自 然 言 語 処 理　134－1
（1999．11．25）

中日対訳コーパスの作成状況と今後の課題

曹 大峰[I II III]　中野 洋[II]
徐 一平[III]　隈井 裕之[IV]

[I]山東大学　[II]国立国語研究所
[III]北京日本学研究センター　[IV]日立製作所

概要

筆者らは複数言語の対照研究に利用できる「中日対訳コーパス」を中国社会科学基金助成の共同研究プロジェクトとして作成している。これはコンピュータ技術の飛躍的発展とマルチ言語環境の迅速な改善により実現可能になった。本稿では1300万字の規模をもつその基礎版の内容と多言語処理機能を披露し、現在の進捗状況と今後の課題を紹介するとともに、特にアラインメントの簡易化と効率化、中日両国語の対応表示と検索、情報付与基準の標準化、言語対照研究への応用などの問題を取り上げ、現場からの報告と提言を行う。
キーワード：中日対訳コーパス　多言語処理　アラインメント　タグセット

Development of "Chinese-Japanese Bilingual Corpus" and Its Remaining Tasks

Dafeng Cao[I II III]　Hiroshi Nakano[II]　Yiping Xu[III]　Hiroyuki Kumai[IV]
[I]Shandong University　[II]The National Language Research Institute
[III]Beijing Center for Japanese Studies　[IV]Hitachi Ltd.

Abstract

With the rapid development of computer technologies and multilingual environments, a "Chinese-Japanese Bilingual Corpus", which is supposed to be useful for multilingual contrast researches, is being constructed as a collaborate research project sponsored by China Social Sciences Fund. Here, with a scale of 13 million characters, we will introdnce the content of the basic edition of this corpus and its function in multilingual processing as well as its development and remaining tasks. Especially, we will deal with problems in alignment's simplification and efficiency; problems in the corresponding expression and reference between Chinese and Japanese; problems in the standardization of principles in adding new information and problems in this corpus's application in language contrast researches and so on. Finally, we also provide reports and suggestions coming from real practice in dealing with these problems.
Key Words: Chinese-Japanese Bilingual Corpus, multilingual disposal, alignment, tag set

1．はじめに

最近、コンピュータ技術の飛躍的発展とマルチ言語環境の迅速な改善により、単一言語の研究のみならず、複数言語の対照研究にもパソコンの利用が可能になり、現実化されつつある。そのような情勢の中で、北京日本学研究センターでは昨年秋から、「中日対訳コーパスの構築と応用研究」という共同研究プロジェクトが実施され、これまで1300 万字規模の中日対訳コーパス基礎版が作られたのである。

同プロジェクトは、これまでにない規模と開発目標を有することにより、今年度中国社会科学基金(自然科学基金と並ぶ国家助成基金)プロジェクトと選定されて、今後3年間で完成していくことになっているが、中国の日本語学や日本語処理の分野においてまったく新しい出来事なので、大方の研究者の関心と期待を受けている一方、困難もたくさん予想された。これまでにいくつかの難題を乗り越えてきたが、ここにその状況を報告し、問題点と今後の課題を中心に検討する。

2．基礎版の作成状況

基礎データの準備と検索機能や対応付けなどの検討を目的に、一年間をかけて基礎版を試作した。本節でその状況を報告する。

2.1 使用OSとアプリケーション

文科系の研究者をユーザー対象とし、中国や日本で最も普及している WINDOWS95/98 と OFFICE97/2000 を作動OSとアプリケーションに採用した。作業は中国語と日本語の WINDOWS において行った。

2.2 収録内容と文字コード

初めての中日対訳コーパスとして、言語・文学・翻訳など幅広い研究領域に資することを考慮し、その汎用性と二次開発性を重んじて、表1のように内容デザインをした。

表1

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	近代語	文語

▨ 一次的　▨ 二次的　□ 三次的

基礎版では、まず中国の文学作品 23 篇、日本の文学作品 22 篇とその訳本を合わせて 98 篇(複数の訳文を含む)の作品全文と出典情報を収録した。文字数では日本語 704.9 万、中国語 590.0 万、合計約 1294.9 万字に達した。収録作品の件数は表2のように時代差を考慮に入れた。

表2

	現代 (解放後・昭和以来)	近代(口語) (解放前・昭和以前)
日本	13	9
中国	16	7
計	29	16

電子テキストの文字フォントはユニコードに即した漢字コードで対応表示を図るために、中国語は GB、日本語は JS を使用した。それによって中国語 OS でも日本語 OS でも中日二言語の対応表示が容易になっている。

2.3 アラインメントの基準と状況

対訳コーパスは性質上、原文と訳文の対応が必須条件として要求されているが、基礎版では中国の作品 12 篇と日本の作品 10 篇に対し、次の基準によって人手作業で試験的に対応づけを施した。

a.段落(論理行)を基本単位とする。しかし、見やすくするために、長すぎる段落(日本語 220 字、中国語 176 字超えたもの)に対して分割加工を施す。

b.基本的には原本を元とし、それに対応するように訳文に対して段落の分割と合併の加工を施

す。

c.やむを得ない場合、訳文に対応するように原文に対して段落の分割と合併の加工を施す。

d.分割と合併の加工標識はそれぞれ暫定的に /// と +++ を使用して文中に記入する。

その結果、入力ミスや欠落のないテキストでは100%の対応率が得られた。

2.4 検索ツールの開発と利用

対訳コーパスには、複数の言語を検索しその結果を表示できる検索ツールが不可欠である。コードの違いによる文字化けを解消し、中日両言語の文字を同じウィンドウに対応的に表示できるように、専用の検索ツールを開発しているところである。今のところ、日本語 WINDOWS での対応表示機能はすでに実現し、日本語と中国語の検索条件で対訳付きの用例を抽出できるようになっている①。また、複数文書指定、検索言語指定、ダブルキーワード指定、候補語追加、間隔文字数指定、グループ指定、準正規表現、デリミタ、命中語統計など言語検索に必要な機能が揃うようにした。

そのほか、対応表示機能を要求されない限り、フォント切り替え機能と grep 機能付のエディタでも、基礎版のデータを検索することができるようになっている。つまり、対訳コーパスから単一の中国語または日本語コーパスとしての使い道も考慮してあるのである。

3．問題点と今後の課題

本節では、基礎版で残った問題を含めて、今後3 年間中国社会科学基金助成プロジェクトとして、解決すべき課題について、考えてみたい。

3.1 対訳テキストの追加

基礎版では人手と経費の限度で、文学作品しか収録できなかった。収録文体の比率を上の内容デザイン通りに実現するために、他のジャンルの文章で 700 万字ほど追加する予定であるが、下表を参照して選定していくことが合理的だろうと考えられている。具体的には、まず既存の電子テキストを優先的に利用して、電子化されていないテキストの入力はこれまでのように中国語は外注、日本語はスキャナ、OCR 識別、校正という方法でやる予定である。

表3

文体＼時代		現代（解放後・昭和以来）	近代（口語）（解放前・昭和前）	近・古代（文語）（民国前・明治前）	%
創作文	小説	48(99)	17(101)	5	70(92)
	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
%		63(75)	28(61)	9	100(65)

（ ）内は基礎版で入力済作品の字数パーセント

また、基礎版でまずかったのは、長過ぎる作品も収録されてしまい、言葉の重複が多かったということである。教訓としては、なるべく短い文章をたくさん収録したほうが言語内容として重複が少なくて済むことであろう。

3.2 テキストの外字補足と正確率向上

基礎版のテキストには、外字が今、「■」の記号となっており、作業メモを参照にして外字補足をしていかなければならない。そのために、通用性の良い外字フォントと入力プログラムを探しているところである。また、ミス率は今の 0.1%から0.05%に落としていかないと、公開には適格しないだろうと思う。しかし、外字補足および二次校正には、人手と経費のかかる作業なので、今後、公開 CDROM 版の出版に連動して、出版社の力で解決することを望むものである。

3.3 アラインメントの細密化研究

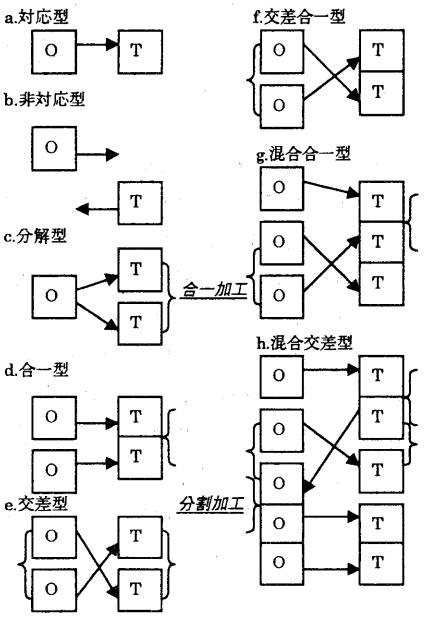
基礎版では、上述 2.3 のように、暫定的な方法で段落を基本単位として対応づけを進めているが、この基礎作業は来年 3 月に完成する予定である。その時、すべての対訳テキストに対する対応的検

① 付録資料をご参照

索が可能となるのである。

しかし、言語研究や機械翻訳への効果的利用を考えれば、さらに、段落から文へ、文から語へと細密化する必要があろう。そのために、対応づけのためのタグセットの研究と自動処理プログラムの開発が重要な課題だと思う。これまでは、字数や語数に基づく方法、辞書と統計を用いた方法②、翻訳の類似性による方法③、漢字対応の利用による方法④など研究されてきたが、いずれも大規模テキストの処理にはそのまま実用することが難しいのではないかと思われる。そこで、本稿では簡易処理と有効処理の考え方を主張したい。そのために、まず、段落対訳の類型を調べ、基礎版で使われた簡易対応付け法を含め、表4に示した。そして、文対訳の類型に関しても、サンプル調査を行い、その結果を表5と表6で示した。これによって、対応づけのポイントをクロースアップしていこうと考えたのである。

表4　段落対訳の類型と簡易対応付け法



② 参考文献[1]
③ 参考文献[2]
④ 参考文献[3]

表5　文対訳の類型

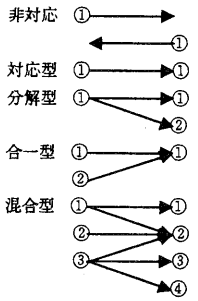


表4と表5で示したように、段落と文の対応状況が類型的に近似性をもっているのである。対応付けの重要な対象は混合型や交差型の類型であることが明らかである。

表6　文対訳状況に関するサンプル調査

	ファイル/段落	原文数	訳文数	対応型	分解型	合一型	混合型
日本語↓中国語	koya/28	1	3		1*3		
	/231	3	4		2*2	1	1
	/589	1	4		1*4		
	kuroiame/80	5	6	4	1*2		
	/810	4	3	2		1	
	/179	3	3	3			
	sunano/297	4	4	4			
	/12	4	5	3	1*2		
	/387	5	5	5			
	yukiguni/36	2	2	2			
	/270	4	4	1	1*2	1	
	/1302	3	5	1	2*2	1	1
	syayo/682	9	9	6	1*2	1	1
	/1259	3	3	3			
	小計	51	60	34	10	5	3
	%	100	117	66	19	9	6
中国語↓日本語	qiwang/66	4	6	1	1*4	1	1
	/75	4	4	4			
	/244	5	6	4	1*2		
	chadui/13	4	6	2	2*2		
	/428	5	5	5			
	/1188	9	9	7	1*2	1	1
	gaiuan2/93	9	8	5	1*2	2	1
	gaiuan1/62	7	10	4	3*2		
	gynr/234	6	10	5	1*5		
	小計	53	64	37	9	4	3
	%	100	120	70	17	7	6
総計		104	124	71	19	9	6
%		100	119	69	18	7	6

また、表6で示した文対応の状況から分かるように、対応付け作業の主な対象である「混合型」は、わずか全体の6%しかないのである。人手作業でも機械作業でもそれを重点に解決していくことができれば、良い効率が得られるのではないかと思われる。

もちろん、この調査結果を問題の解決に結びつけるためには、まだまだ言語学者と言語工学者の共同の努力によらなければならない。

3.4 情報付与基準の研究とその制定

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せて情報付与など加工する必要がある。中日対訳コーパスに関しても、情報付与(タグ付け)はその加工の基本的課題であり、付与する情報は出典情報、言語情報、対応情報と文書情報の四種類である。出典情報は書名、著者、訳者、版元、出版時間などを記載し、テキストの先頭に付与するものだが、言語情報は形態素/品詞、構文、意味など、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である。対応情報は原本と訳本のそれぞれにおける言語情報や文書情報の対応関係を示すものとして、前節で述べた対応づけの細密度に関わる対訳コーパス専用の情報である。文書情報はコーパステキストの多言語的共存と流通を考慮に入れた、文章構造情報やフォント情報などを記載する情報であるが、HTML、SGML、XML など国際通用のマークアップ言語がその記載に使われる。

基礎版に付けた情報は、出典情報と段落対応加工情報だけであったが、今後の目標は形態素/品詞情報 100%、構文情報 10%、意味情報 10%なのである。また、対応情報と文書情報も付けていく予定である。そのために、まず単位切りと情報付与の基準を設定しなければならないが、対訳コーパスの性質と通用性を考えれば、国家基準や国際基準に基づいていくのが理想的である。しかし、これまで調べたところでは、下表のように基準はまだ未定か、若しくは不十分なようである。そこで、なるべくそれに近いような基準を参考にして、対訳コーパスの特性に合う見直しと補足を当面の課題として考案しているところである。

表7

	国際	中国	日本
語の分割		○	
形態素/品詞		△	
構文/係受け		△	
意味		△	

○GB13715　△国家語委，北京大学

UNL　国研　奈良先端大　京大　NTT　GDA

3.5 情報付与プログラムの導入と統合

昔、カード作業でコーパスが作られたのだが、いまやコンピュータ処理による情報付与がないとコーパスとはいえない時代になってきた。大規模テキストの処理効率を図るためにも、自動付与プログラムが不可欠である。幸いに日本でも中国でも言語工学分野の皆様がその開発に励んでいるので、最新のプログラムと成果を導入させていただくことが可能である。ただし、今のところ、WINDOWS で使用可能なプログラムはまだ少ないのである。

そこで、今後とも中国の語言文字応用研究所、北京大学計算語言学研究所、清華大学、日本の国立国語研究所、日立製作所中央研究所、奈良先端科学技術大学院大学松本研究室、通産省電子技術総合研究所などの研究機関の学者·研究者と、協力関係を強化し、その技術と成果を導入して、中日対訳コーパスの情報付与システムに統合していくように、努力を重ねていく考えである。

具体的な付与作業は、機械作業と人手修正を繰り返しその過程を機械に学習させて、だんだんプログラムを改良し人手作業を減らして行く方法が考えられているが、来年5月から実施していくことになると思われる。

3.6 検索ツールの改良と機能拡充

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が言語の調査と分析に不可欠である。基礎版の対訳コーパスでは、対応検索の機能は実現しているが、スピードはまだ、理想的とはいえず、中国語 WINDOWS 版もこれから出さなければならない。また、コーパスの加工度が深まり、大量処理が必要になるに連れて、たえずその対応と改良を図らなければならず、応用研究からの要求に応じて、機能を拡充していくことも必要であろう。今後、大事な課題は検索のスピードアップと付与情報への対応であるといえよう。

中日対訳コーパスの検索ツールは北京日本学研究センターと日立製作所中央研究所との協力提携

による開発体制を確保している。これまでいいチームワークによって、目にみえる成果が上がってきた。今後とも協力関係を強化し、優れた検索ツールを開発していく考えである。そのような協力と努力は多言語処理技術の発展と他の対訳コーパスの開発へも寄与するようなことになれば、まさに望外の喜びである

4．試用状況と計画

実用的で良い対訳コーパスを作成するためには、基礎版の段階から試用をしながら、改良と発展を図らなければならないと思う。

そこで、今は試験的に基礎版のコーパスを使って断定保留表現のモダリティ形式のうち、代表的といわれる日本語の「だろう(でしょう)」と中国語の「吧」を中心に、その用法と対訳表現を調べているが、対応付きのあるテキストからまず下記の用例統計を求めようと思っている。

だろう

	原本	訳本
吧1	168	
吧2	115	
呢	86	
吗	101	
啊(呀)	33	
其他1	37	
φ	164	
大概	66	
可能	21	
会	56	
也许	54	
恐怕	32	
说不定	7	
一定	8	
难道	26	
难怪	1	
是否	8	
其他2	51	
計	763	795

吧

	原本	訳本
推測	80	
確認要求	40	
問いかけ	32	
意志	140	
誘いかけ	93	
勧告	52	
依頼	58	
命令	116	
許可	18	
祈願	4	
呪詛	19	
仮定	21	
間投	3	
呼掛け	13	
計	692	1174

「だろう(でしょう)」については、これまで認識的モダリティ形式として、「推量」「判断保留」という意味的側面が深く研究されてきた(奥田 1985、仁田 1991、益岡 1991、森山 1992、鄭 1994、宮崎 1995 など)。最近、その伝達的側面の性質も留意されて、「表出」という特徴が指摘される(安達 1997)など、研究は活発な様相を呈しているといえよう。　一方、対照研究では、中国語の「吧」に対応するという説が普通であるが、しかし、中日対訳コーパスで調べたところによれば、「だろう」の対訳部には「吧」の不対応例が多いようである(上表と付録資料)。対照研究の必要性が感じられるところである。

「吧」については、判断保留表現だけではなく働きかけ系の表現などにも広く使われることは周知の通りであるが、文末モダリティ形式としてその統一的説明と諸用法の共起関係の究明が大事な課題だと思う。対訳コーパスから得られた表現例から傍証的に「だろう」の意味機能を確認するとと

もに、「吧」の意味機能とその周辺の「呢」「吗」「啊(呀)」や副詞表現「大概」「可能」「会」などとの相違を考察しているところである。

対訳コーパスによる対照研究の最大な利点は、文脈付の対照情報を大量かつ高速に得られることにあるだろうと思う。そのために、良い訳文を選び適切な言語情報(タグ)をつけて、信頼度の高いコーパスを作ることは重要であるが、大量の「生」の対照情報を目の前にして、いかに迷わずに研究に活用していくかということも重要であろう。そのために、対訳コーパスで出来ることと出来ないこと、分かりやすいことと紛れやすいことを模索していくのも今後の課題だと思い、出来れば、適宜なモニターを増やし、随時フィードバックでその意見と状況を把握していきたいものである。

5．終わりに

本稿では中国で実施中の共同研究プロジェクト「中日対訳コーパスの構築と応用研究」について、進捗状況と今後の課題を報告した。その中に、特にアラインメントの簡易化と効率化、情報付与基準の標準化、中日両国語の対応表示と検索、言語対照研究への応用などの問題を取り上げ、調査報告と提言を行った。

よい対訳コーパスを作り上げるためには、関係国の言語学者と情報工学者の学術交流と共同研究が不可欠である。この認識に基づいて、今後とも研究者と専門家の意見を聞くために、広く本プロジェクトの進捗状況を報告していく考えである。

謝辞

関係情報を提供して頂いた国立国語研究所加藤安彦研究室長、柏野和佳子研究員、山崎誠研究室長、奈良先端技術大学院大学松本祐治教授、通産省電子技術総合研究所橋田浩一主任研究官のご好意に謝意を表す。なお、本稿は国立国語研究所外国人研究員招聘期間においてまとめられたものである。

参考文献

[1] 春野雅彦 山崎毅文「辞書と統計を用いた対訳アライメント」情報処理学会研究報告 96-NL-112

[2] 黄道三 長尾真「類似性に基づいた日韓対訳テキストの文対応」情報処理学会研究報告 94-NL-99

[3] 陳樹霖 長尾真「漢字対応の利用による日中対訳テキストの文対応付け」情報処理学会研究報告 94-NL-128

[4] 奥田靖雄 1984·85「おしはかり(1)(2)」『日本語学』3-12, 4-2.

[5] 仁田義雄 1991『日本語のモダリティと人称』ひつじ書房

[6] 益岡隆志 1991『モダリティの文法』くろしお出版

[7] 森山卓郎 1992「日本語における「推量」をめぐって」『言語研究』101

[8] 宮崎和人 1995「「だろう」をめぐって」『広島修大論集』35-2

[9] 鄭相哲 1994「推し量りのメカニズム」『阪大日本語研究』

[10] 安達太郎 1997「「だろう」の伝達的な側面」『日本語教育』95

「中日対訳コーパスの作成状況と今後の課題」付録資料

kuroiame.txt	K:でしょう	74	「みなさんは、おうちのことが御心配でしょう。もし御希望でしたら、今から広島市へ送り届けて差上げます。私の家内は子供の安否が心配で、今すぐ広島へ帰りたがっています」	///“大家可能担心自己的家里吧。如果你们愿意的话，我现在就送大家回广岛市去。我妻子惦着小孩的安全，很想现在马上回广岛去。”
etizen.txt	K:だろう	970	と玉枝はいった。その忠平の世話してくれる宿へいって、なんとしても忠平に躯のことをうちあけねばならない。この男はどんな顔をするだろう。玉枝は、早く宿へ行って忠平を待っていようと思った。	///玉枝回答。+++ 玉枝想，到了忠平替自己找的旅店后，无论如何也要把受孕的事向他说清楚。那时，这个男人将会有一副什么表情呢！玉枝真想早点上旅店去等忠平。
syayo.txt	K:でしょう	708	///まじめに、+++「その必要は、ございませんでしょう。おかぜでございますから、しずかにしていらっしゃると、間もなくおかぜが抜けますでしょう。」+++とおっしゃった。	“大概无此必要。因为是上感，只需静养当可痊愈，”医生一本正经地回答道。
tyshcq.txt	K:でしょう	94	///『今日は冬至でしょう』彼は一言言い添えたわ『我々は身内の墓参りをしに行くのです』	“‘今天是冬至节。’他解释了一下，‘我们是看亲人坟墓去的。’”

xbz.txt	K:吧	282	“拾来，你过年就十八了吧！”	「拾来、おまえは年が明ければ十八だね」。
tyshcq.txt	K:吧	80	我没有吱声，只用眼色示意她再讲下去，我心里忽然隐隐感到一阵不安，我莫名其妙地意识到，这个姑娘大雪天跑来跟我讲这些，可能和我的某一段生活有关吧。	私は口をつぐみ、つづけて話してくれるよう目で伝えた。この時ふいに胸の中にかすかな不安が生まれたのだ、まだはっきりと意識したわけではないが、このひとがこんな大雪の日にやってきて私に聞かせようという話は、私自身の過去と何か関係があるのかもしれない、という……。
jia2.txt	K:罢	288	“该不会又有巷战罢，”瑞珏惊惧地说。	「まさか市街戦じゃないでしょうね」瑞■は驚いていう。
jia2.txt	K:罢	282	“你自己看罢，”觉英得意地说着，就把手里捏的一张纸递过去。	///「自分で見てごらんよ」覚英は得意になって、手につかんだ紙片を渡す。
gaiguan2.txt	K:吧	76	“你干什么吃的！……我不要了！不吃了！……这叫炸糕吗？成元宵了！……”辛小亮气得喊起来，“退你吧，我不要了！”	「こんなものを食わせる気か、もういらないよ。食いたくもない！これでも揚げ饅頭と言えるのかね？まるで元宵［一月十五日の元宵節に食べる米の粉で作った餡入り団子。茄で汁といっしょに食べる］じゃないか！］+++辛小亮はわめきたてた。+++「これ返すよ、もういらないよ」
gaiguan2.txt	K:吧	56	“熊？我出气了！临走，趁屋里没人，顺手把身边的暖气给他关了！把旋钮摘下来，出门又扔回他家报箱了！别看你是煤炭部的，冻一宿吧！……”	「フニヤフニヤなもんか、俺はうっぷん晴らしをしたさ。帰りがけ、部屋に誰もいないすきにスチームをとめてやったよ。ネジをはずして、門を出てから郵便受けの中に投げこんどいた。石炭部のお役人であろうと一晩中凍えるがいいさ！」