 形式ファイル [3] であり、各行が文字ボックス 1 個に対応している (図 4)。1 行は 5 つのフィールドからなり、順に、文字ボックス左上の X 座標\*、文字ボックス左上の Y 座標\*、文字ボックスの幅、文字ボックスの高さ、文字の UTF-16 による 10 進数表現となっている。

実際のデータベースでは、筆者が作成したシェルスクリプト csv2djvuxml により、CSV 形式ファイルを DjVuXML 形式ファイル (図 5) に変換して用いている。

<sup>¶</sup> ちなみにハイライト表示を消すには、URL から ?DJVUOPTS 以降を削除すればよい。

<sup>||</sup> Microsoft Windows 版の DjVu プラグインにはバグがあり、ミニボックス中の文字がしばしば文字化けすることが報告されている。

\*座標原点は拓本画像の左上。

1471, 178, 64, 57, 22823  
1480, 240, 58, 56, 21776  
1479, 304, 62, 67, 27931  
1476, 367, 62, 67, 24030  
1477, 438, 56, 53, 21029  
1481, 499, 56, 53, 39381  
1474, 562, 56, 53, 22823  
1477, 642, 56, 53, 23559  
1475, 700, 56, 53, 36557  
1478, 772, 56, 53, 23828  
1473, 838, 56, 53, 20844  
1475, 897, 56, 53, 22971  
1470, 971, 56, 53, 24235  
1473, 1031, 56, 53, 29380  
1474, 1098, 56, 53, 22827  
1471, 1161, 56, 53, 20154  
1471, 1236, 56, 53, 22675  
1472, 1307, 56, 53, 35468  
1473, 1365, 56, 53, 37528  
NaN, NaN, NaN, NaN, 12290  
NaN, NaN, NaN, NaN, 13  
NaN, NaN, NaN, NaN, 10  
1411, 178, 56, 53, 22827  
1410, 245, 56, 53, 20154  
1412, 308, 56, 53, 35569  
1408, 376, 56, 53, 30494  
1408, 441, 56, 53, 30456  
NaN, NaN, NaN, NaN, 12290  
1410, 501, 56, 53, 24658  
1409, 564, 56, 53, 24030  
1409, 631, 56, 53, 20195  
1409, 697, 56, 53, 37089  
1404, 760, 56, 53, 20154  
1409, 829, 56, 53, 20063  
NaN, NaN, NaN, NaN, 12290  
1402, 901, 56, 53, 31062  
1404, 967, 56, 53, 24178  
NaN, NaN, NaN, NaN, 12290  
1405, 1030, 56, 53, 40778  
1411, 1103, 55, 49, 22826  
1408, 1170, 55, 49, 23561  
1410, 1239, 55, 49, 20844  
1405, 1301, 55, 49, 22826  
1405, 1365, 55, 49, 23472  
1403, 1423, 55, 49, 31456  
1400, 1486, 55, 49, 27494  
NaN, NaN, NaN, NaN, 13  
NaN, NaN, NaN, NaN, 10  
1343, 177, 55, 49, 29579  
NaN, NaN, NaN, NaN, 12290  
⋮

図 4: 「唐洛州別駕大將軍崔公妻庫狄真相墓誌銘」の座標付釈文情報

```

<?xml version="1.0" ?>
<!DOCTYPE DjVuXML PUBLIC "-//W3C//DTD DjVuXML 1.1//EN"
  "pubtext/DjVuXML-s.dtd">
<DjVuXML>
<HEAD>tou0001x.djvu</HEAD>
<BODY>
<OBJECT data="tou0001x.djvu" type="image/x.djvu"
  height="2078" width="1695" usemap="tou0001x.djvu" >
<PARAM name="DPI" value="400" />
<PARAM name="GAMMA" value="2.200000" />
<HIDDENTEXT><WORD>
<CHAR coords="1471,178,1535,235" sep="no">##22823;</CHAR>
<CHAR coords="1480,240,1538,296" sep="no">##21776;</CHAR>
<CHAR coords="1479,304,1541,371" sep="no">##27931;</CHAR>
<CHAR coords="1476,367,1538,434" sep="no">##24030;</CHAR>
<CHAR coords="1477,438,1533,491" sep="no">##21029;</CHAR>
<CHAR coords="1481,499,1537,552" sep="no">##39381;</CHAR>
<CHAR coords="1474,562,1530,615" sep="no">##22823;</CHAR>
<CHAR coords="1477,642,1533,695" sep="no">##23559;</CHAR>
<CHAR coords="1475,700,1531,753" sep="no">##36557;</CHAR>
<CHAR coords="1478,772,1534,825" sep="no">##23828;</CHAR>
<CHAR coords="1473,838,1529,891" sep="no">##20844;</CHAR>
<CHAR coords="1475,897,1531,950" sep="no">##22971;</CHAR>
<CHAR coords="1470,971,1526,1024" sep="no">##24235;</CHAR>
<CHAR coords="1473,1031,1529,1084" sep="no">##29380;</CHAR>
<CHAR coords="1474,1098,1530,1151" sep="no">##22827;</CHAR>
<CHAR coords="1471,1161,1527,1214" sep="no">##20154;</CHAR>
<CHAR coords="1471,1236,1527,1289" sep="no">##22675;</CHAR>
<CHAR coords="1472,1307,1528,1360" sep="no">##35468;</CHAR>
<CHAR coords="1473,1365,1529,1418" sep="no">##37528;</CHAR>
</WORD><WORD>
<CHAR coords="1411,178,1467,231" sep="no">##22827;</CHAR>
  :
</WORD></HIDDENTEXT>
</OBJECT>
<MAP name="tou0001x.djvu">
<AREA coords="1471,178,1535,235" alt="##22823;" href="/djvuchar?5927" />
<AREA coords="1480,240,1538,296" alt="##21776;" href="/djvuchar?5510" />
<AREA coords="1479,304,1541,371" alt="##27931;" href="/djvuchar?6D1B" />
<AREA coords="1476,367,1538,434" alt="##24030;" href="/djvuchar?5DDE" />
<AREA coords="1477,438,1533,491" alt="##21029;" href="/djvuchar?5225" />
<AREA coords="1481,499,1537,552" alt="##39381;" href="/djvuchar?99D5" />
<AREA coords="1474,562,1530,615" alt="##22823;" href="/djvuchar?5927" />
<AREA coords="1477,642,1533,695" alt="##23559;" href="/djvuchar?5C07" />
<AREA coords="1475,700,1531,753" alt="##36557;" href="/djvuchar?8ECD" />
<AREA coords="1478,772,1534,825" alt="##23828;" href="/djvuchar?5D14" />
<AREA coords="1473,838,1529,891" alt="##20844;" href="/djvuchar?516C" />
<AREA coords="1475,897,1531,950" alt="##22971;" href="/djvuchar?59BB" />
  :
</MAP>
</BODY>
</DjVuXML>

```

図 5: 「唐洛州別駕大將軍崔公妻庫狄真相墓誌銘」の DjVuXML

変換の例を挙げると、「1471,178,64,57,22823」という文字ボックスに対しては、テキストビハインド内の透明文字を表す CHAR タグ

```
<CHAR coords="1471,178,1535,235" sep="no">&#22823;</CHAR>
```

と、ミニボックスおよびハイパーリンクを表す AREA タグ

```
<AREA coords="1471,178,1535,235" alt="&#22823;" href="/djvuchar?5927" />
```

とに変換をおこなう。釈文中の句読点に対しては CHAR タグや AREA タグの生成はおこなわないが、句読点ごとに「</WORD><WORD>」という WORD タグ上の区切りを入れることで、句読点をまたいだ検索を抑制している。この DjVuXML 形式ファイルを、`djvuparsexml`<sup>†</sup>を用いて、拓本 DjVu 画面の DjVu ファイル中に埋め込み、透明テキスト付き画像を実現している。さらに、DjVuXML 形式ファイルの CHAR タグと WORD タグを、そのまま OpenText でインデクス化することで、文字列検索エンジンを実現している。

集字結果画面の各画像は、DjVu ファイルから `ddjvu`<sup>†</sup>と `pnmcut` で文字画像を抽出し、`pnm-scale`と `cjpeg` で幅 50 ピクセルの JPEG 画像としている。集字結果が複数の文字に渡る場合、それらが拓本で同一行にあるときには単一の JPEG 画像としているが、複数行に渡るときには複数の JPEG 画像を CSS の縦書きモード<sup>§</sup>で上下に配置している。また、拓本 DjVu 画面におけるハイライト表示は、DjVu プラグインの機能を用いて実現している。実際には URL 中の `DJVUOPTS` パラメータと `HIGHLIGHT` パラメータがそれである。ただし、DjVu プラグインでは座標原点が画像の左下となっており、たとえば図 3 でハイライト表示されている「墓」は、CSV 形式ファイル中では「1471,1236,56,53,22675」（図 4）、DjVuXML 形式ファイル中では「coords="1471,1236,1527,1289"」（図 5）だが、URL 中では「HIGHLIGHT=1471,789,56,53」（図 3）となる。

### 3 「縁」モデルの導入

拓本文字データベースの集字結果画面に、漢字の「縁」を基にした検索モデルを 3 種類、導入した。この章ではそれについて述べる。

#### 3.1 「熟語」という「縁」

漢字が連続して生起する際に、それぞれの漢字の「義」を超える意味が生じる場合、それは伝統的に「熟語」という概念で扱われてきた。情報学的には、一定以上の生起確率を持つ N-gram が連続している場合、そこに「熟語」や「決まり文句」が隠れていることが多い [4]。

<sup>†</sup>LizardTech 社の SPARC Solaris 版『Document Express with DjVu』中のコマンド。フリーの『djvulibre 3.5.14』中のコマンド `djvuxmlparser` は、残念ながら CHAR タグに対応していない。

<sup>‡</sup>『djvulibre 3.5.14』中のコマンド。

<sup>§</sup>HTML 上は `STYLE="writing-mode:tb-rl;width:50px"` の SPAN タグ。縦書きをサポートしていないブラウザでは、残念ながら上下に配置されない。



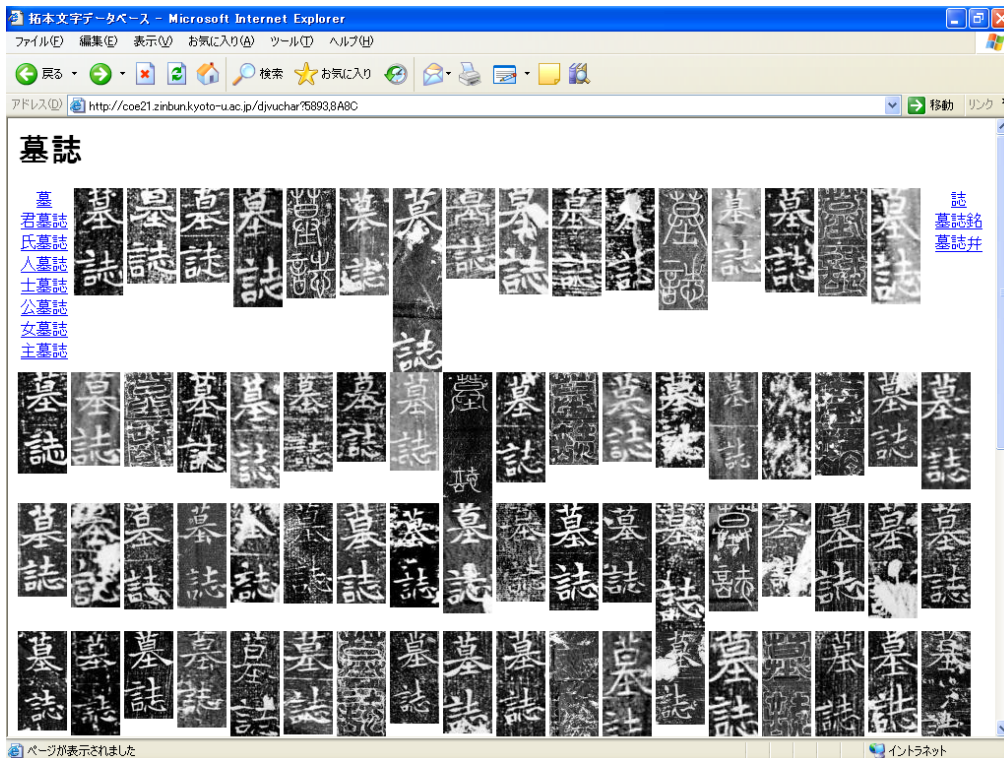


図 6: 「墓誌」の集字結果

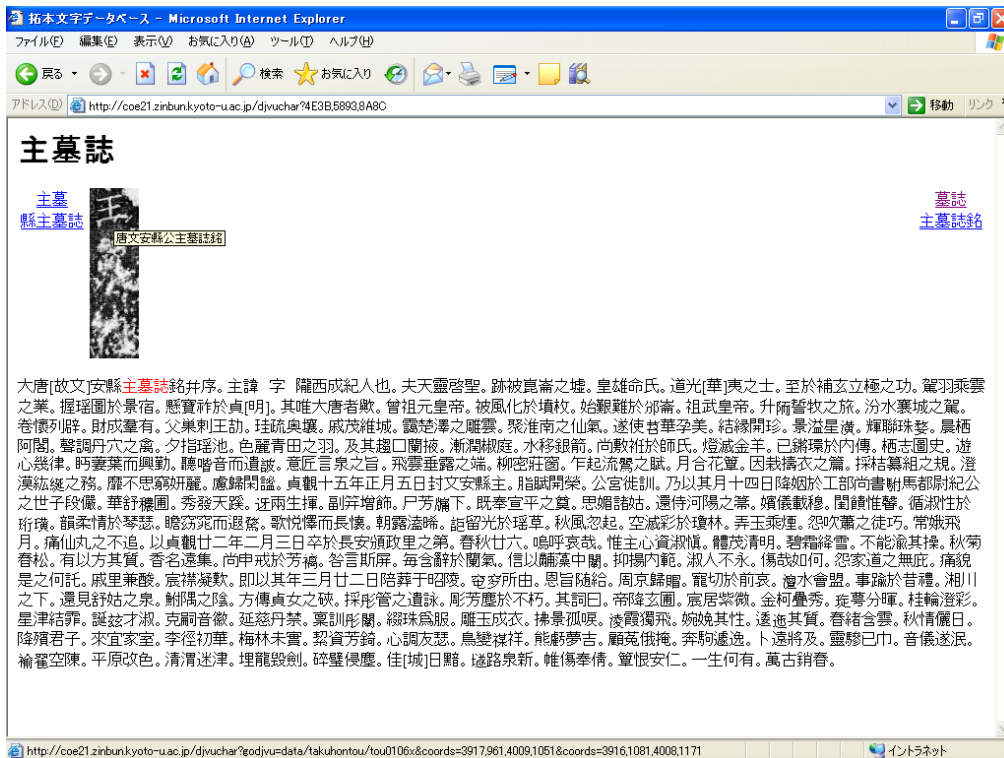


図 7: 「主墓誌」の集字結果

このモデルを集字結果画面に導入するにあたり、意味上の「熟語」の判断は検索者にゆだねることとし、画面上は生起確率の高いN-gramから順に並べて表示することにした。たとえば図2の集字結果画面では、「墓」を含む2-gramのうち、「君墓」と「墓誌」が上位にあることが一目でわかるようにした。各N-gramは、そのままその文字列の集字結果画面にリンクしており、たとえば図2の画面で、右上の「墓誌」をクリックすると、図6の画面にジャンプする。あるいは図6の画面で「主墓誌」をクリックすると、図7の画面にジャンプする。ただし、孤立用例<sup>1)</sup>においては、それを含むN-gramの生起確率を論じられないため、当該拓本の釈文を示して検索者の判断をあおぐようにした(図7)。

### 3.2 「異体字」という「縁」

「異体字」も漢字の「縁」の1つである[1]。筆者は「全国漢籍データベース」で、文字列検索における異体字の展開手法を実現しており[5]、拓本文字データベースにも同様の手法を導入した。平均的な日本人に対しては常用漢字による検索を、平均的な中国人に対しては簡化字による検索をそれぞれおこなえるようにするため、筆者の異体字対応表は非常に現代的なものとなっている。たとえば「弁」を検索した場合には、「弁」「辨」「瓣」「辯」などが同時に集字される(図8)。

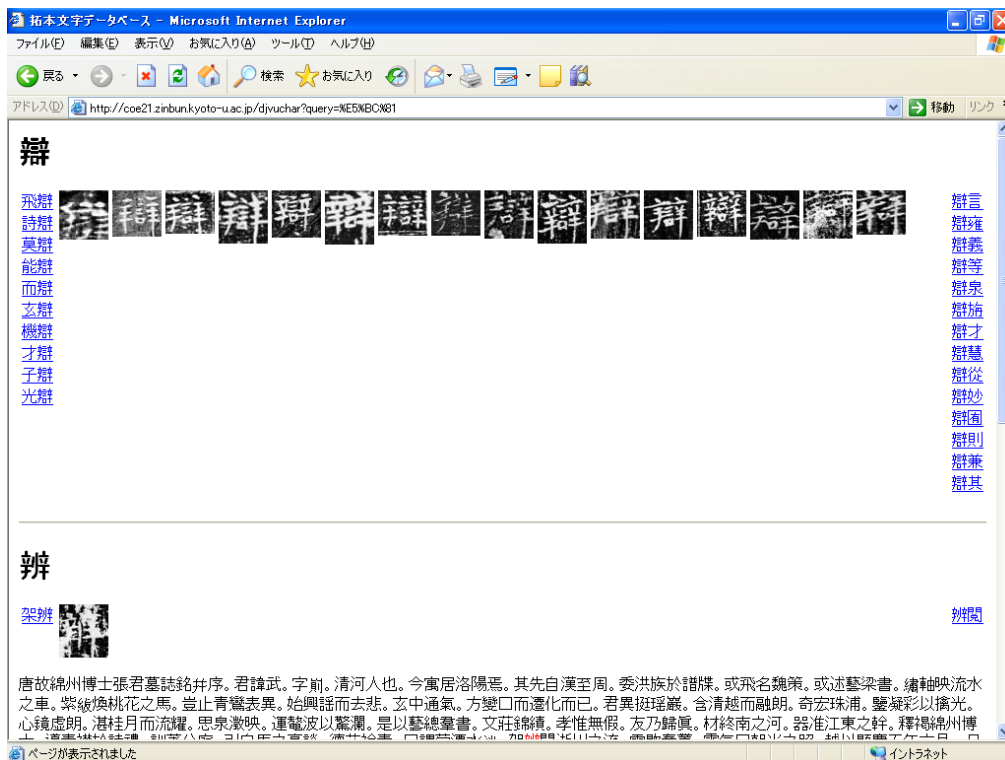


図8: 「弁」の集字結果

<sup>1)</sup>複数の同じ用例が単一の拓本のみに見える場合を含む。



### 3.3 「同音」という「縁」

「異体字」が「同音」かつ「同義」の「縁」ならば、それを緩めた「同音」という「縁」も考えられるだろう。拓本文字データベースに「同音」という「縁」を導入するにあたり、筆者は「現代韓国音」による「縁」を考え、ハングルによる検索を実現した。平均的な韓国人は、漢字による検索をめったにおこなわないため、ハングルによる検索は必須だと考えられるからである。実際には検索画面において、たとえば「각」を入力すると、「刻」「却」「各」「恪」「脚」「覺」「角」「閣」などを同時に集字する(図9)。検索画面においては、複数のハングルによる文字列検索も可能としている。

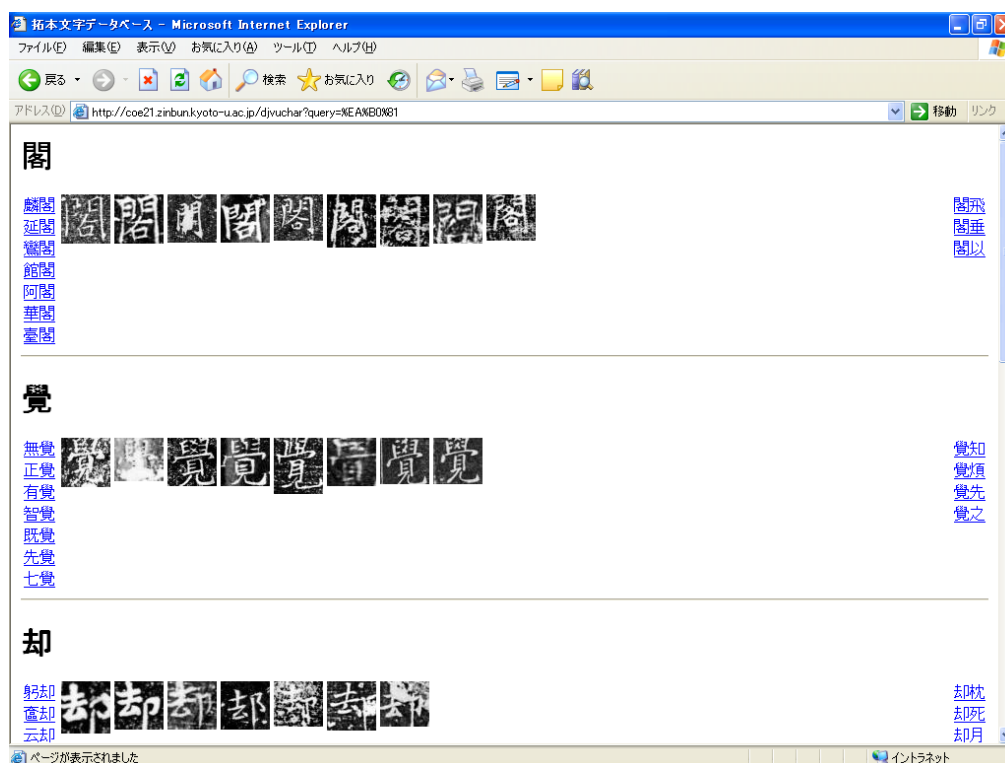


図 9: 「각」の集字結果

## 4 おわりに

透明テキスト付き画像による拓本文字データベースを構築するにあたり、漢字の「縁」モデルを適用することを考え、それを実装した。この拓本文字データベースは、京都大学 21 世紀 COE「東アジア世界の人文情報学研究教育拠点」の成果物の 1 つであり、現在 <http://coe21.zinbun.kyoto-u.ac.jp/djvuchar> で公開中である。この拓本文字データベースは、日本語環境のみならず、中国語環境(簡体字でも繁体字でも)や韓国語環境からも容易に検索可能である。現時点で拓本文字デー

データベースに収録されているのは、約 300 点の唐代拓本である。今後は他の時代の拓本も収録していくことで、時代による漢字字体変遷などの研究に資したい。

## 参考文献

- [1] 高田時雄: 現代における漢字の新要素, 東洋文化, 第 79 号 (1999 年 3 月), pp.1-7.
- [2] 安岡孝一: 透明テキスト付き画像へのいざない, 第 14 回「東洋学へのコンピュータ利用」研究セミナー (2003 年 3 月), pp.31-42.
- [3] 安岡孝一: 透明テキスト付き画像作成ツールの開発, 第 15 回「東洋学へのコンピュータ利用」研究セミナー (2004 年 3 月), pp.9-16.
- [4] Makoto Nagao, Shinsuke Mori: “A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese”, COLING 1994: 15th International Conference on Computational Linguistics (August 1994), pp.611-615.
- [5] 安岡孝一: 全国漢籍データベースの設計と WWW での運用, 全国文献・情報センター人文社会科学学術情報セミナーシリーズ, No.12 (2002 年 11 月), pp.45-57.