日韓中同形二字漢字語の品詞性ウェブ検索エンジン [1]

于 劭贇 [2]
玉岡 賀津雄 [3]

要約: 本検索エンジンは，同形二字漢字語の日韓中の3言語間の品詞性およびそれに関連する情報（朴・熊・玉岡，2014a，2014b；熊・玉岡，2014）を，Web 上で誰でも自由に閲覧・検索できるように開発した。この検索エンジンは，中国語あるいは韓国語を母語とする日本語学習者や研究者を対象としている。エンジンの構築には，MongoDB，Flask，AngularJS などの最新の技術を使用した。朴・熊・玉岡（2014a）は，日韓中3言語の8冊の国語辞書をもとに，日本語能力試験の4級から2級までの 2,060 語の二字漢字語の品詞情報を中心としたデータベースを作成している。しかし，このデータベースは 23 列（ID 番号を除く）の情報からなる Excel の表形式のデータであり，そのままでは閲覧・検索がし難い。そこで，このデータベースを広く一般に利用してもらえるように，多様な検索機能を備えた検索エンジンを開発し，Web 上で公開した（http://kanjigodb.herokuapp.com）。本検索エンジンでは，データを6種類のデータのグループに分け，分かりやすいカードの形で表示した。そして，(1) 単語を個別に確認するのに適した「シンプル検索」および(2)一定の特徴を持つ単語を選び出すのに適した「詳細検索」，の2つの検索機能を構築した。本稿では，この検索エンジンの使用法を説明した。

キーワード: 同形漢字語　品詞性　ウェブ検索エンジン　データベース　日本語教育

---

[1] English Title: A Web-accessible search engine for grammatical category for orthographically-similar two-kanji compound words between Japanese, Korean and Chinese

[2] YU, Shaoyun (Graduate Student, Graduate School of Languages and Cultures, Nagoya University, E-mail: eric.rongmu@gmail.com)
[3] TAMAOKA, Katsuo (Professor, Graduate School of Languages and Cultures, Nagoya University), E-mail: ktamaoka@lang.nagoya-u.ac.jp)

1．はじめに

日本語では漢字表記の語彙が多い。とりわけ，漢字2字で構成される語彙は日本語の国語辞書の見出し語の約 70% を占めており（Yokosawa & Umeda, 1988），頻繁に使われている。その比率を反映して，『日本語能力試験出題基準【改訂版】』（国際交流基金・日本国際教育協会，2007）に掲載されたレベル別の語彙一覧でも多くの漢字表記語が挙げられている。中国語を母語とする日本語学習者は，母語の漢字知識を活用して，これらの語彙の認知的な処理が促進させることが知られている（玉岡, 1997, 2000; 大和・玉岡, 2013a; 大和・玉岡・初，2013，先行研究の紹介は，Tamaoka，2014）。また，韓国語を母語とする日本語学習者については，現代韓国語では，漢字があまり使われていないものの，漢字由来の語彙の知識を活用して，音韻的に（ある程度は書字的に）日本語の漢語を習得できるとされている（大和・玉岡, 2013b）。

ところが，同じ意味の漢字語でも，日韓中3言語における品詞がそれぞれ異なることで，学習者が誤用を起こすことが少なくない。例えば，日本語では「乾燥している」のように，「乾燥」を動詞として使うのに対し，中国語では「乾燥」は形容詞である。そのため，中国語母語の学習者が，「乾燥だ」と表現してしまうことがある。これらの誤用は，なかなか回避することが難しく，日本語能力の上級になっても続く傾向がある。また，日本語についても，1つの漢字語の品詞の定義が必ずしも統一されていない。例えば，「学問」という語について，この品詞を「名詞」とする国語辞書もあれば，「名詞・自動詞」とする国語辞書もある。加えて，「名詞・自動詞・他動詞」とする国語辞書も見られる。このように，中国語・韓国語を母語とする日本語学習者の漢字語の習得を検討する際は，語彙の意味だけでなく，言語間の複雑な品詞性の非対称性も考慮しなくてはならない。

朴・熊・玉岡（2014a）は，こうした言語間の品詞性の違いを一覧にして，「同形二字漢字語の品詞性に関する日韓中データベース」を構築した。このデータベースでは，日韓中三言語の合計8冊の国語辞書をもとに，品詞情報を掲載したものである。これにより，複数の国語辞書の定義を同時に確認し，明確な基準を持って二字漢字語の各言語における品詞性を判断することができるようになった。しかし，このデータベースは，ID 番号を含んで 24 列で構成された Excel の表の形を取っており，閲覧・検索するのに不便である。そこで，日本語学習者から研究者まで幅広い利用者が簡便にアクセスできるように，朴・熊・玉岡（2014a）のデータベースを Web 上で自由に検索できるように豊富な検索機能を備えたシス

テムを開発し，検索エンジンを http://kanjigodb.herokuapp.com の URL で公開した。

2. データの構成と検索結果の表示

オリジナルのデータベース（朴・熊・玉岡，2014a，2014b；熊・玉岡，2014）の Excel の表形式のデータは，表1に示したように，ID 番号を含んで 24 列で構成されている。ただし，番号は ID を除いて，1から 23 までとした。現時点で本検索エンジンに収録されている語彙は，『日本語能力試験出題基準』（2007，改訂版）の＜文字・語彙＞から抽出した4級から2級までの漢字2字で作られている 2,060 語である。これらには，「叔父」または「伯父」で「おじ・おじさん」と読ませている語も含まれている。これは，父・母の兄には「伯父」，弟には「叔父」と書く違いを反映している。また，「叔母」または「伯母」は，「おば・おばさん」と読ませる。これも，父・母の姉は「伯母」，妹には「叔母」と漢字表現を変えることからきている。また，「祖父」を「おじいさん」，「祖母」を「おばあさん」と読ませている。これらの 6 語は，出題基準の語彙として含まれているものの，それぞれの漢字に対応した発音というより，2つの漢字の組み合わせで慣習的に親族関係を表現しているだけであり，漢字表現そのものは，日本語学習者にとっての基本語彙というわけではない。しかし，出題基準の語彙として掲載されているので，そのままデータベースにも含まれている。そのため，本検索エンジンでもこれらの二字漢字語も含むことにした。

また，朴・熊・玉岡(2014a)のデータベースは，もともとは日韓中の3言語間の品詞性の違いを示すことを目的としていた。その後，2014 年の出版時点のデータバースに，文化庁(1978)と張(1987)による同形漢字語の日中両言語間の意味的関係の情報が新たに追加された（表1の最下2行）。意味的関係は，2字漢字語の情報として非常に貴重である。

(1)「書字」：日韓中の3言語における漢字表記（列番号：1，11，17）
(2)「発音」：日韓中の3言語における発音 （列番号：2，12，18）
(3)「難易度」：語彙の難易度を評価する指標（列番号：10）
(4)「頻度」：コーパスにおける使用頻度（列番号：8，9）
(5)「品詞性」：日韓中の3言語の国語辞書による品詞情報
（日本語は番号列で 3-7，韓国語は番号列で 13-16，中国語は番号列で 19-21）
(6)「日中意味的関係」：日中両言語間の意味的関係（列番号：22，23）

これらの6つの分類により，検索後の情報が，分かり易く提示できる。

表 1 オリジナルのデータベースの構造

#	列ヘッダー	説明
	No	見出し語の番号
1	表記	日本語の見出し語の漢字表記
2	読み	日本語の読み表記（ひらがな）
3	新明解	新明解国語辞典(第7版)【机上版】の品詞情報
4	角川	角川新国語辞典(初版) の品詞情報
5	岩波	岩波国語辞典(第6版) の品詞情報
6	明鏡	明鏡国語辞典(第2版) の品詞情報
7	旺文社	旺文社詳解国語辞典(初版) の品詞情報
8	朝日新聞	1985-1998 年の朝日新聞の使用頻度, 天野・近藤 (2000)
9	毎日新聞	2000-2010 年の毎日新聞の使用頻度
10	級	『日本語能力試験出題基準』（2007, 改訂版第4刷）の配当級
11	韓・漢字表記	韓国語の漢字表記
12	ハングル	漢字のハングル表記
13	~*hata*(動)	能動態の動詞を作る接辞-*hata*
14	~*toyta*(動)	受動態の動詞を作る接辞-*toyta*
15	~*hata* (形)	形容詞を作る接辞-*hata*
16	~的	接辞「 - 的（-*cek*）」
17	中国語表記	中国語の漢字表記(簡体字)
18	中国語読み	中国語の読み表記(ピンイン)
19	漢語辞典	現代漢語辞典(第5版)の品詞情報
20	規範辞典	現代漢語規範辞典(第1版)の品詞情報
21	一致性	『漢語辞典』と『規範辞典』に掲載された品詞情報の一致性
22	文化庁	文化庁(1978)による日中同形語の意味的関係の分類
23	中日対比辞典	張(1987)による中日同形語の意味的関係の分類

表 1 に示したように，もとものとデータベース（朴・熊・玉岡，2014a）では，ほぼ日韓中の3言語の順にしたがって，二字漢字語についての 23 列（ID番号を除く）の情報を並べられていた。しかし，本検索エンジンでは，情報の種類に基づき，オリジナルのデータの列を図1の例のように，6つのグループに分けた。なお，表1の番号の1から23は，そのまま図1の①から㉓と一致するように表示した。

図 1 検索結果の画面

図1のように，ヘッダーの見出し語部分には，「未来」と表示された日本語の書字を示した。その横に，「難易度」の指標として使われることが多い二字漢字語の⑩の JLPT レベル，さらに，⑧の『朝日新聞』14 年分と⑨の『毎日新聞』の 11 年分の使用頻度を表示した。これらは二字漢字語に関する研究をするにおいては，基本情報である。次に，語彙の基本情報として，日韓中の３言語における①書字と②発音の情報を目立つ形で提示した。中国語も日本語と同じ２字の漢字で表記されるので⑰の「未来」となり，その発音をピンイン(pinyin)で，⑱に weilai と示した。韓国語の漢字表記は現在ではほとんど使われていない。しかし，漢字を使用する場合には，日本や中国と異なり漢字の韓国では簡略化を行っていないので，字体は、日本語のいわゆる旧字体、中国語の繁体字に近いものが使われている。韓国語の漢字表記にしたがい⑪のように「未來」と表記する。韓国語の発音の情報については，もちろんハングル表記を使い「미래」と表記するが，ハングルに馴染みのない利用者のために，イェール式のローマ字表記の milay も併記できるようにした。

基本情報に続くのは，同形二字漢字語の各言語における「品詞性」の情報である。日本語については図１の③から⑦に示した５冊の国語辞書，中国語については⑲と⑳に示した２冊の国語辞書による品詞情報を提示するようにした（詳細は，熊・玉岡, 2014 を参照）。さらに，中国語の２冊の辞書で品詞の記述が一致しているかどうかを，一致していれば✓のチェックマークを，一致でなければ×のバツマークで示した。韓国語については日本語・中国語と違い，品詞性を記載する代わりに，『標準国語大辞典』（国立国語院，2002）に基づいて，独自に４種の接尾辞の接続が可能かどうかを調べ，品詞性を判断する基準とした（詳細は，朴・熊・玉岡, 2014b）。まず，動詞については，図１の⑬と⑭に示したように，日本語のサ変動詞に相当する軽動詞の能動態の接尾辞-hata あるいは受動態の接尾辞-toyta が二字漢字語につくかどうかで判断した。形容詞については，⑮のように形容詞を作る接辞の-hata が付くかどうかで判断した。なお，能動態動詞の-hata と形容詞の-hata は同じであるが，使い方が違うので，違いを判断することができる。⑯は，日本語でいえば「的」に相当する接辞であり，韓国語でも「的」で적または-cek と表記される。接尾辞の적（cek）が付加される場合には✓で，付加されない場合には×で示した。

さらに，カードの最後には，図１の㉒と㉓に「日中意味的関係」の情報を記載した。文化庁（1978）も張（1987）も，日中同形漢字語の意味的関係を，S(Same)，O(Overlap)，

D(Different)，N(Nothing)の 4 つのカテゴリーに分類している。Sは，たとえば「音楽」のように日中で意味が同じ（または極めて近い）語，Oは，たとえば「兄弟」のような意味が一部重なる語，Dは「手紙」のような意味が著しく異なる語，Nは「財布」のようにどちらか一方の言語にしか存在しない語を表している。また，検索された語について，前述のデータに欠損値が存在する場合は，すべて「⊘」のマークで示される。たとえば，「財布」という語は，日本語にしか存在しないため，その語の中国語と韓国語における情報はすべて「⊘」で示されることになる。

このように，本検索エンジンは，多くの列で構成された，もとの Excel の表形式のデータを，図１のようなカードの形式で提示するようにした。これにより，日韓中の３言語の書字，発音，品詞性，そして意味の異同が確認しやすくなり，もとのデータベースの閲覧の不便が解決された。

3. 検索機能

本検索エンジンの検索機能は「シンプル検索」と「詳細検索」の2つを準備した。「シンプル検索」は，書字あるいは発音が，検索ボックスに入力された内容と一致した単語を探すシンプルな検索機能である。一方，「詳細検索」は，前節に述べた6種類の情報のすべてを任意に組み合わせて，検索条件に指定できる高度な検索機能である。なお，詳細検索は，Internet Explorer では，フロントエンドに使われている Angular Material の技術にまだ完全に対応していないため，本検索エンジンの使用においていろいろな問題がある。そのため，Chrome あるいは Firefox による閲覧を推奨する。以下，それぞれについて詳細に使用法を説明する。

3.1 シンプル検索

シンプル検索では，書字と発音のいずれかの情報に対する完全一致検索である。そのため，語の書字と発音を知った上で，それについての情報を個別に確認するのに適している。図2に示したように，検索エンジンのページの最上部に位置する検索ボックスに，書字あるいは発音の情報を入力し，エンターキーあるいは検索ボックスの右側の検索アイコンを押すことで，シンプル検索が実行できる。日韓中の3言語のいずれの書字あるいは発音でも検索することができる。

図 2　検索ボックス

例を挙げると，「学問」という語についての情報を，漢字から検索する際には，日本語の書字の「学問」，中国語簡体字の「学问」あるいは韓国語の「學問」のいずれからでも同じ検索結果が得られるようにした。また，発音から検索するに際も，日本語ひらがなの「がくもん」，中国語ピンインの「xuewen」，韓国語ハングルの「학문」あるいはそのイェール式の「hakmwun」のいずいれでも「学問」の情報を検索することができる。検索で一致した二字漢字語が見つかれば，その語の情報が検索結果として，第2節で説明した図1のような形で表示される。

3.2　詳細検索

ある種の情報あるいは複数種の情報の組み合わせに基づいて，一定の特徴を持った語を選び出す場合も考えられる。そのために開発されたのが詳細検索機能である。この検索は，日本語の語彙能力テスト作成における二字漢字語検索，漢字語学習のカリキュラムの開発などに利用できると思われる。検索アイコンの右側にある図3に示した「詳細検索」ボタンをクリックすることで詳細検索を始めることができる。

図 3　詳細検索を開くボタン

詳細検索をクリックすると，図4に示した詳細検索のための検索条件を入力・選択するフォームを開くことができる。誤った操作を避けるため，入力フォームが開いた状態では，シンプル検索の検索ボックスと検索アイコンが利用できないように設定した。シンプル検索を有効にするためには，「詳細検索」ボタンをもう一度クリックして，入力フォームを閉じなくてはならない。詳細検索では，書字と発音の情報のほか，第2節に述べた「難易度」「頻度」

「品詞性」および「日中意味的関係」の全ての情報が検索条件に指定できる。入力フォームの最下の「検索」ボタンで，検索条件に一致した語を検索することができる。そして「リセット」ボタンで，前に入力された検索条件をクリアすることができる。その使い方を具体的な検索例を使って説明する。

図４　詳細検索の入力フォーム（一部）

４．詳細検索の使用法

4.1　難易度に基づく検索

オリジナルのデータベース（朴・熊・玉岡（2014a）は，『日本語能力試験出題基準』（2007，改訂版）を使って旧 JLPT のレベルを入力し，それを難易度としている。旧日本語能力の出題基準では，4級から1級までの4つのレベルに分かれている。しかし，データベースでは，もともと2級レベルまでの 2,060 語の二字漢字語を選択しているので，4級から2級までしかない。本検索エンジンも JLPT のレベル4からレベル2までの3つのレベルが選択できるようにしてある。これらのレベルは，入力フォームの右上に位置する図5の「難易度の指定」から選択できる。

図5　難易度の指定

JLPT のレベルでは，2級レベルの「JLPT2」，3級レベルの「JLPT3」および4級レベルの「JLPT4」の3つの選択肢から，検索される単語の難易度を指定することができる。たとえば，図5のように，「JLPT4」を選択して検索を行うと，まず，図６に示されているような結果リストが表示される。左下に，「172 件の結果が見つかりました。」と表示され，172 語見つかったことが分かる。

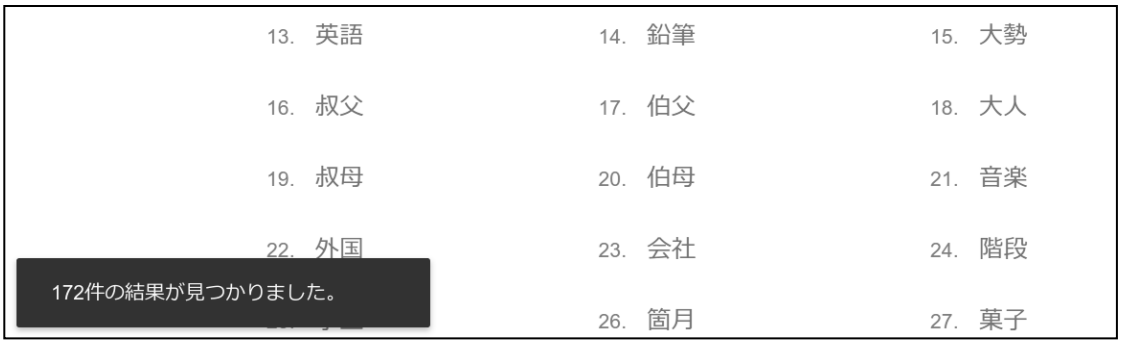

図６　詳細検索の結果リスト

結果リストの各項目は，日本語の漢字で表記されており，検索結果に対して通して番号が付けられている。特定の単語をクリックすると，前述の図１で示したような「単語カード」が，ポップアップウィンドウの形で，結果リストの上に表示される。図７は，14 の「鉛筆」をクリックした場合の例である。全体は，図１にように表示されるが，ここでは，上の方の一部だけを示している。この方法で，個々の二字漢字語についての詳細な情報を確認することができる。ポップアップウィンドウの右上の「閉じる」アイコンをクリックすると，画面が結果リストに戻り，他の操作に移ることができる。

図7　詳細検索における「単語カード」の表示（一部）

4.2 日中意味的関係に基づく検索

本検索エンジンでは,『中国語と対応する漢語』（文化庁，1978）と『中日漢語対比辞典』（張，1987）の２つの参考資料の意味類似性の分類を，検索条件に指定することができる。図８に示した入力フォームの「日中意味的関係」セクションで， Same，Overlap，Different，Nothing の４つから，いずれのセレクトボックスをクリックすることで，日中両言語間の意味的関係に基づいて，二字漢字語を選択することができる。

たとえば，中国語と日本語の意味が異なっている二字漢字を調べたいとすれば,「文化庁（1978）」のセレクトボックスで「Different」を選択すればよい。検索を実行すると，本検索エンジンに収録された日中同形二字漢字語のなか，文化庁(1978)によって「Different」に分類された語の 35 語がすべてリストアップされる。これらの語数からも分かるように，日中で異なる意味を持つ語は極めて稀であることが分かる。あるいは，中国語にはない日本語を選ぶ場合には，「Nothing」で選択すると，256 語がリストアップされる。こちらも全体の 2,060 語から考えれば 10%くらいである。なお，詳細の日中の意味分類については，熊・玉岡（2014）を参照のこと。この選択機能で，中国語と意味が異なっている語，中国語には存在しない語を容易に選べるので，日本

図８　日中意味的関係の検索条件の指定―文化庁（1978）の Different の例

語教育における授業や語彙理解テストでの語彙選択にも活用できるであろう。

4.3 複数の条件を組み合わせた検索

詳細検索の良さは，2つ以上の条件を組み合わせて検索することができることである。つまり，複数の条件に一致した語に絞って，二字漢字語を選択することができるのである。特定の特徴を共有する語彙を選択できるので，より細かい考察が可能になる。たとえば，「JLPT レベルが4級で，なおかつ文化庁（1978）における分類が Different になっている」という条件（図5と図8の両方の条件を組み合わせ）で検索を行ってみる。レベル４（図5）と「Different」（図8）の条件での単独検索だと，それぞれ 172 語と 35 語の結果が得られるが，両方を組み合わせると，共通部分，$A \cap B = \{x | x \in A, x \in B\}$，$A$=レベル４，$B$=Different，に当たる「段階」「結構」「丈夫」「新聞」などの９語が選出される。これが，旧 JLPT４級に存在する日中同形異義の二字漢字語が結果リストに相当する。

4.4 品詞情報に基づく検索

品詞情報は，本検索エンジンの中心的な語検索条件である。詳細検索の入力フォームの「品詞性」セクションでは，日韓中の3言語の品詞性に関する情報を検索条件に指定することができる。日本語については，５冊の国語辞書から品詞情報を指定することができる。中国語については，2冊の国語辞書から品詞情報を指定できる。1つの漢字語に複数の品詞性が含まれることが多いので，これらの国語辞書に対応するセレクトボックスでも，複数の品詞性を指定することができるようにしてある。セレクトボックス（選択肢を選ぶボ

図９ 日本語・中国語での品詞性を複数選択できるセレクトボックスの例

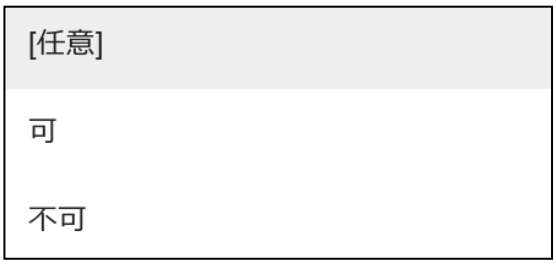

図 10 韓国語で接辞の接続の可・不可を選択するセレクトボックス

ックス)をクリックすると，該当する辞書で選択可能な品詞の種類が図９のようにリストされる。品詞の名称をクリックすることで，品詞性の選択および選択の解除を行うことができる。検索を実行すると，指定された品詞性を有する二字漢字語が選び出される。

韓国語の品詞性は第2節で述べたように，『標準国語大辞典』（国立国語院，2002）に基づいて，４種の接尾辞の接続が可能か否かで表されている。そのため，韓国語については，図10に示したように接尾辞ごとにセレクトボックスを設けた。これらのセレクトボックスでは，該当する接尾辞を付けることが可能(「可」)か不可能(「不可」)かを選択できる。検索を実行すると，接尾辞の接続上の特徴が検索条件と一致した二字漢字語が選出される。

ここに，検索例を２つ挙げる。第１の例は，日本語と中国語の両方の品詞性を指定するケースである。まず，日本語ついては，図 11 のように『明鏡国語辞典』(第2版)での品詞性が「自サ変」「形動」を含むように設定する。そして，中国語については，図 12 のように，『現代漢語辞典』（第5版)での品詞性が「形」(形容詞)を含むように設定する。この２つの条件で検索すると，日本語では『明鏡国語辞典』(第2版)により，自動詞と形容詞として使え，中国語では『現代漢語辞典』（第5版)により，形容詞として使える語が選び出される。具体的には，「妥当」「安定」「共通」「相当」「不足」「迷惑」という６語がリストされる。

図 11　日本語の品詞性の設定例

図 12　中国語の品詞性の設定例

図 13　韓国語の品詞性の設定例

第２の検索例は，日本語と韓国語の品詞性を共に指定したケースである。日本語の品詞性は，例１と同じく図 11 のように設定する。そして，今度は韓国語の品詞性を，図 13 のように，単語が「하다(hata)」形容詞として使えるという条件を設定する。検索を実行すると，指定された特徴を持つ語が選ばれる。具体的には，「妥当」「凸凹」「無理」「相当」「不足」「乱暴」の６語である。

4.5 頻度に基づく検索

入力フォームの最も下に位置する「頻度」のセクションでは,『毎日新聞』の 2000 年から 2010 年までの 11 年分, または『朝日新聞』の 1985-1998 年の朝日新聞の使用頻度（天野・近藤, 2000）の 14 年分の新聞コーパスにおける使用頻度を検索条件に指定できる。なお,『毎日新聞』については, Tamaoka, Makioka, Sanders and Verdonschot (2015)が独自に算出ものである（http://www.kanjidatabase.com/を参照のこと）。使用頻度効果（word frequency effects）については, 多くの研究（Balota & Spieler, 1999 など）で証明されているので, 実験の統制条件としては必須である。例えば,『朝日新聞』のコーパスにおいて使用頻度が 234 回から 345 回までの語に限定したい場合には, 図 14 のように頻度下限の値を 345 と, 頻度下限の値を 234 とそれぞれの入力欄に入力すればよい。

図 14 頻度を検索条件に指定する

4.6 正規表現を使った検索

入力フォームの最初に位置する「書字」と「発音」の欄は, シンプル検索と同じく, 日韓中の３言語のいずれの書字と発音でも検索を行うことができる。ただし, シンプル検索での単純な完全一致検索とは異なり, 詳細検索のこれらの２つの欄では, 正規表現を使うことができる。正規表現とは, 複数の文字列を１つのパターンでマッチできる表記方法である。それを「書字」と「発音」の検索条件に利用すると, 特定の書字あるいは発音上のパターンを持つ単語を選び出すことができる。正規表現の詳しい説明は本稿の扱う範囲を超えるが, 本検索エンジンでの活用例を以下に紹介する。

4.6.1 部分一致検索

「書字」と「発音」の欄で, 入力した文字列に部分一致した語を検索できる。これは正規表現のデフォルトの動作である。たとえば, 漢字表記に「時」の文字が含まれる語を検索したい場合は, 図 15 のように,「書字」の欄に「時」を入力して検索を実行

すればよい。そうすると，「時期」「時刻」のような「時」が語の左側の頭に来る語と「当時」「同時」のような「時」が語の右側に来る語がすべてリストされる。

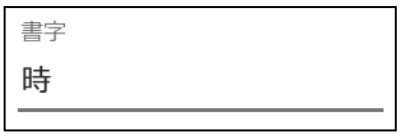

図 15 部分一致検索の例

4.6.2 特殊文字を使った検索

前述の「時」の部分一致検索では，二字漢字語の左右の位置を指定することができなかった。しかし，正規表現の特殊文字を使えば，どちらか一方の単語に絞ることができる。「書字」の入力欄の内容を図 16 のように「^時」にして検索を実行すると，「時」が語の左側に来る語だけが選出される。それに対し，図 17 のように「時$」で検索すると，「時」が語の右側に後に来る語だけがリストされる。

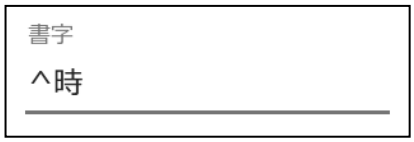

図 16 「^」記号の使用例

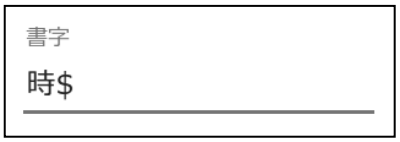

図 17 「$」記号の使用例

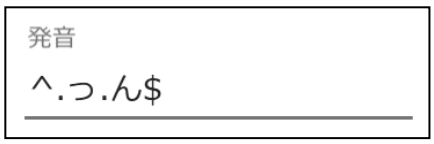

図 18「.」記号の使用例

文字列の先頭（左側）あるいは末尾（右側）に固定することのできる「^」「$」の他に，任意の1文字に一致できるドット記号「.」もよく使われる。例えば，「じっけん（実験）」や「いっぱん（一般）」のように，日本語のひらがな表記が「かな＋っ＋かな＋ん」というパターンになっている単語をすべて検索したい場合は，図 18 のように詳細検索の「発音」の欄に「^.っ.ん$」を入力して検索すればよい。そうすると，「じっかん（実感）」「せっきん（接近）」「はっけん（発見）」など，入力されたパターンに一致した15語がリストされる。

以上は，本検索エンジンのすべての検索機能である。「シンプル検索」と「詳細検索」により，もとのデータのあらゆる情報が検索に使用でき，利用者の目的に合った多様な検索

が可能となるであろう。

5. 開発の概要

本検索エンジンは，「同形二字漢字語の品詞性に関する日韓中データベース」（朴・熊・玉岡，2014a）の検索・閲覧サービスを提供するウェブ・アプリケーションである。このアプリケーションは大きく「データベース」，サーバ側で動作する「バックエンド」，そしてブラウザー側で動作する「フロントエンド」の 3 つの部分によって構成される。主として，MongoDB，Python 言語の Flask フレームワーク，JavaScript 言語の AngularJS フレームワークおよび UI（ユーザーインタフェース）フレームワークの Angular Material を用いて，SPA (single-page application)として開発された。

5.1 データベースについて

朴・熊・玉岡（2014a）のデータベースは，エクセルに基づく表形式のデータであるが，プログラミングのために MongoDB のデータの形に変換した。MongoDB は伝統的な表形式の RDB（関係データベース）ではなく，データを階層構造で保存できるデータベースである。本検索エンジンでは，個々の単語のデータが主に「情報の種類 > 日韓中３言語のそれぞれの値」という階層構造になっている。また，MongoDB のデータが JavaScript のオブジェクトの表記法に近い形式で表現されているため，フロントエンドで動作する JavaScript との整合性がいいという利点もある。

5.2 バックエンドについて

アプリケーションの中枢に当たるバックエンドは，Python 言語の Flask フレームワークに基づいて開発した。バックエンドの職能は主に次の３つに分けられる。第１には，ユーザーのアクセスに応じて，フロントエンドを動かせるための静的ファイルやスクリプトをブラウザー側に配信することである。第２には，検索機能のためのウェブ API（application programming interface）を提供し，フロントエンドからの検索リクエストを受け取り，検索された結果をレスポンスとして返信することである。第 3 には，MongoDB サーバにつなげ，データベースに対する実際な検索操作を実行し，結果を獲得することである。

5.3 フロントエンドについて

ユーザーが直接操作するのがフロントエンドである。本検索エンジンのフロントエンドは，ビジネスロジックの面では主に AngularJS を，ユーザーインタフェースの面では主に Angular Material を利用して開発した。フロントエンドは，ユーザーに操作しやすい使用環境を提供し，バックエンドを介して，ユーザーの見つけたい同形二字漢字語の情報をデータベースから取り出し，ウェブページに分かりやすく表示する。

6. おわりに

母語と日本語に共通した□字漢字語の特性が，中国語または韓国語を母語とする学習者の日本語習得にどのような影響を与えるかを検討する際は，語彙の意味のほか，品詞性の違いも視野に入れる必要がある。「同形□字漢字語の品詞性に関する日韓中データベース」（朴・熊・玉岡，2014a）は，日韓中三言語の品詞性の異同を比較するための基準を構築する試みである。本検索エンジンの開発と公開は，こうした研究成果へのアクセスのハードルをできるだけ下げる試みである。本検索エンジンは，もともとのデータの複雑な内容を，情報の種類によって分かりやすく表示し，「シンプル検索」と「詳細検索」の機能を通じて，データベースを多様に活用できるようにした。日本語教育に携わる研究者だけでなく，教育現場にいる日本語教師や一般の学習者にも利用していただけることを切に願っている。

［参考文献］

Balota, D. A., Spieler, D. H. (1999). Word-frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General, 128,* 32-55.

于劭贇・熊可欣・早川杏子・玉岡賀津雄 (2015)「同形□字漢字語の品詞性に関する日韓中データベースのオンライン検索エンジンの構築」『2015 年度日本語教育学会秋季□会予稿集』, 329–330.

国際交流基金・日本国際教育協会 (2007)『日本語能力試験出題基準【改訂版】』（第4版）東京: 凡人社.

張淑榮 (1987)『中日漢語対比辞典』東京: ゆまに書房.

朴善娰・熊可欣・玉岡賀津雄 (2014a)「同形□字漢字語の品詞性に関する日韓中データベ

ース」『ことばの科学』27, 53–111.

朴善娟・熊可欣・玉岡賀津雄 (2014b)「同形二字漢字語の品詞性に関する日韓中データベースの概要」『ことばの科学』27, 3–23.

熊可欣・玉岡賀津雄 (2014)「日中同形二字漢字語の品詞性の対応関係に関する考察」『ことばの科学』27, 25-51.

文化庁 (1978)『中国語と対応する漢語』東京: 大蔵省印刷局.

玉岡賀津雄 (1997)「国語と英語を母語とする日本語学習者の漢字および仮名表記語彙の処理方略」『言語文化研究』17, 65-77.

玉岡賀津雄 (2000)「中国語系および英語系日本語学習者の母語の表記形態が日本語の音韻処理に及ぼす影響」『読書科学』44, 83-94.

Tamaoka, K. (2014). The Japanese writing system and lexical understanding. *Japanese Language and Literature (The American Association of Teachers of Japanese, AATJ), 48,* 431-471.

Tamaoka, K., Makioka, S., Sanders, S., & Verdonschot, R. G. (2015). The new 2136 Japanese Jōyō kanji web-accessible database. http://www.kanjidatabase.com/からアクセスできる。

大和祐子・玉岡賀津雄 (2013a)「中国人日本語学習者による外来語処理への英語レキシコンの影響」『レキシコンフォーラム』6, 229-267.

大和祐子・玉岡賀津雄 (2013b)「中国語母語話者と韓国語母語話者の日本語テキストの読み処理における言語的類似性の影響」『小出記念日本語教育研究会論文集』21, 61-73.

大和祐子・玉岡賀津雄・初相娟 (2013)「中国人日本語学習者のテキストのオンライン読みにおける語彙と文法能力の影響」『日本教科教育学会誌』36(1), 33-43.

Yokosawa, K., & Umeda, M. (1988). Processes in human Kanji-word recognition. *Proceedings of the 1988 IEEE international conference on systems, man, and cybernetics* (pp. 377-380). August 8-12, 1988, Beijing and Shenyang, China.

于 劭贇 – 名古屋大学大学院 国際言語文化研究科・大学院生

玉岡 賀津雄 – 名古屋大学大学院 国際言語文化研究科・教授

A Web-accessible search engine for grammatical category of orthographically-similar two-kanji compound words between Japanese, Korean and Chinese

YU, Shaoyun (*Graduate Student, Graduate School of Language and Cultures, Nagoya University, Japan*)
TAMAOKA, Katsuo (*Professor, Graduate School of Language and Cultures, Nagoya University, Japan*)

Abstract: The paper present a search engine developed for Korean and Chinese speakers learning Japanese, which allows users to search grammatical category and related information about 2,060 basic orthographically-similar two-kanji compound words between Japanese, Korean and Chinese (Park, Xiong2 & Tamaoka, 2014a, 2014b; Xiong & Tamaoka, 2014). The 2,060 words were taken from the Japanese Language Proficiency Test (JLPT), levels 4 to 2. This search engine was built using a Flask (a framework for Python) and MongoDB back end, with an AngularJS (JavaScript) front end. The MongoDB backend is an improvement on an existing, 23 column Microsoft Excel spreadsheet (excluding ID), which was not easily searchable. The search engine is freely available on line at http://kanjigodb.herokuapp.com. Each of the 2060 words are visually represented as a card in the search result and each card contains six types of information. The application allows two search functions; (1) 'simple search' to discover information for each word, and (2) 'detailed search' do list all words with specific features. This paper explained how to use the search engine.

Keywords: two-kanji cognate compound words, grammatical category, Web-assessable search engine, database, Japanese language education