2001年11月1日（木）16時10分～55分
データベース2001東京，プレゼンテーションセミナー
第7回「ことばフォーラム」（国立国語研究所・紀伊國屋書店共催，日本書籍出版協会後援）

日本語情報の海外提供

横山詔一・笹原宏之・熊谷康雄・エリク＝ロング（国立国語研究所）

1. 日本語による情報提供の必要性

1-1.日本語の国際化

国際交流基金の調査結果によると，海外の教育機関で学ぶ日本語学習者数は，最近10年の間に約73万人（1988：昭和63年）から約210万人（1998：平成10年）へと急増しており，海外における日本語学習需要は今後も伸びていくものと思われます。1998年には海外115か国・地域で日本語教育が実施されています。

このような状況を見ると，日本語学習者の手元に，インターネットを介して「日本語で」さまざまな情報を円滑に届けられるシステムが求められる時代になったと言えましょう。日本語による情報提供の仕組みは，日本語学習者のためばかりではなく，日本の文化や社会そしてビジネスについて深く理解してくれる人々を海外に増やすことにもつながっていくであろうと期待されています。

1-2.日本語Webページの問題

ところが，現状では，海外のインターネット閲覧ソフト（以下，ブラウザと呼ぶ）から日本語で書かれたWebページにアクセスすると，日本語の文字が意味不明の記号などに置き換えられ，まったく「読めない」状態になっているのが普通です。つまり，いつも「文字化け」した状態なのです。

海外のブラウザで日本語のWebページを読むには，あらかじめ日本語フォントをインターネット端末にセットしておく必要があるのですが，それは以下のような問題を引き起こします。

(1) 海外のインターネット端末に日本語フォントを組み込もうとすると，それなりの知識と手間が必要です。パソコンやブラウザに関する技術的な知識に自信がない人は，二の足を踏んで，結果的にあきらめてしまうことが多いのではないかと思われます。

(2) 海外の大学などでは学内LANに接続されている端末のパソコンに使用者が勝手に日本語フォントを組み込むことを許していないケースも多いようです。そのため，日本語のWebページを読めないことがよくあると聞きます。

1-3.日本語表示の基本的なアイデア

では，世界中のブラウザで日本語を確実に表示できるようにするには，どうすればよいのでしょうか。現在いくつかの方法が開発されていますが，国立国語研究所が中心になって開発しているシステムは，日本語の一つ一つの文字を「画像（イメージ）」で画面に素早く表示できるような工夫をしています。以下にその実例を2つ紹介しましょう。

2. 実例その1：出版情報データベースの提供

2-1.日本語の資源を世界に

本のタイトル（書名）などを英訳すると，本来のニュアンスがどうしても伝わりにくくなります。海外へ我が国の出版情報を提供する場合，少なくとも書名や著者名は日本語で表示されるのが望ましいと考えられます。そこで，国立国語研究所は，我が国の出版情報データベースを海外からも「日本語で高速に検索できる」システムの研究に取り組んでいます。

このシステムは，海外の日本語学習者にも役立つ可能性があります。日本は，外国語で書かれた文献の自国語への翻訳点数において世界有数の地位を占めており，その結果，諸外国の文物に関する日本語による豊富な蓄積が生じています。これらは，世界の文化資産の一つとして活用し得るものです。現に，アジアからの留学生がヨーロッパの文献を，ア

メリカからの留学生が中国の文献を日本語で学んでいるような例も見られます（国語審議会答申，2000）。

2-2.データベースの中身

海外に提供するデータベースは（社）日本書籍出版協会が構築している出版情報データベース「Books」です。そこには現在入手可能な書籍約60万件の書誌情報が収められています。

2-3.検索システムの使い方

ブラウザを起動し，入力エリアに検索したい書籍の書名や著者名の一部をローマ字表記（半角英文字）で入力すると検索結果が表示されます。

2-4.検索と文字配信の連携

このシステムの特徴は，検索システムと文字配信システムが連携して動くところにあります。システム全体は図1に示すような仕組みになっています。

■検索段階：まず，ブラウザは「書誌情報検索サーバー」にアクセスし，検索結果がブラウザに返送されます。日本語の部分には，その文字に対応するGIFファイルの所在（リンク先）が記述されています。

■文字配信段階：ブラウザはリンク先をたどって「文字配信サーバー」にあるGIFファイルにアクセスし，日本語が利用者端末に表示されます。

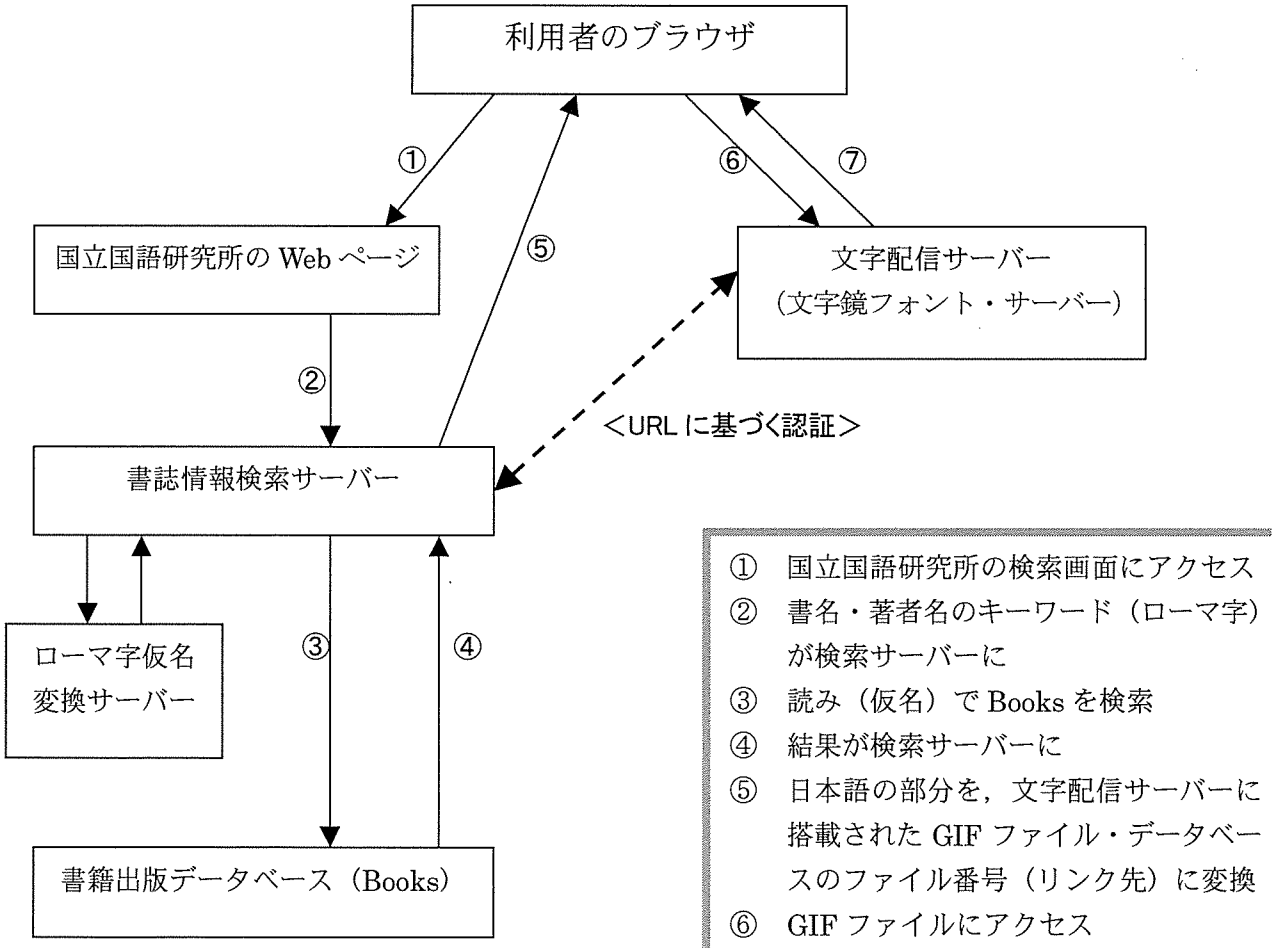


図1　検索と文字配信の連携

2-5.今後の予定

2003年3月までにローマ字日本語変換の機能を付け加え，検索したい書名や著者名を漢字仮名交じりの日本語で入力できるようにする予定です。ちなみに，ホームページの日本語表示にGIFを用いたシステムとして米国で開発された「Shodouka（注1）」がよく知られていますが，日本語入力の仕組みは持っていません。

（注1）http://web.lfw.org/shodouka/

3．実例その2：用字用語データベースの提供

3-1.データベースの中身

国立国語研究所は，1956年（昭和31年）当時に出版された代表的な一般雑誌九十種類から約100万字を抽出して，どのような文字や語が，どのくらい使われているのかを調査しました（以下，雑誌九十種と呼ぶ）。このシステムは，雑誌九十種に掲出された用字用語データベースを海外からも検索できるようにしたものです。

3-2.検索システムの使い方

国立国語研究所のホームページ（注2）にアクセスすると図2に示すような画面が表示されます。入力エリアに検索したい語のローマ字表記を半角英文字で入力し「Find」のボタンをクリックすると図3に示す検索結果が表示されます（これは「いし」という読みの語を検索した例）。Written form（表記）は日本語表示，Origin（語種：和語，漢語，混種語，外来語など）やPart of Speech（品詞）などの情報は英語表示となっています。

（注2）http://www.kokken.go.jp/public/zassi90syu_e.htm

Complete Vocabulary from a Survey of Ninety Contemporary Japanese Magazines

Enter word to search for:

ishi　[Find]　[Help]

図2　入力画面の例

いし			
Written form	Form count	Total count	Origin / Part of Speech
石	28	28	Native noun
意志 意思	28 12	40	Sino noun
遺志	1	1	Sino noun
医師	33	33	Sino noun
縊死	1	1	Sino noun

図3　検索結果の例

3-3.データベースを利用した研究の例

現在，国立国語研究所は，1994年（平成6年）に出版された月刊雑誌七十種類の用字用語調査を進めています（以下，雑誌七十種と呼ぶ）。これは調査対象を月刊誌に限定している一方で，サンプルとして抽出した文字数を約210万字と雑誌九十種の2倍に増加させているなどの点で，雑誌九十種との単純な比較はできません。しかし、ここでは一つの試みとして，両者を比べてみることにします。なお，雑誌七十種（1994）については，今回はサンプルを拡大した約290万字分のデータを用いましたので，雑誌九十種の3倍近い文字数になっています。（以下の説明は，昨年のデータベース2000東京で述べたものと一部重複するところがあります。）

・増加した漢字の例

	1956		1994
麺	0	→	17

雑誌九十種用字用語データベースで「めん」という読みの語を検索してみると，頻度がゼロで，サンプルには出現しなかったことが分かります。1956年当時の記事を調べてみると，「そうめん」などの語が使われていますが，すべて仮名表記でした。そもそも「めん類」について，雑誌で話題になることが少なかったようです（1994年のうち拡張新字体［簡易慣用字体］は2，康熙字典体［印刷標準字体］は15）。

・減少した漢字の例

	1956		1994
糎	124	→	0

1994年でもサンプル箇所以外で出現はしていますが，仮名表記（センチメートル）や記号（cm）で書かれる傾向が強まり，使用頻度は明らかに下がっています。

・変化した表記の例

＜コーヒー＞

	1956		1994
珈琲	3	→	1

現在，街中で目にする表記としては，コーヒーは50%程度が「珈琲」ですが，雑誌においてはほとんど仮名表記しか現れません。雑誌は，各分野にわたる表記が，比較的自由に使われていますが，人間が街中で目にしている表記の割合と一致するものではないという可能性があります。

＜たまご＞（「たまごやき」「ゆでたまご」などの複合語を含む）

	たまご		タマゴ		玉子		卵
1956	6	:	2	:	26	:	34
1994	1	:	1	:	3	:	99

「たまご」は，意味やニュアンスに幅を持つ語です。「卵」は，カエルの「たまご」のような生物的な意味や，学者の「たまご」のような派生的な用法の用例を多数含んでいます。一方，「玉子」は，語源的な表記であるとともに，食材として料理に用いるための婉曲的な表記と言えます。「玉子」は，雑誌においては仮名表記とともに減っているようです。ただし，この結果は雑誌によるものであって，チラシなどに使われる表記の一般的な傾向とは異なる可能性があります。

4. 国内に蓄積された言語資源の海外提供に向けて

4-1.まとめ

今回紹介した日本語情報検索システムは，日本語で書かれたさまざまなデータベースを海外へ提供する際に役立つのではないかと考えられます。

・図書館蔵書データベース
・特許情報データベース
・官報情報データベース

4-2.その他

国立国語研究所は，「日本語研究の文献目録」や日本語教育の教材作りに利用できる「素材データベース」（日本語教育支援総合ネットワークシステム）など，いろいろなデータベースをインターネットで公開しています（注3）。ただし，海外のブラウザからこれらのデータベースにアクセスすると，日本語で書かれた部分が読めない状態になっています。この問題を解決するため，ここで紹介した文字配信システムなどの利用を検討していく予定です。

（注3）http://www.kokken.go.jp

参考文献（アルファベット順）

国立国語研究所（1962～1964）『現代雑誌九十種の用語用字』（国立国語研究所報告21,22,25），秀英出版

国立国語研究所（1997）『現代雑誌九十種の用語用字：全語彙・表記【FD版】』（国立国語研究所言語処理データ集7），三省堂

日本書籍出版協会（2001）『日本書籍総目録2001』

横山詔一・エリク＝ロング・江川清・笹原宏之・古家時雄（2000）「海外WWWブラウザ対応の日本語データ検索システム―『現代雑誌九十種の用語用字：全語彙・表記』を例に―」電子情報通信学会技術研究報告TL2000－16，17-24，電子情報通信学会

附記

出版情報データベース「Books」の利用においては，（社）日本書籍出版協会のご協力をいただきました。記して厚く感謝申し上げます。

プレゼンテーションセミナー
国立国語研究所「ことばフォーラム」

日本語情報の海外提供

共催：国立国語研究所・紀伊國屋書店
後援：(社)日本書籍出版協会

日本語による情報提供の必要性

1.日本語の国際化
　海外の日本語学習者は210万人

2.日本語Webページの問題
　海外のブラウザ→文字化け

3.日本語表示の基本的なアイデア
　一つ一つの文字を画像で表示させる

開会の辞

司会
国立国語研究所・吉岡泰夫

（ハンドアウトとアンケート用紙の確認）

実例1：出版情報データベースの提供

1.日本語の資源を世界に
　本の書名や著者名などは日本語で

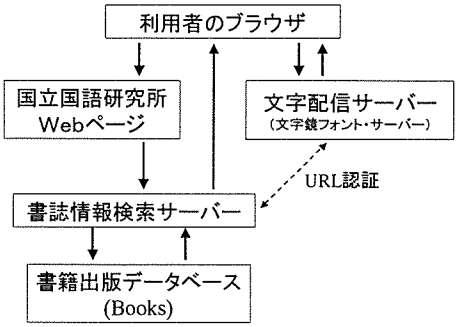
2.データベースの中身
　『日本書籍総目録』データベース「Books」(書協)
　現在入手可能な書籍60万件の書誌情報

3.検索システムの使い方
　＜ダイアルアップでイーストに接続＞

ご挨拶

国立国語研究所長・甲斐睦朗

4.検索と文字配信の連携



まとめ

今回のシステムを応用すれば
日本語で書かれた以下のようなデータベースを
世界中で検索できるようになるだろう

例えば
図書館蔵書検索システム
特許情報検索システム

5.今後の予定
ローマ字を仮名や漢字に変換
そのためのシステムを来年度に開発

その他

国立国語研究所データベースの一つを紹介

「日本語教育支援総合ネットワーク」のデモ

実例2:用字用語データベースの提供

1.データベースの中身
1956年当時の一般雑誌90種類を調査
どのような文字や語がどのくらい
出現したのか

2.検索システムの紹介
＜『雑誌九十種』のデモ＞

閉会の辞

国立国語研究所・吉岡泰夫

（アンケート記入のお願い）