ドストエフスキー『罪と罰』コンコーダンスについて

安藤 厚

このたび北大スラブ研究センターから「ドストエフスキー『罪と罰』コンコーダンス」(A Concordance to Dostoevsky's *Crime and Punishment*. Ed. by A. Ando, Y. Urai and T. Mochizuki. 3 vols. Slavic Research Center, Hokkaido University, Sapporo, 1994) が刊行された。これはスラブ研究センターの望月哲男、松田潤、文学部の栗原成郎、灰谷慶三、安藤厚、大西郁夫のチームによる、平成5年度北海道大学教育研究学内特別経費による研究プロジェクト「19世紀ロシア小説のコンピュータによる分析」(研究代表者 望月哲男) の最初の成果である(1)。

本書は『罪と罰』のロシア語テキスト (*Достоевский Ф. М.* Полное собрание сочинений в тридцати томах. Т.6. Л., 1973) の用語索引で、この小説に使用されている全単語25,794語形、169,890件 (句読点・引用符・記号類は除く) のうち、本文では、25,778語形、130,597件をアルファベット順に配列し、左右に文脈をつけて、出現場所を示し、別表では、使用頻度が1,000回を越える16語形 (и, не, в, что, он, на, я, с, а, как, это, его, так, но, же, да)、39,293件について、文脈なしで出現場所のみを示した。別表には、そのほか、使用頻度順の語形一覧、イタリック体表記の単語と「-с, -то, -ка」付きの単語のコンコーダンスも収録してある(2)。

長編小説の翻訳や文体研究をしていると、個々の単語の用例一覧があったらと思うことがある。古典語・英独仏語の世界では、1970年代からコンピュータによるコンコーダンスの作成が盛んに行われており、最近ではFD、CD-ROM等による文学テキストの電子出版が普及しているが、ロシア語の世界では、壮大な手仕事とも言うべきプーシキン辞典(3)や聖書のコンコーダンス(4)を別にすれば、米国でマンデリシタム、プーシキンの詩のコンコーダンス(5)が作られているくらいで、国内では電気通信大の岡本哲也教授がかつてチェーホフの短編のコンコーダンスを試作した例があるだけである(6)。ドストエフスキーに関しても、テキストの細部の研究が進むにつれ、コンコーダンスの必要性が話題になってはきたが、自分でそれを作ることになろうとは夢にも思わなかった。

4年前、新しいパソコンが買えたらと、深く考えもせず応募した平成2年度教育研究学内特別経費による研究プロジェクト「19世紀ロシア文学の計量的研究へのアプローチ」(研究代表者 灰谷慶三) が思いがけず採択され、急に話が進んだ。その前年ロシア文学会でカラムジーンの言語についてデータベースを利用した研究を発表した(7)福井大の浦井康男氏に教えを乞い、欧文用光学的文字読み取り装置 (OCR) Kurzweil 5200を購入し、『罪と罰』(約400ページ、100万字) のテキストデータの入力をはじめた。

Kurzweilは、英独仏語ほかローマ字表記の西欧の主要言語については、それぞれ数万語の参照辞書を内蔵し、高度な学習機能を備えているので、自動運転でもきわめて精度の高い読み取りデータを高速 (1ページ1～2分) で作ってくれるが、キリル文字については、フォントも参照辞書もないので、その学習機能を利用し、キリル文字に図形の似たローマ字・数字・

記号を当てて記憶させ、対話的な運転により入力を行った。キリル文字とローマ字・数字・記号の対応は、試行錯誤の末、б → 6、л → ji、п → n、р → p、ы → bi/ bI/ bl/ b1、ь → b、ю → IQ、я → & など、かなりユニークなものになったが、それはそれでモニター画面上の読み取り結果の点検には十分役に立った。作業速度も、最終段階では、対話的運転で 1 ページ約 15 分と、十分実用的なスピードが得られ、読み取り精度もほぼ 100% に達した。夏休み、春休みに集中的に仕事をして、1992 年春にはデータの入力が終わった。

読み取ったデータは浦井氏考案の「Advanced bits を利用したロシア語入力システム」のコードに変換し、同氏制作のスペルチェッカー(8)で誤りを点検した。このスペルチェッカーはカラムジンの『ロシア人旅行者の手紙』のデータをもとにした特殊なものだが、それにより、カラムジンとドストエフスキーの用語を比較した「新出単語表」（形容詞約 1,200 語、その他の品詞約 7,900 語形）が得られた。

スペルチェッカーのおかげでつづりの誤りはほぼなくなったが、до → да、на → не といった誤りや句読点・引用符・記号類の誤りは発見できないので、学生諸君の手も借りてプリントアウトによりテキストの点検を行った。

1992 年秋テキストデータの点検が一段落し、浦井氏が定評のある Oxford Concordance Program (9)の NEC パソコン版 Micro-OCP を使ってコンコーダンスを 1 部試作してくれた（A4 版、5 分冊、約 2,200 ページ）。

テキストデータベースの利用法は、コンコーダンスのほかにもいろいろ考えられる。

文章解析プログラムの文字列検索を利用すれば、アイディアさえあれば、小さなテーマの論文がほとんど無限に作れるだろう。

また、浦井氏は岡本教授らと共同で、ロシア語の文法規則をプログラム化した「ロシア語形態生成プログラム」を制作し(10)、カラムジーンの作品について各単語の形態情報を持ったデータベースを作成している。この手法を利用すれば『罪と罰』についても同様のデータベースが作れるので、その準備として上記『罪と罰』の「新出単語」の基本形・文法的属性等を確認する作業をはじめた。

各単語の形態情報を持ったデータベースが得られれば、基本形のもとに各変化形の全用例を記述した「辞典」の作成も可能になる。また「語形」ではなく「単語」を単位としたさまざまな数量的分析の道も開ける(11)。

1993 年春、平成 4 年度教育研究学内特別経費による研究プロジェクト「ゴンチャローフの文体の数量的・総合的研究」（研究者 大西郁夫）のおかげで、露文研究室に Macintosh のパソコンと Oki のポストスクリプトプリンタが入り、札幌 NJK システムの松本義昭氏の協力で、浦井氏の「ロシア語入力システム」によるデータを Mac のキリル文字フォント Trans Cyrillic のコードに変換するための「ロシア語変換プログラム」が完成し、私たちのデータを Mac に移して高品位でプリントアウトできるようになった。試みに、このシステムを使って、DTP の手法で文章解析プログラムによる文字列検索を利用した小論文を書いてみたところ、写植にも負けない印刷面が得られた(12)。

昨年末、これまた思いもかけず、平成 5 年度教育研究学内特別経費配分の通知が届いた。年度末まで 3 カ月余りしかないので迷ったが、思いきって『罪と罰』のコンコーダンスを出版

することにした。

まずテキストデータの再点検が必要だった。スペルチェッカーと学生諸君の協力のおかげで単語についてはほぼミスのない自信はあったが、句読点・記号類についてはいろいろ点検整備が必要だった。たとえば、左右の引用符の数が一致せず、検索プログラムを使って約1,000個の引用符を端から点検することになった（300個ほどのところでミスが見つかって救われたが）。版下を印刷所に渡したあと、全集の正誤表に載っている誤植の訂正を忘れていたのに気づき、あわてて関係の数十ページを差し替えるといった失敗もあった。

OCPによるコンコーダンスの出力は浦井氏が引き受けてくれたが、①テキストデータの文字・記号の定義と、②OCPの処理命令と、③データをMacに移すための「ロシア語変換プログラム」の仕様とを首尾一貫したものにするため、主に記号類の扱いについて、さまざまな調整が必要だった。たとえば、今回のOCPの処理命令では引用符（«»„“）をpunctuation（句読点――単語の切れ目となる）に定義したため、«вояж»-то が出力ファイルで «вояж» と -то の2語に分割されてしまい、手作業で訂正した。引用符・かっこ類はpadding（暫定文字――書体情報等を示し、単語の切れ目にはならない）に定義しておくべきだったと、あとで気づいた。

Micro-OCPには、データを読み込んだあと、コンコーダンス作成の際、作業ファイルへのアクセスが約10,000回を越えるとプログラムが止まってしまう欠陥があり、『罪と罰』のデータ（約1MB）を一度に処理することはできないことがわかり、浦井氏が東大か京大の大型計算機センターで汎用機上のOCPを動かすことも考えたが、これはコンコーダンスの出力を3回に分割することで解決した。

ページのレイアウトについても、Macのシステムによるプリントアウトを前提に、さまざまの工夫が必要だった。9ポイント・2段組にすることで、A4変形版・3分冊・約1,500ページと、手ごろな大きさの本になった。参照の便を考え、同一のキーワードごとに見出し語と使用頻度数を入れることにしたが、Micro-OCPの仕様では各見出し語の前に空白行が1行ずつ入り、無駄が多いので、浦井氏が新たに書いたレイアウト用プログラムで必要な整形を行った（不要な空白行を削除し、1段60行ごとに肩見出しをつけ、各行に3カ所タブを入れ、肩見出し・見出し語の前後にゴシック体の指示の記号をつける）(13)。

さらに、松本氏の協力で「ロシア語変換プログラム」を改良し、データをMacに移す際、独仏文字・イタリック体・ゴシック体の部分は、前後に一定の記号をつければ機械的に必要なフォント・字体に設定されるようになった。この「ロシア語変換プログラム」改訂版は、浦井氏の「ロシア語入力システム」、NEC機＋テクノメイト、OASYS機＋ヨールカで入力したテキストファイルのキリル文字を、MacのTrans Cyrillic、あるいはHaxoткa フォントのコードに変換することができ、パソコン通信ネットNIFTY-Serveに公開されている。

OCPの仕様では1行の長さは固定長（今回は1行70字）だが、Macでは半角文字の幅がそれぞれ異なるプロポーショナル・フォントを使うので、固定長のデータのままでは文脈の両端が不揃いになる。この点も、文脈の両端を一定の寸法でブロック指定し削除することで、整然としたレイアウトが実現した。

今回の経験から、このようなコンコーダンスの作成には大型計算機も電算写植機も不要で、

多少強力なパソコン（NECとMac）とポストスクリプトプリンタがあれば、そのほうが、印刷まで含めて、はるかに小回りがきき、柔軟な作業ができることを実感した。

データ入力を思い立ってからコンコーダンスの刊行までの4年間に技術の進歩はめざましいものがあった。パソコン・プリンタ類の飛躍的なパワーアップはもちろん、ソフト面でも、ロシア製の、強力なスペルチェッカーソフトORFOや、きわめて精度の高いOCRソフトAutorが出回りはじめた。これらをうまく利用すれば、語学・文学ばかりでなく、歴史・経済・社会などの分野でも手軽にデータベースが構築できるようになると思われる。

テキストに関しても状況が変化している。データの入力が終わった段階で、ペテルブルグから新たに旧正書法のテキストによる35巻全集刊行の計画が伝えられた。その広告には、30巻全集のテキストの問題点として、①Бог等が小文字表記に変えられている、②句読点に変更がある、などが指摘されている。そのほか私たちは、合成語におけるハイフンの使用が、30巻全集では初出の «Русский вестник» 版(14)等より広く行われているのに気づいた。これらについては、もちろん、再検討が必要だが、テキストの校訂自体は私たちの仕事ではないし、私たちは句読点やハイフンの用法に特に関心があるわけでもないので、これらについて問題が残るとしても、コンコーダンスの本質には影響ないと考え、今回は30巻全集のテキストをそのまま用いた。全集の正誤表に記されている誤植、及び私たちの気づいた数個の誤植は訂正した。30巻全集のテキストに改善の余地のあるのはたしかだが、このようなコンコーダンスの作成がテキストへの理解を深め、データの信頼度を高めるという面も強調しておきたい。

本書の準備の過程では、逆引きの語形一覧も用意したが、結局スペースの都合で割愛した。しかしこれは語学的な研究には不可欠のものなので、もし機会があれば、別冊のかたちででも刊行したいと思っている。反省点の一つである。

注

(1)本書については「スラブ研究センターニュース」58（1994.7）、「えうゐ」25（1994.8）にも紹介記事を書いたが、書き漏らした点もあるので、この場を借りて改めて紹介させていただくことにした。

(2)『罪と罰』のイタリック体の用法については、池田和彦「『罪と罰』のイタリック」: RUSISTIKA 10 (1993.6), pp.201-212を参照。

(3)Словарь языка Пушкина. 4 тт. и приложения. Гос. изд-во Иностранных и национальных словарей, М., 1956-1961.

Новые материалы к словарю языка А. С. Пушкина. «Наука», М., 1982.

Bartoszewicz, A., Komendacka, I. *Indeks a tergo do słownika języka Aleksandra Puszkina.* Wyd. Uniwersytetu Warszawskiego, Warszawa, 1985.

(4)Симфония или алфавитный указатель к Священному Писанию. Изд. 3-е. Изд-во Миссионерского Союза „Свет на Востоке", Корнталь, 1971.

Симфония или словарь-указатель к Священному Писанию Ветхого и Нового Завета. Под ред. митрополита Питирима. Том 1: А — Г. Изд. Московской патриархии, М., 1988.

(5) Koubourlis, D. J. (ed.) *A Concordance to the Poems of Osip Mandelstam.* Cornell University Press, Ithaca, NY, 1974.

Shaw, J. T. (ed.) *Pushkin : A Concordance to the Poetry.* 2 vols. Slavica Publishers, Columbus, Ohio, 1984.

(6)岡本哲也, 坂本義行「ロシア語テキストのコンコーダンスの自動編集」: ロシヤ語ロシヤ文学研究 6 (1974) , pp.94-95.

中世ロシア語の分野では、「ノヴゴロド第一年代記 (シノド本) 語彙集」: 古代ロシア研究19 (1994.3) のような成果が出てきている。

(7)浦井康男「カラムジンにおける強調の大文字の使用について——パソコンによる文字列検索を利用して」: ロシア語ロシア文学研究 22 (1990) , pp.28-40.

(8)浦井康男「キリル文字の光学的読み取りとスペルチェッカーの作成」: 福井大学教育学部紀要 I-44 (1992.7) , pp.55-75.

(9) Hockey, S. and Marriot, I. *Oxford Concordance Program, Version 1.0, Users' Manual.* Oxford University Computing Service, Oxford, 1980.

長瀬眞理, 西村弘之『コンピュータによる文章解析入門—— OCP への招待』 オーム社, 1986.10.

山縣宏光「文章解析プログラム OCP の使い方」: 東京大学大型計算機センターニュース 19-9 (1987. 10) , pp. 76-84.

小澤義明「文章解析プログラム OCP (Oxford Concordance Program)」: 京都大学大型計算機センター広報 27-1 (1994.2) , pp.13-17.

(10)岡本哲也, 山本佳代子, 浦井康男「ロシア語の形態生成と Prolog 系 (I) ——名詞・形容詞・代名詞」: 電気通信大学紀要 6-1 (1993.6) , pp.29-38.

岡本哲也, 山本佳代子, 浦井康男「ロシア語の形態生成と Prolog 系 (II) ——動詞」: 電気通信大学紀要 6-2 (1993.12) , pp.157-169.

(11)浦井康男「カラムジンにおけるエピテットの自動分析——パソコン上のワードフォーム辞書を利用して」: ロシア語ロシア文学研究 24 (1992) , pp.75-87.

浦井康男「数量化理論によるエピテットの性格付けの試み——ロシア語形態処理の発展」: 福井大学教育学部紀要 I-46 (1993.7) , pp.21-33.

(12)安藤厚「『罪と罰』における「古風な表現」について——女性名詞単数造格語尾「-ию/-ью」の使い分け」: RUSISTIKA 10 (1993.6) , pp.187-200.

安藤厚「『罪と罰』における「血」の用法」: えうゐ 24 (1993.12) , pp.34-40.

(13) Micro-OCP による作業の詳細は、浦井氏が「福井大学情報処理センターニュース NETWORK」 8-2 に紹介を予定している。

(14) «Русский вестник» (1866) (IDC Micro-Edition) . International Documentation Centre, Tumba.