コーパス言語学から見た日本語研究と辞書編纂

——中日対訳コーパスの構築とその応用研究をめぐって——

徐　一平

はじめに

二〇世紀九〇年代から、コンピューター・テクノロジーの飛躍的な発展により、自然科学はもとより、人文科学の諸分野でも思考革新や手段更新が急速に進んできている。特に自然科学に近いと言われる言語研究の分野においては、コンピューター・コーパスによる新しい言語研究の方法、つまりコーパス言語学がついに言語学界に認められ、数多くの研究者に愛用されるようになり、二一世紀の言語研究はコーパス言語学抜きでは語れないという現象さえ生まれてきているのである。そして、多くの研究者の努力により、大型コーパスも次々と完成されている。その代表的なものをあげると、例えば、時代の早いものとして、欧米では二〇世紀の六〇年代からコーパスが構築され、八〇年代になるとすでに大規模なコーパスが完成されている。Brown コーパスやイギリスの国家コーパスなどはそのすばらしいものである。そして、中国や日本あるいは韓国においても欧米に次いで大型コーパスが開発されている。

本稿ではコーパス言語学の理論と立場を踏まえ、中国と日本における日本語と中国語コーパスの作成と利用の状況および成果を紹介し、また筆者が所属している北京日本学

研究センターが開発した「中日対訳コーパス」の構築とその応用研究をめぐって、それがどのように日本語研究に関わっているのかということを考えながら論を進めていきたいと思う。

一　コーパス言語学とコーパスの構築

(一)　世界におけるコーパス言語学の発展

コーパス（Corpus）という術語は、最初の定義は「言語分析のための言語資料の集積」であったが（中国語訳の「語料庫」は正にその本来の意味が訳し出されている）、最近はコンピューター処理の可能な言語資料の集積という意味、つまり「コンピューター・コーパス」や「電子コーパス」と同義に使われるようになっている。このコンピューター処理の可能なコーパスに基づいて言語研究を行う学科は、「コーパス言語学」と称され、言語研究の新しい研究法と研究分野の一つとして発展しつつある。その特徴はLeech［1992］によれば次の通りである。

言語能力よりも言語運用能力に中心を置く。
言語の普遍的特性の解明よりも個別言語の言語記述に中心を置く。
質的な言語モデルのみならず、数量的な言語モデルにも中心を置く。
言語研究における合理主義的な立場よりも経験主義的な立場に中心を置く。

つまり、コーパスを駆使して行われるコーパス言語学の研究はチョムスキーを代表とする合理主義的な言語研究法とは対照的なもので、一種の経験主義的な言語研究法だという解釈である。情報化社会の到来により、ますます大規模な言語データを処理しなければならないという状況から、自然言語処理の研究により緊迫した課題が出されている。また言語理論の研究においても、研究者の内省による言語データに頼ってしかできない研究もますます難しくなってきているため、最近、いわゆるコーパス言語学的な研究法は、記述言語学や機能文法など経験主義的な立場をとる研究者だけでなく、構文文法や生成文法などの合理主義的な学者にまで援用されるほど、言語学界での一つの流れになりつつある。コンピューターが日増しに普及しつつある今日の情報化社会においては、コーパスとコーパス言語学はもはやわれわれの言語研究や語学教育に縁遠いものではなく、むしろ大いに取り入れるべきものになっているのである。

ただ、他の言語学に比べれば、コーパス言語学は現在、

まだコーパスの作成法と利用法に研究の重点を置く段階で、少し遅れていると言えよう。

作成法に関する研究には、コーパスの規模や種類（サンプル型とモニター型、汎用型と特殊型、共時型と通時型、話し言葉型と文章語型、文字型と音声型、単一言語型と複数言語の対訳型など）、構造デザイン、言語の処理機能と情報付与などの問題解決がその課題になっている。

利用法に関する研究には、利用目的、分野別（語彙、文法、文体、言語史、対照研究）利用法、関連分野（辞書学、語学教育、翻訳学、文化学、インターネット）への応用などの課題がある。特に最近注目されているのは、学習者コーパスの利用で、それを使って学習者の言語使用状況を把握することができ、第二言語習得の教育現場で応用されつつある。日本語教育者にとっては、常に日本語の実態を把握する必要があり、大量の言語データを調査しなければならない。コーパスがなかった時代には、それはすべて手作業のカード式で行われていて、短時間で対象の必要に応じた用例収集はほとんど不可能に近い作業であった。また、中国人日本語教育者のように、自分の母語ではない者にとっては、反省に頼れない用例の収集はさらに困難な作業であった。コーパスの出現により、これらの作業がいとも簡単にできるようになった。豊富な実例をもたらすコーパスの完成は、まさにテクノロジーの進歩であり、言語研究と言語教育にとってはすばらしい福音になるであろう。

上記のような課題に応えることができるコーパスを作るには、言語工学者のみならず、一般の言語学者や教育者の参加と互いの協力が必要となってくる。それがあって初めて理想的なコーパスを立派に完成できるのであろう。

コーパス言語学は初期のSEU（Survey of English Usage, 一九五九年）など先駆的なものから、最初の機械可読なBrown大型コーパス（一九六四年）を経て、現在は大規模化、情報付与の再加工、多様化と複数言語化の時代を迎えているが、いずれもコーパスの作成法と利用法がそれぞれのコーパス開発の研究課題の中心であった。もちろん、コーパスによって新しい言語規則と理論を構築するという試みも大変重要な研究課題だと研究者たちも分かっているのだが、現段階ではまだ十分な条件が整っておらず、多少の研究が行われてもわずかの成果しか得られないのが現実であろう。しかし、コーパス言語学は、言語学という以上、言語理論の発展への貢献もなければならない。おそらくコーパス言語学に努力している全員がそう願っているであろう。今後作成法と利用法の研究がさらに進み、コーパスの普及と利用が拡大するにしたがって、研究の中心も必ず徐々にそちらのほうへと移っていくのであろう。

(二) 日本と中国におけるコーパスの構築と利用

日本語と中国語のコーパス構築や応用は欧米に遅れを取ってはいるものの、コーパス言語学の発展とともに、それなりの経験と成果が見られた。特に最近その発展は著しく、欧米に劣らない大型コーパスの開発構築とその応用研究が盛んに行われている。

まず、日本では、国立国語研究所（現独立行政法人国立国語研究所）の言語資料集・データ集とそれに基づいた研究成果および機械処理の成果が先駆的な貢献としてあげられる。日本で最初に電子計算機を言語研究に導入したのは国立国語研究所であるが、そこで「新聞記事データベース」「高校教科書データ集」「話し言葉データ集」や「国定読本用語総覧 CD-ROM 版」「日本語学習者による日本語作文データベース」などのコーパスが開発されている。特に最近公開されている「現代日本語書き言葉均衡コーパス」は、書籍約三千万語（一万四二三サンプル、プレーンテキスト／XML ファイル）、白書約四八〇万語（一五〇〇サンプル、プレーンテキスト／XML ファイル）、Yahoo! 知恵袋約五二〇万語（四万五七二五サンプル、プレーンテキスト）、国会会議録約四九〇万語（一五九サンプル、プレーンテキスト）などを収録し、ファイルの形式は、プレーンテキスト（タグなし）および XML ファイル（タグあり）になっており、現段階では日本語として最も大規模なコーパスになっているのである。そのほかに三省堂から出版された「国定読本用語総覧 CD-ROM 版」には、一九〇四年四月から一九四九年三月までの間に使用された文部省著作の小学校用国語教科書六種の全文内容が電子化されており、用例 KWIC、語彙表、用例データベース検索プログラムなどが載せてあり、日本語研究や教科書研究への利用が期待されている。また、文部科学省助成の大型プロジェクトである日本方言データベースも多数の学者の共同研究により完成し、その成果は二二枚もの CD-ROM で公開されている。

さらに中国語に関しては、大阪外国語大学中国語学科（現大阪大学外国語学部）で公開された中国語コーパスが知られている。

その他、情報処理振興事業協会が公開した IPA コーパス、新情報処理開発機構が公開した RWC テキストデータベース（形態素解析済み日本語コーパス）、国際電気通信基礎技術研究所が開発した ATR 対話データベース（日英対訳コーパス）、CD-ROM 毎日新聞（データ集）、朝日新聞全文記事 CD-ROM、読売新聞、中日新聞、日本経済新聞など各種新聞コーパスも知られているが、一般研究者に多く使われているのは朝日新聞記事 CD-ROM であった。その成果は遠藤［1990］、近藤［1993］、後藤［1993］、荻野

[1994]、荻野・塩田[1994]、田野村[1994、2009]、後藤[1996]、石井正彦[2009]などがあるが、その内容は語彙、文法、修辞、文体にわたるものである。さらに文学作品に関連するものとして「新潮文庫の100冊」やインターネット上で公開されている「青空文庫」も非常に利用しやすいコーパスになっている。

また、非公開の自家製コーパスを利用した研究は最近、日本の大学、特に関係学科の学生や院生の間で盛んに行われており、彼らの学会論文や学位論文には大量の実例データと綿密な統計結果が発表されており、その成果はきわめて目立ってきている。つまり、コーパスによる言語研究の方法は、すでに日本の語学研究者の間では日増しに一般化していると言えよう。特に『日本語教育』一三〇号に、特集「コーパスと日本語教育――現状と課題」が組まれ、その成果が大きく注目されている。

一方、中国では、ここ十数年来いくつかの大学や研究所で中国語コーパスの開発が進んでいる。その中で、特に清華大学が開発したタグ付きコーパスや北京大学の計算語言学研究所が開発した中国語研究用コーパス、中国社会科学院語言応用研究所の大型中国語コーパスなどが有名である。そして英語教育の面では、北京外国語大学が開発した英語の話し言葉コーパスや学習者コーパスなどもその先端を走っている。それに対して、中国の日本語学界では、二、三年前まではまだコーパスに関する認識が薄く、日本語コーパスの利用は一部の帰国研究者とその学生にしか見られず、研究者によってはコーパスの術語さえ分からない人もいるような状態であった。

しかし、九〇年代の後半から、注目すべき研究が少しずつ出始めた。例えば、戴宝玉[1997]の「複合辞データベースの構築と応用研究」、徐一平・施建軍[1999]の「日中作文コーパスから見た日本語の否定表現」、曹大峰・森山卓郎[1999]の「感動詞に関する日中対照研究」など、いずれも既成のコーパスや自家製コーパスを使った研究である。一方、そのような状況を打開し、世界のコーパス言語学の流れについていこうと、北京日本学研究センターが「中日対訳コーパス」の開発に乗り出し、以下に述べる成果をあげたのである。

二　中日対訳コーパスの構築とそれをめぐる問題点

以上、紹介したような単一言語のコーパスがどんどん開発されている中、二言語対訳あるいは多言語対訳のパラレルコーパスも、コーパスの開発に従って注目されてきている。しかし、現在開発されている対訳パラレルコーパスは、どの国においても英語との対訳パラレルコーパスで

（例えば中国では英中対訳コーパス、韓国では韓英対訳コーパスなど）、日本語と中国語の対訳コーパスは未だに稀有に近い状態である。そのような状況を受け、特にインターネットによる多言語情報交流の要請に応じて、北京日本学研究センターでは、一九九六年一〇月から、中日双方の専門家やスタッフを集め、検討に検討を重ねて、ほぼ二年間の準備を経て、「中日対訳コーパスの構築と応用研究」という研究プロジェクトを企画、一九九八年九月からその実施を始めた。このプロジェクトの代表者は北京日本学研究センターの徐一平で、正式に参加したメンバーとしては、山東大学、洛陽外国語学院、上海外国語大学、北京大学、北京第二外国語学院、広東外語外貿大学などの日本語研究者が中心となっていた。特別参加者としては、日本から国立国語研究所の故中野洋先生、京都橘女子大学の宮島達夫先生、協力参加者としては、福建師範大学、訳林出版社、東京外国語大学、和光女子大学などの研究者もいた。そのほか、センター客員研究員、プロジェクト開始以来のセンター派遣教授、特に言語コースの先生方、修士課程、博士課程の学生たち多数がこのプロジェクトの開発に参加した。

協力機関として、日立中央研究所、日本大学、奈良先端科学技術大学院大学、北京大学計算語言学研究所などから、技術的な支援・協力をいただいた。その後、国際交流基金の資金援助と日立中央研究所の技術協力を得、中国の国家社会科学基金研究プロジェクトとしても認定されたこの研究プロジェクトは、世界初の「多用途・大型中日対訳パラレルコーパス」の構築を目標に、中日両国の研究者により、着実に進められ、二〇〇二年の一一月に、ほぼ当初の目標を達成し、CD-ROM化された。このコーパスは、中国において日本語研究と日本語教育に携わり、コーパスに対して強い関心をもつ学者の集結と中日双方の協力によって、初めて完成させることができたと言えよう。以下、われわれの開発経験に基づき、いくつかの問題点を報告していきたいと思う。

(一) 汎用性を重んじた内容構成

まず、今回の中日対訳コーパスの第一次計画としては、二千万字（実際に完成したコーパスの文字数は、約二〇一三万字）からなる文章データを入力し、その上で、アラインメントや文法情報を入れたタグ付けなどの加工を考えていた。そして、構築されたコーパスが多種多様の研究目的に適応できるものにしていくために、まず汎用性を重んじて、表1のような内容構成を目標にした。

この構成案は、次の表2が示すように言語研究を中心に、翻訳や文学、文化研究へも利用できるように図るものであるが、収録テキストは従来のサンプル型コーパスとは

表1　中日対訳コーパスの内容構成案

内容＼時代・文体		現　代（解放後・昭和以降）	近　代（口語）（解放前・大正～昭和前期）	近代・古代（文語）（民国以前・明治以前）	%
創作文	小　説	45	20	5	70
	シナリオ	6	3	1	10
	エッセイ	2	2	1	5
情報文	論説文	5	4	1	10
	説明文	2	2	1	5
%		60	31	9	100

表2　中日対訳コーパスの文章構成案

多言語（中日対訳）	特定言語（中／日）	
全文型		サンプル型
文章語		会話文
創作文		情報文
現代語	近代口語	文語

表3　中日対訳コーパスの実際の構成内容

（単位：万字）

	中国語データ		日本語データ		合　計
	原文	訳文	原文	訳文	
小　説	287	422	244	224	1177
散　文			18	16	34
伝　記	105	157	30	23	315
詩　歌			11	9	20
論　説	132	205	20	16	373
法　律			1	0.78	1.78
その他	5	7	45	36	93
小　計	529	791	369	324.78	2013.78

違って原本・訳本ともに全文テキスト、一部の名作（例えば、夏目漱石の『坊ちゃん』や川端康成の『雪国』など）には複数の訳本を平行して収録した。中日対照研究はもちろんのこと、単一言語の研究にも利用できるようにしたいと考えたからである。また、文章語情報の共時的研究と利用を中心に考える一方、会話文の研究や通時的研究をも、ある程度配慮したものである。

入力の方法としては、まず中日の原本と訳本をそれぞれ中国語版の Windows と日本語版の Windows で処理し、GB コードと Shift-JIS コードという異なった文字コードを使って基礎テキストファイルを作り、対訳コーパスを構築する段階で改めて同一文字コードに変換させる方法をとっている。

実際には、おびただしい翻訳作品の中から、専門家を集めて選択を行い、さらにテキスト入手の問題も絡み、最終的に完成したコーパスの内容は、表3に示す結果となっている。

(二) 中日通用型の同窓モニターとアラインメント

中日対訳コーパスということは、二言語のテキストをそのまま同じ画面で見比べることが必要であるが、そのためには、中日両言語の同窓モニターとアラインメントが前提条件となろう。

しかし、開発当初のパーソナル・コンピューターにおける漢字コードは国によって異なるものであった。例えば、「中」という漢字の区点コードをみると、中国のコード（GB2312-80）では5448だが、日本のコード（JIS90）では3570であるので、中国コードの「中」は日本語版 Windows では半角カタカナの「ﾖﾒﾒ」に、日本コードの「中」は中国語版 Windows では「面」に化けてしまって一致しなかったのである。そのために、中日両言語の同窓モニターは、それまでほとんど日本語版の DOS や Windows 上で日本語フォントの代用や擬似中国語フォントの使用によるものであったが、精確さに欠けるばかりでなく、中国語版の Windows への通用性もないのである。

その後、この問題を根本的に解決するために、中日韓統合漢字拡大コード集（Unicode, ユニコード）が制定され、Microsoft Office97/2000 に実装されるようになったが、OS（Windows95/98）のコード体系はやはり、中国の GBK と日本の JIS が異なっているので、問題すべてが解決されたとは言えず、それに関する応用研究もまだ少なかったのである。そのような状況のもと、今回われわれが開発した中日対訳コーパスは日本語版 Windows のみならず、中国語版 Windows でもそのまま対訳テキストをモニターできるものでなければならないという当初の目的があるので、漢字コード変換技術や多言語表現技術などの応用研究が重要

な課題となっていた。

また、対訳パラレルコーパスには、文レベルの二言語平行テキストが一番理想的であるが、言語表現の異質により完全に実現することは不可能である。そのため、まず段落レベルのアラインメントから進めることが、第一歩として重要であるが、課題として中日両言語の分段規則の研究と分段差異の対策研究などがあった。

いわゆるアラインメントは、当然文レベルのアラインメントが最も理想である。しかし、実際の対訳作品を見た場合、必ずしも一文対一文に翻訳されているとは限らない。特に意訳という文学手段が使われた場合、以下のように少なくとも四種類のケースが見られる。

(a) 原文が一文で訳文も一文に訳されている。
(b) 原文が一文で訳文が二文以上の文（一段落になる場合をも含む）に訳されている。
(c) 原文が二文以上の文で訳文が一文に訳されている。
(d) 意訳の方法がとられ、原文と訳文の間は文と文の対応関係が見られない。しかし段落全体の意味はほぼ同じである。

以上のように、実際の対訳コーパスにおいては、原文と訳文の間で文レベルのアラインメントが実現できるのは一部分に過ぎないのである。このような状況に基づいて、今回われわれが開発した中日対訳コーパスにおいては、以下のような原則でアラインメントの方法をとったのである。

(1) 原文を基準にし、訳文を忠実に合わせてアラインメントを実現し、原則的には訳文を直してはいけない。
(2) 原文一文と訳文一文が対応する場合、文レベルでアラインメントをする。
(3) 原文一文に対して訳文が二文以上の文に訳されている場合、一文対多文でアラインメントをする。
(4) 原文二文以上の文に対して訳文が一文に訳されている場合、多文対一文でアラインメントをする。
(5) 原文二文以上の文に対して訳文も二文以上の文をもって意味上対応している場合、段落レベルでアラインメントをする。

大規模な対訳コーパスのアラインメントをする場合、もしコンピューターを駆使して自動的にそれができれば、もちろんそれに越したことはないが、しかし、前述の状況から、機械的にアラインメントをすると確かに効率は上がるが、その反面、正確さはかなり下がる結果がもたらされるに違いない。このような状況に基づいて、このたび中日対訳コーパスを開発するにあたって、われわれはまず手作業

によってアラインメントをしたが、今後はそれをベースにしてアラインメントの自動化ソフトを開発し、徐々にコンピューターによるアラインメントを実現していきたいと考えている。

㈢ 適宜な情報付与を施した加工型コーパス

高度の処理効果と分析効率を図るためには、電子テキストを使用目的に合わせて加工する必要がある。中日対訳コーパスに関しても、まず情報付与が基本的な課題である。

テキストに付けておく情報は大体二種類に分かれる。テキスト情報と言語情報である。テキスト情報は出典情報（例えば、書名、著者名、訳者名、版元、出版時期など）と文字コード情報（例えば、GB、GBK、JIS、CJKなど）を含み、普通テキストの先頭に簡単に付与できるものだが、言語情報は形態素、品詞、構文、意味とコンテキストなど、全テキストの細部に付与する情報なので、情報の正確性、有効性、通用性などが大きな課題となる。特に漢字圏言語は英語などのように語と語の間にはっきりした標識がないために、形態素や語の分割作業が情報付与の前提となる。また、通用性を考えて、分割と付与の基準を勝手に設定することができない。さらに、大量処理の能率を図るためにも、自動標識付与プログラムの開発が必要である。

これらの課題をいくつかの段階に分けて解決していかなければならないが、これまでの日本語コーパスと中国語コーパスの長所を集成するとともに独自の特色を示すものの開発を進めたのである。

構築されたコーパスをより有効的に言語研究に活かすためには、取り入れたデータをいかに分割するか、またいかにタグ付けをするかにかかってくる。開発された中日対訳コーパスについては、二千万字のすべてのデータに対して、品詞性、構文、語意などすべての情報を入れたタグ付けを行うのは無理なので、二千万字の全データについて品詞性付きのタグ付けをした。が、そのうちの三〇万字については、さらに構文や語意情報付きのタグ付けをしたのである。

いわゆる文法情報付きのタグ付けというのは、つまり単語を単位にし、入力されたテキストファイルについて、品詞性や他の文法的な情報を一々付けていくことである。このようなタグ付けは、手作業による方法と機械で自動的に付けていく方法と二種類あるが、中国語についても日本語についてもいずれもすでに開発されたタグ付けシステムがあり、われわれのタグ付けのための基本的な条件は用意されていたわけである（例えば、日本語については奈良先端科学技術大学院大学が開発したChaSen、中国語については北京大学計算語言学研究所が開発したシステムがそれで

ある。今回開発された中日対訳コーパスの品詞タグ付け作業は、いずれもこれらのシステムを利用させてもらった)。

しかし、どのシステムにもそれを開発した作者の言語観が反映されていて、必ずしもわれわれのねらいと一致しない場合がしばしばあった。例えば、「北京日本学研究センター」という単位について、それを一語と認定するのか、それともさらに細かく分割するのか、また、「考え」という単位について、それを名詞とするのか、それとも動詞の連用形とするのか、システムによってそれぞれ違うタグ付けをしているのである。

以上のような問題を解決するために、われわれは、今後さらに中日対照言語学の立場から、現存のSGML基準について分析を行い、中日対訳コーパスにふさわしいタグ付けシステムを開発していきたいと考えている。このシステムを使って、まず手作業でタグ付けをし、その過程の中で、さらに機械による自動的なタグ付けの方法を模索していきたいと考えているのである。

(四) 多様な情報処理機能を持つ知能型コーパス

コンピューター・コーパスの長所と効率を活かすためには、多様な情報処理機能を備える必要がある。特に検索機能と統計機能は言語の調査と分析に不可欠である。今回開発された中日対訳コーパスには、次の機能を持つ知能的な検索ツールを付けておいた。

KWIC検索：
命中語を一列に並べて、その左右の文字列を一行だけ切り出す検索法

文型検索：
複数の語による文型を含めた文を選り出す検索法

候補語検索：
複数の候補文字を検索語にし、その中の一文字から命中した文を選り出す機能

複雑検索：
正規表現による包括、除外、組み合わせなどの処理機能を含む検索法

平行検索：
命中語の中日対訳文も平行に取り出す機能

度数統計：
言語単位（語、句、文、段など）の出現度数を統計して出す検索法

それまで、われわれが知っているすでに開発済みの中国語と日本語の検索統計ツールは数種類あったが、上記の多種多様の機能を一体に備えたものはなかった。また、検索統計のスピードや精確さも大きな課題になっていた。これ

らの問題をすべて解決できる検索ツールを開発するために、われわれは、日本の日立中央研究所に協力してもらい、コンピューター工学の専門家の協力の下、目指す検索機能を実現していったわけである。

三　中日対訳コーパスの応用研究

言語研究が進むにつれて、研究者の内省による研究よりも、大規模言語データに基づく経験主義的な研究がますます重要視されるようになってきた。非母語話者が当該言語を研究するにあたって、そのような研究はなおさら重要な意味をもつと考えられる。こういった研究の必要性から、コーパスによる言語研究はますます脚光を浴びるようになってきた。

一方、対照研究をする上で、対訳のある言語データを集めにくいということは、対照研究を束縛する大きな枷でもあった。そのため、われわれの中日対訳コーパスは、開発当初から中国国内はもちろんのこと、日本の研究者にも注目され、開発と同時に応用研究として利用されてきた。以下にあげる例は、いずれもわれわれの中間成果を利用して行われた応用研究の成果である。

(一)　中国語の〝吧〟と日本語の「だろう」の対照研究例

今までの中日対照研究の領域では、中国語の〝吧〟と日本語の「だろう」は基本的に対応していると考えられていたが、われわれの中日対訳コーパスを利用して研究した結果、必ずしもそうではないということが分かった。

曹大峰［2001］によれば、日本語の「だろう」に中国語の〝吧〟が対応しているのは三六％弱に過ぎず、一方、中国語の〝吧〟に日本語の「だろう」（「だろう＋が・がね・か・な・ね」の複合式も含めて）が対応しているのもわずか九％に過ぎなかった。

つまり、中国語の〝吧〟が多用される意志文、行為文、軽い問い掛けの文などに日本語の「だろう」は対応していないのである。

一方、日本語の「だろう」も、中国語の〝吧〟だけでなく、〝呢〟〝吗〟〝啊／呀〟〝大概〟〝可能〟〝会〟〝也许〟〝恐怕〟〝说不定〟〝一定〟〝难道〟〝是否〟など、多種多様の言語表現形式が対応していることが分かった。

(二)　中国語の〝呢〟の研究例

現代中国語の〝呢〟については、従来の研究では主に二つの観点に分かれると見ている。一つは疑問文の文末形式の〝呢〟のスコープが疑問詞にあると見る考え方であり、

もう一つは〝呢〟のスコープが命題全体にあると見る考え方である。

程遠巍［2002］が、中日対訳コーパスを利用して、それまでと違う以下のような〝呢〟の表記機能を明らかにしている。

(1)「W＋V(O)＋呢」の心理文は二種類に分かれる。一つは、話し手が命題に疑念を感じ、判断をする疑いの文（非反語表現と称する）であり、もう一つは、話し手の命題に対する判断を表す反語表現の形をとった判断文である。

(2) 非反語表現の場合は、〝呢〟が付加されることによって、話し手の回答を求めようとする意味から疑いを表す意味に変化する。反語表現の場合は、話し手の強い判断をあらわす意味から〝呢〟の機能によって疑いを含んだ話し手の判断を表すことになる。

(3) 非反語表現においても、反語表現においても、〝呢〟の上位的意味は、命題に示された内容に対して、話し手自身の疑念を表すものであり、その下位的意味は、判断を諦めた疑念と、疑念の表出を前提にする判断である。

(4)「W＋V(O)＋呢」における〝呢〟の標記機能は、話し手心中の疑いを表し、聞き手の介入を求めようとしないことを表すものである。〝呢〟は対自的であり、対外的なものではない。

(三) アスペクトに関する中日対照研究例

現代日本語の「している」という同時の形態論的な形式は、「継続性」（〈動作の継続〉と〈結果の継続〉）という基本的なアスペクト的な意味だけでなく、「パーフェクト」という派生的なアスペクト的な意味をも表すと、日本語研究者によって明らかにされている。一方、中国語ではアスペクト的な意味を表す言語形態としては、〝着〟と〝了〟があるが、しかし「パーフェクト」を表す場合には、中国語ではどうなるのだろうかという問題は、必ずしも明らかにされていない。

彭広陸［2002］が、中日対訳コーパスを利用して研究した結果、以下のように指摘した。

(1)〈パーフェクト〉を表すのに〝着〟より〝了〟がよく利用されており、両者間にはある程度の使い分けが見られる。

(2)〝了$_2$〟だけではなく、〝了$_1$〟も〈パーフェクト〉を表すことができる。

(3)〈パーフェクト〉を表す手段として、〝着〟〝了〟だけでなく、他の形式も利用、あるいは併用される。中

国語にはもっぱら〈パーフェクト〉を表現するための文法形式は存在しないということになる。

(4) 中国語では、無標の形も〈パーフェクト〉の表現手段として多用される。

(5) 日本語の〈パーフェクト〉の表現形式が中国語に訳される時は、〈パーフェクト〉を表さない形式になることがある。

徐京梅［2002］は、中日対訳コーパスを使って、特に中国語の〝了〟と日本語の「している」形式について対照研究を行った。その結果を以下のように指摘している。

(1) 中国語の助詞〝了〟は、完成相として、その根本的なアスペクト的意味は「完成性＝限界到達性」である。一方、日本語の「している」形式のパーフェクトは、工藤［1995］が指摘しているように、その一番本質的なアスペクト的意味は〈完成性＋結果・効力〉だから、この〝了〟と「している」形式の理論上の接点は「完成性」である。

(2) 「している」形式の文は、パーフェクト的な意味で、よく〝了〟を使った文と対応する。〝了〟文は、結果状態性の意味で、「している」文に対応する。この結果状態性の意味は、単に〝了〟によるものではなく、ほかに特殊な文型や、場面・文脈との共同作業によって表される場合が多い。

(3) 「している」形式は、内部分化の意味で〝～了〟に対応するのに対して、〝～了〟は、外部要素との融合による意味――結果状態意味指向の意味で、「している」形式に対応する。

(4) 中国語には、パーフェクト相の専用形式は存在しない。

(四) 教育上の応用例

このように完成された中日対訳コーパスは、研究者の研究に資するだけでなく、学生教育上でも大いに利用されている。当センターは、大学院修士課程以上の教育を行っているが、学生の修士論文の作成にあたっては、中日対訳コーパスが大きな役割を果たしている。中国人学生は、しばしば一つの言語と他の言語との対照を通して研究テーマを決定する。その場合、対訳コーパスは、論文テーマに沿った言語データの収集に非常に威力を発揮すると考えられる。例えば、二〇〇二年度当センター言語コースの学生の修士論文作成にあたり、四人のうち三人がこの中日対訳コーパスを使って第一資料を調査したのである。その意味で、この対訳コーパスの開発は、正に教育の観点からしても、非常に重要なインフラ整備だと言えよう。

そのほかに、中日対訳コーパスを使った翻訳の研究や、他のコーパスと関連させて研究した成果もある。詳しいことは、文末の参考文献にあげた『中日対訳語料庫的研制与応用研究論文集』を参照されたい。

四　中日対訳コーパスと辞書編纂

従来、辞書は言語学者が自分の作例あるいは少数の実例で、ことばの表現を解釈し、説明するものが多く、それらの解説が必ずしも言語事実に完全に一致するとは限らない。それは大規模なコーパスが作成されていない事情からして、やむを得ない一面もあるだろう。しかし、パソコンの性能向上と大規模なコーパスの構築にしたがって、最近では大型コーパスを基にして、客観的に言語の使用実態を反映する辞書の編纂が可能になり、しかも言語学界でも日に日に重要視されてきた。例えば、すでに公にされている、HarperCollins Publishers Limited で出版されたコーパスに基づいた英語辞書のシリーズや北京大学計算語言学研究所編『現代漢語語法信息詞典』などがその例である。ただし、中国では、いまだにコーパスを基にした日中・中日の辞書が編纂されていないのである。それは、やはり今まではそれにふさわしいコーパスが開発されていなかったためだと思われる。

ところで、今回開発された中日対訳コーパスの完成により、コーパスに基づく日中・中日辞書の編纂も可能になったわけである。このコーパスに基づけば、用語の選択、用例の採集、語彙の解釈のいずれも、客観的なデータの裏付けが得られ、このような辞書が完成されれば、われわれが開発した中日対訳コーパスの二次開発になるだけでなく、きっと今後の日本語教育、日本語研究にも大きな刺激が与えられるに違いないと考えられる。

中日対訳コーパスに基づく辞書には、以下のようないくつかの特色が考えられる。

・日常よく使われる語彙が厳選収録される。
・語彙の理解を助けるために、適切な実例がコーパスから抽出される。
・全項目に使用頻度や必要な文法情報が明示される。
・掲載語彙をより広くより深く理解してもらうために、関連の類義語や関連語もコーパスより抽出され、掲示される。

おそらく、このような日本語辞書は、中国においても、日本においても初めての試みになると思うので、是非、関係者の皆様にもご協力とご指示をいただきたいと心から願っている。

結び

中日対訳コーパスの構築とその応用研究は、中国ではもちろんのこと、日本あるいは世界でも新しい応用型の研究プロジェクトである。理論的な意味からも、実践的な意味からも、その成果は必ずや言語学、第二言語習得、翻訳学、情報工学またはコンピューター科学に大きく寄与するに違いない。それに基づいた辞書編纂も、紙媒体の辞書が実現できない機能も実現されるに違いない。そればかりでなく、さらに多言語コーパスの開発あるいは中日両言語の機械翻訳に貴重な経験とデータを提供してくれるであろう。また、多くの中日言語研究者、教育者、学習者に、より便利な言語資料と新しい言語手段を提供してくれるであろう。そして、このようなコーパス言語学の発展にしたがって、より多くのコーパスもさらに構築されていくに違いない。現在中国国内で進行している日本語関係のコーパスとしては、上海外国語大学が中心になって開発している日本語学習者作文コーパス、北京日本学研究センターが開発している日本人講演者による話し言葉コーパスなどがある。さらに、中日対訳だけでなく、中日韓、あるいはもっと多くの言語を対象にした多言語パラレル対訳コーパスへと、今後構築が発展していけば、正に国際的な研究に結びつく研究成果が完成するに違いないだろう。

今後の日本学研究は、より学際的な研究、あるいは国際的な研究が要求される時代になると思われる。このコーパス言語学を通しての日本語・中国語の対照研究は、日本語学研究者だけでなく、翻訳研究者、コンピューター自然言語処理研究者、言語工学研究者などが協力して、初めてできる研究である。

本稿は、二一世紀の日本語研究のために、コーパス言語学という新しい課題と可能性を示し、この「情報革命」や「知的経済」時代に、われわれ言語学者や教育者が担うべき、もしくは担わざるを得ない仕事を予想しながら、筆者の経験したコーパス構築のプロジェクトを通して、日本語研究と周りの状況を報告したものである。同分野の学者もしくは近隣分野の学者や専門家の皆様からいささかでもご関心やご指摘が得られれば幸甚に思う。

主要参考文献

石井正彦編『日本語コーパス語彙研究――単語使用の諸相』私家版、二〇〇九年。
遠藤仁「「親類」と「親戚」の語誌」『国語学研究』三〇、東北大学文学部「国語学研究」刊行会、一九九〇年。
荻野綱男編『日本語の文法の構造 2』私家版、一九九四年。

荻野綱男・塩田雄大「朝日新聞データベースを使用した言語研究」『日本語学』一三―五、一九九四年。
工藤真由美『アスペクト・テンス体系とテクスト――現代日本語の時間の表現』ひつじ書房、一九九五年。
国立国語研究所『日中作文コーパスの作成とその利用　論文とデータ』一九九九年。
後藤斉「「神話」の比喩的用法について――コーパス言語学からのアプローチ」『東北大学言語学論集』二、一九九三年。
後藤斉「言語研究のためのデータとしてのコーパスの概念について」『東北大学言語学論集』四、一九九五年。
後藤斉「コーパスとしての新聞記事テキストデータ――終助詞『かしら』をめぐって」『東北大学言語学論集』五、一九九六年。
近藤泰弘「文法研究における大量言語データ――副助詞研究を例にして」『武蔵野文学』四〇、一九九三年。
齊藤俊雄ほか『英語コーパス言語学』研究社、一九九八年。
徐一平・施建軍「日中作文コーパスから見た日本語の否定表現」『日中作文コーパスの作成とその利用　論文とデータ』国立国語研究所、一九九九年。
徐一平・曹大峰『中日対訳語料庫的研制与応用研究論文集』外語教学与研究出版社、二〇〇二年。
徐京梅「「～了」とシティル形式の対照研究」『中日対訳語料庫的研制与応用研究論文集』外語教学与研究出版社、二〇〇二年。
曹大峰・森山卓郎「感動詞に関する日中対照研究」『二一世紀的日語教育』大連理工大学出版社、一九九九年。
曹大峰「中日対訳コーパスとその対照研究への援用――「吧(ba)」と「だろう」の研究例」『中国日語教学研究文集』大連理工大学出版社、二〇〇一年。
戴宝玉「複合辞データベースの構築と応用研究」『日語学習与研究』三、一九九七年。
田野村忠温「丁寧体の述語否定形の選択に関する計量的調査――「～ません」と「ないです」」『大阪外国語大学論集』一一、一九九四年。
田野村忠温『電子資料と日本語研究・続』私家版、二〇〇九年。
中国日語教学研究会編『二一世紀的日語教育』大連理工大学出版社、一九九九年。
陳原『漢語語言文字信息処理』上海教育出版社、一九九七年。
程遠巍「W＋V(O)＋呢」における「呢」の標記機能――「中日対訳コーパス」の用例を利用して」『中日対訳語料庫的研制与応用研究論文集』外語教学与研究出版社、二〇〇二年。
日本語教育学会『日本語教育』一三〇号、二〇〇六年。
苗楓林ほか『信息網絡時代与日本研究』山東大学出版社、一九九九年。
彭広陸「パーフェクトを表す「している」と対応する中国語の表現――「中日対訳コーパス」を資料として」『中日対

訳語料庫的研制与応用研究論文集』外語教学与研究出版社、二〇〇二年。

Leech, G., "Corpus Processing," W. Bright ed., *International Encyclopedia of Linguistics*, New York: Oxford University Press, 1992.