# テキストタイプのロシア語データベースの共同構築について（研究事例と提案）

外国語学部　山　田　　隆

日本語ベースのコンピュータを使って漢字とロシア文字が混在する文書を作ったり、計算するのは、作業自体の難しさよりも、まず環境を整えることの方がはるかにやっかいである。どんな機種でも、またどんなプログラムでもいいわけではない。細かい仕様と神経質な調整を要求するのである。言い換えれば、ロシア語混在環境を実現するには経済的な負担と精神面でのストレスを覚悟しなければならない。

しかしながら日本語とロシア語を混在して、実用目的に使うことは十分に可能である。マックは基本的には、キーボードドライバーとフォントさえ組み込んでやれば、すぐさまロシア語入力が可能となる。フォントの調達だけで片づけられるのなら、タイ文字でも、ヘブライ文字だって問題はない。

一方ウインドウズマシンでは「Windows95」から、システムにロシア語入力ドライバーが付属して、漢字とロシア文字の混在が可能になった。ようやくマックに肩を並べたといえよう。したがって従来から2つのOSの間にあった入力と混在が可能かという能力の差は、一応解消されたことになる。この意味では、どちらのタイプの機種でも日本語とロシア語の混在文書を作成することができる。

しかし解決を要する問題はまだ尽きていない。どちらのシステムとも、プログラムとの相性がデリケートで、結果としてロシア語入力ができない場合が多い。システムはロシア語入力装置をもっていても、必ずしも入力のパスが通るとは限らないという難点が頻発する。ワープロを例に挙げても、市場に出回っているものの中でロシア語の実用になるのは、それの何分の一かにすぎない。

文書作成の場合、ロシア語入力の他に、ワードラップや両端そろえ、アクセ

ント入力など学習教材が果たさねばならない役割や仕上がったときのプレゼンテーション効果を考えた仕様が、最低限でも求められると思うのだが、現実はなかなか厳しい。この3つの機能を備えたワープロに巡り会えただけでも、その研究者は幸運だったといわねばならない。

部分的にせよ、日本語やロシア語などの自然言語をコンピュータで処理する時代は、今や現実のものとなっている。電算処理が、もちろんすべての研究分野を覆い尽くすものではない。しかしコンピュータを利用する研究が増えているのも周知の事実である。個人研究のレベルに限らず、さらに共同研究が発案され、そして異機種間でデータのやり取りが行われるところまで発展する。そのときにまた新たな問題が発生する。それが、データの互換性云々である。これは、詰まるところコンピュータシステムが使用するコード体系の問題に帰せられる。研究者各人に自分が使用する字体に寄せる思い入れが強ければ強いほど、相互のデータ互換性がますます困難になってしまう。ワープロフォーマットによる互換性はおろか、テキスト形式のデータでさえ、文字化けが激しいという結果になる。確かにロシア語を打ち込んだはずなのに、何が書いてあるのか読めないのである。それもそのはずだ。ロシア語フォントには代表的なもので4種類のコード表がある。DOS 規格、WINDOWS 規格、MAC 規格、そしてKOI 規格のことであるが、これらの間にロシア語コードの一致はまったくみられない。

このようなカオスは、コンピュータを専門としないロシア語研究者にとってはとてつもなく大きな障害であり、この1件だけでコンピュータの導入にためらいを感じさせる程である。これに対する解決策はいくつかある。その1。研究者集団が同一のキーボードドライバーとフォントを使用すること。たとえば、本来ユニックス用の規格であったが、今ではインターネットの標準といわれるKOI8 規格のフォントを使用することである。その2。テキスト形式のデータで保存し、必要なときにデータコンバートをすること。これには多少ともコンピュータの知識と技術を必要とする。だから万人向きの解決方法とはいえないかもしれない。しかしロシア語のコード表を知っているのなら、どのようなフォ

ントにでも組み換えることができる。ロシア文字を自在に操る観すらある。福井大学の浦井康男氏を始めとするロシア語研究者は、ストリームエディタを使って、目的とするタイプのデータに変換して、相当の研究成果をあげている。

次に、蓄積したロシア語テキストの再利用例を紹介しよう。教材作成や例文列挙のときにこれがデータベースに早がわりするのである。これにはワープロの検索機能や単体の検索プログラムが必要である。ロシア語コードが入力できれば、検索プログラムはどんなものでも構わない。単語や語形による検索が一般的な利用法と思うが、文法的な構文をテーマにするときは、この手法がベストとはいい難い。接頭辞や接尾辞をキーワードにして検索すると、なるほど、余計な例文も拾ってしまうが、それなりの結果は得られる。不定人称文に特徴的な語尾とか、目当てとする格語尾を頼りに用例を検索していく。ワイルドカードも当然のごとく使用する。

検索プログラムの想定を裏切るような使い方をするので能率のあがらない方法ではあるが、研究者の記憶にある、経験則的な引用をはるかに凌ぐ分量をこなすことができるし、時間の節約を図ることもできると思う。ハヴローニナ著の『ロシア語を話しましょう』全体を検索するのに10秒とはかからない。全体で350頁の『みんなのロシア語』教科書から例文を引用するのにも、やはり20秒を費やすことはない。この時『File検索犬ポチ』や『GripGrop』のように検索プログラムにエディタが連動していれば、快適な作業空間が保障され、能率がとてつもなくあがる。但し、どのようなツールを使っても、それが目的に適ったロシア語文なのかを判定するのは研究者自身であり、データの適否を検証するのに長い時間がかかることの方がむしろ、積み残されている深刻な課題である。

北海道大学を中心とする研究者が、ドストエフスキィの『罪と罰』の電子テキストを完成させている。カラムジンの『哀れなリーザ』の電子テキストも知られている。研究テーマに関連するテキストは通常、苦労を伴って研究者自身の手によってコンピュータに打ち込まれてきた。現在でも状況に大きな変化はないと思う。そのような労作を他の研究者も利用できるとすれば、研究の上で

大きな助けとなるに違いない。

今日インターネットの利用によって電子テキストの収集や蓄積が容易になってきた。ホームページが提供する情報、研究機関や個人と取り交わす情報がすべて、電子メディアであり、蓄積を含めて再加工の対象にすることができる。目的を絞り込んだデータの収集にあたるなら、利用価値の高いものになると期待される。ここで一言。本稿で取りあげているいかなる種類のデータも、使用目的としては研究だけを想定しており、営利目的の利用はまったく考えていない。しかしいずれの場合にも、著作権や版権の問題には充分な配慮を必要とすることは言うまでもない。

この他にCD版の電子テキストや電子辞書も利用の候補にあげられよう。入手ルートを日本市場に限定したとき、出版点数はまだ数例を数えるにすぎない。ロシアの国内市場ではこのようなCDロムをかなりひんぱんに見かけるという。ロシアで購入するのでもないかぎり、日本国内では将来のCDロム販売と品揃えを待つことになろう。

ロシア語の入力作業は、これまでキーボードによる手作業に頼るのが主流であった。この作業形態は今後とも消えることはない。その一方で、福井大学の浦井康男氏が精力的な研究を続けてきて、その後北海道大学の安藤厚氏や北海学園大学の寺田吉孝氏が継承発展させているOCR装置を使う入力方法が、テキストの大量生産に適していると思われる。電子機器の性能向上により読み取り精度が、実用的な水準に達している。これはコピー感覚で電子テキストを作りだすのだから、きわめて簡便で、生産力を伴う方法である。作業の手順は、テキストの読み取りとスペルチェックが連動しており、その後プリンタ出力などをして人による校正作業で締めくくられる。OCRによる入力作業については、たとえばリングヴォ社『スタイラス』のような、現在のロシア市場で市販されているプログラムを使用するのが、もっとも経済的だと考える。ロシアのパソコン市場では9割方をIBM互換機が占有しており、ウィンドウズマシンが花盛りである。Windows95で稼働するプログラムでは確認していないが、Windows3.1用のプログラムを走らせるには英語版のウィンドウズをロシア語

化して、その上でプログラムを稼働させる方法がある。ロシア語版 OS を使用するにせよ、これは Windows3.1 においては必要不可欠な条件である。

ロシア語版 OCR プログラムで作成したテキストは、必要なコード変換を経て、ウィンドウズとマッキントッシュのいずれの機械でも利用することができる。コード変換用ツールには『SED』や『XTR』、またマックでは『SED』や『KONVERT』などが代表的である。

またウィンドウズマシンにライフボート社の『システムコマンダー』を組み込めば、1台のウィンドウズマシンで日本語とロシア語などの複数の OS が共存可能になるから、ロシア語システムで生成したテキストデータをハードディスクからそのまま日本語システムで使用できるという利点が生まれる。コンピュータの専門家から見れば非常に危険な使い方かもしれないが、今のところトラブルは発生していない。

最後にこうして蓄積されていくロシア語データを研究者の共有資産として活用することを、控えめに提案したい。なぜ控えめにかというと、1つに、データの総容量が、リームバブルメディアでなければ収納できないほど大きなものになることが予想され、時代のトピックとはいえ普及度の点で一般的とはいえない側面をもつからである。2つめとして、集積ルートと共通するが、供給ルートとして電子メールやネットサーバーを想定していて、これらはすべての研究者に無制限に開放されているとはいい難いからである。

実は電子テキストと一口にいっても、定本を定め、内容校正を済ませるまでに相当の労苦がつきまとう。コンピュータにデータを「打ち込むのは、末代の宝」ということばを、かつて黒崎氏は雑誌『ASCII』の連載で述べている。一度できあがったデータを利用するのは、データ総量が多ければ多いほど用途が広がると思うのだが、自分の文章の場合とは違い、どこにどれだけあるのかわからない資料を集積するのであるから、さまざまの困難が予想される。文学作品ほどではないにしても、ロシア語教材や新聞資料についても著作権や校正の問題が残っている。早い話しが、スキャナーの読み取り作業を誰が担当するのかなどは、たぶん誰もが敬遠したい工程に違いない。

残されている、そしてこれから発生する課題は、山積みの状態である。しかしこの種のデータベースが曲がりなりにでも構築される気運ができあがれば、日本におけるロシア研究の支援体制に新たな局面を加えることができるのではないか、と期待される。末筆になってしまったが、今回の発表と提案にあたって、ロシア語システムの動作確認の環境づくりを支援してくださった北海学園大学の寺田吉孝氏には記して、心からの謝意を表わしたい。

〈参考資料〉　ロシア語編集からみた日本語版ソフトウエアの状況について。これまで試すことのできたソフトウェアについて私見を交えながら判断してみた。文字どおりの参考である。

Mac OS の場合。

1．日露フォントの混在が可能であり、編集機能もある程度まで有効である。

| | | ワードラップ | 両端揃え | |
|---|---|---|---|---|
| ワープロ系 | Mac Write II | ○ | ○ | |
| | Nisus Writer 4.06J | ○ | ○ | |
| | MS Word 6.0 | ○ | ○ | (英語系スクリプトに限る) |
| | Word Perfect 3.1J | × | ○ | |
| | EGWord　6.7 | × | ○ | |
| エディタ系 | Tex-Edit Plus | ○ | ○ | |
| | Simple Text | ○ | × | |
| | Dimple Text | ○ | × | |
| | LightWay Text | × | × | |

2．1書体のみ編集可能で、ロシア語入力も可能なもの。

| | | | | |
|---|---|---|---|---|
| エディタ系 | MBB－Text | ○ | × | |
| | Jedit | ○ | × | |
| | 褌エディト | × | × | |

| | | |
|---|---|---|
| Yoo－Edit | × | × |

3．その他。次のソフトウエアにもロシア語入力が可能である。

Hyper Card 2.3　（ハイパーテキスト作成ツール）

クラリスワークス 4.0　（DB 機能で、ロシア語ソートが一部不正確になる）

スティッキーズ　（メモランダム）

4．入力が不可能らしい。

一太郎 for Mac

Windows　95 の場合。

1．日露フォントの混在が可能であり、編集機能もある程度まで有効である。

| | ワードラップ | 両端揃え | |
|---|---|---|---|
| MS Word for Windows 7.0 | ○ | ○ | （マック版とデータの互換性あり） |
| MS Word Pad | ○ | × | |

2．1書体のみ入力が可能で、ロシア語入力も可能である。

| | | |
|---|---|---|
| 秀丸エディタ 2.09 | ○ | × |

3．その他

クラリスワークス 4.0　（マック版とデータの互換性あり）

4．入力が不可能らしいもの。

一太郎 7.0 for Windows

Mifes　2.0

ワープロソフトには多機能と高性能を誇るところがあって、一般的には高価な商品が多い。研究予算に限度のある研究者には評価の対象に選定しにくい側面がある。その点で経済的な負担の軽いフリーウエアやシェアウエアの方にテスト品目が多くなってしまった。

ワープロ系のソフトウエア、特に日本語編集に特化したものは、力まかせに両端揃えをやるので、ロシア語の綴りがあらぬ所で切られる。その結果、どうにも判読に耐えない箇所が続出する傾向にある。一方エディタ系のソフトウエアは、文字入力の効率化に力を注ぐ向きがあるので、両端揃えのできないものが多い。ただし表音文字を基礎にしたエディタの場合には、句読点やスペースを区切りにしてワードラッピングが有効になる。教材作成としては見た目に多少格好悪いが、スペルを勝手に分断されるよりも、ましかもしれない。

なおアクセント記号を備えたフォントは、Linguist's Software Inc.社から「Cyrillic II」シリーズが提供されている。また一つひとつを書き記すことはしないが、上に挙げた商品名はすべて、各々の開発、供給元の登録商標である。これらのソフトウエアを含めて、ロシア語を扱うことのできるソフトウエアについてコメントを持ち合わせている方のご教示を切にお願いしたい。連絡は、yamada-t@sapporo-u.ac.jp.