書誌情報データベースにおける非ローマンアルファベット系言語の原綴り・翻字相互変換システム

望月 源[†] 大和 加寿子[††]
前嶋 淳子[††] 林 俊成[†]

日本語，韓国語，中国語の各語を除く非ローマンアルファベット系言語で書かれた図書・資料を扱う大学図書館では，書誌情報を ALA-LC 方式の翻字規則によりローマンアルファベットに翻字したうえで登録する必要がある．この翻字作業には，知識と経験が必要であり人材が限られる，原綴りの入力に比べて時間がかかるという問題がある．また，検索時も翻字が必要であり，検索結果も翻字で表示されるなど，図書館の一般利用者にとっても利便性に問題がある．本研究ではこうした問題に対処する機能を提供する，原綴り・翻字相互変換の自動化について述べる．ALA-LC 翻字規則は，人手での作業を前提に，異なる言語ごとに定義されたものであり，広範囲な言語に対する自動化はこれまであまり検討されていない．我々は，翻字を必要とする主要な言語について，自動化の観点から難易度と実現の可能性を検討した．その結果から本論文では，最初の実装として，ロシア語とヒンディー語を選択しシステムを構築した．実際の書誌情報データを用いた評価実験を行い，実用的な精度で変換が行えることを確認した．

Automatic Transliteration and Back-transliteration System for Languages That Use Non-Roman Alphabets in Bibliographic Citation Database

Hajime Mochizuki,[†] Kazuko Yamato,[††] Junko Maejima[††] and Chun Chen Lin[†]

University libraries will be required to use the ALA-LC romanization tables as transliteration schemes when they register the bibliographic citation for books written in non-roman alphabets except for Japanese, Korean and Chinese. There are some difficulties for transliteration; registrars who have basic knowledge about the target language and experience of transliteration are limited, much time for registration is required compared with the input of original scripts and to retrieve a book catalog with transliteration and to back-transliterate the results of retrieval are inconvenient for all users. In this paper, we describe an automatic tranliteration and back-transliteration system for languages that use non-roman alphabets in bibliographic citation database. We will discuss the difficulties and possibilities of automatic transliteration and back-transliteration for each non-roman alphabet language preceding to construct the system, and we will adopt Russian and Hindi for our first implementation. The effectiveness of the system is also evaluated by experiments that use actual bibliographic citation data.

1. はじめに

近年，計算機での多言語処理環境が向上し，扱える言語の種類も増えている．しかし，計算機利用に関して長い歴史のある図書館の書誌情報データベース（以降「書誌 DB」と記す）の登録では，非ローマンアルファベット系言語をそのまま用いずに，ローマンアルファベット[☆]で翻字することが多い．ここでいう翻字とは「原綴り[☆☆]のローマンアルファベットによる転写」のことであり，元の言語の発音を考慮したうえで，文

† 東京外国語大学外国語学部
Faculty of Foreign Studies, Tokyo University of Foreign Studies
†† 東京外国語大学附属図書館
Tokyo University of Foreign Studies Library

☆ 日本では英語に代表されるローマ文字（ラテン文字ともいう）のみを「アルファベット」と呼ぶこともあるが，アルファベットは正式には言語によらず「文字」のことをさす．本論文では本来の呼び方に従い，ローマ文字のアルファベットを「ローマンアルファベット」と呼び，それ以外のアルファベットを「非ローマンアルファベット」と呼んで区別する．

☆☆ 「原綴り」とは，ある文字や単語，文などが，その言語の元々の文字を用いて綴られている状態をさす．

字レベルだけでなく文脈などの情報も加味した発音をローマンアルファベットで書き表すことをいう．日本の大学図書館では，共有の書誌 DB として国立情報学研究所の目録作成システム，NACSIS-CAT [1] を利用しているが，ここでも翻字が行われている．NACSIS-CAT では，非ローマンアルファベット系言語のうち，日本語と韓国語は原綴りのみを扱うが，それ以外の言語（ロシア語，ヒンディー語，アラビア語，ペルシア語，タイ語など多数）は翻字での登録を行う．また，翻字規則は，中国語（ピンインを用いる）を除いたすべての言語で ALA-LC 方式（以下「ALA-LC 翻字規則」と記す）を用いなければならない．ALA-LC 翻字規則[2] は，元の言語の発音を再現することに重点を置いた翻字であり，54 種類の，言語ごとの原綴りとローマンアルファベットの対応表からなる基本規則と，文脈に応じて発音が変化する場合などに対応した例外規則によって定義されている．そのため，ALA-LC 翻字規則を理解するには元の言語の基本的知識が必要になる．また，中国語の場合は日常的に用いられているピンインを翻字に使用するので問題はないが，他の言語の ALA-LC 翻字は日常的に用いられているわけではないので，その言語の話者であっても改めて学習する必要がある．

こうしたことから原綴りを直接用いる言語に比べ，非ローマンアルファベット系言語の書誌 DB 登録作業は効率が良くない．また，蔵書検索も翻字で行わなければならず，一般利用者にとっての利便性も悪くなる．さらに，検索結果も翻字で表示されるため，翻字から元の言語で何と書いてあるかを理解しなければならない．仮にすべての言語が原綴りのみで扱われるようになれば，こうした翻字の問題はなくなる．しかし，実際には原綴り用の文字コードやフォントの整備が不十分な言語の存在や，これまで蓄積されてきた既存データの有効活用などの観点から，今後も翻字はなくならず，原綴りと併用されるものと思われる．実際に，アラビア語では最近になって書誌 DB に原綴りを追加する方針が定まっているが，翻字も引き続き使用される．そのため，新たにデータを登録する場合は，原綴りと翻字の両方を使用し，すでに登録されているデータには原綴りでの情報を追加するという作業が行われることになる．将来的には他の非ローマンアルファベット系言語に対しても同様な作業が行われると考えられる．また，一部の図書館では NACSIS-CAT のような共有の書誌 DB に先駆けて，自館専用のローカルな書誌 DB に原綴りによる情報を追加する作業が行われている．こうしたことから，非ローマンアルファベット系言語を扱うことの多い外国語学部を持つ大学の附属図書館などでは翻字の処理の問題が今後もあり続けると考えられる．また，既存の翻字による書誌 DB に原綴りの情報を加えるための効率の良い手法の開発も重要になると考えられる．

本論文ではこうした問題に対処する以下の 2 つの機能を提供する，原綴り・翻字相互の自動変換システムについて述べる．

- 原綴りから翻字を自動作成する機能
- 翻字から原綴りを自動作成する機能

前者の利用により，原綴りを元にして翻字での書誌登録が行える「書誌 DB 登録支援システム」と，蔵書検索を原綴りのままで行える「検索支援システム」が実現できる．また，後者の利用により，「検索支援システム」の翻字での検索結果を原綴りに変換し，利用者の利便性の向上が図れる．また，既存の書誌 DB 内の翻字から原綴りを作り，書誌 DB に追加する機能を「書誌 DB 登録支援システム」上に実現できる．

ALA-LC 翻字規則は，文字や文法，発音などの異なる言語ごとに個別に定義されている．また，人手での作業を前提にしており，計算機による自動処理は考慮されていない．そのため，原綴り・翻字の相互変換システムを実現するには，個別の言語ごとに ALA-LC 翻字規則の記述を調べ，どの程度の自動化が可能かも含め，実現方法を検討する必要がある．一方，日常的に非ローマンアルファベット系言語を扱う図書館はそれほど多くないこともあり，広範囲な言語に対する翻字の自動化はこれまであまり検討されていない．本論文では，翻字を必要とする非ローマンアルファベット系言語について，計算機での自動化の観点から，原綴り・翻字相互変換システム実現の難易度，可能性を検討する．次に，検討結果から最初の実装に最適な言語を選択し，我々の最初のシステムを構築する．また，実際の書誌情報データを用いた実験を行い，構築したシステムの有効性を評価する．

以下，2 章で，翻字規則と言語の違いによる翻字の難易度について述べ，実装対象とする言語を決定する．3 章で，本論文で扱う原綴り・翻字相互変換システムについて説明し，4 章で，原綴り・翻字相互変換システムの精度を測るための実験を行う．

2. 翻字規則と変換の難易度

本章では，ALA-LC 翻字規則について簡単に説明し，言語ごとの難易度について調査し，実装対象とする言語を選定する．

2.1 ALA-LC 翻字規則

ALA-LC 翻字規則（ALA-LC Romanization Tables）は，全米図書館協会と米国議会図書館が定めた翻字規則の集合である．ALA-LC 翻字規則による翻字への変換は，基本的には言語ごとの文字レベルの対応表に従って原綴りを音標符号付きのローマンアルファベットに置き換えることで行う．たとえば，ヒンディー語で「ヒンディー語」を意味する「हिंदी का」（ヒンディーカーと読む）は，音標符号付きのローマンアルファベットにより「hindīkā」と翻字する．ただし，ALA-LC 翻字規則は元の言語の発音を再現することを重視した翻字であるため，単純な文字レベルの対応表だけでなく，文脈などによる発音の変化に合わせた例外的な規則がどの言語の場合にも存在し，より複雑なものとなっている．たとえば，非ローマンアルファベット系言語である日本語の「は」をローマンアルファベットで翻字する場合を例にとると☆，基本的な対応表では「ha」となる．しかし,「私は」のような文脈では「は」の発音は変化するので,「wa」と翻字する必要がある．こうした複雑さの度合いは，その言語に必要な例外の種類や数の違いによって異なり，翻字化の難易度にも関係する．一般的には同じ綴り字を何通りにも読み分ける言語や，母音を補って読む必要がある言語の翻字は，より複雑で難しい．

言語によって翻字規則の内容が異なるため，原綴り・翻字自動変換システムの主要な部分は個別の言語ごとに実装する必要がある．次節で，実装の候補となる言語における原綴り・翻字変換の難易度について述べ，最初のシステムとして実装する言語を選定する．

2.2 変換の難易度と実装言語の決定

本節では，原綴り・翻字変換システムで実装対象とする言語を選定する．本研究では実装の候補とする言語をロシア語，モンゴル語，ヒンディー語，アラビア語，ペルシア語，ウルドゥー語，タイ語，ラオス語，カンボジア語，ビルマ語の 10 言語とする．これらの言語は，日本国内で非ローマンアルファベット系言語の蔵書割合がきわめて多い東京外国語大学の附属図書館で日常的に扱っている言語である．実際に実装する言語は，これらの候補を対象に翻字規則の難易度や翻字自動化の需要などを考慮して，実装する言語とその実装順を決定することとする．具体的には次の方針で選定する．

- 各候補言語について，原綴り・翻字間での変換の難易度を調査する．
- 各候補言語について，現時点での計算機での原綴りの表示や入力方法の一般性を調査する．
- 上記 2 点をふまえ，現時点で実装が可能であるものについて難易度の順位付けをする．

2.2.1 難易度の調査

実装の候補となる 10 言語について，各言語の特徴と ALA-LC 翻字規則の記述を参照し，次の点に注目して原綴り・翻字変換の難易度を調査する．各言語についての調査項目とその結果を**表 1** に示す．

- ALA-LC 翻字規則に記された原綴りと翻字対応表での原綴りの数と翻字の数（表 1 の「綴：翻」欄）．
  元の言語の文字数と翻字数が同数に近い方が，原綴りと翻字の 1 対 1 対応の割合が増え，曖昧性が少なくなるため，自動化処理の難易度が下がると考えられる．また，翻字の数が少ないものは，異なる綴り字が同じ翻字になる割合が高いため，翻字から原綴りへの変換時に曖昧性が大きくなる．原綴りの文字数が少ないものは，原綴りから翻字への変換時に曖昧性が大きくなると考えられる．
- 例外的な扱いの必要な翻字規則の多さや複雑さ（表 1 の「例外規則」欄）．
  一般に翻字規則中の例外の数が多いほど自動化処理は難しくなる．また，例外の内容が複雑なものであれば，数が少なくても難しい場合もあるので，例外の複雑さも難易度に影響を与える．
- ALA-LC 翻字規則内での記述ページ数（表 1 の「頁」欄）．
  一般に ALA-LC 翻字規則内での記述ページ数の多い言語は例外も多く自動化処理が難しい．
- 元の言語（原綴り）における分かち書きの有無（表 1 の「分かち書き」欄）．
  ALA-LC 翻字規則では分かち書きをしない言語に対しても，意味の切れ目で分かち書きを行ったうえで変換を行う必要がある．そのため，分かち書きのない言語では，それだけで自動化処理が難しくなる．
- Unicode の整備状況と計算機での原綴り入力の可否およびその他特記事項（表 1 の「その他」欄）．
  文字コードの整備状況を Unicode 4.01 [3] を基準にして，調べる．言語によっては，文字コードが規定されていても実際には必要な文字が不足している場合や，フォントが十分に整備されていない場

☆ 日本の図書館では日本語の翻字は行わないため本研究では日本語の翻字は対象としないが，ALA-LC 翻字規則としては「ヘボン式」に準拠した基本的な対応表と多くの例外によって定義されている．ここでは文脈による発音の変化の説明のため，日本語を例とした．

表 1 翻字の難易度に関する調査結果
Table 1 Difficulties of transliteration.

言語	文字	綴：翻	頁	例外規則	分かち書き	その他
ロシア	キリル	76:76	2	単純，少	あり	
モンゴル	モンゴル	126:78	2	やや複雑	あり	Unicode 文字不足，Windows で入力困難
ヒンディー	デーヴァナーガリー	88:94	4	やや複雑	あり	
アラビア	アラビア	132:46	10	複雑，多	あり	母音表記なし，大・小文字区別なし
ペルシア	アラビア	130:41	7	複雑，多	あり	母音表記なし，大・小文字区別なし
ウルドゥー	アラビア	181:65	8	複雑，多	あり	母音表記なし，大・小文字区別なし
タイ	タイ	102:86	16	複雑，多	なし	
ラオス	ラオ	79:83	4	やや複雑	なし	Unicode 文字不足，Windows で入力困難
カンボジア	クメール	118:78	3	やや複雑	なし	Unicode 文字不足，Windows で入力困難
ビルマ	ビルマ	91:85	3	やや複雑	あり	Unicode 文字不足，Windows で入力困難

合などもある．また，現在最も普及している OS である WindowsXP において，特別なソフトウェアを必要とせずにその言語の文字入力が可能かどうかも実質的な利用のしやすさに関係するため，難易度の判断材料となる．

表 1 より，モンゴル語，ラオス語，カンボジア語，ビルマ語は現時点で文字コードの整備が完全とはいえない．これらの言語では計算機上で原綴りを正確に表すことができない状況にあるといえ，翻字の自動化以前の問題から，現時点では不完全な実装にならざるをえないという問題がある．また，分かち書きがされないタイ語，ラオス語，カンボジア語では，原綴りに対する正確な発音を得るために，まず意味の切れ目で分かち書きをする必要がある．ALA-LC 翻字規則では発音に応じて翻字が異なるため，この作業は必須である．分かち書きの自動化には自然言語処理技術が必要になるが，これらの言語に対して正確な分かち書きを行うのは現時点では困難であるため，実装は難しい．残る 5 言語は，文字としては 3 つのグループ，キリル文字（ロシア語），デーヴァナーガリー文字（ヒンディー語），アラビア文字（アラビア語，ペルシャ語，ウルドゥー語）に分けることができる．この中で，アラビア文字には，母音が独立した文字としては存在せず，原綴りには母音が記述されない☆．ただし，発音としては母音は存在するので，読み手が文脈から母音を補って読む決まりになっている．そのため，アラビア語，ペルシャ語，ウルドゥー語の翻字を自動化するには，発音上母音が必要な位置に，どの母音が入るかを子音の配列や文脈から推定する機構を組み込む必要がある．また，アラビア文字を用いるどの言語においても，翻字規則の中に例外に関する複雑な記述が多く，相対的な難易度が高い．総合的に検討すると，キリル文字，デーヴァナーガリー文字，アラビア文字を使う言語の順に翻字の相対的な難易度は低いといえる．この結果から，我々は最初の実装言語として，ロシア語（キリル文字）とヒンディー語（デーヴァナーガリー文字）を選び，システムを構築することにし，アラビア文字を用いる言語は今後の課題とする．

☆ 初学者向けの文章や厳密な発音を明示する必要がある場合に対応するため，アラビア文字にも母音を表す補助記号は存在するが，一般的な文章ではこの母音記号は記述されない．

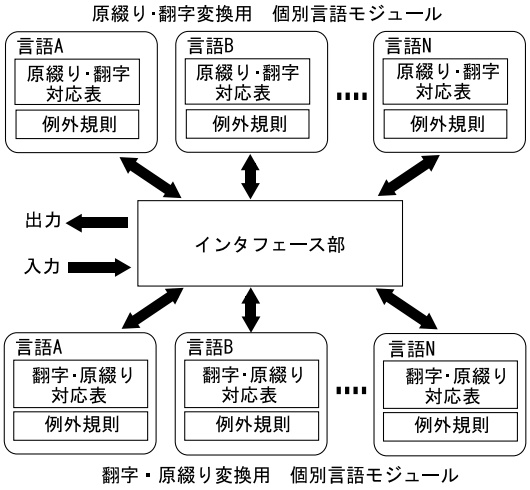


図 1 自動変換システムの構成
Fig. 1 The system architecture.

3. 原綴り・翻字自動変換システム

3.1 自動変換システムの概要

我々が開発する非ローマンアルファベット系言語の原綴り・翻字相互の自動変換システムのイメージを**図 1** に示す．

本システムは，システムの共通インタフェースを与えるインタフェース部と，各実装言語に依存して実装されるモジュールから構成される．インタフェース部では，入力文字列をその言語に対応した変換モジュールに渡し，その結果を受け取りユーザに返す．一方各言語モジュールは「原綴りから翻字への変換用モジュール」と「翻字から原綴りへの変換用モジュール」に大

きく分かれ，それぞれ各言語用に機能する．「原綴りから翻字への変換用モジュール」は，各言語ごとに翻字化の基本規則（原綴り・翻字対応表）と例外規則からなる．「翻字から原綴りへの変換用モジュール」は，各言語ごとに原綴り化の基本規則（翻字・原綴り対応表）と例外規則からなる．このシステムでは，各言語用のモジュールを変換方法別に作成し，追加することによって対応する言語を増やせるようになっている．

本システムは以下の 2 つの自動変換機能を提供する．

- 原綴りから翻字への自動変換．
  ある言語の原綴りを，それに対応する ALA-LC 翻字に変換する機能．
- 翻字から原綴りへの自動変換．
  ある言語の ALA-LC 翻字を，それに対応する原綴りに変換する機能．

3.2 原綴りから翻字への自動変換

原綴りから翻字への変換は，各言語ごとに ALA-LC 翻字規則に則って行う．翻字規則の中で，原綴り側の各文字を翻字に対応させるだけで処理できる基本規則の部分については，「原綴り・翻字対応表」としてまとめる．一方，前後の文字の種類や音韻などを考慮して場合分けしながら翻字を決定する必要のある例外部分については，個別に「例外規則」を作成する．

この対応表と例外規則による原綴りから翻字への自動変換の基本的な処理手続きは次のようになる．まず，原綴りの先頭から 1 文字を取り出し，その文字を含む例外規則が存在するかどうかを調べる．もし，例外規則が存在する場合は，その文字の位置や，前後の文字も調べ（言語や例外の種類によりどれだけの範囲を見る必要があるかは異なる），例外に該当するかどうかを判定する．例外に該当する場合は，その例外規則によって翻字を行う．該当しない場合，および，例外規則に関連がない場合は，対応表に基づいて文字レベルでの翻字を行う．該当する翻字のない文字はその言語以外の文字か記号であると判断し，翻字しない．

なお，ALA-LC 翻字規則はその言語の基礎的な知識がある人間向けに書かれているため，記述内容を解釈しながら，計算機上で実現できる形で規則化する必要がある．本研究では参考書[4] と各言語の書誌 DB 登録作業者の助言を得ながら，具体的な規則を作成している．

3.2.1 ロシア語

ロシア語の ALA-LC 翻字規則には，大文字，小文字それぞれ 38 文字，計 76 文字に対応する翻字が定義されている．本研究では，これらの文字について，原綴り・翻字対応表を作成した．

また，「ъ が単語の最後に出現する場合は，翻字しない」という例外が 1 つある．本研究ではこの例外を扱うため次の例外規則を作成した．

- ъ が出現する場合，その文字の単語内の位置を調べ，単語の最後でない場合のみ翻字化する．

なお，規則の動作確認のため，約 200 の原綴り・翻字データを作成し，繰り返し規則の修正を行った．

3.2.2 ヒンディー語

ヒンディー語の ALA-LC 翻字規則では，音節の先頭にくる母音・二重母音 19 字，子音に続く母音・二重母音 19 字，子音 46 字およびサンスクリット語からの外来語で用いられる文字である，アヴァグラハ（Avagraha）「ऽ」および他の文字に付与されるヴィサルガ（Visarga）「◌ः」☆に対応する翻字が定義されている．本研究では，これらの文字について，原綴り・翻字対応表を作成した．

また，単純な対応表では処理できない例外として以下のものがある．

- 鼻音化記号であり主に子音の前に来るアヌスワーラ（Anusvara）「◌ं」は，続く文字の種類に応じて 6 通りに翻字し分けなければならない．
- 鼻音化記号であり母音と子音に付くアヌナーシカ（Anunasika，チャンドラ・ビンドゥ）「◌ँ」は，続く文字の種類に応じて 2 通りに翻字し分けなければならない．
- 次の場合を除いて，すべての子音の後ろには母音「a」を付けなければならない．
  - 他の母音が子音に続く場合．
  - 母音を省略する特別記号であるハル記号（halant）「◌्」が続く場合．

本研究では，これらの例外に対する例外規則を作成した．なお，規則の動作確認のため，約 600 の原綴り・翻字データを作成し，繰り返し規則の修正を行った．

3.3 翻字から原綴りへの自動変換

ALA-LC 翻字規則は，原綴りを翻字に変換する点だけを考慮しており，翻字から原綴りへの変換については，定義された規則は存在しない．そのため，本研究では，翻字と原綴りを参照しながら ALA-LC 翻字規則を逆に変換する規則の作成を行う．原綴りを翻字に変換する場合と同様に，翻字側の各文字を原綴りに対応させれば処理できる部分と，例外的な処理の必要な部分が存在する．そこで，基本的には原綴りから翻

☆ ヴィサルガ「◌ः」と後述するアヌスワーラ「◌ं」，アヌナーシカ「◌ँ」およびハル記号「◌्」は文字の右，上，下につく記号であるため，ここでは破線円とともに記しているが，実際の綴り字では，それぞれ破線円の位置に他の文字が入る．

字への変換で作成した「原綴り・翻字対応表」を逆にして,「翻字・原綴り対応表」を作成する．そのうえで，対応表では処理できない部分については，個別に「例外規則」を作成する．

対応表と例外規則による翻字から原綴りへの自動変換の基本的な処理手続きは次のようになる．まず，翻字の先頭から 1 文字を取り出し，その文字を含む例外規則が存在するかどうかを調べる．もし，例外規則が存在する場合は，その翻字の前後（言語や例外の種類によりどれだけの範囲を見る必要があるかは異なる）の翻字も調べ，例外に該当するかどうかを判定する．例外に該当する場合は，その例外規則によって原綴り化を行う．該当しない場合，および，例外規則に関連がない場合は，対応表に基づいて文字レベルでの原綴り化を行う．該当する原綴りのない文字は翻字ではないと判断し，原綴り化しない．

3.3.1 ロシア語

ロシア語では，1 種類の翻字が複数のキリル文字に対応するということはない．そのため，翻字から原綴りへの変換も大部分は対応表で対処できる．ただし，異なる翻字間で同じ文字が重複するものがあるため，その点を処理する例外規則が必要になる．たとえば，キリル文字の「**ъ**」の翻字は「ˈE͡」であり,「**я**」の翻字は「ˈA͡」となるため，翻字の中で「ˈ」と「͡」の各文字が重複する．本研究では，76 種類の「翻字・原綴り対応表」を作成した．うち 50 種類については対応表だけで対処可能であるが，26 種類の翻字は，**表 2** に示す文字の重複がある 10 グループに分けられるので，例外規則を作成した．

また,「Shch」と「shch」に関しては，他の翻字「Sh」,「sh」および「ch」との間で以下の曖昧性が存在する．

- Shch として「**щ**」に変換するか，Sh と ch に分けて「**ш ч**」に変換するか．
- shch として「**щ**」に変換するか，sh と ch に分けて「**ш ч**」に変換するか．

この 2 点について，ロシア語話者に確認したところ，ロシア語では，**ш**と**ч**，あるいは**ш**と**ч**が連続することはほとんどありえないため，それぞれ**щ**と**щ**であると考えて差し支えないという解答を得た．そのため，本システムでは，Shch，shch を 1 つの単位として原綴りに変換することとする．

なお，規則の動作確認のため，原綴りから翻字への変換で用いたものと同様の約 200 データを用いて，繰り返し規則の修正を行った．

3.3.2 ヒンディー語

ヒンディー語の翻字から原綴りへの変換は，ロシア語に比べて複雑である．基本的には，原綴り・翻字対応表を逆にした，翻字・原綴り対応表を作成するが，以下の例外について対応する必要がある．

- 文字の重複がある翻字が，母音 8 種類 3 グループ，子音 30 種類 13 グループあるため，これらについては場合分けをする必要がある．
- 母音・二重母音全 19 種は，原綴りでは単独の場合と子音に接続する場合で文字が異なるが，翻字では区別がないため，当該翻字の単語内の位置と他の翻字の並びから原綴りを判断する必要がある．
- 子音の後に母音を表す翻字が存在しない場合，母音が省略されているので，原綴りにハル記号「◌्」を加える必要がある．
- アヌスワーラ「◌ं」を表す 6 種類の翻字のうち，4 種類「ṅ，ṇ，ṅ，m」は子音「ङ，ण，ङ，म」の翻字と同じであるため，アヌスワーラであるか，子音であるかを区別する必要がある．ALA-LC 翻字規則においてアヌスワーラは後に続く子音により翻字の種類が決まるので，翻字の並びからアヌスワーラになる可能性があるかどうかは判別できる．

ただし，最後の例外において，アヌスワーラになる可能性があると判別された翻字の並びは，子音+ハル記号でも原綴り化できるという曖昧性が存在する．たとえば，nandana という翻字の 3 文字目の「n」を，アヌスワーラだと考えれば「नंदन」となり，子音+ハル記号だと考えれば「नन्दन」となる．この点について，ヒンディー語の書誌 DB 登録作業者に確認したところ，意味的にはどちらも同じであるが，辞書ではアヌスワーラを用いるとの解答を得た．そのため，本システムでは，アヌスワーラと解釈して変換することとする．

なお，規則の動作確認のため，原綴りから翻字への変換で用いたものと同様の約 600 データを用いて，繰り返し規則の修正を行った．

表 2 文字の重複があるロシア語の翻字
Table 2 Romanized scripts of Russian in which character overlaps.

翻字	原綴り（キリル文字）
Z, Zh	**З**, **Ж**
I, ˈE͡, ˈU͡, ˈA͡	**И**, **ъ**, **Ю**, **Я**
K, Kh	**К**, **Х**
S, Sh, Shch	**С**, **Ш**, **Щ**
T,T͡S	**Т**, **Ц**
z, zh	**з**, **ж**
i, ˈe͡, ˈu͡, ˈa͡	**и**, **ъ**, **ю**, **я**
k, kh	**к**, **х**
s, sh, shch	**с**, **ш**, **щ**
t, t͡s	**т**, **ц**

4. 実　　験

3 章で述べた方法によりロシア語とヒンディー語の原綴り・翻字相互変換システムを作成した．本章では，システムの精度を確かめるため，未知のデータによる以下の 2 種類の実験を行う．

実験 1　原綴りから ALA-LC 翻字への変換

実験 2　ALA-LC 翻字から原綴りへの変換

4.1 実　験　1

実験 1 では，原綴りから ALA-LC 翻字への変換の精度評価実験を行う．

実験の評価用データとして，原綴りとその翻字の対を用意し，原綴りを，システムに入力する問題，翻字を正解として正解率を計算する．今回の実験では，東京外国語大学附属図書館のロシア語とヒンディー語の書誌情報データの中からそれぞれ 283，270 のデータを無作為に抜き出して評価用データとする．東京外国語大学附属図書館では，非ローマンアルファベット系言語の書誌情報を原綴りと翻字の両方で独自に登録しているため，本実験のデータとして利用することができる．

各データは主に図書の書誌タイトルの原綴りと翻字の対になっているため，分かち書きされた複数の単語を含んでいる．そのため，正解率の計算は 1 レコード全体でなく，分かち書きされた単語ごとに計算することとする．実験結果を**表 3** に示す．

結果から，今回の実験ではロシア語についてはほぼ正しく翻字に変換することができた．システムの出力が誤りだった 2 例は，単純な対応表の記述ミスであったため，すぐに修正が可能である．

ヒンディー語についても，ほぼ正しく翻字に変換することができた．特に今回のデータでは，原綴り・翻字対応表の範囲での誤りはなかった．16 例の誤りはすべて例外規則に関するものであり，子音の後ろに母音「a」を補う規則とアヌスワーラの翻字における場合分けの規則に不具合があることが原因だった．今後，この点を修正することで精度は向上するものと思われる．

4.1.1 実　験　2

実験 2 では，ALA-LC 翻字から原綴りへの変換の精度評価実験を行う．

実験の評価用データとして，実験 1 と同じデータを用いる．ただし実験 2 では翻字を，システムに入力する問題，原綴りを正解として正解率を計算する．また，正解率の計算は，実験 1 と同様に 1 レコード全体でなく，分かち書きされた単語ごとに計算することとする．実験結果を**表 4** に示す．

表 3　原綴りから翻字への変換の実験結果
Table 3　Experimental results of transliteration.

言語	件数	分かち数	正解数	正解率（%）
ロシア	283	1,365	1,363	99.9
ヒンディー	270	2,518	2,502	99.4

表 4　翻字から原綴りへの変換の実験結果
Table 4　Experimental results of back-transliteration.

言語	件数	分かち数	正解数	正解率（%）
ロシア	283	1,365	1,334	97.7
ヒンディー	270	2,518	2,294	91.1

結果から，今回の実験ではどちらの言語の場合も実験 1 に比べて正解率は低いものの，90%以上の精度を得ることができた．

ロシア語については，かなりの精度で正しく原綴りに変換することができているといえる．今回のデータでは翻字・原綴り対応表に誤りはなく，システムの出力が誤りだった 31 例は以下の原因によるものだった．

- 原綴りで ъ が単語の最後にくる場合は翻字されないという翻字規則により，翻字側に情報がないため，原綴りで ъ が再現できない誤り（2 例）
- 翻字にまざっていた英語部分が翻字で用いるローマンアルファベットと一致したため，キリル文字に変換された誤り（3 例）
- 翻字にまざっていた記号が翻字で用いる文字と一致したため，キリル文字に変換された誤り（26 例）

上記のうち，1 つ目の誤りは，元の翻字規則において非可逆な変換となるため，完全な自動化は不可能である．対策としては，ъ が最後に付く単語を収集しておき，ъ が省略されている可能性の高い語が出現した場合に，利用者に確認を促すことなどが考えられる．他の 2 つの誤りは，ロシア語に他の言語の文字や記号がまざることが原因であり，純粋なロシア語の問題ではないが，実際の書誌情報では頻発するため，間違いになりやすいパターンを登録しておき，変換時に利用者に確認をするなどの対処をする必要がある．

ヒンディー語については，誤りが 224 例あった．その原因は以下のようである．

- 母音の省略を示すハル記号が適切に挿入されていない誤り（17 例）
- 3.3.2 項で述べたとおり，アヌスワーラか子音+ハル記号かで曖昧性がある場合，本システムではアヌスワーラで原綴り化しているが，実際の図書では子音+ ハル記号が用いられていたことによる誤り（196 例）
- 翻字にまざっていた記号が翻字で用いる文字と一致したため，原綴りに変換されてしまった誤り

（11 例）

1 つ目の誤りは，例外規則を再度見直す必要がある．2 つ目の誤りは，現実の図書では「子音+ハル記号」が相当数存在することを意味している．そのため，このような曖昧性がある場合には，他の候補も利用者に提示する必要がある．3 つ目の誤りは，ヒンディー語の問題ではないものの，頻繁に見られるため，利用者に確認をするなどの方法を検討する必要がある．

5. 関 連 研 究

原綴りからの翻字作成は，機械翻訳や言語横断検索などでの応用を目的に研究されることが多い[5)~8)]．これらの研究では，主に固有名詞を対象として，異なる言語間で単語対を獲得するアライメントの要素が強い．そのため，既存の翻字規則に則った翻字よりも，実際のテキスト対に現れる単語対を自動的に発見することに重点が置かれる．本研究では，特定の単語だけでなく書誌 DB に必要なすべての文字列を翻字する必要がある点，翻字規則には ALA-LC 翻字規則を用いなければならない点，完全な自動化よりも正確な翻字を優先しなければならない点がこうした研究とは異なる．

ALA-LC 翻字規則での翻字を行う他のシステムとしては大阪外国語大学附属図書館のシステム[9)] が存在するが，翻字への変換精度や方式は明らかでない．また，蔵書検索としての利用に特化しているため，翻字から原綴りへの変換は対象に入っていない．

6. まとめと今後の課題

本論文では，我々が開発している非ローマンアルファベット系言語の原綴り・翻字相互の自動変換システムについて述べた．最初の開発言語として，ロシア語とヒンディー語によるシステムを実装し，実験による評価を行った．結果から比較的高精度で変換が行えているといえる．本システムをさらに良くするため，課題として次のことがあげられる．

- 例外規則などの修正．
  今回実装した 2 言語について，実験により明らかになった例外規則の問題点を修正し，より精度の高い相互変換を実現する必要がある．
- 他の言語への対応．
  2.2 節で難易度を検討し，今回実装を見送った他の言語についても，今後モジュールを作成し，本システムで扱える言語を拡大していく必要がある．現在，我々は次の開発言語をアラビア語に定め開発にとりかかっている．また，分かち書きが必要な言語の中で，タイ語では分かち書きと他の翻字を同時に行う研究が行われている[10)]．日本国内のタイ語の蔵書数は非常に多いため[11)]，今後こうした研究を参考にタイ語への対応も検討する必要がある．

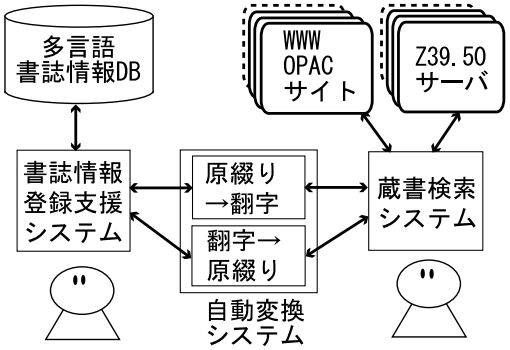


図 2 応用システムへの拡張
Fig. 2 Extension to application systems.

また，本システムによって提供される「原綴りから翻字への変換」と「翻字から原綴りへの変換」機能を**図 2** に示すような枠組みで応用することも今後の課題である．「書誌 DB 登録支援システム」においては，翻字作成の作業を効率化するだけでなく，将来の書誌情報の本格的な多言語化を見据えて，既存の書誌 DB 内の翻字から原綴りを作成する支援も視野に入れる．

また，「検索支援システム」では，世界中の ALA-LC 翻字を用いる図書館の中で，インターネット経由の Z39.50[12)] や HTTP-OPAC での接続が可能な図書館に対するサイト横断検索の実現を視野に入れている．

謝辞 本論文に対して有益なご意見・ご指摘をいただきました査読者の方に感謝いたします．

参 考 文 献

1) 国立情報学研究所：NACSIS-CAT/ILL，目録所在情報サービスホームページ．
http://www.nii.ac.jp/CAT-ILL/contents/home.html
2) Barry, R.K. (Ed.): *ALA-LC Romanization Tables—Transliteration Schemes for Non-Roman Scripts*, Library of Congress (1997).
3) Unicode Inc.: The Unicode Character Code Charts (2005).
http://www.unicode.org/charts/
4) 町田和彦：ヒンディー語研修テキスト—1 文字と発音，東京外国語大学アジア・アフリカ言語文化研究所 (1994).
5) Knight, K. and Graehl, J.: Machine Transliteration, *Proc. 8th Conference on European chapter of the Association for Computational*

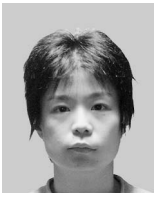
Linguistics, pp.128–135 (1997).
6) Kang, B.J. and Choi, K.S.: Automatic Transliteration and Back-Transliteration by Decision Tree Learning, *Proc. 2nd International Conference on Language Resources & Evaluation (2000)*.
7) 阿玉泰宗，橋本泰一，徳永健伸，田中穂積：日英言語横断情報検索のための翻訳知識の獲得，情報処理学会論文誌：データベース，Vol.45, No.SIG10 (TOD 23), pp.37–48 (2004).
8) AbdulJaleel, N. and Larkey, L.S.: Statistical transliteration for English-Arabic cross language information retrieval, *Proc. 12th International Conference on Information and Knowledge Management*, pp.139–146 (2003).
9) 大阪外国語大学多言語検索システム．http://wwwlib.osaka-gaidai.ac.jp/
10) Aroonmanakun, W. and Rivepiboon, W.: A Unified Model of Thai Romanization and Word Segmentation, *Proc. PACLIC 18*, pp.205–212 (2004).
11) 国立情報学研究所：付録 外国語図書の所蔵状況等に関する調査結果，NACSIS-CAT/ILL ニュースレター，No.5, p.6 (2001).
12) Z39.50 Maintenance Agency: Information Retireval (Z39.50): Application Service definition and Protocol Specification, the National Information Standards Organization (2003). http://www.loc.gov/z3950/agency/Z39-50-2003.pdf

(平成 17 年 6 月 20 日受付)
(平成 17 年 10 月 4 日採録)

(担当編集委員　中挾 知延子)

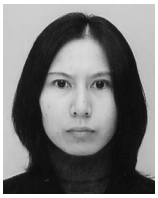
望月　源（正会員）

1993 年金沢大学経済学部経済学科卒業．1999 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．同年より，北陸先端科学技術大学院大学情報科学研究科助手を経て，2002 年より，東京外国語大学外国語学部講師，現在に至る．博士（情報科学）．自然言語処理，知的情報検索，コーパス言語学等の研究に従事．言語処理学会，教育システム情報学会，AECT 各会員．



大和加寿子

1995 年東京外国語大学外国語学部中国語学科卒業．1997 年東京都立大学大学院人文科学研究科修士課程中国文学専攻修了．1997 年 4 月より東京外国語大学附属図書館に勤務，現在に至る．日本図書館協会会員．



前嶋　淳子

1999 年関西大学文学部哲学科卒業．同年 4 月より，滋賀県立図書館臨時的任用職員を経て，2000 年 4 月より，東京外国語大学附属図書館職員．

林　俊成（正会員）

1991 年早稲田大学理工学部電気工学科卒業．1996 年同大学大学院理工学研究科博士後期課程単位取得退学．同年より早稲田大学理工学部助手，1998 年より東京外国語大学外国語学部助手．同大学講師を経て，2004 年より助教授，現在に至る．博士（工学）．文書画像構造解析，マルチメディア語学教材開発，e-learning 等の研究に従事．教育工学会，SICET，AECT 各会員．