漢字索引の効率性の比較分析

ヴォロビヨワ ガリーナ
キルギス民族大学コンピュータ技術・インターネット学部

キーワード　漢字索引　効率指数　選択性　比較分析　漢字のコード化

１．はじめに

漢字辞典の使い方は非漢字系日本語学習者にとって複雑であることは周知のとおりである。漢字辞典を引く際に一般に活用されている部首索引, 総画索引と音訓索引はよく知られている。しかし, 部首による検索法は,「巨は工部」などのように部首の抽出が分かりにくいこともあり,「間，閉，開は門部」,「問は門部でなく口部，聞は門部でなく耳部」などのように不統一でもある。総画索引を使用する場合は，画数を数える際に間違えたり，同画数の漢字がたくさんあり，複雑である。音訓索引を使用するためには読み方を知る必要があるが, 読み方を知らない非漢字系の人には音訓索引は利用しにくい。

より効率的な検索方法の開発を目指す先行研究があり，漢字辞典を引く際には，一般に活用される上記の部首索引, 総画索引と音訓索引以外にも漢字圏でも非漢字圏でも多様なタイプの索引が構築され使用されている。例えば，ロシアの研究者によって開発された五段排列漢字表（ロゼンベルグ（1916)），中国で構築された四角号碼（王雲五（1925)），主な意符の索引（白石（1971/1978)），カタカナ字形分類索引（加納（1998)），書き出しパターン索引（加納（1998)），意味記号索引（加納（1998)），筆順索引（若尾&服部（1989)），Key Words and Primitive Meanings Index （Heisig（1977/2001)），Index by Radicals（Hadamitzky&Spahn （1981））, System of Kanji Indexing by Patterns “SKIP”（Halpern（1988)），Kanji Fast Finder（Matthews（2004)）などである。筆者も漢字の検索に関する研究を進めてきた（ヴォロビヨワ（2005/2007，2009，2011，2012，2013)）。

２. 研究目的と研究方法

研究目的は次の通りである

(1)既存の漢字索引の記述と効率の比較分析

(2)漢字字体を適切に表す漢字のコード化に基づいた効率的な漢字索引の開発

研究方法は次の通りである

(1)多様な既存の漢字索引については，各々の効率の評価と比較分析が必要であると考えた。漢字索引の効率の比較にあたって，コンピュータデータにおける処理の効率を表す「選択性」（Selectivity）という概念を用いることにした（ヴォロビヨワ（2009)，p.72)。それから漢字索引に対して「選択係数」（Coefficient of Selectivity）という概念を導入し，選択係数の計算をもとに既存のタイプの漢字索引の効率を比較評価した。

(2)字体を適切に表す漢字のコード化に基づく新しいタイプの漢字のアルファベット・コード索引，シンボル・コード索引，セマンティック・コード索引，バイナリセマンティック・コード索引及び部首と画のコード索引を開発し，その効率を比較評価した。3 節では 12 種類の既存の漢字索引の特徴について紹介する。

３. 既存の漢字索引の記述

３．１　五段排列漢字表

19 世紀にロシアの研究者は漢字字体に着眼し，統一分類することを主義とし，未見の漢字配列と検索法を試案した。それは部首がどれか分からなかったり，画数の数え方に困ったり，読み方が分からなかったりしても，簡単な原則を覚えるだけで使える，字体で配列する方法である。

ロゼンベルグ（1916，p.自序一）は漢字辞典の使用の困難点，部首順による漢字配列について次のように記している「非常なる不便と困難とを感じたり。そは主として漢字にアルファベットの如き順序なきによれるなり。」ロゼンベルグは Васильев（ 1867）によって構築された「グラフィックシステム」を土台にし，従来の部首索引，総画索引，音訓索引とまったく違う文字の形だけに基づいて検索ができるようなシステムを構築し，斬新なる漢字配列法と検索法を含めた日本語の『五段排列漢字典』（ロゼンベルグ（1916））（図 1）を日本で出版した。ロシア人はキリル文字

やローマ字の表記に慣れていて，漢字表記でもアルファベットのような体系化の必要性を感じている。それを考慮に入れ，ロゼンベルグは一番最後に書く書記素（ストローク）によって漢字を分類し，配列した。ロゼンベルグ（1916）は 5 つのグループに分けた 24 種類の書記素を採用している。漢字に書記素順のシステムを適用するには「本」という漢字の書記素を例にして，その書記素の 5 つの方向をもとに，漢字の 24 種類の書記素を抽出した。その書記素を形と方向によって 5 つのグループに分け，各グループの 1 種類の代表的な書記素を決めた（図 2）。ロゼンベルグ（1916）, p.二〇）のシステムの基本的な書記素は下記の 5 種類の書記素である。

「↗ － ／ ，↘ － ㇏ ，↙ － ノ ，↓ － 丨，→ － 一」。

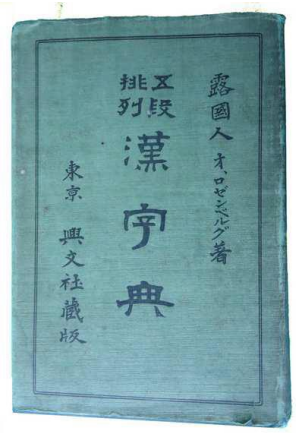

図 1 ロゼンベルグによる
『五段排列漢字典』

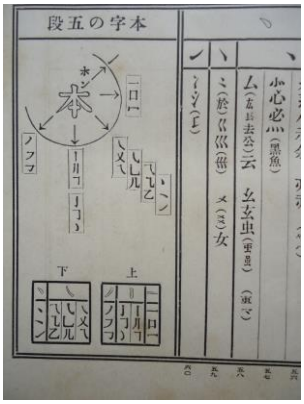

図 2 ロゼンベルグのシステムの
基本的な書記素

ロゼンベルグ（1916）の字典による「漢字の字母表」は図 3 に提示してある。字母表には 5 つの基本的な書記素，それに所属した，形によって分類された 24 種類の書記素，また 60 の欄に分けたそれぞれの書記素に所属した 567 種類の字母（漢字と漢字のパターン）が入っている。五段漢字表という索引に入っている漢字はその書記素と字母によって配列されている。

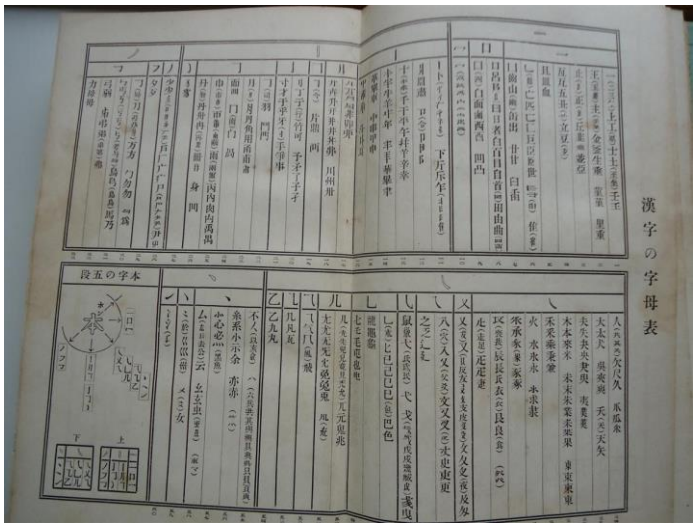

図3　ロゼンベルグによる字典の漢字の字母表

３．２　四角号碼

漢字の書記素の特色による配列と検索の「ロシアのグラフィックシステム」は，1920年代に中国で開発された四角号碼という漢字検索システムの構築に当たって参考になった（王雲五（1934），p.38）（図4，図5）。

図4『四角號碼檢字法·附檢字表』
（王雲五1934）

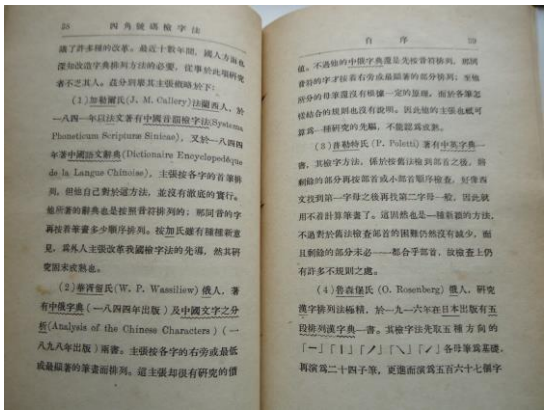

図5『四角號碼檢字法·附檢字表』
（王雲五（1934），p.38）

四角号碼（しかくごうま）は漢字の検索方式の一つであり，日本語の文脈では「四隅コード」という意味である。王雲五（Wong Yunwu）により中国で考案された。1925年に『號碼檢字法』（王雲五（1925））, 翌年には『 四角號碼檢字法』（王雲五（1926））, それから『四角號碼檢字法·附檢字表』（王雲五（1934））が出版された。四角号碼は, 部首・画数・筆順・音・意味などにとらわれず，漢字の四隅の字画の形に基づくコードによって検索できるものである。漢字の四隅を形により 0 から 9 まで番号を付与し, 更に同一番号となる漢字を区別するために「附角」番号を付与し，5 桁の数値で漢字を配列する。

例「**法**」漢字のコード化，四角号碼：34131

3　　4
　法　1
1　　3

日本で出版された『大漢和辞典』（諸橋（1960））が四角号碼を採用した。『大漢和辞典』で採用されているコードの規則は諸橋（1984，p.一〇三八）に記載されている。

３．３　カタカナ字形分類索引

カタカナ字形分類索引（加納 1998，p.1007）は常用漢字 1945 字のすべてを「アイウエオ」五十音のカタカナ字形に分類し，五十音順に並べた索引である。共通する部位の下にその漢字が提示してある。カタカナの形の漢字の中の位置は問われていない。例えば，ア部の漢字　「了子孔好」など，イ部の漢字　「仙代付休」などである。

３．４　書き出しパターン索引

書き出しパターン索引（加納 1998, p.1020）は一番最初に書く書記素を扱っている。下記の 6 種類の書き出しパターンを決め，その書き出しパターンが同じ漢字を画数順に並べた索引である。

1 - 一，2 - 丨，3 - ノ，4 - 丶，5 - フ，6 - レ

例えば,「一」の部に入っている漢字は「一」から書き出す「一二丁三天」などである。

３．５　筆順索引

筆順索引は若尾＆服部（1989）によって作成された『くずし解読字典』に入っている。その字典のくずし字は，筆の運び（筆順）とその方向によって配列されている。筆の動きが8つの方向の矢印で示されて，それぞれの方向には決まっている0から7までの番号（コー

ド）がつけてある（若尾&服部1989, p.466）。
↑ － 0, ↗ － 1, → － 2, ↘ － 3, ↓ － 4, ↙ － 5, ← － 6, ↖ － 7
そして各々の漢字の最初の 4 つの筆（起筆，第二筆，第三筆，第四筆）の方向を表す 4 桁の数字のコードによって漢字が配列されている。例えば，「仙」という漢字の「イ」の筆順は「↙ － 5,↗ － 1,↓ － 4」（↗ － 1 が入っている理由は漢字をくずして書くとき線をひき，いったん筆をもとまで戻すからである）（若尾&服部（1989），p.469）。

３．６　Key Words and Primitive Meanings Index

Key Words and Primitive Meanings Index （Heisig （2001）, p.506）の中では各々の漢字とその構成要素には唯一の字義（漢字の意味）がつけてある。それからその字義を表す英単語はアルファベット順で並べてある。その字義によって漢字をひくことができる。そのためには予め英語の字義を覚える必要がある。

３．７　System of Kanji Indexing by Patterns（SKIP）

Halpern（1988/1990, 1999）は System of Kanji Indexing by Patterns（SKIP）の開発にあたって辞典に入っている各漢字に数字のコードをつけた。そのためにはまず漢字字体の 4 つの基本的なパターンを確定し，それぞれのパターンに 1 から 4 までの数字を当てた。
1 － ◧ 左右に分ける漢字　　2 － ⬒ 上下に分ける漢字
3 － ◻ 構えを含めた漢字　　4 － ■ その他の漢字
この数字は漢字コードの最初の数字となる。タイプ 1～3 の漢字は 2 つの部分に分ける。それぞれの部分の画数がコードの 2 番目と 3 番目の数字となる。例えば，漢字「相」は左右に分けるタイプ 1 で，左の「木」の画数は 4，右の「目」の画数は 5 である。SKIP コードは 1-4-5 となる。タイプ 4 のコードの 2 番目の数字は漢字の画数，3 番目の数字は 1～4 の数字で，漢字の形のコードである（そのコードの説明を省略する）（表 1）。

表 1　System of Kanji Indexing by Patterns (SKIP) のコードの例

タイプ	漢字	画数	SKIP コード
1	相	9	1-4-5
2	父	4	2-2-2
3	間	12	3-8-4
4	女	3	4-3-4

コードのデータを昇順に並べ替え，SKIP という索引が構築された。

３．８　Fast Finder

Matthews（2004）は Halpern（1988/1990, 1999）と同様に漢字のパターンを確定しているが，漢字のコード化をしない。漢字は構成要素の位置によって 8 種類のパターンに分けてある。左右の左，左右の右，上下の上，上下の下，構えの 3 種類と分けない漢字である。Fast Finder の各々のグループが掲載された最初のページから漢字の構成要素の形とそれに所属する漢字が提示されている。同じ構成要素に所属している漢字の字体が複雑な場合は，グループの漢字が字体によって細かく分類され，配列されている。

３．９　Index by Radicals

Hadamitzky&Spahn（1981）は部首による漢字の検索を簡単にするために部首の数を減少させることにした。一般に使用されて，国際規格であるユニコード（Unicode 6.1.0）（http://www.unicode.org/versions/Unicode6.1.0/）によってスタンダード化された 214 種の部首の中から 79 種だけの Radical を抽出し，それに基づき Index by Radicals という漢字索引を構築した。そして形が複雑だと思った部首を使用せずに，その部首に所属している漢字を

79 種の Radical の中のある Radical に所属させた。表 2 ではこのような部首変更の例を提示する。

表 2　Hadamitzky＆Spahn（1981）による部首変更の例

一般に使用される 214 部首の中の		Hadamitzky＆Spahn（1981）による 79 Radical の中の	
部首番号	部首	Radical 番号	Radical
176	面	3s	囗
177	革	3k	艹
178	韋	3d	口

部首の数を減少させた結果，1 種の Radical に所属している漢字の数が増加した。

３．１０　意味記号索引

意味記号索引は，加納（1998）に出てくる 495 個の意味記号を，画数順に並べた索引である。加納（1998, p.6）は「（前略）漢字の意味を表している部分を,「意味記号」として示しました。（中略）「意味記号」は「部首」と同じ形のものもありますが，たとえば（みず）と（さんずい）というように，呼び方がちがいます。（中略）意味記号が，その漢字の中にない場合があります。その場合，その漢字のもとになった漢字を示しました。」と説明している。例えば,「斎」の部首は「斉」, 意味記号は「示（祭だん）」である（加納 1998, p.952）。

３．１１　主な意符の索引

主な意符の索引（白石 1971/1978）は部首索引に似ているが，部首の代わりに 243 字の主な意符を採用した索引である。主な意符の中に部首も，漢字も，部首ではない漢字の構成要素もある。主な意符は画数によってグループ化されている。

漢字索引のタイプの調査結果, 索引は様々な種類があることが明らかになった。上記の, 漢字の構成要素と書記素に基づく，索引の中に漢字の数字のコードを利用した索引もある。

３．１２　字形索引

字形索引は 512 字の漢字を含めている教科書（坂野, 池田, 品川, 田嶋, ＆渡嘉敷（2009））についていて，部首索引に似ている。しかし 215 個の「字形」には部首も，部首ではないパターンも入っている。「字形」（漢字の部分）を画数順に並べ，各々の漢字部分を含む漢字とその教材の中の漢字番号を示している。

４. 既存の漢字索引の効率の比較評価

４．１　既存の漢字索引の共通点

上記の様々な既存の漢字索引の共通点は，漢字の構成要素か諸性質の中から一つの要素か性質だけを取り出し，それをもとに作成してあることにあると思う。用いられている要素か性質は漢字の画数，部首，書き出しパターンなどである。

４．２　漢字索引の選択係数の定義

漢字索引の効率の比較評価にあたって，漢字索引に対して，コンピュータデータにおける処理の効率を表す「選択性」(Selectivity) という概念を用いることにした（ヴォロビヨワ (2009), p.72)。http://www.akadia.com/services/ora_index_selectivity.html というサイトには Selectivity of an index の定義がある"The ratio of the number of distinct values in the indexed column / columns to the number of records in the table represents the selectivity of an index."

「選択性」（Selectivity）という概念を用いるため漢字索引の効率指数「選択係数（Coefficient of Selectivity – CS）」という概念を次のように定義した。CS=V/N×100％

ここで漢字の字体に基づく索引の場合，N は索引に入っている漢字の数で，V は索引の中で漢字が所属するグループの数である。グループというのは同じ部首に所属する漢字群，画数が同じ漢字群などである。例えば総画索引の場合，V は索引に含まれる漢字の同様の書記素数の漢字のグループの数，部首索引の場合，V は部首の種類の数である。しかし，音声に基づく音訓索引などの場合 N は索引に入っている漢字の読み方の延べ数（total number）で，V は読み方の異なり数（number of distinct）（読み方の種類の数）である。

これまで開発された索引を比較するため，ここでは新常用漢字ではなく，常用漢字を扱うことにする。分析した結果，1945 字種の常用漢字群に含まれる同様の書記素数の漢字のグループの数は 23 で，採用されている部首の種類の数は 201 である。

例　総画索引　V=23，N=1945，CS=23/1945×100％=1.2％
　　部首索引　V=201，N=1945，CS＝201/1945×100％=10.3％

部首索引の効率は総画索引のおよそ 10 倍であることが明らかになった。

４．３　既存の漢字索引の効率の比較評価

本論文では選択性（Selectivity）という概念に基づき選択係数の計算をもとに 15 種類の既存の漢字索引の効率を比較評価する（表 3）。

表 3　漢字索引の選択係数

索引のタイプ	選択係数（%）
漢字字体に基づく索引	
総画索引（Henshall（1988）） カタカナ字形分類索引（加納（1998）） Index by Radicals（Hadamitzky＆Spahn（1981）） 書き出しパターン索引（加納（1998）） 五段排列漢字表（ロゼンベルグ（1916）） 四角号碼（諸橋（1984）） 部首索引（Henshall（1988）） 筆順索引（若尾&服部（1989）） 主な意符の索引（白石（1978）） Fast Finder (Matthews（2004）) SKIP (Halpern（1988）) 意味記号索引（加納（1998））	1.2 2.6 4.1 6.1 7.1 10.2 10.3 10.7 12.4 14.1 15.4 25.4
漢字の読み方に基づく索引	
主な音符の索引（白石（1978）） 音訓索引（Henshall（1988））	27.6 40.6
漢字の意味に基づく索引	
Key Words Index and Primitive Meanings（Heisig（1977/2001））	100.0

その中に漢字字体に基づく索引と漢字の読み方に基づく索引と漢字の意味に基づく索引を対象とした。分析した結果，漢字字体に基づく漢字索引の選択係数は 1.2～25.4％と低いことが明らかになった。その理由は一般の漢字索引は主に一つのみの漢字の性質か要素に基づいているからであると考えられる。例えば，部首索引は部首だけ，総画索引は画数だけ，書き出しパターン索引は最初の書記素の形だけに基づいている。そして個々の部首や総画数，書き出しの書記素にはたくさんの漢字が所属しているということである。漢字の読み方に基づく索引の選択係数は 27.6～40.6％であり，漢字字体に基づく索引より高いということが明らかになった。しかし，読み方に基づく索引を使用するためには，予め，読み

方を覚える必要がある。漢字の意味に基づく索引 Key Words Index and Primitive Meanings（Heisig 1977/2001）の選択係数は 100%に達しているが，それを使用するためには，予め，それに入っているすべての漢字の意味を覚える必要がある。

５. 新しいタイプの索引の効率の比較評価

上記の評価と比較分析を行い，全体的な漢字字体を表すコードに基づく，選択性が高い漢字索引が極めて必要だと考え，新しいタイプの索引を開発することにした。漢字辞典の検索をより効率的にするためには，非漢字系学習者の考え方に相応しい，字体を適切に表す漢字の文字・数字のコードに基づくアルファベット・コード索引，シンボル・コード索引, セマンティック・コード索引及び部首と画のコード索引を開発した（ヴォロビヨワ (2009, 2011)）。表 4 には常用漢字 1945 字の漢字群をもとに計算した新しいタイプの索引の選択係数が提示してある。選択係数はそれぞれの索引に入っている漢字コードの異なり数を延べ数で割り 100%をかけて，できた数値である。

新しいタイプの索引の効率の比較分析の結果，アルファベット・コード索引，シンボル・コード索引及び部首と画のコード索引の選択係数は 100%に近く, 漢字字体に基づく既存の索引よりはるかに高いことが明らかになった。つまり同じ漢字コードが少なく，ユニークな漢字コードが多いという意味である。例えば，部首と画のコード索引の中の 11 字の漢字だけが他の漢字と同じコードがある。

しかし，セマンティック・コード索引の選択係数は 64.1%であり，その理由は，セマンティック・コードにはすべての構成要素ではなく，先に書く２つの要素のコードのみが入っているために，同じコードの漢字が比較的多いことによる。漢字の意味を利用した，より効率的な索引を構築するにはセマンティック・コード索引を土台に，新たな索引であるバイナリセマンティック・コード索引の開発が必要であると考えた。そのため複雑な漢字でも 2 つのみの意味的単位に分解し，筆順に従い，その 2 つの要素のセマンティック・コードを書いて，バイナリセマンティック・コードとした。

例えば,「露」という漢字を構成する最小意味的単位は「雨」「足」「夂」「口」である。以前，セマンティック・コードを開発した際は，長いコードにしないように最初に書く２つの構成要素のコード（意味）のみを用いた。「露」の場合は,「雨」と「足」の２つであり、「露」のセマンティック・コードは「Rain/ Foot」になる。

それに対し、バイナリセマンティック・コードでは，同じ漢字「露」を「雨」（Rain）と「路」（Road）のように，２つのみの要素に分解した。そして要素の意味を表す単語「Rain/ Road」をバイナリセマンティック・コードとした。従って，上記の要素の中の「路」（Road）のように，バイナリセマンティック・コードは必ずしも最小の意味的単位ではない。

分析の結果，新常用漢字の中に同じバイナリセマンティック・コードは 2 つしかなく，バイナリセマンティック・コード索引の選択係数は一番高く 99.9%であることが明らかになった（表 4)。しかし，セマンティック・コード索引と比較すると，バイナリセマンティック・コード索引の短所は，最小意味的単位だけではなく，その組み合わせの一部を覚える必要があり，学習者にとって負荷になる。

表 4　新しいタイプの索引の選択係数

新しいタイプの索引	選択係数（%）
セマンティック・コード索引	64. 1
アルファベット・コード索引	98. 4
シンボル・コード索引	99. 4
部首と画のコード索引	99. 4
バイナリセマンティック・コード索引	99. 9

６. まとめと今後の課題

本論文では非漢字系日本語学習者の漢字辞典使用の難しさについて述べ，15 種類の既存の漢字索引と筆者が開発した 5 種類の索引の特徴について紹介した。そして漢字索引の効率の比較評価を目指し，漢字索引に対して，コンピュータデータにおける処理の効率を表す「選択性」（Selectivity）という概念を用いることにし，漢字索引の効率指数として「選択係数（Coefficient of Selectivity - CS）」という概念を定義した。さらに既存の索引と筆者が開発した新しいタイプの索引の選択係数を計算し，索引の効率の比較評価をした。分析の結果，アルファベット・コード索引，シンボル・コード索引，バイナリセマンティック・コード索引及び部首と画のコード索引の選択係数は 100%に近くて漢字字体に基づく既存の索引よりはるかに高いことが明らかになった。今後，漢字索引の効率を表すのに選択係数だけではなく，ヴォロビヨワ（(2011)，p.24）で定義された索引の使用のための労力などもはかり，各々の漢字索引の総合効率指数を定義し分析する計画がある。

謝辞

本研究は公益財団法人　博報児童教育振興会の日本語海外研究者招聘事業に参加する機会を得て行った。そして国立国語研究所理論・構造研究系の横山詔一教授からご指導を受けた成果の一部でもある。さらに，津田塾大学の非常勤講師関麻由美氏には原稿の修正をしていただいた。また，キルギス民族大学の准教授ヴォロビヨフ・ヴィクトル氏にご協力をいただいた。ここに記して感謝申し上げる。

参考文献

ヴォロビヨワ・ガリーナ（2007）『漢字物語Ⅰ』ビシケク

ヴォロビヨフ・ヴィクトル＆ヴォロビヨワ・ガリーナ（2007）『漢字物語Ⅱ』ビシケク

ヴォロビヨワ・ガリーナ（2009）「選択性が高い漢字索引の開発」『日本語教育方法研究会誌』Vol. 16, No. 1，pp.72-73.

ヴォロビヨワ・ガリーナ（2011）「構造分析とコード化に基づく漢字字体情報処理システムの開発」『日本語教育』No.149, pp.16-30.

Vorobeva Galina（2013）「漢字検索法の効率性の分析　―ロシアのグラフィックシステムなど―」『JSL 漢字学習研究会誌』第 5 号 JSL 漢字学習研究会，pp.86-94.

加納喜光（1998）『常用漢字ミラクルマスター辞典』小学館

白石光邦（1971/1978）『要素形的漢字学習指導法』桜楓社

諸橋轍次（1984）『大漢和辞典』大修館書店

ロゼンベルグ・オ.（1916）『五段排列漢字典』東京　興文社

若尾俊平＆服部大超（1989）『くずし解読字典』栢書房

Galina N. Vorobeva, Victor M. Vorobev（2012）“An Analysis of Efficiency of Existing Kanji Indexes and Development of a Coding-based Index”OPEN JOURNAL SYSTEMS: Acta Linguistica Asiatica Vol. 2, No. 3, Slovenia,University of Ljubljana, http://revije.ff.uni-lj.si/ala/article/view/180/318, pp.27-59.

Hadamitzky, W.＆Spahn, M.（1981） *Kanji &Kana Revised Edition A Handbook of the Japanese Writing System* Tuttle Language Library.

Halpern, J. （1988/1990） *New Japanese-English Character Dictionary*, Kenkyusha.

Halpern, J. （1999） *Kanji Learner's Dictionary*. Kodansha.

Heisig, J. （1977） *Remembering the Kanji. Vol. 1*. Tokyo: Japan Publications Trading Co. Ltd.

Henshall, K. （1988） *A Guide to remembering Japanese characters*, Tuttle.

Matthews, L. （2004） *Kanji Fast Finder* 漢字早引き辞典 Tuttle Publishing.

Васильев В. П.（ 1867） *Китайско-русский словарь. "Графическая система китайских иероглифов."*

王雲五（1925）『號碼檢字法』商務印書館

王雲五（1926）『四角號碼檢字法』商務印書館

王雲五（1934）『四角號碼檢字法·附檢字表』商務印書館