漢輔：外国人のための漢字検索インターフェース

田中久美子　　Julian Godon
東京大学情報理工学系研究科
kumiko@i.u-tokyo.ac.jp, jgodon@gmail.com

概 要

既存の漢字検索方法は漢字文化に根ざしており、日本人・中国人以外の人が漢字を検索する際には困難を伴うことがある。本発表では、既存の方法とは異なる漢字の検索方法に基づく漢字検索インターフェース「漢輔」を紹介する。本システムでは、現在日本語、中国語それぞれ７０００字弱の漢字を検索することができる。(URL: http://www.ish.ci.i.u-tokyo.ac.jp/kansuke/)

1　はじめに

漢字がわからない時には、日本人・中国人は読み、部首、画数により漢字辞典をひく。しかし、これらはいずれも漢字文化に深く根ざした方法であるため、外国人が漢字を調べたいときには困難を伴うことがある。日本語や中国語に興味を持つ外国人は昨今多いが、このような事態が東アジアからの情報発信を妨げる一要因となっている可能性がある。

外国人のための漢字の入力方法として過去に提案された手法は、Jim Breen のホームページなどにまとめられている [2]。たとえば、Jack Halpern の SKIP コードは実際の辞書で用いられているものであるし [3]、また、Slaven Bilac の発案した FOKS システム [1] は、誤りを含む読みからでも漢字や語を検索できる画期的なシステムである。

しかし、これらは日本や中国の言語や書道文化にそれなりに造詣がある外国人のためにデザインされており、まったくの初心者には用いることはできない。

そこで、本稿では漢字を縦、横、その他の辺の数で検索する新しいインターフェース「漢輔」を提案する。漢輔は、漢字文化に関する特別な知識を必要とせず、漢字の形を見ただけで漢字を検索することができる。たとえば、「東」の字は、横 4 本、縦 3 本、その他（ななめの線）2 本から成る。そこで、漢輔コードとして

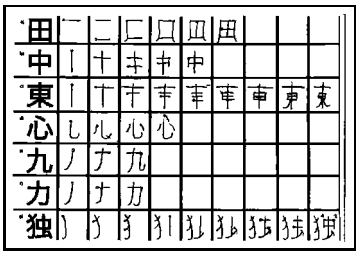

図 1: 漢字初心者による手習い

ユーザが 4-3-2 を漢輔に入力すると、漢輔はこのコードに相当する漢字を列挙する。本稿では、このようなインターフェースに関して、設計とユーザ評価を示す [5]。

2　漢輔コード

図 1 はで、一度も漢字を書いたことがないドイツ人による漢字の手習いである。同じ人であっても、縦や横の書き方の順序などは統一性がないことが伺える。このようなアンケート調査を複数の外国人に対して行った結果、書き方は千差万別ではあるものの、いくつかの点で共通点が見られた。

- 角がある一画は複数辺と考える場合が多い。
- 直線か曲線かの判断は見解が一致する。
- 縦か横か斜めかの判断は一致する。
- なめらかな曲線は一辺と考えている場合が多い。

以上の観察から、伝統的な書道文化に根付いた「画」の概念とは別に、角のない一筆として「辺」を定義する。たとえば、「田」の字は、五画であるが、九本の直線の数から成ることから九辺の字と考える。その上で、辺を横、縦、その他に分類し、それぞれの辺の数で漢輔コードを定義する。その他は縦横以外のすべての辺を含む。たとえば、「東」であれば、辺の数が九辺であり、うち横が 4 本、縦が 3 本、斜めが 2 本なので、漢輔コードは 4-3-2 である。同様に「文」であれ

表 1: 漢輔コードの良さ：横縦その他の辺の数の分布、ならびに既存の手法との候補数の比較

JIS 漢字に含まれる横、縦、その他の辺の数	
横	43295
縦	32307
その他	45260

一入力に対する候補数			
	平均	最大	標準偏差
画数による検索	205.8	565	196.5
SKIP コード	11.4	160	18.4
漢輔コード	7.4	68	10.4

ば、1-1-2 となる。

「辺」の定義にのっとったとしても、漢輔コードのデザインにはさまざまなものが考えられる。たとえば、その他をさらに「斜めの直線」と「曲線」に分けるなどといった案も考えられる。しかし、漢字は横と縦の直線の辺が圧倒的に多い。表 1 の第一ブロックは日本語の全 JIS 第一、二水準の漢字について横、縦、その他の辺数を示しているが、その数はほぼ拮抗している。また、辺の分類を細かくしすぎると、初心者には却って分類が難しくなってしまうことなどを踏まえ、試行錯誤の末にこの漢輔コードを採用した。

このほか、漢輔コードを用いると、ユーザの一入力に対して挙がる漢字の候補数は既存の手法に比べて少なくて済むことが判明している。表 1 の第二ブロックは、画数、SKIP コード[1] 、漢輔コードを一つ入力した際に挙がる漢字の数の平均、最大数、標準偏差である。漢輔コードはいずれも最小となっており、初心者のためのインターフェースとしては望ましいことが統計的にわかる。

以上の漢輔コードを用いて、漢字検索インターフェース「漢輔」を作ることはできる。しかし、これだけでは初心者向きのシステムとしては、まだ問題がある。第一は、曖昧性の問題である。たとえば、「文」の最上部の点相当の部分が果たして縦かその他なのかについては、議論を呼ぶところであろう。このような曖昧性は、ユーザが見るフォントだけでなく、ユーザの漢字の認識の差異によっても起こりうる。第二に、「鬱」などの画数の多い漢字のすべての辺を調べて漢輔コードを作り出すのは、初心者にはかなり大変である。

[1]SKIP コードは、漢字を 2 ブロックで構成されるとみなし、1. 縦型、2. 横型、3. くりぬき型に分類する。「土」のようにどうしても 1 ブロックになってしまうものはを第 4 番目の類型とみなす。その上で、各ブロックの画数を与える。たとえば、「相」であれば、縦型で、左ブロックは 4 画、右ブロックは 5 画なので、SKIP コードは 1-4-5 となる。

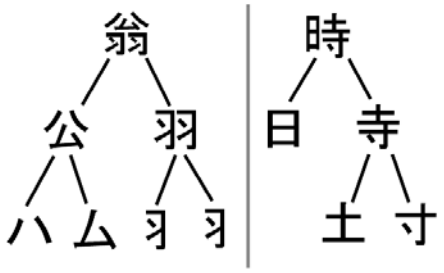

図 2: 漢字構造木の二つの例

3 漢輔インターフェース

3.1 漢字構造木

漢字は視覚的なまとまり部分に再帰的に分解することができる。たとえば、「時」であれば、「日」や「寺」、さらに「寺」は「土」「寸」へと分解することができる。この性質を利用し、漢輔内では、全漢字は図 2 のような構造木により表現されている。

漢輔では、この漢字構造木のすべてのノードに漢輔コードを割り振っており、すべてのノードの漢字部分を漢輔コードを用いて検索することができる。たとえば、たとえば「鬱」であれば、「木」と「缶」の漢輔コードをそれぞれ入力して検索した上で、漢字全体である「鬱」を検索することができる。

漢字構造木を用いると、ユーザは非曖昧な漢字部分の組み合わせだけから漢字を検索することもでき、曖昧性の問題を部分的に解決することができる。とはいえ、「文」のように辺数の少ない漢字の場合には曖昧な漢字部分を避けて通ることができないこともある。その場合には、一つの漢字構造木のノードに複数の漢輔コードを割り振る。たとえば、「文」の上の点のように

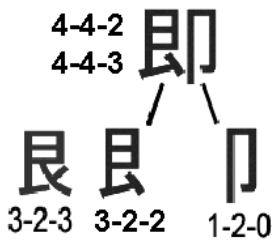

図 3: 曖昧な漢字部分と漢輔コードの伝播

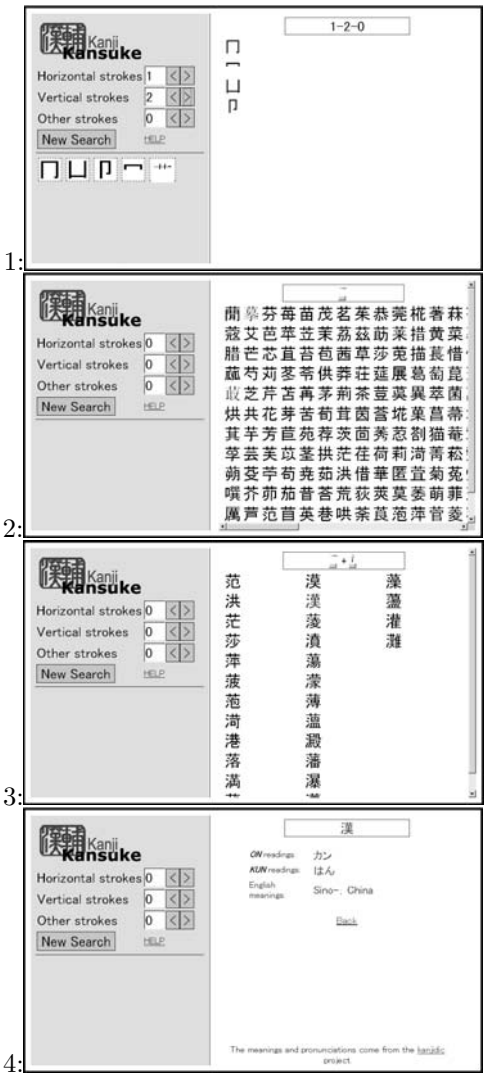

図 4: 漢輔インターフェース

縦ともその他とも判断されうる場合には 1-1-2 のほかに、1-0-3 もコードとして割り振っておく。フォントによって形が異なる「即」の左部の場合も、図 3 に示すように、二種類のコードを割りふる。

データベースが木構造になっているので、複数のコードが割り振られたノードより上位のノードに複数のコードを伝播させて割り当てることができる。その結果、上位のノードも複数の漢輔コードを用いて漢字を検索することができる。

以上のデータベースを基盤として、インターフェース漢輔を実現した。現在漢輔は、すべての JIS X 0208-1990 中の日本語の漢字 (6355 字) ならびに、すべての GV2312 中の中国語の漢字 (6849 字) に対して検索を行うことができる。

漢輔を用いて漢字「漢」を検索する一例を図 4 に示した。まず 1 では草冠の漢輔コード 1-2-0 を入力し、対応する漢字が右に、漢字部分が左に表示されている。漢字も漢字部分もすべてクリックすることができる。2 は、1 の左部分から草冠を選んでクリックした直後を示しており、右側には草冠を含むすべての漢字が表示されている。漢字の数が多すぎるため、さらにさんずい相当の漢輔コード 0-0-3 を入力し、漢字を絞り込むと 3 のようになる。「漢」は右の 2 行 2 列目に表示されているので、これをクリックすると、4 のように edict[2] の内容が表示される。中国語の場合には、web 上のオンライン辞書が表示される。

表 2: 漢字セット、被験者、検索インターフェースの組み合わせ

	A	B	C	D
被験者 1,5	画数	SKIP	漢輔	部首
被験者 2,6	SKIP	漢輔	部首	画数
被験者 3,7	漢輔	部首	画数	SKIP
被験者 4,8	部首	画数	SKIP	漢輔
	E	F	G	H
被験者 9,13	画数	SKIP	漢輔	部首
被験者 10,14	SKIP	漢輔	部首	画数
被験者 11,15	漢輔	部首	画数	SKIP
被験者 12,16	部首	画数	SKIP	漢輔

表 3: 評価実験に用いた 8 つの漢字セット

A	B	C	D	E	F	G	H
白	王	凹	幺	予	叮	月	斗
咲	承	息	垂	狽	蚣	屋	衫
款	射	訪	嵯	惱	覚	跟	等
酵	駿	衝	賊	銹	髷	衛	緬

4 評価

漢輔インターフェースを画数による検索、SKIP コード、部首による検索の三つの既存のシステムと比較した。部首による検索では部首と思われる漢字部分を複数選択することのできる multi-radical 方式を採用した [2]。これは、中級レベルの外国人によく使われている検索方式である。

被験者は 16 人づつ、初心者、学習者（以上外国人)、日本人あるいは中国人の 3 種類のユーザ、計 48 人から成る。検索インターフェースの順番、漢字のセット、

表 4: ユーザ評価実験

初心者			
	秒 (平均) /標準偏差)	失敗率 (%)	好き (%)
画数	107.1 / 107.6	47%	0%
SKIP	62.7 / 47.8	31%	19%
部首	76.7 / 65.0	27%	12%
漢輔	59.6 / 49.0	14%	69%
学習者			
	秒 (平均) /標準偏差)	失敗率 (%)	好き (%)
画数	65.2 / 62.4	26%	6%
SKIP	41.9 / 35.6	16%	32%
部首	57.8 / 79.9	14%	28%
漢輔	53.5 / 48.6	23%	34%
日本人・中国人			
	秒 (平均) /標準偏差)	失敗率 (%)	好き (%)
画数	54.3 / 44.9	6%	0%
SKIP	31.8 / 30.9	11%	13%
部首	55.5 / 71.3	8%	56%
漢輔	55.9 / 55.8	19%	31%

ユーザの熟達レベルはいずれも偏りがないようにラテン・スクエア [4] を利用して、実験を計画した (表 2 参照)。漢字セットは表 3 に挙げた 8 セットを用意した。各行ごとに、概ね同じ画数、辺数の漢字となるように漢字を選択した。各ユーザは、まずインターフェースの説明を読み、その後、試しに一つを検索してトレーニングをした後、続いて本番の漢字 4 つを検索する。これを 4 セット分、異なるインターフェースで行い、全部で 20 の漢字を検索する。検索が不可能と判断される場合には、「skip」ボタンを押すことができ、失敗数が計測される。実験終了後に、4 システムの中から最も好きなシステムを選んでもらった。

結果を表 4 に示す。初心者においては、漢輔は好成績であった。最少の失敗率で最速の検索ができ、最多の人が漢輔システムを好ましいと答えている。漢輔のユーザとしては初心者も想定しているので、本システムの有効性はまずまずであることがこの結果から伺える。一方で、学習者や日本人・中国人の場合には、既存のシステムが優位であった。経験者は既存の検索インターフェースに慣れている一方で、漢輔だけが初めて試すシステムであったことに原因がある。特に、日本人・中国人の画数を利用した検索では、画面いっぱいの漢字の中から目的の漢字を瞬時に探しだす能力には目をみはるものがあった。

検索が失敗した原因は、どのインターフェースにおいても曖昧性のひとことにつきる。どの検索方式であっても、曖昧性を完全になくすことはできない。たとえば、画数の場合には「凸」は曖昧な漢字となるし、部首検索の場合にはどの部分が部首なのか、また類似した形が複数あるなど曖昧となる。また SKIP は最初の 4 分類が曖昧となる。漢輔も、複数の漢輔コードを利用して、曖昧性を緩和しているものの、曖昧性を完全に根絶させることは難しい。このことからも、唯一のシステムに限るよりも、ユーザは複数のシステムを併用することがよい。

たとえば「凸」の書き順や画数を数えられない日本人は多いので、漢輔は日本人や中国人を対象としても有効な場合があるであろう。また日本語と中国語では書き順や画数は完全に一致するものではないので、日本人が中国語を検索する場合やその逆などにおいても役に立つことがある側面はあると考えられる。

今後は漢輔インターフェースの電子辞書搭載を目指し、活動したい。

参考文献

[1] S. Bliac, T. Baldwin, and H. Tanaka. Bringing the dictionary to the user: the FOKS system. In *In Proceedings of the 19th International Conference on Computational Linguistics*, pages 84–91, Taipei, 2002. Association for Computational Linguistics.

[2] J. Breen. Jim Breen's WWWDic Site, 2007. `http://www.csse.monash.edu.au/ jwb /cgi-bin/wwwjdic.cgi`, accessed 2007.

[3] J. Halpern. *Kanji Learner's Dictionary*. Kodansha, Tokyo, 1999.

[4] D.C. Montgomery. *Design and Analysis of Experiments, Student Solutions Manual*. John Wiley and Sons, New York, 2003.

[5] K. Tanaka-Ishii and J. Godon. Kansuke: A Kanji look-up system based on a few stroke prototypes. In *International Conference on the Computer Processing of Oriental Languages*, pages 310–311, 2006.