平成１１年度第１回総合目録データベース実務研修レポート
１９９９．１０．７

日本語ヨミの自動付与および翻字の自動入力

広島大学附属図書館情報管理課
洋書目録情報係　野村和子

１．はじめに

図書館サービスの最終的な目的は、利用者の求める資料情報をいかに早く正確に提供できるかということだと思う。

利用者が求める資料情報にたどり着くのに、もっともよく利用されるのがＷｅｂｃａｔやＯＰＡＣではないだろうか。

もちろん、データはＮＡＣＳＩＳ－ＣＡＴに登録されるのであるから、目録担当者からは、ＷｅｂｃａｔやＯＰＡＣは副産物となるのだが、利用者はその副産物からアクセスする方法しかないということをよくわかっておかなければならない。

そのための検索手段として、検索キーが重要になってくる。

ここでは検索キーの使われ方を通して検索キーの作り方を検討してみたい。

２．ＯＰＡＣはヨミで探すか？

「日本語の資料を探すとき、漢字を入力しますか？それともカナを入れて探しますか？」と学生にきいてみた。閲覧室で勉強している学生のうち、２３人中１７人が漢字で探すと答えた。ちなみに残り６人中カナのヨミで探していた学生は一人もいなかった。その内訳は、一人は日本語の図書は探したことがないという学生だったし、もう一人は留学生で、引用文献等に書いてある通りに入力して探すという返事だった。そして残念なことに残り４人はＯＰＡＣを使ったことがなかったのである。

１９９９年８月２０日午後２時、夏休みの最中、本調査する前の下調査のつもりで行った聞き取り調査であったが、たったこれだけのサンプルで、既にヨミで探すものはひとりもいないという結果が出てしまった。

そもそも、ワープロがこれだけ普及した現在、ＯＰＡＣやＷｅｂｃａｔを使う利用者は、わざわざヨミを考えて検索なんてしないものである。ローマ字入力かカナ入力で「ジョウホウ」と入れ、変換キーを押せば、まず「情報」が出てくるのがあたりまえの時代に、ヨミで検索する利用者がそんなにいるわけもないような気がする。

少し古いが、広大の図書館で１９９５年に利用者を対象にアンケートをとったことがある。広大では今年の４月からシステムを更新したのに伴い、ＯＰＡＣも新しくなっているので、このアンケートは今のシステムを正確に反映しているわけではない。しかし、だいたいの方向は間違っていないと思われる。

このアンケートの中で、ＯＰＡＣに対する要望の中で多かったのが、「うまく検索できない」「あるはずの本が見つからない」といった類のものである。個別に話をきいたわけではないので推測にすぎないが、検索キーに問題はないだろうか。

利用者はどうやって図書を探すかというと、まずタイトルをそのまま入力するのが一般的である。いわゆるフルタイトルキーによる完全一致検索をして見つかればそれでよし。見つからなかったとき、次にどうするか。ヨミで探すとは思えないのである。

先ほどの留学生のように、参考文献や引用文献にでてきた資料を探している場合、ヨミでなくても正しい書名や著者名がわかっているのだから、そのとおりに入力すれば問題なく必要な資料は見つかるだろう。しかし、その文献が旧漢字を使っている古い図書で、なおかつ参考文献の欄には新漢字で記載されていたら、どうだろう。あるはずのものがみつからないこともあり得るのである。

逆に、ヨミで探して見つからないこともある。たとえば、「吾輩は猫である」をカナ入力してみよう。目録の入力基準どおりに「ワガハイ　ワ　ネコ　デアル」と入力する利用者は何人いるだろうか。一般の利用者は「ワガハイ　ハ　ネコ　デアル」とするのがふつうではないだろうか。（本当はもっと簡単に単語だけで検索するのだろうけれど）

また、「トウホクチホウ」と「トウホク　チホウ」が検索条件としては違うのだということがわかっている人ばかりが検索するとは限らないのである。目録の入力基準をわかって図書を探す人なんてほとんどいないのが現状である。

にもかかわらず、各大学のＯＰＡＣをみても、ＨＥＬＰの中にカナ表記の説明はほとんど見あたらない。北大のように、カナは単語で探すというような例をあげているところがあるが、それで充分なのであろう。

固有名詞の読み方がわかるので便利という話があるが、これは全く別の問題である。

３．カナ検索のメリットとデメリット

検索もれを防ぐためには、カナ検索が向いているといわれているが、それは利用者のみならず目録担当者にもあてはまる。

つまりデータ作成の上でもっとも大切なことは、重複書誌を作らないことであろう。総合目録データベースの品質向上のためにも、検索を繰り返してもれのないようにすることである。

検索もれで考えられるものの中に、旧漢字と新漢字の違いがある。旧漢字で検索すれば新漢字の書誌が、新漢字の場合は、旧漢字の書誌が検索もれとなる可能性があるが、カナで検索すれば、両方を探すことができる。

たとえば、「国際法」をＮＣで検索した場合５１７件ヒットしたのに対し、「國際法」では１６７件、「コクサイホウ」では５５７件ヒットした。漢字ではどちらで探しても検索もれは防ぐことができないのである。ちなみに、「国際法」と「國際法」を合わせて「コクサイホウ」と同じ件数にならないのは、「国際法」という図書の別タイトルに「國際法」が入っている等のように重なっているものもあるからである。

このように、カナ検索は検索もれを防ぐにはもってこいなのであるが、そのかわりにヒット件数が増えて、目的の書誌にたどり着くのに時間がかかるという欠点がある。たとえば、「情報検索の基礎」という図書を探す場合、「情報検索＊」では５９件ヒットするが、「ジョウホウ」と「ケンサク」では１５９件もヒットするのである。

ちなみに、「ジョウホウケンサク」では１件もヒットしない。これは「ジョウホウケンサク」という検索キーを作っていないからである。

たとえば、「金融論／新庄博著」という図書を探す場合、「キンユウ　ロン」で検索すると４７件ヒットし、その中にこの図書は含まれているのだが、「キンユウロン」で探すと３３６件もヒットするのにこの図書は含まれていない。これはこの図書のデータを入力するとき、「キンユウ　ロン」というヨミを作っても「キンユウロン」というヨミを作らなかったからである。つまり全くヒットしなければ、検索キーに問題があるのではないかとすぐに思い当たるところだが、これが少しでもヒットすれば検索もれに気づかないまま、重複書誌を作ってしまうおそれがあるという例である。

また、たとえば「新世代のコンピュータ」という図書は、本来ならばヨミは「シンセダイ」と入れることになっているのだが、「シン　セダイ」と入力してしまった。或いは、ＮＣにデータがなく、ＴＲＣにヒットしてヨミの確認を忘れてそのまま入力してしまった。ＴＲＣは独自の分かち書き基準でヨミを入れているので、この場合その基準でいうと「シン」と「セダイ」は分離することになっているのである。

「シンセダイ」と入れたら１０１件もヒットするが、「新世代のコンピュータ」はヒットしない。ここで「シン　セダイ」もあり得ると判断して再検索すれば見つけられるが、気づかないで重複書誌を作ってしまう可能性はとても大きいのである。間違ったヨミを入力した場合、間違いを想定した検索をしない限り、その情報はヨミでは探し出せなくなるという例である。

４．入力できない漢字の処理

もうひとつ大きな問題がある。それは機器の関係で入力できない漢字の処理である。現在、各作成館で入力できない漢字があったときは、文字コードを黒菱で囲む方法をとっているが、これでは一般利用者が検索することはできない。

そこでヨミがあれば、カナ検索で目的の書誌にたどりつける。カナ入力の最大のメリットといえよう。ただし、カナ検索をすることを利用者にもっと宣伝する必要がある。

５．ヨミ不要論

最近の日本語検索システムはとても発達していて、ヨミなど必要としないソフトがたくさん出ている。総合目録データベースも、ヨミ入力を必須としないで検索できるようにすればよいのではないか。

システムが、カナであろうと、或いは新漢字でも旧漢字でも、何でも探してくるようになれば、人間の方は全くヨミを気にせずデータ入力に集中できる。重複書誌の検索も短時間に効率よくできるだろう。そうなれば仕事の省力化もすすむというものである。

今まで、図書館の仕事は蓄積するばかりで、捨てることは論外であった。しかし、これからは捨てることも必要ではないかということを考えてみてはどうだろうか。

とはいうものの、ＮＡＣＳＩＳ－ＣＡＴのような大量のデータを持つところで運用するには、検索速度が遅すぎて無理との意見が大勢を占めるのが現状である。あえてそれを否定するつもりはない。

しかし、利用者のことを考えるならば、ヨミ入力にかける時間を省いて少しでも早く、正確にデータを提供することはできないだろうか。

総合目録データベースに使われているヨミは、各作成館の目録担当者が、個々に入力している。そして、日本目録規則の「片かな表記法」および目録編成規則の「ワカチガキ」に準拠しているとはいえ、全く個人の判断にまかせられているそのヨミが日本語のキーの切り出しに使われている。

これが、日本語のキーがふらつく原因となり、このあいまいなままの日本語検索キーをそのまま使うことで、利用者にわかりにくい検索キーが作られることになった。

たとえば、「要求仕様の探検学」という図書を探す場合、Ｗｅｂｃａｔで「要求仕様」と入力したら見つからないが、「要求」「仕様」と入力したら見つかるのである。これは、ヨミを入力する段階で「ヨウキュウ」「シヨウ」と分けて作ったためである。もちろん利用者に、単語を区切らないで探すだけでなく、区切って探してみることを勧めればよいのだが、利用者の立場になれば、どちらでも探せる方がよいのである。或いは、あやふやなままより規則性のあるほうがまだしも使いやすいと思う。

そういった意味では、日本語のキーの切り出しはあいまいなヨミを元にするのでなく、一般の日本語検索システムに使われている日本語構文解析技術を使って抽出するのがよいのではないだろうか。

６．ヨミの自動付与

ここでは規則性を持ったヨミを入力するために、ヨミの自動付与を考えてみた。

ヨミを辞書に登録しておいて、機械的にヨミが入るようになれば、検索キーのあいまいさも、間違って入力する心配も解決するだろう。

ここでひとつヒントになることは、Ｅｘｃｅｌの自動ふりがな機能である。

自動ふりがな機能とは、入力した漢字に自動的にふりがなを振る機能のことで、Ｅｘｃｅｌではデータを入力した際の「読み」をセルが記憶しており、この「読み」をふりがなとして表示している。もちろんまちがったふりがながつけられた場合は、ふりがなを個別に修正することができる。

市販のソフトにこれだけの機能があるのだから、応用すれば総合目録データベースの漢字にもヨミをあてはめることは可能である。ワープロのように、ヨミの辞書さえ作っておけば自動的に漢字かな混じり文をカナに直せるはずである。

では、そのヨミの辞書の品質維持は、誰が責任を持つのかという問題が出そうである。これは、現在典拠ファイルの維持に責任を持つことと同様、目録担当者であると考える。

データ１件作るたびに機械的にヨミを入れ、そのヨミが間違っていないか確認し、もし間違っていたらその場で修正するのである。読めない漢字が出てきたときには、機械的に「？？」を当てて、確認する段階で人の手で入力する。これら辞書に載っていない漢字も登録する機能をつくっておいて、気づいたときに追加すれば辞書はどんどん充実し、より使いやすくなるだろう。

こうしておけば、分かち書きの問題や間違ったヨミの入力の問題は解決する。ヨミのゆれが気になるなら、ヨミを自動付与したあとで中身を確認しながら、ＶＴ等に採用されなかったものを入力していけばよいだろう。

７．洋書の翻字への応用

洋書の場合はどうだろうか。アルファベットを使わない言語の図書には、翻字という作業があるが、これが結構やっかいである。何よりも、慣れない文字を 1 文字ずつ翻字していく過程で、うっかり 1 文字読み飛ばしてしまったり、間違った文字を入力したりすることがあっても気づかないで、結局翻字した検索キーでは探せないことが起きる可能性がある。こういう場合、自動翻字が有効である.

また、和書と同様、機器の関係で入力できない文字があったとき、やはり黒菱で囲む方法をとるが、これも一般利用者は検索できないことになる。

しかし、これも翻字したアルファベットがあれば検索することができる。

翻字としてデータ入力の対象になるものには、ギリシャ語やロシア語のキリル文字等がある。これらは日本語よりも少ない文字数であるのと、各対応する文字が１対１（日本語の音訓のような複雑さがない）ということから、自動翻字に向いていると思われる。

辞書さえきちんと作っておけば、言語コードと連動させて、各々の言語の翻字用辞書を見に行って翻字させることができる。こちらは日本語と違って、いちいち目で確かめることも必要ないかもしれない。

８．おわりに

現在のシステムでは、ヨミは個人の努力によって整合され、維持されている。しかし、本来検索キーというものはＡＫＥＹのようにシステムの中で一定のルールをもって作られるべきではないだろうか。でなければ、利用者が検索に余計な時間と労力を必要とする。そのあげく、「あるはずの本がみつからない」とか、「欲しい本がない」と思われてしまう。

一定のルールで検索キーを作っておけば、そしてそのルールを利用者に伝えておけば、利用者もそれだけ使いやすいはずである。

そして目録担当者のヨミ入力の負担が減ることで、仕事の省力化が進み、サービスの向上に役立てることができるのではないだろうか。

参考文献

日本目録規則　１９８７年版改訂版　日本図書館協会　１９９４
目録システムコーディングマニュアル　学術情報センター　１９９８
目録情報の基準　第２版　学術情報センター　１９９１
目録情報の基準　第３版　学術情報センター　１９９７
図書館資料の目録と分類　増訂版　日本図書館研究会　１９９７
ＴＲＣ　ＭＡＲＣ　分かち書き基準　ＴＲＣ株式会社図書館流通センター　１９９４
全文検索　技術と応用　学術情報センター編　１９９８