多言語コーパスと日本語研究 ： 「中日対訳コーパス」の利用研究例から

著者	曹 大峰
雑誌名	日本語科学
巻	22
ページ	59-77
発行年	2007-10-25
URL	http://doi.org/10.15084/00002183

多言語コーパスと日本語研究
——「中日対訳コーパス」の利用研究例から——

曹　大峰
(北京日本学研究センター)

キーワード
多言語コーパス，対訳コーパス，利用モデル，日本語研究

要　旨
多言語コーパスに焦点を絞って，まずこれまで多言語コーパスを分類するための基準が不足していたことを指摘する。さらに，多言語コーパスというものにおいては異なる言語がさまざまな関係によって関連付けられていることを示し，その関係を分類するための基準を提案する。その上で，多言語コーパスをどのように選定し，使い分けるべきかについての目安を示す。
また，「中日対訳コーパス」の作成と利用経験を踏まえて，訳文データの特性に気付かず原語と対等に使うなどの利用上の問題点を指摘したうえ，筆者が提示した利用モデルを説明し，「可能だ」という可能表現，終助詞「だろう」の意味用法，日中同形語である「基本」の意味用法などに関する日中対照研究の事例を通して，対訳コーパスを適正に利用する方法とその効果を示す。

1. はじめに

最近，コンピュータ技術の飛躍的発展により，コーパスの開発は書き言葉コーパスをはじめ話し言葉コーパスや多言語コーパス，マルチメディアコーパスなど多様化を呈して進んでいる。一方，利用者層が広がりつつある中で，コーパスの標準化や利用法の適切性などコーパスの品質と利用スキルの向上も求められるようになった。そこで，各種コーパスの特性と目的別の使い分けや，利用可能性と限界性の研究など利用に関する新しい課題が重要視されてきた。

本稿では多言語コーパスに焦点を絞ってその種類と特徴を分析し，筆者がかかわっていた「中日対訳コーパス」の利用研究例を踏まえて，日本語研究への利用モデルと可能性を考えてみたい。

2. 多言語コーパスの種類と特徴

多言語コーパス(multilingual corpora)は，複数言語のテキストデータを含むコーパスとして，これまでに，いろいろな種類のものが挙げられているが，その詳しい分類はまだ見当たらず，一般には次の二種類に分けられることが多い。

a．並列コーパス（parallel corpus パラレルコーパス）

複数の言語が意味の同一性と一定の単位で並列に対応付けされたコーパス。

その典型的なものは，元のテキストと翻訳されたテキストが文単位で対応付けされた「対訳コーパス」であるが，元のテキストとそのまとめが対応付けされた「要約コーパス」や，言い換え関係にあるテキストが対応付けされた「換言コーパス」もある。

b．類似コーパス

複数の言語が同じフレームとバランスで集積されたコーパス。

意味類似性のフレームで構築されたものはコンパラブルコーパス(comparable corpus)，語族近縁性のフレームで構築されたものはコンパラティブコーパス(comparative corpus)と呼んで分けられることがある。

さらに言語変種を考えれば，国家語や民族語による多言語の他に，時代が異なる現代語と古代語，地域が異なる方言と共通語，習得順序が異なる母語と第二言語など，言語とその変種間で構築されるコーパスもある。それらも視野に含めれば，多言語コーパスの種類は実にさまざまであり，上記の二分類には収まらなくなるであろう。

以下では，日本語および英語，中国語に関する多種類の多言語コーパスを挙げてみる。

①「日英対応付けコーパス」 情報通信研究機構自然言語グループ作成・オンライン公開，作品や新聞記事などの英語と日本語の原文と対訳347,234件収録。

②「中日対訳コーパス」北京日本学研究センター作成，2002年完成・限定公開，中国と日本の多ジャンル文章の原文と対訳157件収録。

③「「西京雑記」対訳コーパス」日本大学作成，2003年完成・限定公開，古典原文・読下し記号付きの原文・現代中国語訳・現代日本語訳を収録。

④「全国方言談話データベース」 国立国語研究所作成，『日本のふるさとことば集成』(CD-ROM・CD・書籍全20巻）として公開，共通語訳付。

⑤「BTS多言語話し言葉コーパス—日本語会話」 東京外国語大学作成，2005年完成・公開，日本語母語話者と学習者の自然会話154件収録。

⑥「日本語学習者による日本語作文と，その母語訳との対訳データベース」 国立国語研究所作成，2001年完成・公開，アジア10ヶ国の学習者約1,100名による日本語作文とその母語訳収録。

⑦「ICEコーパス」 ロンドン大学作成・公開，The International Corpus of English，英語を母語または第二言語とする18の国・地域の1989年以降の英語各100万語ずつ（1990-1994年の話し言葉（60%）と書き言葉（40%）のテキスト）収録。

⑧「LIVAC中国語共時コーパス」 香港城市大学作成・オンライン公開，Linguistics Variation in Chinese Speech Communities，香港・台湾・北京・上海・アモイ・シンガポールの代表的中国語新聞や電子メディア上のニュースを材料に継続収集。

⑨「中国語換言コーパス」ATR音声言語コミュニケーション研究所作成，旅行会話の中国語換言コーパス，２万文の原文と４万文の換言文からなる。

表１　多言語コーパスの分類基準に関する試案（○主要特性　△副次特性）

種類＼関係特性	対応		意味		語族		時代		地域		習得	
	並列	包括	同一	類似	同属	類縁	同代	異代	内域	外域	前後	内外
日英対応付けコーパス	○		○	△			○					
中日対訳コーパス	○		○	△			○	△				
「西京雑記」対訳コーパス	○		○				○	○				
全国方言談話データベース	△		△	○	○		○		○			
BTS 話し言葉コーパス				△	○	○	○		△	△	△	○
日本語学習者による日本語作文と，その母語訳との対訳データベース	△		○				○				△	○
ICE コーパス				○		○	○			○		
LIVAC 中国語共時コーパス				○	○		○		○			
中国語換言コーパス	△	△	○		○		○					

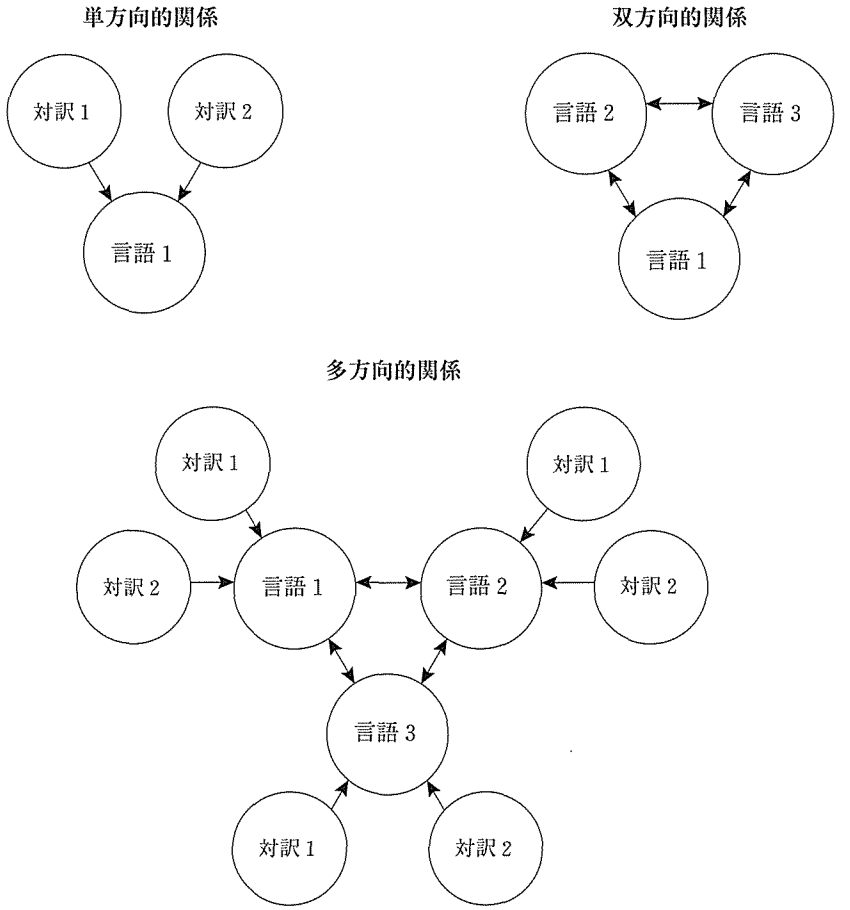

図１　多言語コーパスにおける複数言語の方向性

そこで，①～⑨の多言語コーパスの特徴を考えてみよう。各コーパスを「対応」「意味」「語族」「時代」「地域」「習得」という特性によって分類すると，表１のようになる。

また，多言語コーパスは複数言語のテキストデータが一定の関係でコーパスに入っており，その関係には図１のようにいつも一定の方向性があることが大きな特徴といえよう。

たとえば，言語１とその複数言語の対訳で構築されたコーパスでは対訳が元の言語から独立した言語とはいえず，いつも意味的・文化的に元の言語を指向しているので，単方向的関係といって区別する。また，言語１と言語２や言語３で構築されたコーパスは対等的な関係を成し，互いに独立した言語であり，意味的・文化的に類似性があっても並列性がないので，双方向的関係といえるだろう。さらに，複数の言語でそれぞれの対訳をもって互いに対等的に構築されたコーパスは多方向的関係ということになる。

このように，多言語コーパスはいろいろな種類のものがあり，研究に利用する場合，まず上述の分析基準と相関関係に基づいてコーパスの種類と特性を把握して，自分の目的に合うように選定と使い分けをする必要があるのではないかと思われる。

次章では，「中日対訳コーパス」の開発と利用研究を事例に，多言語コーパスの利用モデルについて具体的に検討していこう。

3.「中日対訳コーパス」と多言語コーパスの利用モデル

3.1.「中日対訳コーパス」の概要

「中日対訳コーパス」は中日両言語双方向並列型の対訳コーパスとして開発されたものである。同コーパスは言語・文学・翻訳など幅広い研究領域に資することを考慮し，表２に示す内容構成と表３に示すジャンルと文字数により，世界的に見ても大規模な並列コーパスとして構築された。また，ユニコードフォントによる並列的表示，多様な検索条件で日本語と中国語の対訳付きの用例抽出，出典・対応・品詞・構文の情報付与など，並列コーパスとしての必須機能が装備されている。

表２　「中日対訳コーパス」の内容構成

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
汎用型	特殊型	
タグ有り	タグなし	

表3 「中日対訳コーパス」のジャンルと字数（単位：万字）

	現代		近代		計（%）
	中	日	中	日	
小説	597.7	305.5	95.8	131.4	1130.4（58.0）
詩歌／散文	11.2	21.4			32.6（2.0）
伝記	256.6	61.4			318（17.0）
政論／白書	329.2	119.4			448.6（22.9）
法律／条約	0.55	1.85			2.4（0.1）
計（%）	1195.25（62）	509.55（26）	95.8（5）	131.4（7）	1932（100）

同コーパスは，日本語や中国語の対照研究に利用できるばかりではなく，技術的には多言語多方向並列型コーパスへと発展できるように開発されたものであるが，研究チームは開発当初からずっと対訳コーパスの利用研究に関心を持って実践的試みをしてきた[1]。その中で，筆者が心がけていたのは利用モデルの分類とその実証研究であった。

多言語コーパスは，二種以上の言語データを研究に利用できるのが大きな特徴である。しかし，上述のように，コーパスにおける複数言語のテキストデータが相互に多様な関係を成しており，それを明らかに認識しなければうまく利用できない。さらに，並列型対訳コーパスの場合は，「原文」と「訳文」間にある関係を正確に把握することが必要になる。

ここで，中国語と日本語の並列型対訳コーパスにおいて，推量文に使われる日本語の助動詞「らしい」と中国語の副詞「好象」を例に，どのような相対関係があるのかを分析してみる。まず，「らしい」には，日本語の原文データ中に出現するもの（以降，このようなものを「原語」と呼ぶ）と，中国語を日本語に翻訳した訳文データ中に出現するもの（以降，このようなものを「訳語」と呼ぶ）とがある。同様に，「好象」にも，中国語の原文と訳文それぞれに出現する原語と訳語の使用がある。

本稿では，まず最初に，それぞれの原文データ中に出現する原語レベルの対応関係を「原語間の対等的関係」と名付けることとする。以下の例文の①，②の場合である。

次に，対訳によって直接対応づけられた文や文章レベルの対応関係を「原文と訳文間の照応的関係」と名付ける。以下の例文の③と④，また，⑤と⑥の場合である。

最後に，日本語原文データと日本語訳文データ中に現れる語レベルの対応関係，あるいは，中国語原文データと中国語訳文データ中に現れる語レベルの対応関係を「原語と訳語間の参照的関係」と名付ける。以下の例文の⑦と⑧，また，⑨と⑩の場合である。

例：①（原語）雨が降っているらしい。
②（原語）好像他也去。
} 対等的（語レベル）

③（原文）雨が降っているらしい。
④（訳文）好像正在下雨。
} 照応的（文・文章レベル）

⑤（原文）好像他也去。
⑥（訳文）彼も行くらしい。　} 照応的（文・文章レベル）

⑦（原語）雨が降っているらしい。
⑧（訳語）彼も行くらしい。　} 参照的（語レベル）

⑨（原語）好像他也去。
⑩（訳語）好像正在下雨。　} 参照的（語レベル）

以上の関係を図示すると，次の図２のようになる。

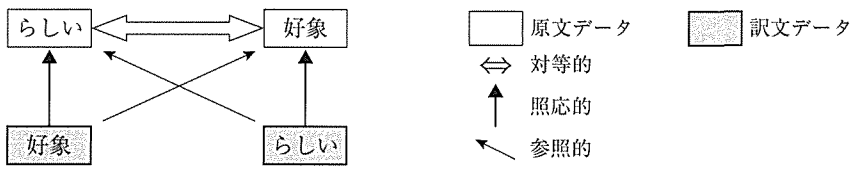

図２　「らしい」と「好象」の相対関係

つまり，対訳コーパスにおいては，二言語間を対比させる時の，原語と原語にある対等的関係，二言語間の原文と訳文の照応的関係に加え，さらに，同一言語を軸にして見るときの，片方の原語と，もう片方の訳語にある参照的関係の，あわせて三種類の関係が存在すると考えられる。

3.2. 多言語コーパスの利用モデル

そこで，どんな研究にどんな種類のデータをどのように利用するかという問題が出てくる。研究目的に適するようにコーパスの使い分けとデータの取り方を工夫しなければ，多言語という利

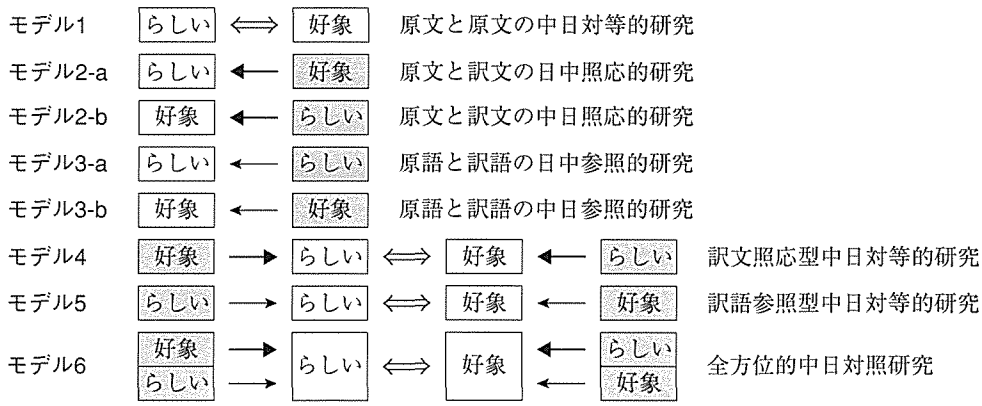

図３　二言語データの利用モデル

点を生かすことができないばかりか，逆効果を招いてしまう恐れがあるかもしれない。このような失敗を防ぐためには，上節の二言語データの相対関係に基づいてその利用モデルを，図３のように定める。

まず，モデル１は複数言語の原文を対象に研究するアプローチである。その特徴は，対象となる原文と原文との関係が，話題は同じでも内容的には必ずしも対応しないということである。たとえば，同一事件に関する複数の言語による報道記事，また同一言語行為における複数言語の異なる表現などである。このモデルは異なる言語の言語表現や言語習慣の違いを知り，発話者の視点や認知スキーマなど知的文化的背景を探索するのに有効であろう。その研究例として，曹・森山(1999)は中日両言語の感動詞を原文コーパスから抽出して，その音声特徴から感情・認知・行動を表現する機能の面で両言語の異同を考察したものであるが，これまで一般言語学，民族言語学，言語類型論，社会言語学などの方法による実証的対照研究は殆どモデル１のような対等的研究であった。このような研究は単言語コーパスでも複数あれば効果的に利用出来るのであるが，その場合，母語でない原文に対する理解が大きな鍵となるので，複数の言語に堪能でない場合，研究者個人で利用しにくい。そこで，対訳を含む多言語コーパスを利用すれば，モデル4，５または６のような複合的視点による研究ができるので，新しい可能性が生まれるかもしれない。

モデル２は訳文との照応で，原文を研究するアプローチである。その特徴は，原文が研究対象，訳文が照応対象であり，利用目的が訳文との照応で原文に関する探索を深めること，また，原文と訳文の並列的対応が文や文章レベルまであり，構文だけでなく場面や文脈情報による対照研究が期待できることである。このように，訳文照応型の研究は原文のみの研究で気づかれない問題を見つけることができるので，最近，対照言語学的方法による一言語または複数言語の研究でも注目されつつあるが，そのような研究は対訳コーパスでモデル２（2-ａ，2-ｂ）を利用すれば，効率的に成果があがることであろう。この場合，上述のように訳文を照応的対象として利用することがポイントであるが，しかし，現実ではそこまで配慮が行かず，つい訳文を原文と同じように対等的に使ったり，研究者個人の「対訳」でその「原語」に関する結論を下したりするような論述があるのではないかと思う。そのような使い方では，客観性と信頼性のある結果が得られない。したがって，対訳コーパスを利用する場合，モデル２でその特徴を明確にする必要性があろう。

モデル３は原語の用例と訳語の用例を対照して訳文を研究するアプローチである。訳語の研究は，これまで翻訳研究ではよく見られるが，一般には訳文とその原点にある原文を比べて「意味転換」のメカニズムや特徴を考察するものが多い。モデル３の場合，訳語とその目標にある言語の原語と比べて訳文の特徴を見るためのアプローチであるので，やや特殊である。たとえば，『雪国』の中国語訳を本場の中国語と対照し（モデル3-ｂ），『ハムレット』の日本語訳を本場の日本語と対照する（モデル3-ａ）ことによって，その訳文としての言語的特徴を見出すことである。訳語は原文の対訳として原語の語調やニュアンスを強く残している一方，訳者の知的生産物としてその言語能力や文化背景をも反映する，一種の「中間言語(interlanguage)」である。多言語コーパスはそのような中間言語の資源を原語付で大量に提供してくれる。言語研究や教育の

立場で考え，翻訳や通訳という多言語情報伝達に不可欠な言語能力や言語活動を研究の射程におく場合，モデル３による訳文研究を試みる必要があろう。また，自然言語処理の分野でも，最近，大量の対訳データによる機械翻訳の革新的進歩が期待されるようになり，対訳パターンを的確に抽出するために，訳文の研究が重要視されるようになったが，そこで，やはりこのアプローチの可能性と有効性を検討する必要もあろう。

モデル4，５，6は上述のモデルを複合的に利用して研究をさらに広げて深めていくためのアプローチである。モデル４はモデル１にモデル２，モデル５はモデル１にモデル３を組み入れたものであり，モデル６はモデル１にモデル２と３を組み合わせたものである。このような組み合わせにより，単一モデルで特定された研究の一側面を総合的に捉えることができ，一言語をメインとする対照研究を双方向的に捉え多言語を対等的に対照する研究ができるようになり，複雑な言語現象を全方位的に考察するのに効果があると思われる。つまり，複数言語の原文を研究の対象に，訳文照応と訳語参照の多側面からその異同を探索するとともに，その訳文に関する研究も期待できるということである。ただし，利用するデータは多層に相関するので，その関係と利用目的をはっきり認識していなければ，全方位的研究は捗らないばかりか，分析を混乱させてしまう恐れもあろう。このような利用モデルは高度な利用能力を要するものである。ただ，基本的には単一モデルがベースとなっている複合型なので，単一モデルから利用の経験を蓄積していけば，自然になれてきて使いこなすようになる。そうなれば，さらに多くの可能性やアイディアが見えてくるであろう。

次章では，日本語の研究で対訳コーパスを利用する場合，特に適する上述モデルの２-a，４，6について，筆者自身の研究事例を紹介し，その可能性と注意点を考えてみよう。

4. 日本語研究への利用例

4.1. 原文と訳文の日中照応的研究

モデル2- a は訳文の中国語との照応で，日本語の原文を研究するアプローチである。その特徴は，日本語の原文が研究対象，訳文の中国語が照応対象であり，利用目的が訳文との照応で日本語に関する認識を深めること，また，原文と訳文の並列は文や文章レベルまであり，構文だけでなく場面や文脈情報による対照研究が期待できることである。

このモデルによる研究例として，最近「可能だ」という可能表現に関する筆者の試みがある。可能表現の研究では，表現形式の意味特定が大きな難点である。たとえば，申(2003)によると，「字を書くことができない」という可能文は次のように多種の意味を読み取れるという。

①（習ったことがなく，文字を知らないから）字を書くことができない。（能力）
②（筆記用具を持っていないから）字を書くことができない。（客観的状況）
③（電気などつかない暗闇の中にいるから）字を書くことができない。（客観的状況）
④（時間の余裕がなく忙しいから）字を書くことができない。（主観的状況）

⑤（病気やケガが原因で書けないから）字を書くことができない。（主観的状況）
⑥（体調が優れなく気分などが悪いから）字を書くことができない。（主観的状況）

この現象に関して，これまでの研究では語・句・文・談話の構成分析による意味特定が多いようであるが，まだ理想的な解決が得られたとはいえない。「可能だ」も，「できる」などの日本語可能表現に類似するところがあるが，これまではその用法と意味に関する記述がほとんどない。そこで，中国語にも同形の表現形式があることから，対訳コーパスを利用して観察したところ，次のように，複数の用法と対訳の用例が多数現れてきた。

(1)「ボクは，自分の足で階段を上ることが可能ですからエレベーターは不要ですし，トイレも車椅子用でなくてけっこうです」と言ってみたが，（乙武広匡『五体不満足』講談社）
対訳：我向他们讲我可以（会／能／能够／？可能）自己上下楼，楼内有没有电梯没有关系，（鄭顕訳　山東文芸出版社）

(2) 日本列島を現在よりももっと豊かで，公害がすくなく，住みやすい国土に改造することは可能である。（田中角栄『日本列島改造論』日刊工業出版社）
対訳：将日本列岛改造成为比現在更为富裕，公害不多的安居乐土是可能（？可以／？会／＊能／＊能够）的。（秦新訳　商務印書館）

例(1)は現在の動作主の能力や状況を相手に伝える表現であるが，例(2)は動作主も相手も一般化され，未来の出来事の実現可能性に関する判断を述べる表現である。対訳の状況をみても，例(1)では「可以」「会」「能」「能够」などの中国語可能表現に対応するが，日本語と同形の「可能」という表現には対応しない。一方，例(2)では日本語と同形の「可能」に対応するが，「可以」「会」「能」「能够」などの表現には対応しない。これで，「可能だ」の意味用法について対訳データを利用して細かく記述する必要性と可能性を感じたので，まず表4のように意味の成立要

表4　「可能だ」の意味の成立要素に基づく分類法

意味要素 / 分類	内的能力・性能 ⟵				外的条件 ⟶
	性能	生(習)得	心身	状況(主)	状況(客)
1－1 能力可能①	○				
1－2 能力可能②		△	○		
2－1 状況可能①			△	○	
2－2 状況可能②			△	△	○
3－1 実現可能①	△	△	△	△	△
3－2 実現可能②	△	△	△	△	△

（○＝明確な場合　△＝渾然とした場合）

素に着眼する分類法を考案した。

この分類法は可能表現の意味に関する先行研究を参考にして，能力・状況・実現という三つのレベルを設定し，内的能力から外的状況まで含む可能成立の要素を基準に「可能だ」の用法を分類するものである。具体的には，能力レベルでは主に可能性の持ち主の内部的属性に注目し，①は客体の性能や特長，②は動作主自身の能力や条件と分けるが，状況レベルでは主に可能性の持ち主を取り巻く外的状況や条件から，①は動作主直接関与の状況，②は動作主周囲の状況と分ける。実現レベルでは主に実現可能性の述べ方に注目し，①は可能性の真偽判断，②は可能性の程度判断と分ける。また，諸要素については，明確に前面に現れた場合（○）と渾然と裏面に隠れた場合（△）に分けるようにしたが，実現レベルは個々の要素を根拠（裏付け）に判断するという性質で，能力や状況レベルと対照的であった。

それによってコーパスでヒットした用例を振り分けて，表５のように意味用法と訳語との関係を観察し整理してみた。

表５　「可能だ」の用法分布と対訳の関係（数字はヒット件数の振り分け）

訳語＼用法	１－１	１－２	２－１	２－２	３－１	３－２	計
会		1					1
可以	5	3	5	7			20
容许				1			1
能／能够／能行	1/ /			4/2/1	1/1/		10
无法／没法				1/	5/1		7
可能					23		23
有可能					5		5
是可能的					28		28
是不行的				1			1
不妨			1				1
得到／得以／得成				1/ /	/1/1		3
不到					3		3
不了				1			1
不过					1		1
不管用				1			1
很难						1	1
容易						1	1
略訳					1		1
計	6	4	6	20	71	2	109

表5の結果に見られるように，「可能だ」の意味用法は「3-1実現可能性の真偽判断」に多く見られ，そのプロトタイプ的な存在が示されているが，その特徴は対訳状況にもはっきりと出ており，前述の例(2)のように，「可能」系の訳語としか対応しないのである。また，前述の例(1)のように，動作主の心身状態と関与状況など現存要素を明確に含んだ用例では，「可能」系の訳語が対応できなくなり，他の複数の対訳で意味の広がりが示されている。

このように，意味の成立要素に着眼した分類法と訳語との対応関係を観察することによって，「可能だ」の表現機能について次の仮説を立てることが可能ではないかと思われる。

○基本機能：事態の実現可能性を確定的に述べ立てる。
○二次機能：現存の能力や実現条件などを含んで述べ立てる場合と未来事態の実現可能性を述べ立てる場合がある。前者は他の可能表現に接し，後者は可能性判断の「あり得る」「かもしれない」に接するが，確定的か未確定的かで「かもしれない」と区別される。

この仮説で考えれば，次のように構文条件だけでは意味の特定が難しい日本語の用例についても，二次機能を示す典型例として説明できるであろう。

(3) 子どもの生活をなにもかも支配しようとしているお母さんに育てられているひとりっ子は自分で考え，ひとりで動くことは不可能です。(中澤次郎・鈴木芳正『ひとりっ子の上手な育て方』産心社)

意味：①自分で考え動く能力が現にない
②母親の支配によって現に許されない
③自分で考え動くようにしたくても，させたくても，成す術がない
④総合的に判断して，実現可能性がない。実現するはずはない。

(4) 有的母亲支配了孩子的整个生活。在这种母亲培养下成长起来的独生子女是不会（不能／不可以／无法／不可能）自己思考，自己行动的。(何明訳　中国国際文化出版公司)

対訳候補：不会→①，④
不能，不可以→②
无法→③
不可能→④

例(3)は文脈によって①～④の意味が読み取れるような例であるが，その複数の意味合いは(4)の対訳においてある程度区別されるようになっているのである。

上述のように，日本語だけでは見えてこない用法や含意が，対訳に見えてくることがあり，そのような現象を利用して原文の研究で認識し難い問題を見つけることができるのである。最近，この種の対照研究が重要視されるようになったが，しかし，研究に使われる訳文は研究者自身の訳によるものが多く，主観性と文脈離脱を免れないという問題が指摘されている。それに対して，コーパスからの文脈付きの対訳データによるアプローチは，質的研究と量的研究で客観性と

信頼性を高めることができよう。

4.2. 訳文照応型中日対等的研究

モデル４は日本語と中国語の原文を対象に，それぞれの訳文と照応して対等に研究するアプローチである。このモデルはモデル1，２の長所を総合し，短所を補おうとするものである。その特徴は，原文間の対等的対照と原文訳文間の照応的対照による複眼的研究が期待でき，両言語の異同を質的・量的に見出すための理想的アプローチであるが，ただ，それには並列データを大量に必要とし，また並列データの活用能力と複眼的観察能力が求められるものである。

曹(2000)は，このモデルの研究事例として位置付けられよう。日本語の助動詞「だろう」と中国語の文末助詞「吧」を対等的に対照し，それぞれの対訳を照応することによって考察を試みたものである。「だろう」と「吧」はいずれもモーダルな文末形式として機能し，これまでの対照研究ではその対応的用法が注目されてきた。しかし，対訳コーパスから両言語の「だろう」と「吧」の原文と訳文を抽出してみたところ，非対応用法も目立って現れてきた。そこで，それぞれの意味用法と対訳状況の関係を調べ，次のような結果が得られた。

表６　「だろう」の用法と対訳[2]

訳語＼用法		焦点推測	非焦点推測	確認要求	事実認識要求	眼前認識要求	中間用法	計	%
文末助詞	吧	2	135	16	8	1	4	166	25
	吧？	1	18	27	19	1	18	84	12
	呢	47	8	0	0	0	6	61	9
	吗	1	2	7	38	1	5	54	8
	啊／呀	8	5	1	2	0	4	20	3
	その他	5	8	1	1	0	0	15	2
	ϕ	20	92	0	23	0	7	142	21
副詞	大概	0	25	0	0	0	0	25	4
	可能	0	12	0	0	0	0	12	2
	会	3	17	1	1	0	0	22	3
	也许	0	22	0	0	0	0	22	3
	恐怕	0	12	0	0	0	0	12	2
	说不定	0	5	0	0	0	0	5	1
	一定	0	3	0	0	0	0	3	0.4
	难道	1	1	0	0	0	0	2	0.2
	是否	1	1	1	0	0	0	3	0.4
	その他	4	16	0	4	0	0	24	4
計		93	382	54	96	3	44	672	100

まず，「だろう」は表６に見られるように，「吧」の対訳率が37%しかなく，中国語文末助詞「呢」に多訳される用法（例５）と「吧」に多訳される用法（例６，例７），「不…吗」に多訳される用法（例８）と「吧」の訳があってもなくてもいい用法（例９）があることが観察された。特に「呢」に多訳される用法には，疑問詞の高い出現率（ほぼ100%）という共起現象が構文的に確認された。それによって「未確定」[3]という「だろう」の基本的意味が形式的にも実証され，「だろう」の意味分類に「焦点推測」という一類を加える可能性が認められたのである。

(5) あいつ，今ごろ，何をしているだろう？（安部公房『砂の女』新潮社）［焦点推測］
家里的“那一位”，现在正干什么呢？（楊応辰訳　珠海出版社）

(6) 女は答えない。答える必要がないほど，分りきったことだったのだろう。逃げられなかったから，逃げなかった……おそらく，それだけのことなのだ。（『砂の女』）［非焦点推測］
女人没有回答。也许她觉得这是个无需回答的问题吧。因为逃不了，所以没有逃走。……恐怕就这么简单。（楊応辰訳　珠海人民出版社）

(7)「これだろう，お兄ちゃん」（井伏鱒二『黒い雨』新潮社）［確認要求］
“是这个吧？哥哥。”（柯毅文等訳　湖南人民出版社）

(8)「だって君の家，病人があるんだろう。」（川端康成『雪国』三笠書房）［事実認識要求］
“可是，你家里不是有病人吗？”（葉謂渠訳　訳林出版社）

(9)「ほら，あすこにあの，ピンク色の洋服を着たお嬢さんと一緒に踊っているでしょう，あれがまアちゃんよ」（谷崎潤一郎『痴人の愛』新潮社）［眼前認識要求］
“你看，那边有个人在和一位穿粉红色洋装的小姐跳舞（吧／φ）。他就是阿熊啊。”（郭来舜訳　陕西人民出版社）

また，「吧」は表７のように用法分布が広く，「だろう」の対訳率が6.2%過ぎず，特に意志文（例10），働きかけ文（例11）と軽い問い掛け（例12）の文では対訳が見られなかった。

(10)“唉，还是睡吧，”鸣凤叹了一口气，没精打采地说，一面解棉袄的钮扣。（巴金『家』人民文学出版社）
「やっぱり眠ってしまおう」彼女は力なくそうつぶやくと，綿入れの上衣のホックをはずす。（飯塚朗訳　岩波書店）

(11)“你要是还没吃了的话，一块儿吧！”虎妞仿佛是招待个好朋友。（老舎『骆驼祥子』人民文学出版社）
「ご飯まだだったら，いっしょにおやりよ」虎妞が声をはずませた。（立間祥介訳　岩波書店）

(12) 一直到十点钟，才剩下我们俩。他这才望了我一眼说：“怎么样，家里还好吧？”（魯彦周『天雲山伝奇』安徽人民出版社）
十時になってやっと我々二人だけになれた。彼は私をみて「どうだ，家の方は？」（田畑佐和子訳　亜紀書店）（＊どうだ，家のほうはいいだろう。）

表7 「吧」の用法分布と「だろう」対訳率[4]

	用法	原語	だろう 対訳率
判定 (21%)	推測	73	2.5
	確認要求	31	2.0
	認識要求	17	1.7
	軽い問掛	24	0
意志 (24%)	意志	65	0
	同意	29	0
	許容	27	0
	提案	41	0
働き掛け (47%)	誘い	81	0
	勧め	32	0
	頼み	46	0
	命令	144	0
	呪詛	12	0
	祈願	5	0
文中 (8 %)	仮定	18	0
	前提	27	0
	例示／提示	2 ／ 4	0
	計	678	6.2

これらの用例では,「対立項（対極性）暗示」という「吧」の基本義[5]が表面化されており,「だろう」が対訳されなくなったのだと考えられる。さらに，下例のように，命題情報に関する把握状況では話者と聴者が同じ程度か，または聴者のほうが多く把握すると思う場合，認識要求の「だろう」は「吧」の基本義と対応しにくい傾向（例13）があり，私的領域の情報表現では配慮と不配慮の相違がみられる（例14)。談話や言語行動のレベルでは,「だろう」の補足挿入文用法（例15）と「吧」の話題提示用法（例16）など周辺的派生用法には両者の対応が見えなくなるのである。

(13)「…私の生れは港なの。ここは温泉場でしょう。」(『雪国』)
"…我出生在港市，可这里是温泉浴场。"（葉謂渠訳　訳林出版社）

(14) 你应该搬到研究所去住。这样，你就有时间了。(謎容『人到中年』百花文艺出版社)
あなたは研究所へ引っ越すべきだと思うわ。そうすれば時間ができるでしょう。(林芳訳　中央公論社)

(15) うどん屋は川岸で，これも温泉場から流れて来る川だろう。尼僧が二人づれ三人づれと前後して橋を渡って行くのが見えた。(『雪国』)
面食店在河岸上。这条河大概也是从温泉浴场流过来的。可以看到尼姑三三两两地先后走

过桥去。(葉謂渠訳　訳林出版社)

(16) 祥子出了曹宅，大概有十一点左右吧，正是冬季一天里最可爱的时候。(『駱駝祥子』)
彼が曹先生の屋敷をでたのは，十一時ごろのことだった。冬の一日でもっともあたたかい時刻だ。(立間祥介訳　岩波書店)

このように，「だろう」と「吧」はそれぞれ「未確定」と「対立項暗示」との基本義から用法が展開し，認識的モダリティ表現において両者は交差し似たような対応を表しているが，最も典型的な用法と周辺的な用法において両者は分かれているということが，対訳データの並列状況から認められたのである。

4.3. 全方位的中日対照研究

モデル6は，モデル1，2，3とをすべてあわせたモデルで，いわば全方位的対照研究を目指すアプローチである。その特徴は，日本語と中国語の原語使用の異同について，訳文照応と訳語参照の多方向から総合的に分析するという点にある。

このアプローチによる研究例はまだ少ないが，曹(2002b)はその試みとして中日近義同形語「基本」を対象に考察を試みたものである。まず，「中日対訳コーパス」で小説と論説文から「基本」の原語と訳語を抽出し，そのヒット件数から使用状況を探った。その結果を表8に示す。

表8　中日同形語「基本」の使用状況

		作品数	字数	件数
小説	中	21	250万	24
	日	22	235万	0
論説文	中	1	13万	25
	日	2	21万	20

表8に見られるように，中国語の「基本」は，論説文，小説ともに同じように使われている。ところが，日本語の「基本」は論説文には使われているが，小説では使用例が見つからない，という結果になった。

そこで，さらに中国語と日本語の「基本」の相違を分析するために，対訳状況を調査した。その結果を表9に示す。

表9　中日同形語「基本」の対訳状況と用法

中国語の「基本」と日本語表現の対応					日本語の「基本」と中国語表現の対応				
「基本」の日本語訳	小説（件）	論説（件）	「基本」と訳された日本語	小説（件）	「基本」の中国語訳	小説（件）	論説（件）	「基本」と訳された中国語	小説（件）
基本	4	12	基礎	1	基本	0	11	基本	4
基礎	1	0	基本的	2	根本	0	3	总	2
最低の	1	0	大体の	1	基礎	0	1	本位	1
基本的	0	7	一通りの	1	核心	0	1		
根本的	1	0	ほとんど	2					
基本的な	3	4	大体	5					
基本的に	2	0	大抵	3					
ほぼ	2	0	一応	2					
ほとんど	1	0							
すっかり	1	0							
一応	2	0							
（略訳，意訳）	6	2			（略訳，意訳）	0	4	（縮訳）	2
計	24	25	計	17	計	0	20	計	9

まず，小説と論説について，それぞれの「基本」が対訳ではどのような表現で訳されているかを調べた。論説においては，中国語から日本語へ，日本語から中国語へ，いずれもほぼ同じ「基本」で対訳されていることが分かる。しかしながら，小説では，中国語から日本語へ対訳される際，「基本」以外の表現が使用されている点が目立つ。特に，「ほぼ，ほとんど，すっかり，一応」といった副詞に対訳されている点が特徴的である。たとえば，次のような例である。

（17）ここに日本人の仕事に対する考え方の基本がよくあらわれている。（岡本常男『心の危機管理術』現代書林）
于此，清楚地反映了日本人対于工作的基本想法。（潘金生・潘鈞訳　北京大学出版社）

（18）那天晚上，佳佳的病基本好了，囡囡的功课也作完了，兄妹俩相继睡去。（王蒙『活動変人形』人民文学出版社）
その日の夜，佳佳の病気はほとんどよくなり，園園の勉強も終わって，兄妹は前後して寝床についていた。（林芳訳　白帝社）

（19）原来拟定三天的日程，两天一晚上就基本完成了。（『活動変人形』）
三日の予定が二日一晩で一応終わった。（林芳訳　白帝社）

また，小説に関しては，日本語のどのような表現が中国語の「基本」に対訳されたか，逆に，中国語のどのような表現が日本語の「基本」に対訳されたかを調査した結果も表9にあわせて示した。この調査でも，日本語の「ほとんど，一応」といった副詞が中国語の「基本」に対訳され

ている点を含め，中国語の「基本」が，日本語の「基本」以外の表現に対応している数の多いことが認められる。

この対応状況で見られるように，「基本」は日本語ではほとんど名詞用法（例17）であるが，中国語では副詞用法（例18，19）にも機能が拡張されている。それに対して，日本語では「基本的に」という副詞的派生形があるが，原語ではヒットがなく，中国語の「基本」ほど広く使われていないのが特徴といえよう。一方，注意を要することは，中国語の原語に対して，日本語から訳された「基本」には，副詞用例が目立って多く，対応率は原語の用例を超えている。これは日本語では中国語「基本」の副詞用法に近い他の副詞用法が多いことによる。つまり，日本語原語の特徴とその訳し方を反映したものとして認識されなければならないのである。

以上のことから，二言語の原語の異同を分析するには，それぞれの訳語と比較するというアプローチが有効であるということができる。

5. おわりに

以上，述べてきたように，多言語コーパスは，コンピュータ技術と言語学の発展と共に様々な種類と多様な特性をもつようになりつつある。研究目的にそって適切な利用法を選定しコーパスを使い分けることにより，日本語の対照研究と言語類型論的研究に貢献することが可能である。

ただ，現在の多言語コーパスはまだ二言語によるものが多く，言語学の分野において対訳を利用する実践的研究も少なく，未熟なところが多いのも現状であろう。

今後，コーパスの多言語化と利用研究が進むにつれて，さらに新しい可能性と新しい課題が生まれてくるに違いない。多数の研究者がこれまでの二言語コーパスの経験と成果を引き継いで，互いに交流と協力を広げていけば，良い多言語コーパスを作り上げるとともに良い利用方法を創り出すに違いないと期待される。

筆者も「中日対訳コーパス」や日本語教科書コーパス[6]の構築と利用研究の経験により，二言語コーパスの不足を感じており，今後の課題として中日韓・中日英または中日韓英仏露などの多言語コーパスの共同構築と利用研究を提言している。その目標を実現するためには，多国の言語学者と情報工学者の学術交流と協力姿勢が不可欠であろう。本稿がその小さな礎の一つとなれば至上の喜びである。

注

1　その成果は『中日対訳語料庫的研製与応用研究論文集』で公刊されている。同書では32編の収録論文のうち，利用研究に関するものは16編あり，日本語研究に関しては語彙や文法，翻訳など多方面に及ぶものである。

2　「か，な，ね，よ」等の助詞が付かない「だろう」の各用法の対訳件数。用法の分類は曹(2000)もあわせて参照されたい。

3　奥田(1984，1985）の「おしはかり—未確証」，森山(1992)の「判断形成過程—未決定」などの諸説に通ずる意味。

4　(n%）は原語での分布率。

5　曹大峰(2000，2002a）で提示した仮説。

6　国際交流基金の助成研究プロジェクトとして北京日本学研究センターで開発，中国大学日本語専攻主幹科目「精読（総合日本語）」で広く使われる四種の初中級教科書を収録している。

参考文献

奥田靖雄(1984)「おしはかり(1)」『日本語学』3（12)，54-69，明治書院

奥田靖雄(1985)「おしはかり(2)」『日本語学』4（2)，48-62，明治書院

徐一平・曹大峰(2002)『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

申鉉竣(2003)『近代日本語における可能表現の動向に関する研究』絢文社

曹大峰・森山卓郎(1999)「感動詞に関す日中対照研究」『中国日語教学研究文集』8，333-341，大連理工大学出版社

曹大峰(2000)「認識モダリティの日中対照例――「だろう」と「吧」」国立国語研究所編『認識のモダリティとその周辺』，101-112，凡人社

曹大峰(2002a)「中日対訳語料庫応用研究初探」『日本学研究』11，学苑出版社

曹大峰(2002b)「パラレルコーパスの特徴と可能性研究」『中日対訳語料庫的研製与応用研究論文集』，49-60，外語教学与研究出版社

曹大峰(2006)「「可能」の可能表現の意味と機能について」日中対照言語学会2006年秋季大会発表要旨（未刊行）

森山卓郎(1992)「日本語における「推量」をめぐって」『言語研究』101，64-83，日本言語学会

付　記

本論文は，国立国語研究所第一回博報海外研究者招聘プログラムによる研究期間中にまとめられたものであり，その内容は筆者の一連の先行研究と日中対照言語学会2006年秋季大会における口頭発表の一部が土台となっている。投稿にあたっては，査読者ならびに編集委員の方々と庵功雄氏から有益なご助言をいただいた。心より感謝申し上げたい。

（投稿受理日：2007年１月31日）

（最終原稿受理日：2007年７月10日）

曹　大峰（そう　たいほう）

北京外国語大学北京日本学研究センター

100089　北京市西三環北路２号　北京外国語大学216信箱

cdfeng2005@163.com

Multilingual corpus for Japanese studies:

Based on the example of Japanese-Chinese Parallel Corpus

CAO Dafeng
Beijing Center for Japanese Studies

Keywords

multilingual corpus, translation corpus, application models, Japanese studies

Abstract

Recent developments of the computer-readable corpus and their applications have become increasingly diversified, and the type of users has also expanded to a wider community. These changes have created new problems in application, and studies on these problems have attracted much attention. The problems include ascertaining the characteristics, potentiality, and limitation of each corpus, and complying with the user's specific requirements.

Based on the series of research results from building and using the Japanese-Chinese parallel corpora for Japanese studies, the author points out the importance of fully understanding the characteristics of each corpus and the problems of treating sentences in parallel corpora as the equivalent of the original sentences. As an illustration of these issues and to demonstrate the best of use of parallel corpus, he presents his study using particular examples, including the Japanese expression *kanou-da* denoting the possibility, the sentence final particle *daroo*, and *kihon (jiben)* that has the same form in Japanese and Chinese.