『中日大辞典』データベースの機能

齊藤正高
〈愛知大学・岐阜大学非常勤講師、中日大辞典編纂所研究員〉

2015年度より準備してきた『中日大辞典』データベースが、2019年度に公開の運びとなったので、現在までの経緯を報告する。

1．基本システムの機能

中日大辞典ウェブ・データベースは、第三版・増訂第二版の電子辞書データをもとに整理した 3 種の基本データベースからなる。その基本統計は以下である。

表 1　中日大辞典データベース主要 3 種の統計情報（単位：件）

	第三版	増訂第二版
見出し語	127,958	150,221
語釈	179,063	199,430
例文	100,245	98,508
総計	407,272	448,159

この基本統計については説明すべき内容が多岐にわたるので節を分けて述べていきたい。

1.1　見出し語

見出し語データベースは、 1)「形態情報」、 2)「ピンイン情報」、 3)「重

要度情報」を柱とする。中国語辞書においては、少なくとも形態（漢字）と音の二つがないとデータとして特定はできないので、見出し語の形態と発音記号であるピンインは一体とする方針をとった。つぎに、この三つの要素について関連事項を解説する。

1）形態情報

親漢字については基本的に簡体字・繁体字・異体字を紙版のままデータベースに収録する方針をとった。熟語についても紙版と同じである。これらを表現する文字コードはユニコードの「エクステンションＡ」まで対応した。しかし、『中日大辞典』にはこの範囲で扱えない文字もあるので、およそ2000種（第三版比で1.6％）について、「グリフ・ウィキ」で公開されている字形画像を使用して外字とした。将来、より多くの文字セットをふくむ「エクステンションＢ」で形態情報を拡張できるかもしれないが、現状は課題としておく。

「部首引き」は「漢字構造情報データベース」を公開している CHISE プロジェクトのデータを基礎にして「部首索引」をつくった。いわゆる「部首」とは『説文解字』（後漢の字書）以来の配列法に由来し、漢字群をまとめる「部」のはじめ（首）にあげられる文字のことである。紙版辞書では通常一つの文字に対して一つの「部首」が対応しているが、近年構築されている「漢字構造情報」においては文字と構成要素との関係は一対多である。したがって、データベース版『中日大辞典』においても、たとえば“情”という字は「青」と「忄」のどちらでも（あるいは両方でも）検索可能である。このような一対多の「部首引き」については「どこまで細かく字を分解するべきか」という問題がのこる。細かく分解しすぎれば、抽出結果が多くなりすぎ、候補から目標記述に到達することが難しくなる

図 1　親漢字の表示

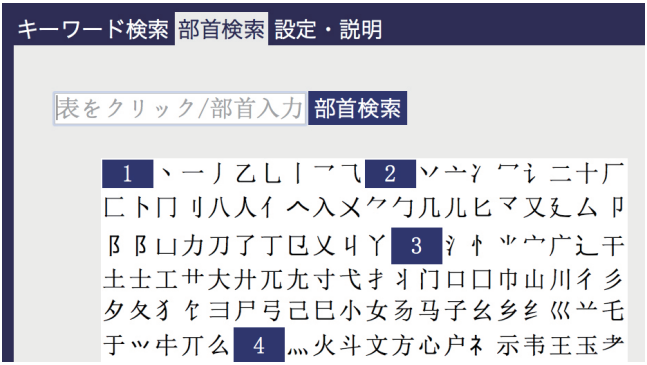


図 2　部首検索画面

ため、今後も実用的観点から調整をすすめていく。

また、電子辞書には手書き入力機能があり、学習者に好評のようである。データベース版『中日大辞典』においては、手書き入力を独自に開発することはできなかった。すでにさまざまな中国語入力アプリ（ソフト）で手書き入力がサポートされているので、それらと組み合わせて使用していただきたい。

なお、グリフ・ウィキ（https://glyphwiki.org）と CHISE プロジェクト

(http://www.chise.org) については、公開時に使用の旨を明記させていただきたいと思っている（自由共有リソースおよび、「自由ソフトウェア」)。『中日大辞典』のウェブ・データベースもこれらネット上の共同作業によって支えられた成果であることをお知らせして感謝の念を表したい。

2）ピンイン情報

「検索用ピンイン」と「表示用ピンイン」の 2 種があり、「検索用ピンイン」はさらに 2 種にわかれる。その内容は声調を付記した「数値つきピンイン」（例：Zhong1guo2）と、声調をのぞいた「数値なしピンイン」（例：Zhongguo）である。検索用ピンインを 2 種に分割したのは検索要求が「数値つき」と「数値なし」のどちらであっても目標記述に到達できるようにするためである。もちろん「数値つきピンイン」から数値を削除すれば、「数値なしピンイン」に変換することはできる。これにより、データベースとしてもつのは「数値つきピンイン」だけでよいという方針も立てうるが、表 1 にみえるように、見出し項目だけでも 12 万件以上のデータがあり、しかも、ウェブ公開をすれば多くのアクセスに備える必要がある。もし「数値なしピンイン」で入力された検索要求を「数値つきピンイン」の情報しか保有していないデータベースから抽出すれば、比較的複雑な作業をくり返すことになり、全般の効率が低下することが懸念された。したがって、ピンインによる検索要求を 2 種に判定し、データベース上の参照部分（「数値つき」と「数値なし」）を変更する仕組みを取りいれた。いずれにしろ候補には「表示用ピンイン」を表示する（例：Zhōngguó)。

また、ピンイン情報に関連して見出し語件数との関係を補足しておきたい。表 1 に示した見出し語の件数は紙版第三版にしたがい、複数の音をもつ同一形態の見出し語をべつの実体として数えた。アル化にともなう音の変化についても同様の基準で数えた（増訂第二版もこれらの数え方に従う)。多音語を別の実体とする方針はデータベースの候補表示を網羅的にするか、絞り込み機能を重視するかという問題を検討した結果である。

候補表示について、つねに内包一致を行って検索要求を「含んだ」、できるだけ多くの候補を網羅的に提供したいという方針は理想的である。しか

し、この場合、ピンインによる検索要求が極端に短い場合（例：a など）、内包一致による抽出結果が極端に多くなり、読者にとっても戸惑いの原因ともなる。つまり、短い検索要求に対しては網羅性よりも絞り込み機能を重視しないと、目標記述に到達することが困難になってしまう。したがって、多音字・多音語は別の実体として扱い、それぞれを別の候補として絞り込める可能性を確保した。

なお、第三版と増訂第二版には多くの共通する見出し語があるが、これについても調査をおこなった。結果として、第三版見出し語総数にたいして77%（約10万件）まで共通する見出し項目を特定できた。二つの版のあいだには軽声の扱いが異なるなどの差異があり、比較・特定を行うにはある程度複雑な処理が必要であった。

3）重要度情報

第三版・見出し語データに対する大きな処理として、HSK（漢語水平考試）の重要度（甲乙丙丁級）を参考に、およそ8000件にたいして「重要度スコア」をつけた（参照文献：《HSK 汉语8000词词典》北京语言文化大学出版社、2000年）。

この重要度スコアは中国語学習に間々みられる問題を一部解決するために付加した。中国語教師は学習者に辞書の引きかたを教えて、ピンインを調べさせるが、ひろく普及している電子辞書は、一部の例外はあるものの、第一声から第四声までの声調の順に候補を表示する機種が多い。これは候補が規則的に並んでいるという点で便利な面もあるが、電子辞書にまだ慣れていない学習者がこの順序にならんだ候補を選択するときに、多音字（破読音）にたいする注意をおこたると、最初に候補にあがるピンインをそのまま書きとってしまうことがある。具体的にいえば、電子辞書で「上」を検索すると、最初に候補にあがるのは shǎng（第三声）である。しかし、これは特殊な音であり、教科書などにおいてはほとんどの場合、shàng（第四声）が正しい。筆者もいくどか「上海」の「上」を shǎng に書いた提出物を訂正したことがあるが、学習者にとっては「せっかく辞書をひいたのに訂正された」という落胆につながらないともかぎらない。この誤りを好機として電子

図 3　重要度の順に表示される候補

辞書の読み方を指導することも一つの方法であるが、もし、紙版辞書をひけば、記述の多さやレイアウトによってよく使う音を選択することも可能であろう。これは紙媒体の辞書がもつ見晴らしの良さや、ブラウジング（パラパラとめくって読む）に適した紙質などによるものであろう。そして、なによりも出版社・編集者が行う版面構成によるのである。

このレイアウトによる「把握」について、ウェブ・データベースの場合を考えてみる。まず、ネットに公開すれば、これにアクセスする機器を限定することは、事実上できないという前提がある。つまり、読者は使用する機器によって広い画面でみることもあれば、電子辞書と同程度の画面で記述を読む場合もありうる。この表示画面の大きさを特定できない問題にくわえ、現状ではデータベースの記述を表示するまえに、候補を表示するという方式をとらざるをえない。とすると、候補選択時にその背後にある記述のひろがりを前もって知ることは難しい。したがって、レイアウトによる把握が読者において働くはずもない。このことはとくに初学者の場合、少なからぬ戸惑いとなるかもしれない。つまり、紙媒体における「把握」のしやすさを補うた

めに、何らかの仕組みをデータベースにも組みこまねばならないと考えた。この仕組みがなければ、データベースの候補表示は平板なデータの羅列となって、使いにくさを残すことになり、辞書をひく楽しみを損なうこともあるかもしれない。

以上をふまえて、HSK を参考にしたスコアを重要度の高い語につけるという方針に至った。検索要求によって表示される候補は、重要度スコアの順に並べかえられて表示される。これにより、shǎng と shàng の錯誤に代表されるような学習の混乱を軽減できると考えた。ただし、候補上位の声調順表示は乱れるので、声調順にデータを見たい方には、かえって不便になる側面もあるかもしれない。この点については検討をつづけていきたい。

1.2　語釈情報

中日辞典や英和辞典など、二カ国語対照辞典における語釈は文字・単語といった比較的小さい単位から成語・諺といった大きな単位までの「訳」が中心である。いわゆる「意味」と「訳」とは異なるものであろうが、読者からみれば「訳」を知るという目的はもっとも重要な要素のように思われる。もちろん、語釈に書かれていることは「訳」だけではない。ほかにも、文法解説・類義語や対義語の参照・言葉のニュアンスや文化的背景・歴史的変遷に及ぶ例もあれば、専門用語においては科学的位置づけや歴史事件の記述に及ぶ例もある。また、ひとつの見出し語に複数の語釈がある場合も非常に多い。つまり、見出し語と語釈の関係は一対多で、「多」の内容は上にあげたように、きわめて豊かである。

そして、「訳」に想定される言語だけで、語釈が記述されているわけではない。いいかえれば、中日辞典について「見出し語は中国語、語釈は日本語」と決めてかかることはできないのである。開発初期に『中日大辞典』のデータを観察した結果、語釈に相当多くの中国語記述をふくんでいることを確認できた。語釈に用いられる中国語は〔～〕によって区別がされている。つまり、語釈は日本語と中国語が並存する文章であるといえる。

では、こうした特徴をもつ語釈を十分に生かすことのできるデータベース

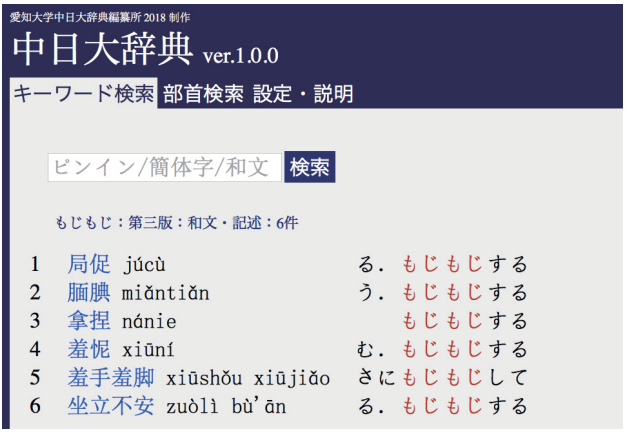


図 4　日中辞典的な使用法 (1)

図 5　日中辞典的な使用法 (2)

とは、どのような仕組みをもつべきであろうか。この仕組みは漢字（日本語・中国語）のみでなされた検索要求の場合、見出し語データベースを参照するだけでなく、語釈データベースも参照しなければならないということになる。

すでに述べたように、見出し語と語釈には一対多の関係がある。したがって、語釈情報のデータには見出し語データとのリンク（連結情報）が欠かせ

ない。このリンクを「見出し語→語釈」の方向に利用するだけでなく、「語釈→見出し語」の方向に利用することで、語釈にマッチした検索要求を見出し語に結びつけることができる。

この仕組みは語釈にふくまれる中国語記述を抽出するという目的から出発したが、副産物としてキーワードに日本語をもちいれば、『中日大辞典』を語釈の範囲で「日中辞典」として用いることもできるようになった。ただし、一般的な日中辞典は見出しに仮名をとる場合が多いように見うけられるが、『中日大辞典』の日本語記述にあらたに読み仮名をつけることは断念せざるを得なかった。仮名のみのキーワードで結果がでないときは、かな漢字混じりのキーワードを再入力して、結果を確認していただけたらと思う。

また、『中日大辞典』の特徴として、専門用語が多く採録されている点をあげることができる。この専門用語は語釈情報に所属するデータで、およそ２万項目、【天】（天文）・【医】（医学）など語釈につけられたマーカーは42分野にわたる。基本的分析として、このマーカーをもとに分野ごとの「百科項目リスト」を作ったが、これを各分野の専門家や大学院生、あるいは通訳の方々に、何らかの方法で提供できれば、中国語による専門分野の交流に役立つことがあるかもしれない。

1.3　例文情報

いわゆる「例文」は語の使用をつたえる情報である。それぞれの辞書の特徴でもあり、言葉の面白さを読者に伝える要素でもあろう。

例文データは基本的に簡体字による表記で統一されているが、一部難読語にピンインがついている。このままではピンインによって例文が分割されて、検索要求にマッチしない可能性があるので、例文記述を「ピンインつき」と「ピンインなし」に二重化し、「ピンインなし」を検索用とし、「ピンインつき」を表示用とした。

例文は語釈と一対多で結びつく情報であり、語釈とのリンクが不可欠である。このリンクを「例文→語釈→見出し語」の順にたどることは可能である。この仕組みによって全例文を検索対象にできた。

一般：第三版：簡体字・例文：52件

1	般 bān	一般大
2	般 bān	哥哥一般高
3	般 bān	跟他一般见识
4	半斤八两 bànjīn bāliǎng	八两一般儿大
5	别 bié	别于一般而言
6	不念旧恶 bùniàn jiù'è	一般反革
7	不要紧 bùyàojǐn	只是一般的感
8	豆浆 dòujiāng	点，一般是油
9	泛指 fànzhǐ	泛指一般的人
10	非同… fēitóng…	非同一般/非

図 6　例文表示　中国語コーパス的使用法

つまり、データベース版『中日大辞典』は例文の範囲で「中国語コーパス」として使うこともできる。そして、『中日大辞典』の例文を分析することで、語のネットワークなども、ある程度抽出することができる（図 7）。

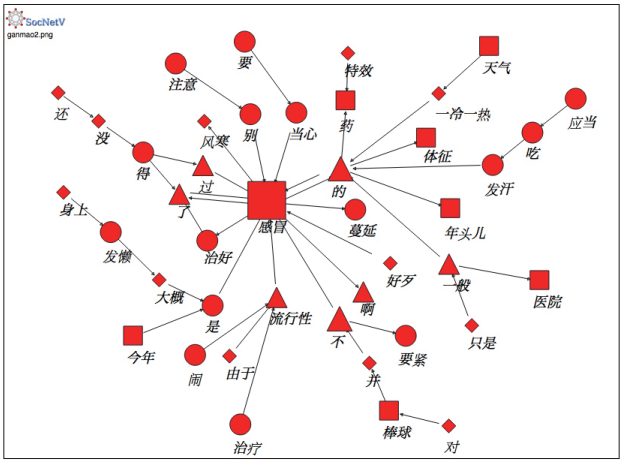


図 7　“感冒”（風邪）のネットワーク

2．今後の展望

まだ実験段階・構想段階のものもあるが、データベースの応用機能について紹介したい。

2.1　編纂機能

一般にデータベースは検索に用いるだけでなく、新たにデータを追加・変更することもできる。これはデータベース版『中日大辞典』の見出し語・語釈・例文のすべてについて言えることである。ただし、無制限に追加・変更ができれば、変更がくり返されて記述が安定しないことが予想される。したがって、変更や追加については「承認」のしくみを導入している。

辞書を編纂するしくみをウェブ・データベース上に展開できれば、新語の登録を行うことができるほかに、語のあり方に即して既収語の記述をより詳細に行うこともできる。

また、百科項目の分野別マーカーを充実させることで、きめ細かく専門用語を抽出できるということも考えられる。このような補訂は記述を充実させていくうえで重要なことであろう。

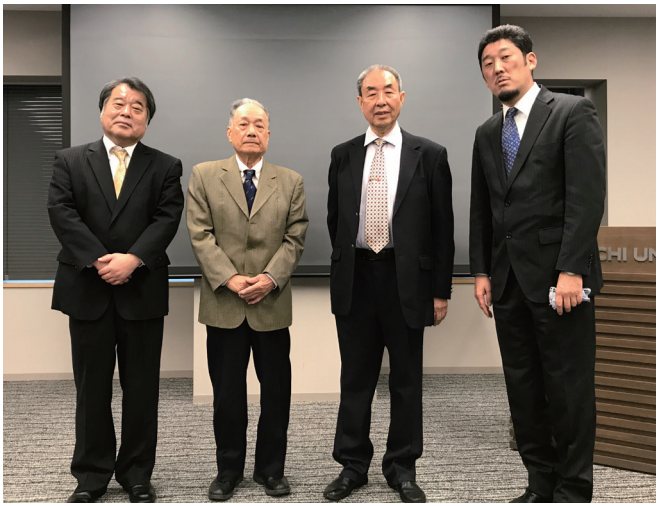
2.2　音声合成

近年のブラウザーによる音声合成（TTS: Text To Speech）は往時より改善されており、第三声の連続やアル化などの変化も「発音」できるようである。しかし、多音字の読み方にはまだ難があり、現状では見出し語や例文にとる音を、まちがえて「発音」することもある。したがって、音声合成についてはいまだ実験段階にとどまるが、今後、技術の進展をみて方針をさだめたい。

2.3　学習コンテンツの開発

辞典データベースの提示方法を変更すれば、穴埋め問題やコロケーションなどの学習コンテンツをつくることはできるだろう。しかし、これらの学習

コンテンツにどの単語をふくめるべきか、「学習上必要な語はまずどれであるか」などの問題を判断する必要もあるだろう。この点については今後、検討を重ねていきたい。



左から安部悟中日大辞典編纂所所長、今泉潤太郎先生、顧明耀先生、齊藤正高先生