

平成26年度

中国特許文献の機械翻訳のための
辞書整備及び機械翻訳の品質評価に関する調査

調査報告書

平成27年3月

一般財団法人 日本特許情報機構

(空白)

目次

1. 調査目的	1
2. 調査概要	4
2. 1 調査実施体制及びスケジュール	5
2. 2 調査結果	6
(1) 対訳辞書データの作成	6
(2) 対訳辞書データの検証及び翻訳品質評価手順に関する調査	7
3. 対訳辞書データの作成	9
3. 1 作成環境	9
(1) ソフトウェア (ツール)	9
(2) ハードウェア	10
3. 2 作成手法	10
(1) 文献対の解析 (対訳コーパス作成)	12
(2) 対訳コーパスの解析 (人手確認用対訳辞書候補データの作成)	14
(3) 人手確認用対訳辞書候補データの解析	22
(4) 対訳辞書データの作成	24
3. 3 作成結果	25
(1) 文献対の解析 (対訳コーパス作成)	26
(2) 対訳コーパスの解析 (人手確認用対訳辞書候補データの作成)	28
(3) 人手確認用対訳辞書候補データの解析	31
(4) 対訳辞書データ作成	40
3. 4 問題点・課題	46
(1) 採否判定作業手順	46

(2) 複合語と複合語以外の採用率.....	47
(3) 複数訳語または複数見出し語を持つ用語の確認	48
(4) 不適切な見出し語への対応	48
(5) 対訳コーパスの整備.....	49
(6) 人手確認作業の支援情報.....	50
(7) 辞書データの機械翻訳システムでの利用方法	50
4. 対訳辞書データの検証及び翻訳品質評価手順に関する調査	52
4. 1 評価の目的と概要	52
4. 2 評価対象文の作成環境	53
(1) 評価用テストセット	53
(2) 評価用機械翻訳ソフト	56
(3) ユーザ辞書の作成	56
(4) 機械翻訳の実行	60
4. 3 評価の実施	61
(1) 評価体制	61
(2) 評価基準（内容の伝達レベルの評価）	62
4. 4 辞書の違いに関する評価結果.....	66
(1) 翻訳の評価結果とその分析	66
(2) 主観評価	69
(3) 優劣のポイントとなった用語.....	72
(4) 二つの辞書の優劣（結果）	72
4. 5 ユーザ辞書の有無に関する評価結果.....	73
(1) 評価点と評価対象文数	73
(2) 機械翻訳 S と機械翻訳 B の比較	74
4. 6 複合語の追加による評価結果.....	76

(1) 複合語の追加により、翻訳が良くなる例（典型的な例）：	76
(2) 複合語の追加により、述語の認識ができなくなる複雑な例：	77
4. 7 評価手順の課題	79
(1) 評価手順の特徴と課題	79
(2) テストセットの課題	80
4. 8 中日機械翻訳に特有の課題	80
(1) 課題の抽出対象	80
(2) 課題の指摘件数	80
(3) 中日の翻訳言語別任意項目	81
(4) 特許庁の仕様書で例示された課題	93
(5) 用語辞書の副作用に関する分析	94
添付資料	95
添付資料 3. 1 対応中国・日本公開特許公報番号リストレイアウト説明	96
添付資料 3. 2 対応中国公開特許公報・和文抄録番号リストレイアウト説明	98
添付資料 3. 3 対訳コーパスレイアウト説明	99
添付資料 3. 4 人手確認用対訳辞書候補データレイアウト説明	102
添付資料 3. 5 人手確認により対訳辞書から除外したデータレイアウト説明	103
添付資料 3. 6 中日対訳辞書データレイアウト説明	104
添付資料 4. 1 機械翻訳結果評価シート（サンプル）	110

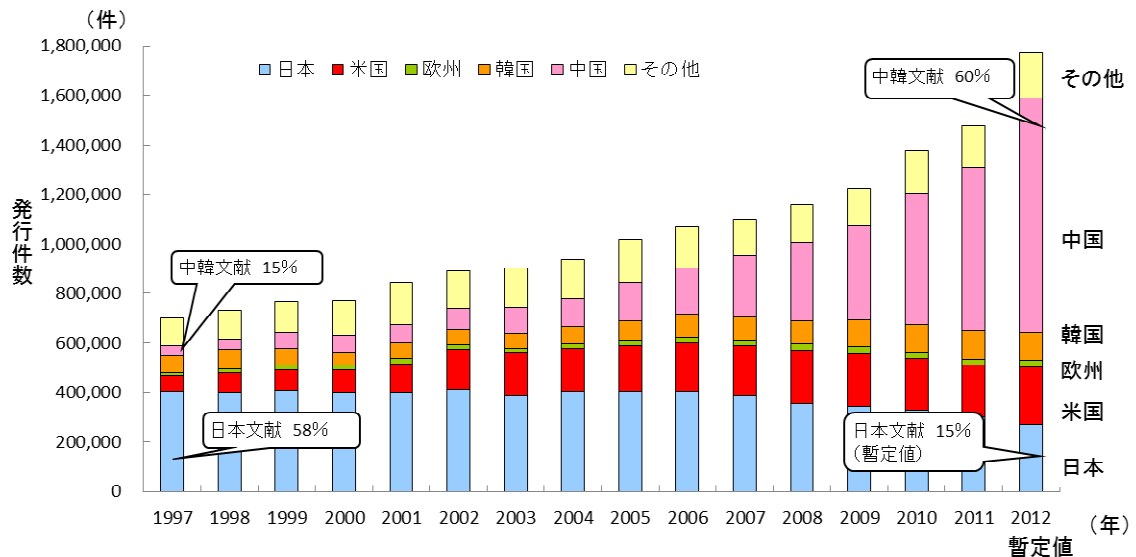
1. 調査目的

従来から特許出願が多くアクセスの必要性が高かった英語などに加え、非欧米諸国の特許出願、特に中国からの出願が近年増大している。

この出願構造の変化により、近年重要性を増してきた中国語については、我が国審査官及び出願人が原文のみから特許内容を理解することは困難であると考えられる。

また、中国の公開特許公報及び実用新案公報などの特許関連文献の年間発行数は、図 1-1 に示されるように、既に米国特許公報、米国公開特許公報、欧州公開特許公報を上回っており、今後さらに増加するものと見込まれている。

図 1-1 世界の特許文献

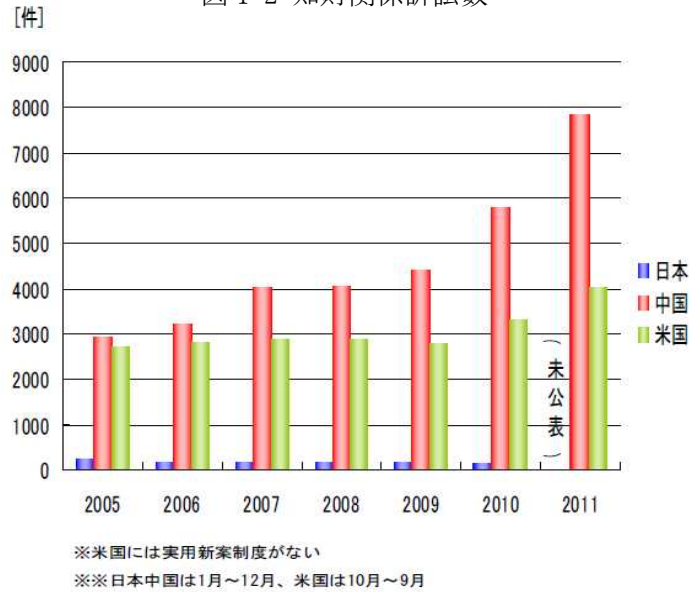


(出典) 特許行政年次報告書 2014 年版 特許庁

そして、世界で通用する安定した権利を設定するためには、日本語、英語はもとより、それ以外の外国文献についても漏れなく調査することが必須となっている。

また、図 1-2 で示されるように、中国では、出願件数の増加を背景に知財民事訴訟件数でも米国を越え、なおも増加傾向にある。

図 1-2 知財関係訴訟数

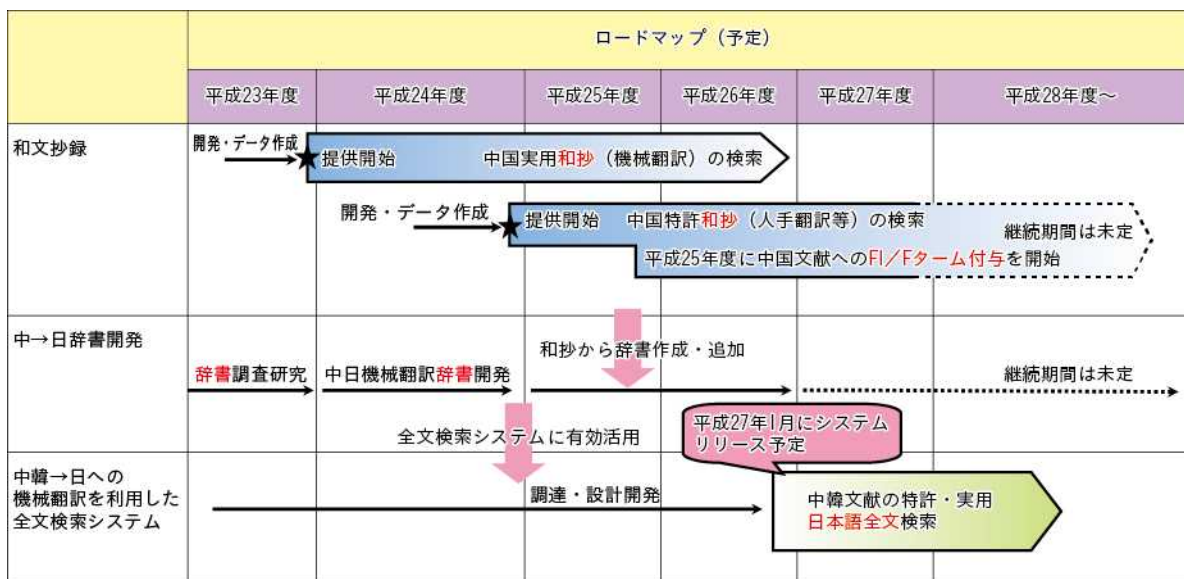


(出典) 産業構造審議会 平成 24 年 6 月 25 日配布資料

また、中国の特許関連文献が引用文献として引用される割合は他の特許庁において既に高く、技術的内容から見ても無視できないものとなっていることから、中国語の特許関連文献へ日本語でアクセス（検索・理解）できる環境を整備することが急務となっている。

このような状況を受け、特許庁は平成 23 年度から中国語の特許関連文献の和抄の提供を開始し、さらに平成 26 年度から機械翻訳により外国文献を日本語により調査・理解を可能とするインフラとして中韓文献翻訳・検索システムをリリースした。

図 1-3 中国文献及び韓国文献の検索システム開発のロードマップ（予定）



(出典) 特許行政年次報告書 2014 年版 特許庁

そうしたなか、特許庁は、平成 23 年度に「特許文献の機械翻訳のための辞書整備に関する調査」を実施し、特許文献の翻訳精度向上に資する辞書データの効率的な作成方法について、外部有識者の専門的知見を活用しながら分析を行い、特許庁が今後実施すべき辞書データの整備方法の検討を行った。当該調査において効率的な辞書作成方法として報告されたパテントファミリーを利用した中日辞書作成方法では、辞書データが作成されるとともに、その過程において中日対訳コーパスが作成された。辞書データはルールベース機械翻訳（RBMT）、対訳コーパスは統計的機械翻訳（SMT）等において活用することが想定される。

特許庁は、平成 24 年度に「中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査」、平成 25 年度に「中国特許文献の機械翻訳のための新語に関する調査」を実施し、2005 年～2012 年の中日パテントファミリーから中日対訳コーパスと 200 万語の中日機械翻訳辞書を作成した。また、機械翻訳の品質を評価するため、平成 25 年度に「特許文献機械翻訳の品質評価手法に関する調査」を実施し、「特許文献機械翻訳の品質評価手順」を作成した。

上述の調査により、効率的な辞書作成方法について一定の知見が得られかつ中日機械翻訳辞書が得られたが、急増する中国の特許関連文献においては、引き続き翻訳品質の向上及び技術の進歩に伴う新しい技術用語への対応のため、中日対訳辞書データ及び中日対訳コーパスをアップデートしていく必要がある。また、機械翻訳の品質評価についてはこれまで作成してきた対訳辞書データを用いた機械翻訳の品質評価を行い対訳辞書データの有効性を調査するとともに、中日機械翻訳の言語固有の課題等について調査の必要があると考えられている。

そこで、本調査では、2013 年の中日パテントファミリーから対訳コーパスと 20 万語の中日対訳辞書データを作成するとともに、「特許文献機械翻訳の品質評価手順」の改訂を目的として、調査を実施した。

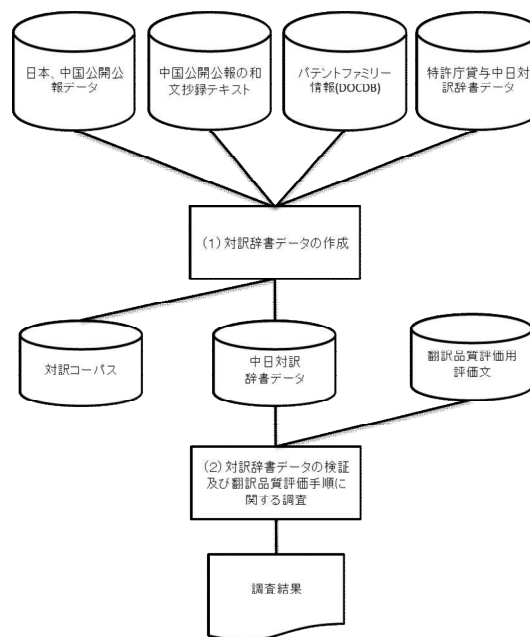
2. 調査概要

特許庁はこれまで、機械翻訳精度の向上のため、平成 24 年度「中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査」、平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」において中日対訳辞書データ 200 万語を作成してきた。また、機械翻訳の精度を評価するため、平成 25 年度「特許文献機械翻訳の品質評価手法に関する調査」を実施してきた。本調査はこれらの調査を基礎とし、中国特許公開明細書（2013 年）と中国公開特許公報の和文抄録（約 25 万件）から中日対訳辞書データ及び対訳コーパスを作成し、当該辞書データを用いた機械翻訳システムによる翻訳文を評価することにより、辞書データによる翻訳品質の効果を把握するとともに、特許庁が作成した「特許機械翻訳の品質評価手順」の改定のための調査を行った。

調査は、2013 年発行の中国公開特許公報と技術内容が対応する日本公開特許公報（公表特許公報及び再公表特許を含む。）を解析するとともに、特許庁が提供する中国公開特許公報の和文抄録と対応する中国公開公報の要約とを解析することにより、中日対訳辞書データ及び対訳コーパスを作成した。

作成した辞書データを中日機械翻訳ソフトウェアに搭載して機械翻訳を実施し、翻訳結果について「特許文献機械翻訳の品質評価手順」に従って品質評価を実施することにより、辞書を用いることによる機械翻訳の品質向上の効果を把握し、今後の辞書作成の課題、改善点等について調査するとともに、評価手順の改善点について調査した。図 2.1-1 に調査概要図を示す。

図 2.1-1 調査概要図



2. 1 調査実施体制及びスケジュール

前述調査概要の各工程を円滑に連携して調査を実施可能とするため、調査は表 2.1-1 に示すチーム体制にて実施した。

表 2.1-1 実施体制

チーム名	主な役割	要員数
統括責任者	本調査全体の統括。	1 名
辞書作成チーム	中日対訳辞書データ及び対訳コーパスの作成の工程を管理・実行。 人手確認用中日対訳辞書候補データの作成。 人手確認結果のサンプルチェック、訳語確認者及び機械処理の改善のフィードバック。	3 名(うち 1 名は中国語を母国語とする者)
対訳辞書検証・翻訳品質評価チーム	作成した辞書データを使い機械翻訳を実施し、翻訳結果について「特許機械翻訳の品質評価手順」に従って品質評価を実施。	管理者 1 名、訳語確認者 30 名、品質評価者 5 名
報告書作成チーム	本調査の調査手法、調査結果、統計情報や実際の作成作業において生じた技術的課題をまとめ、報告書の作成。	3 名
アドバイザー	統括責任者からの問い合わせに対するアドバイス。	大学教授 2 名

本調査は平成 26 年 11 月から平成 27 年 3 月の期間で実施した。実施スケジュールを図 2.1-2 に示す。

図 2.1-2 調査実施スケジュール

項番	作業項目	担当	開始	終了	2014年11月				2014年12月				2015年1月				2015年2月				2015年3月			
1	対訳辞書データ作成	辞書作成チーム	2014/11/3	2015/3/17	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				
2	対訳辞書データの検証及び翻訳品質評価手順に関する調査	対訳辞書検証・翻訳品質評価チーム	2015/1/19	2015/2/28										●	●	●	●	●	●					
3	報告書作成	報告書作成チーム	2015/2/1	2015/3/20												●	●	●	●	●	●	●	●	●
4	納品	統括責任者	2015/3/24	2015/3/24																			●	

2. 2 調査結果

(1) 対訳辞書データの作成

2013 年発行分の中国公開特許公報と技術内容が対応する日本公開公報とを解析するとともに、中国公開特許公報の和文抄録と対応する中国公開特許公報の要約とを解析することにより、2,500 万件の対訳コーパス（中国語と日本語の文と文を対訳にした対訳文対の集まり）を作成した。対訳コーパスの作成においては、中日自動文アライメントツール（対訳となる文書の文分割を行い、文対応付けを行うツール）を用いて作成した。中日自動文アライメントのツールの精度は、中日対訳コーパスの品質に直接影響し、ひいては作成される辞書の品質にも大きく影響することから、中日自動文アライメントに先立ち、特許庁から貸与された約 200 万語の対訳辞書データを予め学習することで、高精度化を行った。

また、中日対訳コーパスの作成過程において、中日対訳コーパスを作成する際の元データとなる、対応中国・日本公開特許公報番号リスト(8 万件)、対応中国公開特許公報・和文抄録番号リスト(25 万件)を作成した。作成した中日対訳コーパスを解析し、日本語の名詞及びサ変動詞になり得るものを抽出し、抽出した日本語に対応する中国語を抽出し、作成した対訳辞書候補データから明らかに解析誤りと思われる用語を機械的に除き、約 40 万件の人手確認用対訳辞書候補データを作成した。この対訳辞書候補データを出現頻度の高いものから人手により辞書登録する用語にふさわしいか否かを判断して、出現頻度上位 20 万語について、複合語であるか否かの情報、特許庁が指定する出現頻度情報を付して対訳辞書データを作成した。

(a) 品質向上策

本調査では辞書データの品質担保を目的に、人手確認用対訳辞書候補データの一部について異なる 2 名の訳語確認者による採否判断を行った。これらの採否結果を分析することで採否判断誤りが多いデータが特定でき、誤りの多いデータを再確認することで、品質向

上策となることが分かった。

さらに人手確認を行い採用した対訳辞書データうち、一つの見出し語（中国語）に対して複数の訳語（日本語）が対応するものや、反対に一つの訳語に対して複数の見出し語が対応するもの、約 2 万件について再度採否判定を行った。再判定の結果、対象の 2 万件の約 11%が不採用とするべきデータであった。不採用と判断されたデータを見ると、複数存在する訳語または見出し語を並べて確認することで単独で判断した際には見つけられなかった誤りを見つけることができ、この複数訳語と複数の見出し語対の再判定が品質向上に有効であることが分かった。

(b) 対訳辞書データの付加情報

本調査にて作成した中日対訳辞書データとこれまでに作成した特許庁の中日対訳辞書データをマージしたデータ（221 万件）に付加情報として、複合語であるか（81 万件、37%）、単語であるか（140 万件、63%）の情報を付した。今年度作成した対訳辞書データ 20 万件の内、複合語が 14%、複合語以外の語が 86%となった。複合語は訳語確認者の採否判定において採用率が 86%となり全体の採用率 60%より高い傾向が見られた。複合語の出現頻度は平均 30 文献となり、複合語以外の出現頻度の平均 42 文献と比べ、出現する文献が少なくなり対訳の汎用性が低いことが分かった。

付加情報としてさらに、対訳辞書データに特許庁が指定する出現頻度情報として文単位出現頻度と文献単位出現頻度とを付した。文単位出現頻度は辞書の見出し語、訳語がともに含まれる対訳コーパスの数であり、文献単位出現頻度は見出し語、訳語がともに含まれる文献対数である。複合語の定義は本報告書 3. 2 章の（4）対訳辞書データの作成を参照のこと。

（２）対訳辞書データの検証及び翻訳品質評価手順に関する調査

分野別の辞書を作成する際、複数の訳語候補を一つに絞るために対訳コーパスにおける出現頻度を使う。従来から使っている文単位出現頻度は一部の文献内のみに繰り返し出現する特殊な訳語の頻度が高くなり、多くの文献で用いられる一般的な訳語が採用されにくくなる課題が存在する。そこで本調査では文献単位出現度に基づく対訳辞書データを作成し、従来の文単位出現頻度に基づく辞書による翻訳結果と比較評価した。

この２種類の辞書の違いは、本調査の評価に用いた辞書では用語 2,147 語（４分野とも同一語数）に対して異なり語数が C 化学：81 語、E 電気：81 語、M 機械：84 語、P 物理：100 語であった。

このように作成した２種類の辞書により特許庁の中日テストセット 150 文を翻訳し、その翻訳結果を特許庁の「特許文献機械翻訳の品質評価手順」に従って評価した。結果として、文献単位出現頻度に基づく辞書を用いた翻訳結果の方が翻訳の品質が優れていること

が分かった。例えば、後述の 4. 4 辞書の違いに関する評価結果 (1) (b) の表 4. 4-3 翻訳品質評価シート 69 では中国語「防皺性」に対して文単位出現頻度では「抗皺性(耐リジング性)」が採用され、文献単位出現頻度では「抗皺性(防皺性)」が採用された。翻訳結果の評価により後者が適切な訳語であることが分かった。

さらに特許庁の「特許文献機械翻訳の品質評価手順」に本調査で得た中日翻訳の知見を追加できるように調査結果を整理した。具体的には、①内容の伝達レベルの評価の 5 段階に対して、評価点 4、3、2、1 に対応する翻訳結果と評価の根拠をあげ、②中日の翻訳言語別任意項目に中日翻訳において現れる課題を分類・整理した。

3. 対訳辞書データの作成

2013 年に公開された中国公開特許公報と技術内容が対応する日本公開公報とを解析するとともに、2011 年に公開された中国公開特許公報の和文抄録と対応する中国公開特許公報の要約とを解析することにより、2013 年発行分の中国公開公報における 20 万語の対訳辞書データを作成した。

また、対訳辞書作成過程において、8 万件の対応中国・日本公開特許公報番号リスト、25 万件の対応中国公開特許公報・和文抄録番号リスト、2,500 万件の中日対訳コーパス、40 万件の人手確認用対訳辞書候補データを作成した。

3. 1 作成環境

(1) ソフトウェア（ツール）

本調査で使用するソフトウェアのうち中日対訳コーパス作成に使用する中日自動文アライメントツールは高い文アライメント精度を有し、かつ、さらに精度を向上するために特許庁の中日対訳辞書データを追加学習する必要がある。そこで本調査では、これら 2 つの条件を満たすツールとして、平成 25 年度調査で使用実績のある中日自動文アライメントツールを使用した。

また、中国語解析ツールは平成 25 年度調査で利用したツールより精度の高いツールを使用した。

その他、新語の抽出では抽出を効率的に行うため既存のソフトウェアツールを活用した。使用した主なツールを表 3. 1-1 に示す。

表 3. 1-1 主要ツール一覧

ツール	ツール名	用途（ツール取得先 URL）
中日自動文アライメントツール	-	中日対訳コーパスの作成
日本語解析ツール	茶筌	中日対訳コーパスから対訳辞書候補データ作成 (http://chasen-legacy.sourceforge.jp/)
専門用語抽出ツール	言選	中日対訳コーパスから対訳辞書候補データ作成 (http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html)
中国語解析ツール	-	中日対訳コーパスから対訳辞書候補データ作成
フレーズテーブル作成ツール	moses	中日対訳コーパスから対訳辞書候補データ作成 (http://www.statmt.org/moses/)
出現頻度集計ツール	Apache Solr	中日対訳コーパスから対訳辞書候補データ作成 (http://lucene.apache.org/solr/)

(2) ハードウェア

対訳辞書作成にかかる処理は大量の特許データを扱うため、長時間の機械処理を必要とする。本調査の限られた期間内に処理を終了するために、複数のハードウェアによる並列処理を行った。並列処理は、クラウド上に提供されているハードウェア資源を利用した。本調査で使用したハードウェア・スペックを表 3.1-2 に示す。

表 3.1-2 主なハードウェア・スペック

処 理	スペック	台数	備考
中日自動文アライメント	CPU:2.5ECU ¹ ×2 Memory:1.7GByte	150	クラウド
対訳辞書候補作成	CPU:Intel Core i7-2600 3.40GHz Memory:16GByte	2	

3. 2 作成手法

対訳辞書データは図 3.2-1 対訳辞書データ作成概要図に示す手順で作成した。手順は下記 (1) から (4) の処理で構成される。

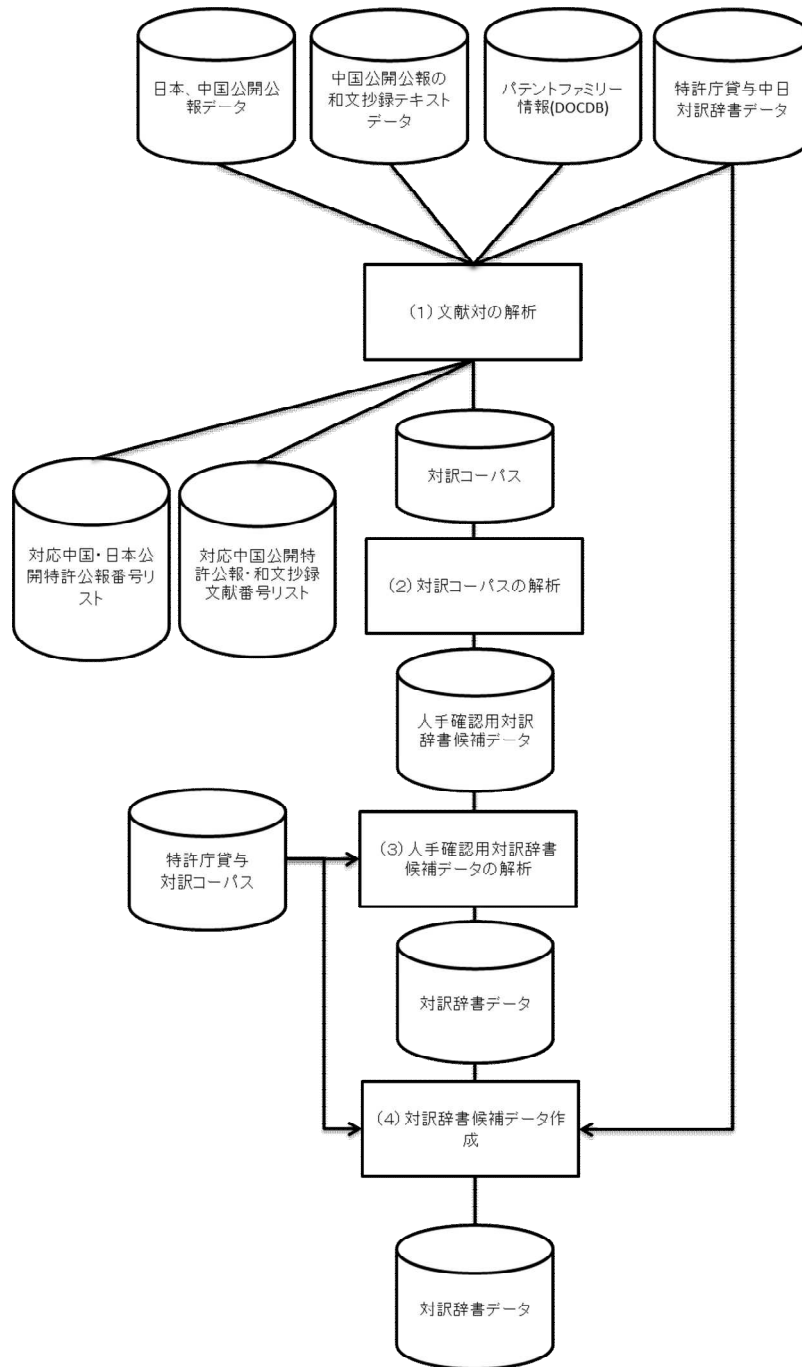
- (1) 文献対の解析 (対訳コーパス作成)
- (2) 対訳コーパスの解析 (人手確認用対訳辞書候補データの作成)
- (3) 人手確認用対訳辞書候補データの解析
- (4) 対訳辞書データ作成

(1) 文献対の解析では DOCDB²に蓄積されているパテントファミリーを利用し、技術内容が対応している中国と日本の公開特許公報の明細書及び中国の文献とその和文抄録から中日対訳コーパスを作成し、(2) 中日対訳コーパスの解析では (1) で作成した中日対訳コーパスを解析することで、中国語と日本語の用語単位対訳データ (以降、人手確認用対訳辞書候補データと呼ぶ) を作成し、(3) 人手確認用対訳辞書候補データの解析では (2) で作成した人手確認用対訳辞書候補データの見出し語 (中国語)、日本語 (訳語) の採否について人手による確認を行い、(4) 対訳辞書データの作成では (3) で人手確認を行い辞書として判断されたデータと特許庁が貸与する辞書データをマージし中日対訳辞書データを作成した。

¹ ECU はクラウド上のコンピュータ資源の性能を表す単位の 1 つ。

² EP0 (欧州特許庁) が発行している世界 70 カ国以上の公開公報を収録したデータベース

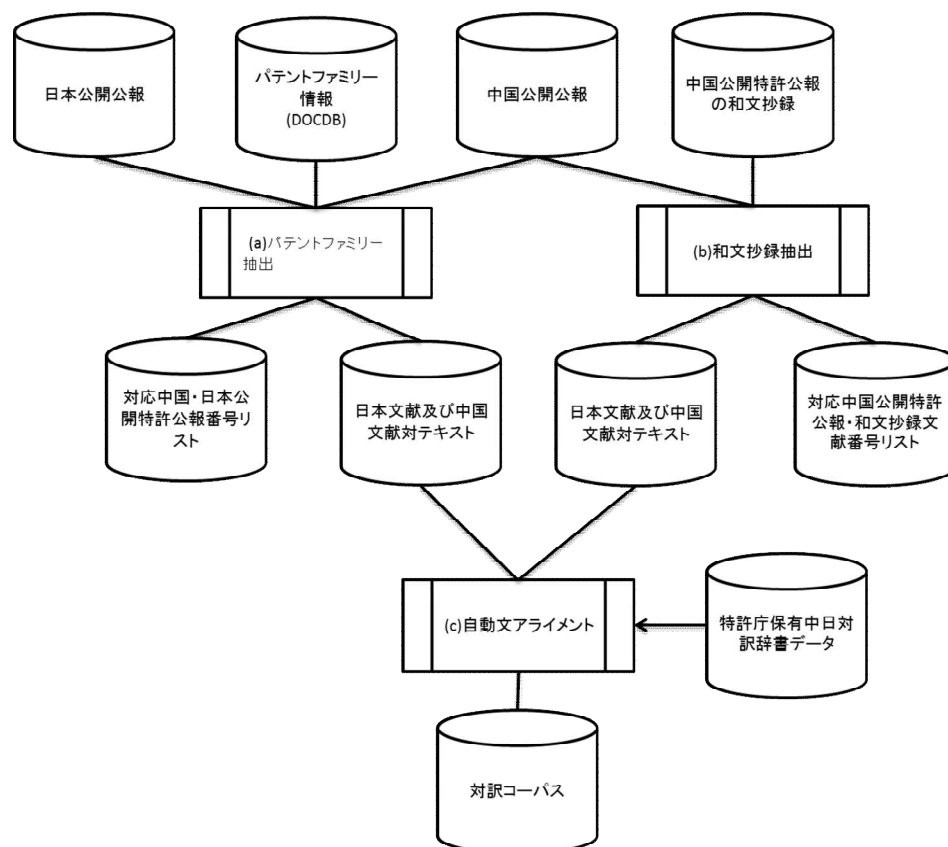
図 3.2-1 対訳辞書データ作成概要図



(1) 文献対の解析 (対訳コーパス作成)

文献対の解析では、パテントファミリー情報に基づいて技術内容が対応する中国公開特許公報と日本公開特許公報の明細書の各文及び中国公開特許公報の要約文とその和文抄録文を対応付けて対訳コーパスを作成した。文献対の解析の処理フローを図 3.2-2 に示す。

図 3.2-2 文献対の解析概要図



本処理は文献対の解析概要図に示すように (a) パテントファミリー抽出、(b) 和文抄録抽出、(c) 自動文アライメントから構成される。

(a) パテントファミリー抽出

パテントファミリー抽出は、DOCDB に蓄積されているパテントファミリー情報を利用し、技術内容が対応する中国の文献と日本の文献対を作成した。具体的には、DOCDB がパテントファミリーを管理するデータ項目 (family id) が同一の中国公開公報と日本公開情報を技術内容が対応するものとし、文献対の番号リスト (図 3.2-3) を作成した。

図 3.2-3 対応中国・日本公開特許公報番号リスト（イメージ）

CN	A	1816909 20060809	33562464	JP	A	2005051217	20050224	H01L	21/683
CN	A	1579344 20050216	34395279	JP	A	2005052652	20050303	A61B	17/88
CN	A	1580634 20050216	34368610	JP	A	2005061688	20050310	F23D	14/06
CN	A	1603645 20050406	34373274	JP	A	2005106105	20050421	F16C	33/14

次に、この対応中国・日本公開特許公報番号リストを基にして日本と中国の公報の明細書から日本語テキスト及び中国語テキストを抽出した。中国公開特許公報と日本公開特許公報の電子データは、ともに XML 形式で作成されている。図 3.2-4 は中国公開特許公報のデータの一部である。図の下線部分は XML タグと呼ばれ、文書データの構造を表している。本調査で仕様する中日自動文アライメントツールはこのような XML 形式をそのまま処理できないため、XML 形式のデータから XML タグを取り除き、日本語テキスト及び中国語テキストを抽出した。

図 3.2-4 中国公開特許公報（XML 形式）

```
<p id="p0006" num="0006">数字视频通信系统可采用一种——且有时候多种——数字视频编
码格式
<br />进行视频的编码、存储和输送。例如，在传统视频会议系统中，使用 H. 261 和
<br />H. 263 视频编码标准两者，而在数字电视系统中，利用 MPEG-2/H. 262 视频编
<br />码。更新近的系统在视频会议和广播设置两者中使用 H. 264 视频。
<br /></p>
```

(b) 和文抄録抽出

本調査では、前述(a)の Patent ファミリーに加えて、中国公開特許公報から作成された中国公開公報和文抄録（図 3.2-5）も使用した。和文抄録の抽出方法は、まず和文抄録データ内に蓄積されている中国の公開番号から対応中国公開公報・和文抄録文献番号リストを作成した。そして、このリストをベースに和文抄録データ及び中国公開公報の要約部分から日本語テキスト及び中国語テキストを抽出した。

図 3.2-5 中国公開公報和文抄録

・・・ティ原動機</P><P>抄録文</P><P>本発明の開示する異型キャビティ原動機はハウジング、回転子および2つのカバープレートからなる。ハウジングは異型面内キャビティ、入り口および出口を有する柱体であり、異型面内キャビティは円弧面と非円弧面によって組み合わせてなる。回転子は回転子本体と2対の組み合わせスライドプレートからなり、回転子本体は伝動軸、センタリング軸および十字交差ガイド溝が加工される円柱体であり、組み合わせスライドプレートはガイド溝内・・・

(c) 自動文アライメント

前述(a)、(b)で抽出した日本文献及び中国語文献テキストから、中日自動文アライメントツールを使い、対訳コーパスを作成した。中日自動文アライメントツールは、平成25年度調査で使用実績のある中日自動文アライメントツールを使用した。なお、このツールは特許データを使った、高い精度の自動文アライメント手法であり、手法が論文³で発表されている。この中日自動文アライメントツールは約15.6万語の中日対訳辞書を使用し、中国語文と日本語文の用語の一致度（文対応スコア⁴）を基に中国語と日本語文の対応づけを行う。

本調査の自動文アライメントでは、自動文アライメントツールの精度を高めるために、中日自動文アライメントツールが備える、約15.6万語の中日対訳辞書に、特許庁から貸与された約200万語の中日対訳辞書データを加え自動文アライメントを行った。

(2) 対訳コーパスの解析（人手確認用対訳辞書候補データの作成）

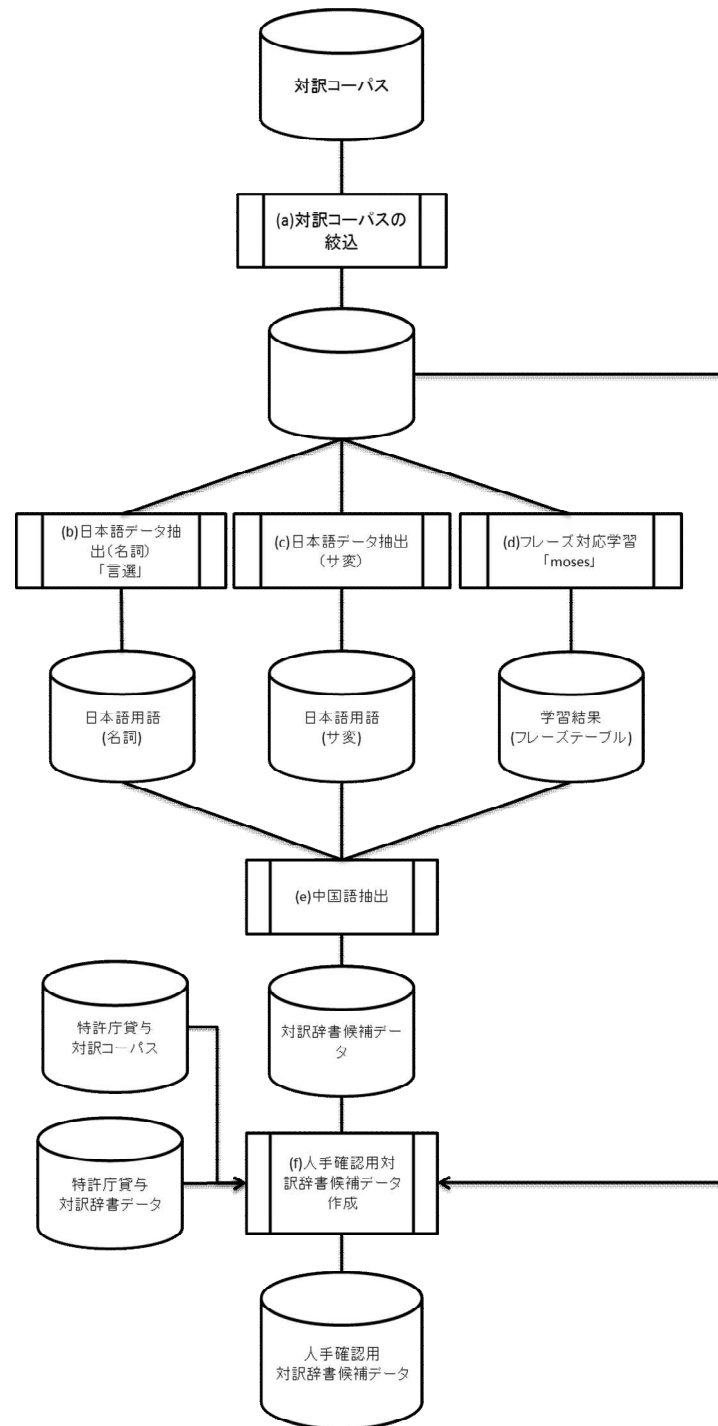
中日対訳コーパスの解析では、前述（1）文献対の解析で作成した対訳コーパスを解析し、人手確認用対訳辞書候補データを作成した。

図3.2-6に中日対訳コーパスの解析の処理フローを示す。解析は(a)中日対訳コーパスの絞込、(b)(c)日本語データ抽出（名詞、サ変）、(d)フレーズ対応学習、(e)中国語抽出を経て中日対訳用語の対を作成した。その後、出現頻度の大きい中日対訳用語対から優先して作業を行うために、(f)中日頻度取得を行い、複合語情報の単語を優先で出現頻度の大きい順に並び替えをし、対訳辞書候補データを作成した。

³ 内山将夫、井佐原均: A Japanese-English Patent Parallel Corpus. MT summit XI, pp. 475-482, 2007.

⁴ 文対応スコアは対応付けの精度となる情報で、対訳文の用語の一致度に基づいて算出される。スコア値が大きいほど精度の高い文対応を意味する。

図 3.2-6 対訳コーパスの解析



(a)中日対訳コーパスの絞込

前述（１）文献対の解析で作成した対訳コーパスの中には中国語と日本語が１文と１文で対応していないものや、中国語文と日本語文の対応が良くないものが存在する。これら

の文は新語抽出の精度を低くする原因となる。そこで本調査では、中日対訳コーパスからこれら文対応度が低いデータを除外した。具体的には、中日対訳コーパス内の中国文と日本語文が1文と1文で対応しかつ、文の対応度合いを示す文対応スコアが0.08以上のものを以降の処理の対象とした。

さらに、対訳コーパス内の中国語と日本語で対応していない括弧書き表記の削除や、文の長さによる中日対訳コーパスの絞込を行った⁵。これら詳細は、後述<<作業における工夫>>で説明する。

(b)日本語データ抽出（名詞）

日本語データ抽出は、前述(a)で絞込んだ中日対訳コーパスの日本語部分から対訳辞書データとなり得る日本語の名詞を抽出した。本調査では名詞の抽出に専門用語抽出ツール「言選」を使用した。なお、同ツールは平成 25 年度調査でも使用され実績のあるものである。

(c)日本語データ抽出（サ変）

前述日本語データ抽出（名詞）で使用したツール「言選」は、名詞を抽出するためのツールであり、本調査で抽出対象となっている「サ変動詞」の抽出には対応していない。そこで本調査では日本語解析ツールを使い、中日対訳コーパスの日本語文を単語分割し、各単語に品詞情報を付加し、この品詞情報を使いサ変動詞の抽出を行った。具体例を図 3. 2-7 に示す。図は日本語解析ツールが処理した結果である。この例では、日本語解析結果の名詞“嵌合”、動詞“する”の単語列をサ変動詞とした。なお、日本語動詞には活用があるため、例にある“する”以外の活用形“し”も抽出対象とした。

図 3. 2-7 サ変動詞抽出例

.....		
INSERT	名詞	一般
を	助詞	一般
<u>嵌合</u>	<u>名詞</u>	<u>一般</u>
<u>する</u>	<u>動詞</u>	<u>自立</u>
こと	名詞	一般
による	助詞	連語
.....		

⁵ 対訳コーパスの絞り込み条件は、平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」と同等とした。

(d)フレーズ対応学習

フレーズ対応学習は、中日対訳コーパスから、統計的手法によりフレーズ対応学習を行い、フレーズテーブルを作成した。フレーズテーブルは図 3.2-8 にあるように、1つもしくは複数の単語の並びからなる日本語及び中国語のフレーズを対応付けたデータである。またフレーズテーブルは、対応する日本語及び中国語のフレーズの他にフレーズの対応度合いを示す確率（翻訳確率）などから構成されている。本調査では、このフレーズテーブル作成に統計的機械翻訳ツールキット「moses」を使用した。なお、同ツールは平成 25 年度調査でも使用されたものである。

(e)中国語抽出

前述(b)、(c)で作成した日本語用語（名詞、サ変動詞）と(d)で作成したフレーズテーブルとを突き合わせ、日本語が一致したものを抽出する。図 3.2-8 に抽出例を示す。

図 3.2-8 中国語抽出例（処理イメージ）

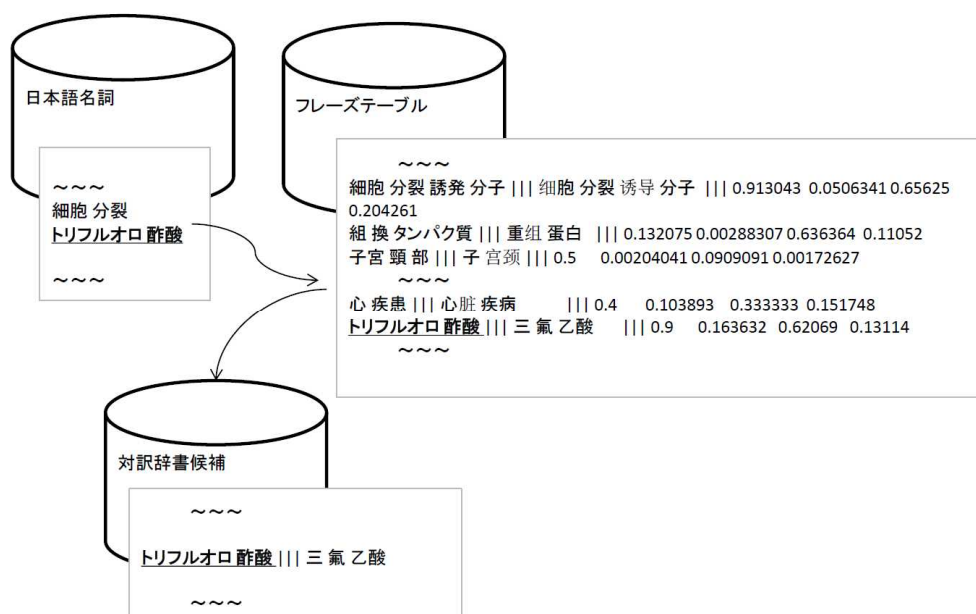


図 3.2-8 は日本語名詞“トリフルオロ酢酸”に対応する中国語“三氟乙酸”の抽出例である。抽出は、図の日本語名詞内の各日本語用語について、同一の用語がフレーズテーブルに含まれているかを検索する。検索の結果、フレーズテーブルに含まれている場合、その内容を対訳辞書候補として出力する。

(f)人手確認用対訳辞書候補データ作成

前述(e)で作成した対訳辞書の候補となるデータについて、特許庁が既に作成済みの対訳辞書データと突き合わせ、重複するデータを排除し、前述(a)で作成した対訳コーパスにお

ける出現頻度を付し、特許庁の既存の中日対訳辞書データと合わせて、複合語であるか否かの情報を付加し、人手確認用対訳辞書候補データを作成した。

<<作業における工夫>>

前述（１）文献対の解析で作成した対訳コーパスの中には中国語と日本語が１文と１文で対応していないものや、文対応スコアの値が低いものの他にも文の対応度合いがよくないものが存在する。中日対訳コーパスの解析における工夫した点を以下に示す。

① １文と１文で対応していないもの

中国語と日本語が１文と１文で対応していないものを取り除いた。

図 3.2-9 中日対訳コーパスの文の対応が１文と１文で対応していないもの

在一些实施例中，设备 110 的总体形状可以与所示出的不同。
<u>色が違っても良い。</u> /// 実施形態によっては、デバイス 110 の全体的な形状は図示したものと異なってもよい。

図 3.2-9 は、中国語文１文に対して日本語文２文が対応している例である。日本語文の下線部“色が違っても良い。”に対応する中国語が存在しないため文の対応が良くない。なお、「///」はツールが合成した文の区切りを示す。

② 対応していない括弧書き表現

文の対応度合いがよくないものとして、対応していない括弧書き表現がある。その例を、図 3.2-9 に示す。

図 3.2-10 対応していない括弧書き表現

図 4 は、本発明に基づく反射投影型 <u>(front-projection-type)</u> 表示装置 10C を示している。
图 4 表示根据本发明的一个前投射型显示器 10C。

図 3.2-10 の日本語文の下線部分“(front-projection-type)”は直前の用語の“反射投影型”の英語表現を括弧書きで表現したものである。一方、中国語文には日本語文にある括弧書き表現が存在しないため、不完全な文対となる。このような表現は、外国から出願された特許文献内に比較的良好に見られる表現である。

上述の例に示したような、中国語文または日本語文どちらか一方にのみ存在する括弧書き表現について、本調査では図 3.2-11 に示すように括弧書き表現を削除することで、中国語文と日本語文の対応度を高めた。なお、対応していない文中の括弧書き表現を一律に削

除すると、過剰に削除してしまう可能性があるため、本調査では括弧書き内に英語表記がされているものを削除対象とした。

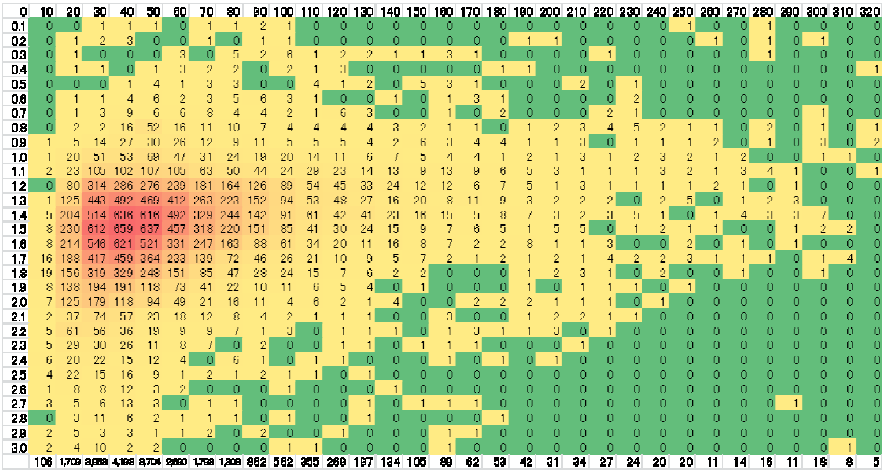
図 3.2-11 対応していない括弧書き表現（括弧書き削除後）

図 4 は、本発明に基づく反射投影型表示装置 10 C を示している。
図 4 表示根据本发明的一个前投射型显示器 10C。

③文の長さによるアライメントの選択

当社は、高精度の中日対訳コーパス作成のため、独自研究で中日自動文アライメントツールの出力結果を調査している。例えば、中日対訳コーパスの日本語文と中国語文の長さの組み合わせ毎に文対応度合いを調査している。図 3.2-12 は、その調査結果の一部であり、横軸に中国語文の文字数、縦軸に[日本語文の文字数÷中国語文]の比率を取ったものである。また、図中の数値は対訳コーパスの出現頻度を示し、色は、出現頻度の高いものから赤、黄色、緑となっている。この図の各部分をサンプルチェックし文対応度合いを調査した。調査の結果、中日対訳コーパスのうち日本語と中国語の文字数がアンバランスのものや、長い文字数のものは文対応度合いが低い傾向にあることが分かった。

図 3.2-12 中国語文と日本語文の文字数比率調査結果(当社調査)



本調査では、課題で説明した当社独自の調査結果を利用し、中日対訳コーパスの文字数や日本語文、中国語文の文字数比率により中国語文と日本語文の文字数のバランスがよくないものや、中国語文または日本語文が極端に長いデータを除外した。その例を、図 3.3-13 に示す。

図 3.2-13 中日対訳コーパスの文の対応がよくない例

令人感兴趣的是，血浆中 RIF 的药代动力学参数不受共同给药 TIM 的影响(表 3)，表明 TIM 可能更特别地起感染组织水平的流出抑制剂作用，增强了 RIF 的活性。

興味深いことに、血漿中の R I F の薬物動態パラメーターは、T I M の同時投与に影響を受けなかった（表 3）。

④日本語専門用語抽出ツールの選定とチューニング

前述(b)日本語データ抽出（名詞）では、名詞を抽出するため日本語専門用語抽出ツール「言選」を使用した。「言選」は、名詞抽出の前処理で日本語解析ツールを使用する。そのため、この日本語解析ツールが正しく日本語を解析した場合に正しい日本語専門用語の抽出が可能となる。しかし、ツールに付属する辞書だけでは図 3.2-14 に示すような名詞“くわえづめ”を正しく解析できないために日本語専門用語として抽出できない場合がある。

当社は日本語見出しで 500 万語を超える特許文書に出現する用語を収集したデータを保有している。このデータ中には“くわえづめ”のような一般の日本語解析では正しく解析できないような用語も含まれている。そこで本調査では、日本語解析ツールにこのような用語を解析用辞書としてチューニングすることにより日本語専門用語の抽出精度の向上を図った（図 3.2-15）。

図 3.2-14 日本語解析結果（チューニング前）

循環	ジュンカン	循環	名詞-サ変接続
運動	ウンドウ	運動	名詞-サ変接続
する	スル	する	動詞-自立
くわえ	クワエ	くわえる	動詞-自立
づ			未知語
め	メ	め	名詞-接尾-一般
キャリッジ			未知語
8	ハチ	8	名詞-数
を	ヲ	を	助詞-格助詞-一般
有する	ユウスル	有する	動詞-自立
			サ変・スル
			基本形
			連用形
			一段
			基本形
			サ変・ースル
			基本形

図 3.2-15 日本語解析結果（チューニング後）

<チューニング後>					
循環	ジュンカン	循環	名詞-サ変接続		
運動	ウンドウ	運動	名詞-サ変接続		
する	スル	する	動詞-自立	サ変・スル	基本形
くわえづめ		くわえづめ	名詞-一般		
キャリッジ			未知語		
8	ハチ	8	名詞-数		
を	ヲ	を	助詞-格助詞-一般		
有する	ユウスル	有する	動詞-自立	サ変・ースル	基本形

⑤中国語解析ツールの選定

前述(d)のフレーズ対応学習では、前処理として中日対訳コーパスの日本語文、及び中国語文を日本語解析ツール及び中国語解析ツールを使い単語に区切る処理を行う。

日本語解析ツールに比べ中国語解析ツールの精度は低く、このことを原因として対訳辞書候補の精度が低くなる課題が想定される。

当社は、独自の研究として一般に入手可能な中国語解析ツールの精度等の比較を行なっている。その結果に基づいて、本調査では、平成25年度調査で使用したツールに比べ解析精度が高いツールを採用した。解析例を以下に示す（図3.2-16）。

図 3.2-16 中国語解析例

原文	根据本发明，通过用形成在薄膜上的 TFT 构成多功能电路而不使用印刷电路板，这样可将重量 <u>轻高性能电路</u> 安装于多种运动器具中。
訳文	本発明によると、薄膜に形成された TFT を用いて多機能回路を構成するがプリント基板を使用せず、 <u>軽量且つ高性能回路</u> を多種類の <u>スポーツ器具</u> に取り付けることが可能である。
以前のツールの解析結果	根据本 发明， 通过 用 形成 在 薄膜 上 的 TFT 构成 多功能 电路 而 不使用 印刷 电路板， 这样 可 将 重量 <u>轻高性能 电路</u> 安装 于 多种 <u>运动器具中</u> 。
本年度使用したツールの解析結果	根据本 发明， 通过 用 形成 在 薄膜 上 的 TFT 构成 多功能 电路 而 不使用 印刷 电路板， 这样 可 将 重量 <u>轻 高性能 电路</u> 安装 于 多种 <u>运动器具 中</u> 。

説明	日本語「高性能回路」に対応する中国語「高性能电路」を抽出する際、以前の解析ツールでは、日本語「軽」を意味する、中国語「轻」と「高性」がつながり、「 軽高性 」と誤って解析されるため、正しい中国語用語「高性能电路」の抽出ができなかった。一方、本年度使用したツールでは「轻///高///性能///电路」（記号///は単語区切り位置を意味する）と正しく単語分割され、「 高性能电路 」の抽出が可能となる。同様に、日本語「スポーツ器具」に対応する「运动器具」も本年度使用したツールでは「 器具 」と「 中 」の間で単語分割されることで抽出可能となる。
----	---

（３）人手確認用対訳辞書候補データの解析

（a）人手確認

前述（２）で作成した人手確認用対訳辞書候補データについて、見出し語及び訳語の人手確認を行い、下記の D1 から D4 のいずれにも該当しないデータを採用して、対訳辞書データを作成した。

- D1 見出し語自体が不適切
- D2 訳語自体が不適切
- D3 見出し語と訳語とが対応していない。
- D4 辞書に登録する用語として不適切

人手確認に当っては、訳語確認書の適切な判断を担保し、作業を効率化するために、自動文アライメント結果である対訳コーパスを訳語確認者に提供した（図 3. 2-17）。

図 3. 2-17 訳語確認者への支援（イメージ）

	A	E	F	H	I
	チェック	日本語	中国語	日本語例文	中国語例文
1					
2	A	排出速度	排放速率	ペレットの 排出速度 は、入手したままの状態で測定し、かつ、それらを状態調節した後に所与温度で所定の持続期間測定する。	原样测量和在给定温度调节它们预定时期之后测量粒料的 排放速率 。
3	A	DNA移行	DNA转移	発現コンストラクトは、それだけでは限らないが、アデノウイルスまたはワクシニアウイルスなどのウイルス感染によってか、またはリボソーム、リン酸カルシウム、DEAEデキストランなどの直接 DNA移行 によって、およびエレクトロポレーションおよびマイクロインジェクションなどの物理的手段によって、哺乳類細胞に移行することができる。	表达构建体可以通过病毒感染(例如但不限于腺病毒或痘苗病毒)或者通过直接DNA转移来转移至哺乳动物细胞。所述直接 DNA转移 如脂质体、磷酸钙、DEAE-葡聚糖以及通过诸如电穿孔和显微注射的物理方法。

また、人手確認を行った人手確認用対訳辞書候補データのうち、対訳辞書データに採用しなかったものについては、上記 D1 から D4 のいずれに該当するかを記録し、「人手確認により対訳辞書から除外したデータ」を作成した。

(b) 人手確認の優先順位及び作成件数

人手確認作業の優先順位は特許庁と協議の結果、複合語より単語を優先し、かつ、出現頻度の値が大きいものから優先して人手確認を行い、人手確認で対訳の妥当性（D1～D4）をチェックして 20 万件の対訳辞書データを作成した。

(c) 技術分野別の作業

人手確認は、中国語・日本語双方に堪能であり、確認対象の用語が属する技術分野についての十分な技術知識を有する者（以下「訳語確認者」）が行う必要がある。このため前述（2）対訳コーパスの解析で付加した技術分野別の出現頻度情報を用いて、人手確認用対訳辞書候補データを技術分野別に分割し作業を行った。

(d) 成果物の品質担保

人手確認結果の品質を担保するため、人手確認を行う人手確認用対訳辞書候補データの内、10%については、異なる 2 名の訳語確認者によって対訳辞書候補データへの採否の判断をそれぞれ行い、2 名の判断の異同を確認し、判断が採用と不採用で異なる用語は再度確認して、いずれの訳語確認者が誤っているかの分析を行った。

また、人手確認を行い採用した対訳辞書データのうち、一つの見出し語に対して N 個以上の対応する訳語が存在するもの、又は一つの訳語に対して M 個以上の対応する見出し語が存在するものについて、N 個以上の訳語又は M 個以上の見出し語の妥当性を再確認した。

<<作業における工夫>>

人手確認用対訳辞書候補データの解析における工夫を以下に示す。

①訳語確認者による作業内容の正確な理解

中日対訳辞書データの最終的な精度は、作業の趣旨及び人手確認の判断基準を正確に理解しているかどうかによって依存する。そのため、各訳語確認者に作業内容を正確に理解させる必要がある。そこで本調査では訳語確認作業の実施にあたり、訳語確認者に本作業の意図を正しく理解させ、また各人の採否基準を揃えるため、専用の手順書を作成し訳語確認者に作業内容を正確に理解させた上で訳語確認作業を実施した。

②支援情報の提供

人手確認用対訳辞書候補データの訳語確認作業を効率的に進めるには、支援情報の提供が不可欠であり、その提供方法についても、訳語確認者が利用し易い方法で提供するように工夫する必要がある。

前述(a)人手確認では自動文アライメント結果である対訳コーパスを訳語確認者に提供した（図 3.2-17）。さらに、訳語確認作業の効率化のため、対訳コーパスの対訳文の、見出

し語と訳語部分を強調表示した。こうすることで、訳語確認者が手間なく支援情報を活用できるようにした。さらに、より多くの例文を参照可能とするために、支援情報として、中日対訳コーパス検索ツールを訳語確認者に提供した（図 3. 2-18）。

図 3. 2-18 中日対訳コーパス検索ツール（画面イメージ）

中日対訳コーパス検索

日本語除外する日本語

中国語除外する中国語

IPC (例) C07H (前方一致)

表示件数

検索

※日本語、中国語、IPCはAND検索です。除外する日本語、除外する中国語はNOT検索になります。

15 件 (15 件中) の検索結果 (1.124 秒)

No.	評価値	番号	J分類	IPC	日本語	中国語
1	0.106	JP2012-196850 CN102689439	M03	B29C65/16	この入射角調整ステップ143は基本的には前記変形例1の入射角調整ステップ142と同様に当該押え板14Aの内面を凹設してテーパ部を形成し、このテーパ部を入射角調整ステップ143の光射出面として構成したものであり、この光射出面を光偏向装置10に向けた方向に傾斜させている。	该入射角调整阶梯143基本上与上述变形例1的入射角调整阶梯142相同,凹设该压板14A的内面,形成斜部,构成为使得该斜部作为入射角调整阶梯143的光射出面,使得该光射出面对准光偏转装置10的方向倾斜。
2	0.080	JP2012-196850 CN102689439	M03	B29C65/16	さらにこの入射角調整ステップ141はテーパ断面形状をしているので、入射角調整ステップ141の内面から出射されるレーザー光Lの出射角、すなわち前面カバー2の外面に入射されるときレーザー光Lの入射角θ2も小さくなる。	该入射角调整阶梯141呈锥形截面形状,因此,从入射角调整阶梯141的内面(背面)射出的激光L的射出角,即入射到前面罩2外面时的激光L的入射角θ2变小。
3	0.119	JP2012-196850 CN102689439	M03	B29C65/16	このとき、押え板14に設けた入射角調整ステップ141は両者の溶着面Rに對向するように位置決めされることは言うまでもない。	这时,当然,设在压板14的入射角调整阶梯141被定位,使其与两者的溶着面R对向。

（４）対訳辞書データの作成

前述（３）に従い作成した対訳辞書データ及び特許庁の既存の対訳辞書データをマージし、「見出し語（中国語）、訳語（日本語）、品詞（見出し語（中国語））、品詞（訳語（日本語））、複合語であるか否かの情報、特許庁が指定する出現頻度情報」が付された形式とした上で、UTX 形式(UTX1. 11)に変換し、中日対訳辞書データを作成した。

本調査における複合語の定義は、見出し語（中国語）が対訳辞書データの中の複数の見出し語のみによって分解でき、かつその訳語が当該複数の見出し語の訳語のみによって分解できる場合を複合語とする。例えば、対訳辞書データに、「見出し語：訳語」の対として「形成：形成」、「区域：領域」という語対がある場合、それらの語対を順に組み合わせた「形成区域：形成領域」は複合語であるとする。但し「形成区域：領域形成」というような語順の異なる語対は複合語でない。本調査では複合語でない語を単語と呼ぶ場合がある。具体的には以下の手順で複合語か否かの情報を作成した。

実際の複合語の判定は特許庁から貸与された中日対訳辞書データ(約 200 万語)に本調査で作成した中日対訳辞書データをマージした約 220 万語の中で行った。下記に処理を箇条書きする。

- ① 中国語の分かち書き：判定対象の見出し語（中国語）を中日辞書の見出しで分かち書きする。分かち書きできない場合は、複合語でない。
- ② 日本語の合成：分かち書きできた場合、分かち書きで得た中国語の見出し語の訳語（日本語）を連結して判定対象の訳語と一致したら複合語とする。

但し、①の中国語の分かれ書きでは、複数種類の分かれ書きが可能なら、その全てについて日本語の合成を行う。②の日本語の合成で辞書の訳語が複数ある場合は、その全ての組み合わせで訳語を判定し、一つでも一致したら複合語とする。以下に複合語と判定した語の例を示す（図 3.2-19）。

図 3.2-19 複合語の例

No.	見出し語	訳語	見出し語の 分かれ書き※	訳語の 分かれ書き※
1	实施方式	実施形態	実施::方式	実施::形態
2	集成电路	集積回路	集成::电路	集積::回路
3	发光二极管	発光ダイオード	发光::二极管	発光::ダイオード
4	合成树脂	合成樹脂	合成::树脂	合成::樹脂
5	复合材料	複合材料	复合::材料	複合::材料
6	制造成本	製造コスト	制造::成本	製造::コスト
7	环氧树脂	エポキシ樹脂	环氧::树脂	エポキシ::樹脂
8	控制系统	制御システム	控制::系统	制御::システム

※複合語を構成する語を::で区切った。

出現頻度情報は、特許庁が仕様により以下を取得した。なお、出現頻度の算出に利用した対訳コーパスは特許庁より貸与された中日対訳コーパスを絞り込んだものと、本調査で作成した中日対訳コーパス（絞込み後）である（表 3.3-6）。

(a) 出現頻度情報：文単位

①全分野、②化学分野、③電気分野、④機械分野、⑤物理分野の各分野に属する対訳コーパスの中で、見出し語及び訳語が共に含まれる対訳コーパス数。

平成 25 年以前に作成した特許庁の既存辞書では文単位の出現頻度を採用している。

(b) 出現頻度情報：文献単位

①全分野、②化学分野、③電気分野、④機械分野、⑤物理分野の各分野に属する対訳コーパスの中で、見出し語及び訳語が共に含まれる対訳コーパス数を文献単位に集約した件数である。例えば、一つの文献から作成した 2 つの対訳コーパスに見出し語及び訳語が共に含まれた場合は 1 件と数える。

3. 3 作成結果

新語の抽出は、前述項番 3.2 抽出方法で示したように、4 つの処理から構成されている。以下に新語データを抽出した際の各処理における統計情報を報告する。

(1) 文献対の解析（対訳コーパス作成）

(a) パテントファミリー及び中国公開公報和文抄録抽出

文献対の解析では、DOCDB に蓄積されているパテントファミリー情報を利用して、2013 年に公開された中国公開特許公報と技術内容が対応する日本公開特許公報（公表・再公表を含む）の番号を取得し、80,087 件の番号リスト（対応中国・日本公開特許公報番号リスト⁶）を作成した（表 3.3-1）。

表 3.3-1 対応中国・日本公開特許公報番号リスト

分 野	文献数
化 学	27,105
電 気	22,299
機 械	15,061
物 理	15,622
合 計	80,087

また本調査では、中国公開特許公報の和文抄録と対応する中国公開特許公報の要約を解析対象とし、252,557 件の番号リスト（対応中国公開公報・和文抄録文献番号リスト⁷）を作成した（表 3.3-2）。

表 3.3-2 対応中国公開公報・和文抄録文献番号リスト

分 野	文献数
化 学	108,394
電 気	48,431
機 械	52,688
物 理	43,040
合 計	252,557

(b) 自動文アライメント

前述(a)で作成した番号リストに基づき、中国と日本の公開特許公報から抽出した発明の名称、要約、請求項、明細書及び中国特許の和文抄録から抽出した中国語と日本語のテキストデータから中日自動文アライメントツールを使い、2,500 万件の中日対訳コーパスを作

⁶ 詳細は、添付資料 3.1 対応中国・日本公開特許公報番号リストレイアウト説明を参照されたい。

⁷ 詳細は、添付資料 3.2 対応中国公開特許公報・和文抄録番号リストレイアウト説明を参照されたい。

成した（表 3.3-3、表 3.3-4）。

表 3.3-3 中日対訳コーパス作成件数（作成箇所別）

作成箇所	公開特許公報から 作成件数	和文抄録から 作成件数	合計件数
発明の名称	80,076	252,557	332,633
要 約	211,949	678,183	890,132
請求項	1,217,645	0	1,217,645
明細書	22,758,538	0	22,758,538
合 計	24,268,208	930,740	25,198,948

表 3.3-4 中日対訳コーパス作成件数（分野別）

分 野	公開特許公報から 作成件数	和文抄録から 作成件数	合計件数
化 学	10,111,925	406,934	10,518,859
電 気	6,246,182	175,717	6,421,899
機 械	3,033,463	189,255	3,222,718
物 理	4,876,638	158,834	5,035,472
合 計	24,268,208	930,740	25,198,948

（２）対訳コーパスの解析（人手確認用対訳辞書候補データの作成）

（a）対訳コーパス（絞り込み後）

前述（１）のデータには文対応の精度が低いデータが混在しているため、全てを利用すると新語抽出の精度を低くする原因となる。そのため、中日対訳コーパスの絞り込みを行った。その詳細は項番 3. 2 抽出手法の（２）中日対訳コーパスの解析を参照されたい。絞り込み後の件数を表 3. 3-5、表 3. 3-6 に示す。

表 3. 3-5 中日対訳コーパス（絞り込み後）（作成箇所別）

作成箇所	公開特許公報から 作成件数	和文抄録から 作成件数	合計件数
発明の名称	0	239, 372	239, 372
要 約	0	406, 373	406, 373
請求項	360, 097	0	360, 097
明細書	13, 201, 326	0	13, 201, 326
合 計	13, 561, 423	645, 745	14, 207, 168

表 3. 3-5 において公開特許公報からの発明の名称と要約を 0 件としている理由は、発明の名称は語数が少なく一部の単語の一致に影響されて誤った対応付けが排除できない可能性があること、要約は日本と中国で記述されている内容が対応しない場合が多いことから、共に処理対象から除外した。

表 3. 3-6 中日対訳コーパス（絞り込み後）（分野別）

分 野	公開特許公報から 作成件数	和文抄録から 作成件数	合計件数
化 学	5, 216, 338	290, 774	5, 507, 112
電 気	3, 780, 555	120, 445	3, 901, 000
機 械	1, 735, 264	123, 789	1, 859, 053
物 理	2, 829, 266	110, 737	2, 940, 003
合 計	13, 561, 423	645, 745	14, 207, 168

(b)日本語データ抽出（名詞／サ変）

前述(a)の中日対訳コーパス（絞り込み後）の日本語文から対訳辞書データとなり得る日本語の名詞及びサ変動詞を抽出した。抽出件数を表 3. 3-7 に示す。

表 3. 3-7 日本語データ抽出（名詞／サ変）件数

分 野	日本語用語（名詞） 抽出件数	サ変動詞 抽出件数
化 学	1, 738, 165 語	3, 578 語
電 気	976, 089 語	2, 916 語
機 械	676, 073 語	2, 905 語
物 理	927, 191 語	3, 057 語
合計 (分野間重複無し)	3, 607, 918 語	4, 591 語

(c)対訳辞書候補データ

前述(b)で抽出した日本語データ（名詞／サ変）に対応する中国語をフレーズテーブルから抽出し、426 万件の対訳辞書候補データ⁸を作成した（表 3. 3-8）。

表 3. 3-8 対訳辞書候補データ作成件数

分 野	対訳辞書候補数
化 学	1, 815, 128 語
電 気	1, 126, 292 語
機 械	757, 012 語
物 理	1, 056, 179 語
合計(分野間重複無し)	4, 264, 418 語

⁸ 詳細は、添付資料 3. 4 人手確認用対訳辞書候補データレイアウト説明を参照されたい。

(d)人手確認用対訳辞書候補データ

前述(c)で作成した対訳辞書候補データから明らかに誤っている用語を除去し、特許庁から貸与された対訳辞書候補データと突き合わせ、重複するデータを排除し、特許庁から貸与された対訳コーパス及び前述(a)で作成した対訳コーパス（絞り込み語）を用いて出現頻度情報を付加した。そして付加した出現頻度の降順（出現頻度の値の大きい順）に並び替え出現頻度上位のおよそ 40 万件⁹について複合語であるか否かの情報を付加し、人手確認用対訳辞書候補データを作成した（表 3.3-9）。

表 3.3-9 人手確認用対訳辞書候補データ件数

分 野	人手確認用 対訳辞書候補数
化 学	171,714 語
電 気	90,758 語
機 械	51,667 語
物 理	85,701 語
合計(分野間重複無し)	399,840 語

⁹ 本調査で作成する対訳辞書データの件数 20 万件と、人手確認用対訳辞書候補データの採用見込み率(60%)と人手確認の 10%を異なる 2 名による採否判断する点を考慮し 40 万件とした。

(3) 人手確認用対訳辞書候補データの解析

(a) 人手確認結果

前述(2)(d)で作成した人手確認用対訳辞書候補データについて、人手による確認を行った。その結果、約40万件の人手確認を行い採用と判断されたデータの優先順上位20万件を対訳辞書として採用した。

また、対訳辞書の作成までに採用とならなかったデータについては不採用理由を記録し、「人手確認により対訳辞書から除外したデータ」を作成した(表3.3-10)。不採用理由毎の件数を以下に示す(表3.3-11)。なお、人手確認の判断フローは、図3.4-1を参照されたい。

表 3.3-10 除外したデータの例

不採用理由	例(上段:見出し語、 下段:訳語)	備考
D1 見出し語(中国語)自体が不適切	地址引 アドレスピン	見出し語「 地址引脚 」の「 脚 」が欠落
D2 訳語(日本語)自体が不適切	表面区 面エリア	訳語「 表面 エリア」の「 表 」が欠落
D3 見出し語と訳語が対応していない	外周側 周囲側面	見出し語「 外周側 」、訳語「 周囲側面 」ともに用語としては問題ないが、下記中国語例文の「表面」が対応していないため不採用。 (中国語例文)「 两个端表面 66 之间的外周側表面 68。 」 (日本語例文)「～の間に延在する 周囲側面 68 を有する。」
D4 辞書に登録する用語として不適切	平行光束光学 平行ビーム光学系	見出し語「平行光束光学」、訳語「平行ビーム光学系」は対応しているが、中国語「平行光束光学系」を機械翻訳する際に、この対訳が使用されると「平行ビーム光学系 系 」と翻訳される悪影響が考えられるため不採用。

表 3.3-11 人手確認により対訳辞書に不採用としたデータ

不採用理由	件数	構成比
D1 見出し語（中国語）自体が不適切	60,838	47.0%
D2 訳語（日本語）自体が不適切	6,428	5.0%
D3 見出し語と訳語が対応していない	40,352	31.1%
D4 辞書に登録する用語として不適切	21,943	16.9%
合計	129,561	100.0%

不採用理由の件数は「D1 見出し語（中国語）自体が不適切」が約半数を占めて最も多く、次いで、「D3 見出し語と訳語が対応していない」、「D4 辞書に登録する用語として不適切」、「D2 訳語（日本語）自体が不適切」であった。

但し、判定手順は人手確認の判断フロー（図 3.4-1）によるため、見出し語が不適切と判断した場合、訳語が不適切かどうかの判断が行わないこととしているため、下段になるほど不採用理由は件数が少なめとなる。

(b)複合語情報と採否判断

人手確認用対訳辞書データ作成では複合語であるか否かの情報を付加した。人手確認の結果を複合語と単語の件数と人手確認の採用率を集計した（表 3.3-12）。

表 3.3-12 複合語情報とその採用率

複合語情報	採用 (A)	不採用 (B)	採用率 $A \div (A+B)$
単語	172,448	125,170	57.9%
複合語	27,552	4,391	86.3%
合計	200,000	129,561	60.7%

集計結果から、単語の採用率は 58%、複合語の採用率は 86%となり、人手確認の複合語の採用率が高い結果となった。

(c)異なる 2 名による採否判断

人手確認作業結果の品質担保のために、人手確認用対訳辞書候補データのうち約 4 万 4 千件について異なる 2 名の訳語確認者によって採否の判断を行った。具体的には、人手確認用対訳辞書データを 1,000 件単位に分割し、これを 1 ファイルとして各訳語確認者が確認作業を行うこととし、この各ファイル内の 10%に当たる 100 件を他のファイルと重複さ

せることで異なる 2 名による採否判断を行った。異なる 2 名による採否判断結果を以下に示す（表 3. 3-13）。

表 3. 3-13 異なる 2 名による採否判断

項目	件数	構成比
採否判断が一致したもの	36, 267	82%
採否判断が一致しないもの	8, 085	18%
合計	44, 352	100%

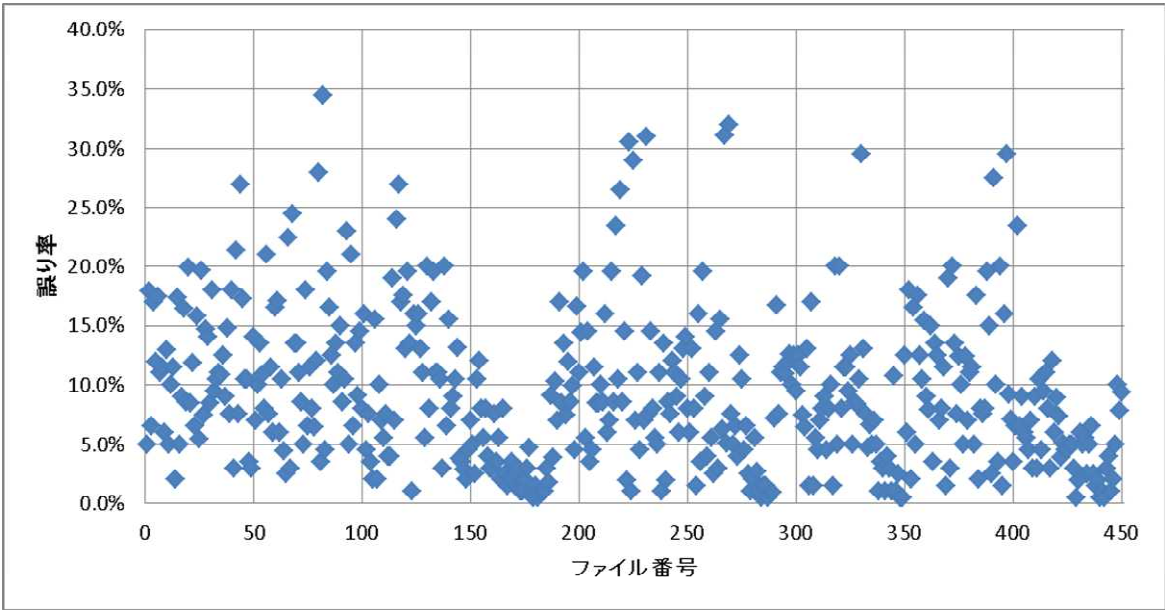
①誤り率の分布

2 名による採否の判断が一致しないものについては、2 名のうちの主担当の訳語確認者が確認結果の見直しを行い、最終的な判定結果を得た。人手確認作業のファイル毎、主担当の訳語確認者による最終採否判断と、当該訳語確認者による最初の採否判断から誤り率を集計した。誤り率は下記に示す式で求めた。誤り率の平均は 9. 2%である。各ファイルの誤り率の分布を図 3. 3-3 に示す。

誤り率
=

最初の採否判断結果と最終採否判断が一致しない数
÷
(
最初の採否判断結果と最終採否判断が一致した数
+
最初の採否判断結果と最終採否判断が一致しない数
+
2 名の採否判断が一致した数
)

図 3. 3-3 人手確認用ファイル毎の誤り率の分布



②誤り率が高いファイル

表 3.3-14 誤り率が高いファイル

ファイル ¹⁰	判定結果納品日	誤り率	フィードバック後の当該訳語 確認者の誤り率（平均）
C.add	2014/12/22	34.5%	11.5%
E.abp	2014/12/25	31.0%	5.8%
E.abh	2015/2/2	30.5%	-（2 度目担当無し）
P.abs	2015/1/22	29.5%	7.7%
E.abj	2015/2/2	29.0%	-（2 度目担当無し）

上記集計で誤り率が高いファイル（表 3.3-14）について、最初の採否判断結果と最終採否判断の不一致の要因を分析した。分析は、最初の採否判断結果と最終採否判断が一致しないものについて、明らかな誤りのものと、その他のものとに分類した。そして明らかな誤りについて、本来採用となるべきものが不採用となったものか、本来不採用となるべきものが採用となったものかを集計した（表 3.3-15）。なお、本調査では下記に示すような分析の結果を元に、気づいた点を訳語確認者へフィードバックし、誤りの減少に務めた。その結果、当該ファイルを担当した訳語確認者の誤り率は比較的低い値になった。表 3.3-14 の誤り率が高いファイルを参照されたい。なお、フィードバック後の当該訳語確認者の誤り率が“-”のセルは、2 度目以降の担当が無かったことを示す。

③誤りの種類

表 3.3-15 誤り内容の内訳

ファイル	誤り数	明らかな誤り	その他の誤り	採用を不採用と誤り	不採用を採用と誤り
C.add	75	64	11	54	10
E.abp	62	51	11	48	3
E.abh	62	32	30	32	0
P.abs	59	14	45	8	6
E.abj	58	28	30	27	1

誤りの内訳数を見るとファイル C.add は明らかな誤りが多数を占め、その他のファイルでは明らかな誤りとその他の誤りがほぼ同様の割合となった。明らかな誤りの判断結果を見ると採用とすべきものを不採用と判断するものが多い傾向が見られた。つまり全体的

¹⁰ ファイル名の先頭 1 桁目は技術分野（C:化学, E:電気, M:機械, P:物理）、ピリオド以降は通し番号（.aaa、.aab、…）を示す。

に判断が厳しい傾向であることが分かった。

④英数字で始まる用語を不採用とした誤り

誤りの内容で、採用を不採用とした誤りは英数字で始まる用語に関する判断ミスが多く見られた。ファイル C. add では明らかな誤り 64 件のうち 41 件、ファイル E. abp では明らかな誤り 51 件中 47 件であった。表 3.3-16 に英数字で始まる採用とすべき用語を不採用とした例を挙げる。

表 3.3-16 英数字で始まる用語を不採用とした例

見出し語	訳語
MRI 数据采集	MRI データ収集
MSE 数据	MSE データ
MYC 蛋白	MYC タンパク質
MnO 浓度	MnO 濃度

⑤見出し語（中国語）が不適切なのに採用

一方、不採用を採用と判断したものは、本来見出し語が不適切なものを採用した誤りが多く見られた。見出し語の用語として中国語の“的”を含む用語は、フレーズに近い単位であるため対訳辞書の見出し語として不採用とすべき用語である。また、見出し語“不透过材料”は用語内の構造が中国語用語の単位としては適切でない構造であるため不採用とすべきだが、採用と判断された（表 3.3-17）。

表 3.3-17 見出し語の誤り例

見出し語	訳語	備考
不透明的液体	不透明液体	見出し語内に“的”を含む
介导的电流	介在性電流	見出し語内に“的”を含む
不透过材料	不透過材料	見出し語の構造が用語として不適切 不(副詞)//透过(動詞)//材料(名詞)

⑥中国語と日本語が対応しないのに採用

上記以外で、不採用とすべき語を採用としたものとして、見出し語“乙型肝炎病毒”、訳語“乙型肝炎ウイルス”があった。これは下記の対訳コーパスから作成された辞書候補であるが、適切な訳語は“B 型肝炎ウイルス”であり不採用とすべき語である。

一种 乙型肝炎病毒 亲和吸附柱及其制备方法	乙型肝炎ウイルス 親和性吸着カラム及びその製造方法
------------------------------	----------------------------------

⑦中国語と日本語の用語の末尾表現を限定した誤り

その他の誤りとして、以下の様な用語末尾の表現に関する判断の相違がみられた（表 3.3-18）。見出し語末尾表現“**元件**”に対する訳語解釈の違いにより、適切な訳語“素子”でないため不採用としたものが見られた、同様な末尾表現として“区域”や“单元”などが存在した。

表 3.3-18 対訳辞書候補の末尾表現

見出し語	訳語
可编程 元件	プログラマブル 要素
交叉阴影 区域	クロスハッチ 部
信标生成 单元	ビーコン生成 部

(d)1つの見出し語に複数の訳語、1つの訳語に複数の見出し語を持つ語の再確認

前述(a)の人手確認で採用した対訳辞書データのうち、1つの見出し語（中国語）に対して2個以上の対応する訳語（日本語）が存在するものは20,906語（見出し語数）あり（表3.3-19）、反対に1つの訳語（日本語）に対して2個以上の対応する見出し語（中国語）が存在するものは11,602語（訳語数）あった（表3.3-20）。

表 3.3-19 1つの見出し語に対して2個以上の対応する訳語が存在するもの

訳語数	見出し語数	訳語数×見出し語数	累計	再確認
24	1	24	24	対象
23	1	23	47	対象
19	1	19	66	対象
18	3	54	120	対象
17	1	17	137	対象
14	4	56	193	対象
13	2	26	219	対象
12	2	24	243	対象
11	5	55	298	対象
10	8	80	378	対象
9	12	108	486	対象
8	21	168	654	対象
7	44	308	962	対象
6	99	594	1,556	対象
5	281	1,405	2,961	対象
4	812	3,248	6,209	対象
3	3,119	9,357	15,566	出現頻度の値が大きいものから再チェック対象が全体でやく2万件になるよう抽出した。
2	16,490	32,980	48,546	対象外
合計	20,906	48,546	－	

表 3.3-20 一つの訳語に対して2個以上の対応する見出し語が存在するもの

見出し語数	訳語数	見出し語 数×訳語 数	累計	再確認
5	2	10	10	対象
4	21	84	94	対象
3	388	1,164	1,258	対象
2	11191	22,382	23,640	出現頻度の値が大きいものから再チェック対象が全体で約2万件抽出した。
合計	11,602	23,640	-	

これら対応する訳語又は対応する見出し語が複数あるデータについて、1つの訳語に対して2語以上対応する見出し語が存在するものについて、訳語数が最も大きいものから4までの6,209語、一つの訳語に対して2語以上の対応する見出し語が存在するものについて、見出し語数が最も大きいものから3までの1,258語、訳語数3と見出し語数2のデータについて、出現頻度の値が大きいものから15,003件の合計22,501万件について妥当性の再確認を行った。再確認を行った結果、19,947件が採用となった（表3.3-21）。

表 3.3-21 複数訳語・複数見出し語の再確認結果

判定結果	1つの見出し 語に複数の訳語	1つの訳語に 複数の見出し語	合計
A 採用	11,073	8,874	19,947
D1 見出し語（中国語）自体が不適切	107	505	612
D2 訳語（日本語）自体が不適切	208	20	228
D3 見出し語と訳語が対応していない	852	329	1,181
D4 辞書に登録する用語として不適切	305	228	533
合計	12,545	9,956	22,501

再確認結果の判定結果を見ると、全体では見出し語と訳語が対応していないものを多く見つけることができた。これは下記例のように対訳コーパス中に類似表現が複数含まれるデータから誤った辞書候補が作成されたものがあった（表3.3-22）。

表 3.3-22 複数訳語・複数見出し語再確認結果例 1

判定結果	見出し語	訳語	対訳コーパス (上段：中国語、下段：日本語)
不採用	管理结构	管理部	结构信息管理部 104 管理结构信息。
			構成情報管理部 104 は、構成情報を管理する。
不採用	変更控制	可変制御	可変控制参数例如是在销售车辆后变更控制参数时使用的参数。
			可変制御パラメータは、例えば、車両が販売されたあとに制御パラメータを変更する場合に用いるパラメータである。

上記例で不採用となった見出し語“管理结构”は、対訳コーパスの中国語“管理结构信息”から作成された。これは対訳コーパスの日本語“構成情報を管理する”の部分に対応するが、誤って“管理部”を訳語として対訳辞書候補としたものである。

1つの見出し語に複数の訳語を持つものの判定結果、1つの訳語に複数の見出し語を持つものの判定結果それぞれを見ると、前者では訳語の誤りを多く見つけることができ、後者では見出し語の誤りを多く見つけることができた。これは以下の例に示すように複数の訳語または複数の見出し語を並べて確認することができ、単独では見逃されたデータの誤りを発見できたと考えられる（表 3.3-23）。

表 3.3-23 複数訳語・複数見出し語再確認結果例 2

判定結果	見出し語	訳語	備考
採用	液压供给系统	油圧供給システム	
不採用		油圧供給系	訳語“統”の欠落
不採用	冷却空间	過冷却スペース	見出し語“过”の欠落
採用	过冷却空间		

(4) 対訳辞書データ作成

(a) 複合語件数

前述(2) 対訳コーパスの解析で作成した対訳辞書データ及び特許庁から貸与された対訳辞書データについて、複合語であるか否かの情報を付した。それぞれの対訳辞書データの複合語の語数を表 3.3-24 に示す。なお、平成 26 年度の複合語の割合が低いのは単語を優先して採用したためである。3.2 の(2) を参照されたい。

表 3.3-24 複合語件数

対訳辞書データ	複合語	複合語以外	合計	複合語 の割合	複合語 以外の 割合
平成 24 年度	404,857	596,390	1,001,247	40.4%	59.6%
平成 25 年度	380,450	620,108	1,000,558	38.0%	62.0%
特許庁追加分	133	8,356	8,489	1.6%	98.4%
平成 26 年度	27,043	172,957	200,000	13.5%	86.5%
合計	812,483	1,397,811	2,210,294	36.8%	63.2%

(b) 出現頻度

前述(a)で複合語であるか否かの情報を付したデータについて、特許庁から貸与された対訳コーパス¹¹及び本調査で作成した対訳コーパス¹²における出現頻度を付した。出現頻度は文単位と文献単位 2 種類の出現頻度を付した。文単位出現頻度は、全分野 と 化学分野、電気分野、機械分野、物理分野の各分野に属する対訳コーパスの中で、見出し語及び訳語がともに含まれる対訳コーパス数となる。文献単位出現頻度は、文単位出現頻度と同様に全分野及び各分野の出現頻度となり、それらの出現頻度値は見出し語及び訳語がともに含まれる中国公開特許公報と日本公開公報又は和文抄録との文献対数となる。

また、本調査で作成した対訳辞書データについて、複合語以外を優先し、文献単位出現頻度値の大きいものから 20 万件を対訳辞書データとした。以下に文献単位出現頻度毎の辞書データ語数の分布を示す(表 3.3-25、図 3.3-4)。

¹¹ 2010～2012 年発行の中国公開特許公報の中日対訳コーパス(約 6,600 万文対)から絞り込んだ 3,808 万文対を利用した。

¹² 対訳コーパスについては、(2) 対訳コーパスの解析(人手確認用対訳辞書候補データの作成)(a)対訳コーパス(絞り込み後)を参照されたい。

①出現頻度と語数の関係

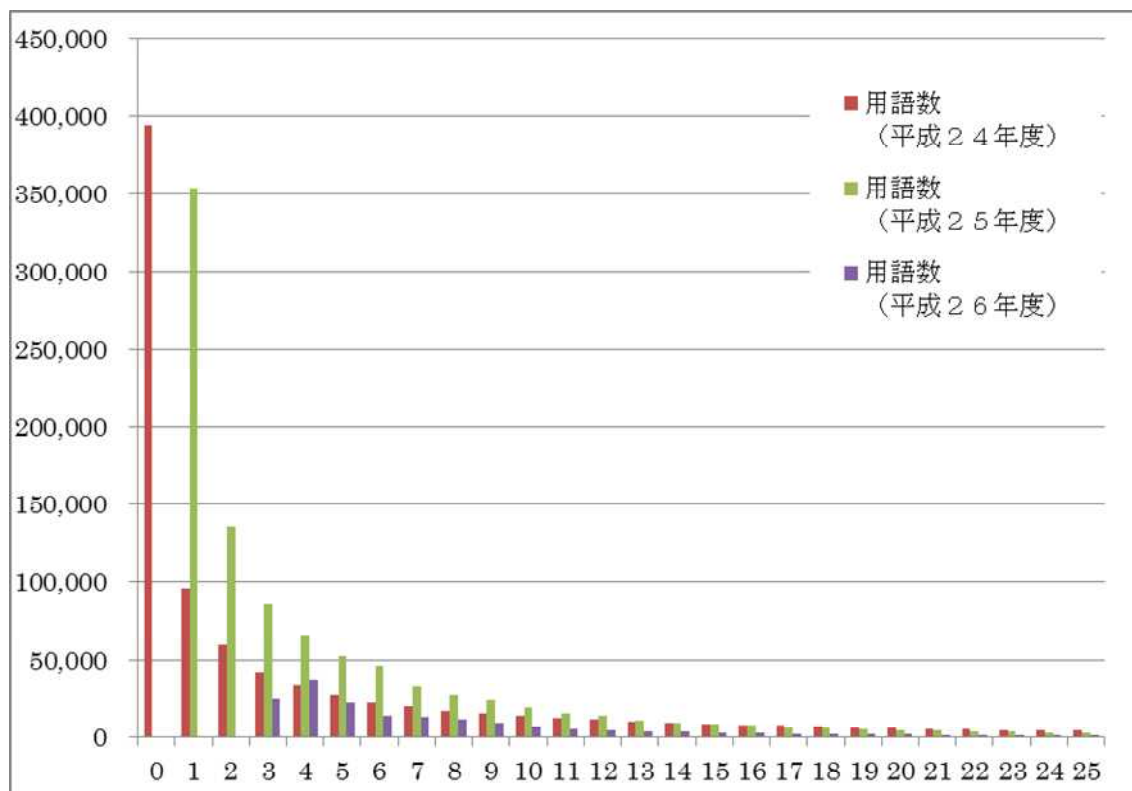
表 3. 3-25 出現頻度分布（文献単位）

出現頻度 （文献単位）	用語数 （平成 24 年度）	用語数 （平成 25 年度）	用語数 （平成 26 年度）
501～	16, 161	1, 192	1,527
451～500	1, 444	147	191
401～450	1, 724	160	241
351～400	2, 225	242	319
301～350	2, 943	330	431
251～300	3, 778	526	615
201～250	5, 472	772	878
151～200	8, 583	1, 430	1,365
101～150	15, 657	3, 458	2,773
51～100	38, 962	13, 320	7,291
41～50	16, 407	8, 008	3,298
31～40	24, 443	14, 291	5,209
21～30	40, 447	27, 866	9,551
11～20	82, 787	86, 033	28,140
1～10	346, 077	842, 783	138,171
0	394, 137	0	0
合計	1, 001, 247	1, 000, 558	200,000

②作成年別の出現頻度と語数の関係

図 3.3-4 対訳辞書データ出現頻度分布（文献単位）

（横軸：出現頻度 0 から 25、縦軸：語数）



平成 24 年度辞書データは出現頻度ゼロの用語数が 4 割弱を占めた。これは本調査で使
用した対訳コーパス（2010～2013 年）と平成 24 年度調査で使
用した対訳コーパス（2005～2009
年）の範囲が重ならなくなったためと思われる。つまり、平成 24 年度調査で使
用した対訳
コーパスのみに出現する用語が多く含まれていたと考えられる。

実際に出現頻度ゼロ用語内容を確認してみたところ、中国語“冷冻箱”、日本語“冷凍
ケース”の辞書データの出現頻度は、平成 24 年度の辞書データで 183 件であったものが、
本調査で使
用した対訳コーパスにおいては 0 件であった。なお、中国語“冷冻箱”単独の
出現頻度は 317 件存在し、日本語との組み合わせでは表 3.3-26 の件数があった。

表 3.3-26 中国語“冷冻箱”と日本語の組み合わせ

冷凍庫	133 件
冷蔵庫	152 件
フリーザ	15 件
その他	17 件

その他には、表 3-3-27 に示すように、中国語“冷冻箱”に対応する日本語が存在しない対訳コーパスが存在した。

表 3. 3-27 中国語“冷冻箱”に対応する日本語が存在しない対訳コーパスの例

因此在 冷冻箱 内, 内壁 16 形成一个支架面 52。
したがって、区画の内部では、内壁 16 は棚面 52 を形成する。

“冷冻箱”の訳語として“冷凍ケース”は比較的マイナーな訳語であり、本年度調査で
使用した対訳コーパスには含まれなかったと考えられる。

③作成年別の辞書の特徴

表 3. 3-25 出現頻度分布（文献単位）の出現頻度値が大きいものを見ると、出現頻度 1, 001
件以上の用語が 9, 185 語あった。このうち 5, 015 語は平成 24 年度に作成した辞書の基本辞
書に含まれる用語であった。

平成 25 年度辞書データは、出現頻度 0 は無いものの、出現頻度 1 から 50 の比較的低頻
度の用語が多数を占めた。これら低頻度の用語のうち出現頻度 1 の用語を確認すると、中
国語“更新模組”、日本語“更新モジュール”が平成 25 年度の辞書では 649 件であったが、
本調査の文献単位の出現頻度では 1 となった。中国語“更新模組”、日本語“更新モジュ
ール”を含む対訳コーパスを確認すると、公開番号 JP2010-092830 と、CN102272770 の文献対
から作成された対訳コーパスであった。つまり、文単位の出現頻度では 649 件と比較的高
頻出の用語となるが、文献単位出現頻度で見ると 1 となり出現頻度の低いデータとなる（図
3. 3-5 の「番号」列を参照）。

図 3. 3-5 「更新模組：更新モジュール」の検索結果

中日対訳コーパス検索

日本語 更新モジュール

除外する日本語

(例) エンジン

中国語 更新模組

除外する中国語

(例) 引擎

IPC

(例) C07H (前方一致)

表示件数 100

検索

※日本語、中国語、IPCはAND検索です。除外する日本語、除外する中国語はNOT検索になります。

100 件 (649 件中) の検索結果 (0.518 秒)						
No.	評価値	番号	J 分類	I P C	日本語	中国語
1	0.081	JP2010-092830 CN102272770	P05	G06F11/00	そのため、次の相互監視処理において、一度不正な更新モジュールと特定された更新モジュールを、正常な更新モジュールと判定する更新モジュールは、不正な更新モジュールである。	因此,在接着的相互监视处理中,将一次确定为非法的更新模組的更新模組判断为正常的更新模組的更新模組是非法的更新模組。
2	0.082	JP2010-092830 CN102272770	P05	G06F11/00	更新モジュール群130に含まれる更新モジュール(更新モジュール133を除く)は、更新モジュール133を更新用更新モジュールに更新する。	包含在更新模組群130中的更新模組(除了更新模組133以外)将更新模組133更新为更新用更新模組。
3	0.115	JP2010-092830 CN102272770	P05	G06F11/00	更新モジュール群130は、複数の更新モジュールから構成される。	更新模組群130由多个更新模組构成。
4	0.086	JP2010-092830 CN102272770	P05	G06F11/00	また、更新モジュール131は、更新モジュール132を監視し(3028)、更新モジュール132は、更新モジュール133を監視し(3029)、更新モジュール133は、更新モジュール131を監視(3030)しているものとする。	此外,假设更新模組131监视更新模組132(3028)、更新模組132监视更新模組133(3029)、更新模組133监视更新模組131(3030)。

文単位の出現頻度と文献単位の出現頻度による影響を見るために、その他の出現頻度 1 の辞書データについて確認したところ、中国語“介入位点”、日本語“介在部位”も文献対（JP2011-528275, CN102159257）から作成された対訳コーパスであり、文献単位の出現頻度では 1 であった。このような特定の文献対のみに出現する用語は汎用性の低い用語と見ることができる。平成 25 年度の辞書データについて、このような文単位の出現頻度が 2 以上かつ文献単位の出現頻度が 1 となる用語数を集計したところ、約 35 万件存在した。

ここまでの調査で、平成 24 年度の辞書データは対訳コーパスの範囲を超えて出現頻度が大きい語を含みかつ出現頻度が小さい語（出現頻度 1～50 文献）が 51.0% であるのに対して、平成 25 年度の辞書データは出現頻度が大きい語が少なく、出現頻度が小さい語が 97.8% を占めることが分かった。

平成 26 年度辞書データは表 3.3-25 の表を見ると、平成 25 年度の辞書データと同様の傾向が読み取れる。なお、辞書データの語数が、平成 25 年度の 100 万語に対して平成 26 年度は 20 万語と大きく異なるので、出現頻度の分布の割合を円グラフにした。図 3.3-7、図 3.3-8 に示すように、出現頻度が小さい語が大半を占め、平成 25 年度の辞書データと同様の分布傾向となった。また、前述表 3.3-25 の出現頻度が比較的高い語（出現頻度 101 以上）の語数を見ると、平成 25 年度の辞書データ（8,257 件）と平成 26 年度の辞書データ（8,381 件）は同等の件数となった。

図 3.3-7 平成 25 年度の辞書データの出現頻度分布

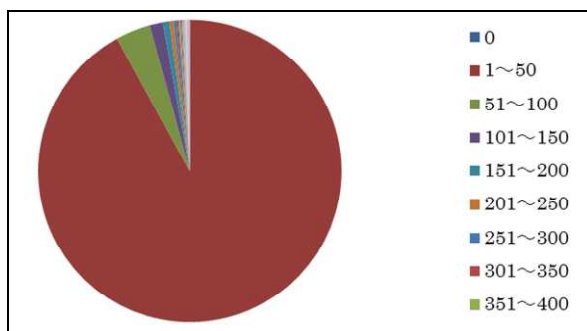
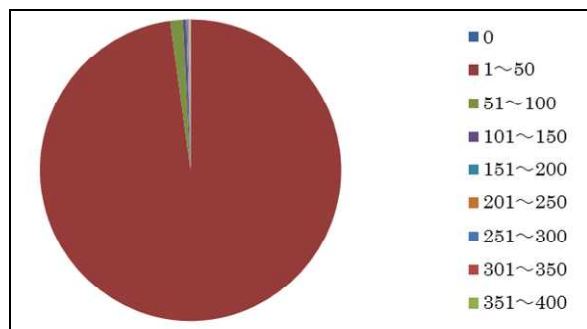


図 3.3-8 平成 26 年度の辞書データの出現頻度分布



④技術分野別の文献単位出現頻度

本調査で付与した出現頻度は各技術分野に属する対訳コーパスの中での出現頻度を集計している。今年度作成した辞書データについて技術分野別の文献単位出現頻度¹³（出現頻度1以上）の分布を集計した（表 3. 3-28）。

表 3. 3-28 技術分野別の文献単位の出現頻度

文献単位の 出現頻度	化学	電気	機械	物理
501～	696	388	154	198
451～500	99	44	19	25
401～450	103	66	22	47
351～400	155	93	38	42
301～350	186	104	42	60
251～300	297	143	59	76
201～250	426	219	90	149
151～200	659	369	185	249
101～150	1,314	767	347	550
51～100	3,427	2,301	1,032	1,611
41～50	1,535	1,052	513	691
31～40	2,516	1,535	864	1,377
21～30	4,832	2,887	1,552	2,370
11～20	13,098	7,723	4,387	6,474
1～10	106,928	77,684	65,106	83,061
合計	136,271	95,375	74,410	96,980

¹³ 同一の用語が複数の分野の対訳コーパスに存在するため、技術分野毎の文献単位の出現頻度は技術分野間で重複している。

3. 4 問題点・課題

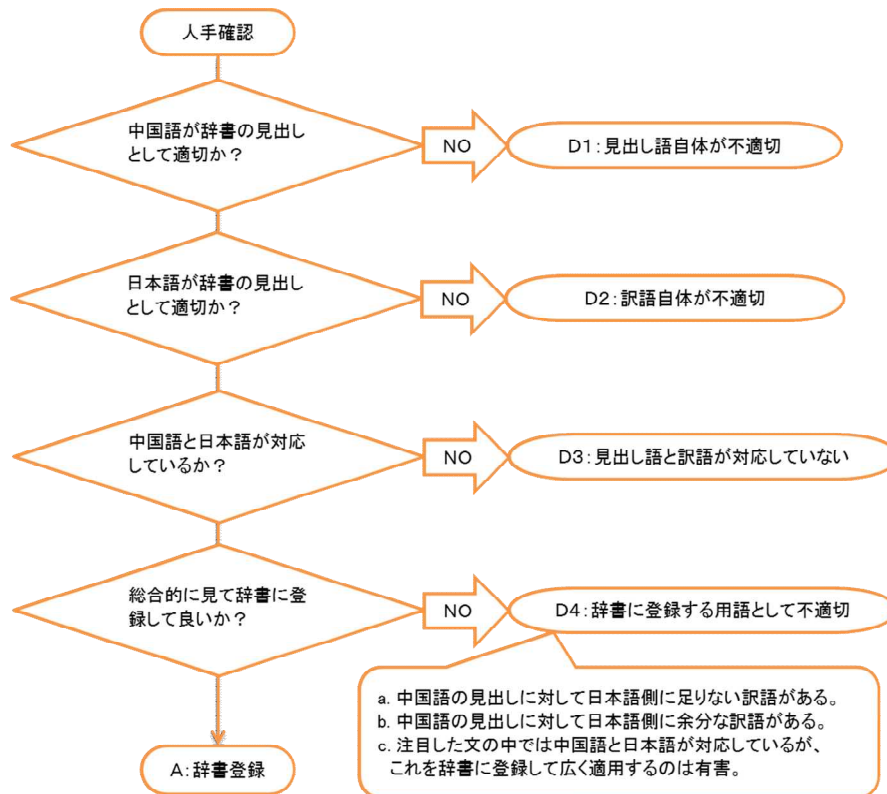
対訳辞書データ作成の問題・課題は主に、人手確認用対訳辞書候補データの手確認に関わる部分で生じた。(1)の採否判定作業手順では、今年度の調査から始めた不採用理由の細分化に関する手順の課題、(2)の複合語と複合語以外の採用率、(3)複数訳語または複数見出し語をもつ用語の確認は複合語情報や複数訳語または複数見出し語をもつ用語に対する人手確認作業への提案、(4)不適切な見出し語への対応では不採用理由の多数を占めた不適切な見出し語に対する課題、(5)対訳コーパスの整備では辞書候補に適さない対訳コーパス、(6)人手確認作業の支援情報では効率的な人手確認を行うための課題、(7)辞書データの機械翻訳システムでの利用方法では、辞書データを機械翻訳システムで使用する際の課題を記載した。

(1) 採否判定作業手順

人手確認用対訳辞書候補データ解析では、訳語確認者が見出し語及び訳語の手確認を行い、以下の D1 から D4 のいずれにも該当しないデータを採用として、対訳辞書データを作成した。作業上、A を採用(辞書登録)、D1～D4 を不採用とした。

- D1 見出し語（中国語）自体が不適切
- D2 訳語（日本語）自体が不適切
- D3 見出し語と訳語とが対応していない
- D4 辞書に登録する用語として不適切

図 3. 4-1 人手確認判断フロー



本調査では一組の見出し語と訳語の確認を一人の訳語確認者が図 3. 4-1 のフローの手順に従って判定した。作業後、一部の訳語確認者より「一度に不採用理由の判断を行うのは効率が悪く感じた。例えば、採用・不採用を初めに判断し、その後、再確認作業として不採用と判断したものの不採用理由判断を判断したほうが効率的である。」とのコメントがあった。さらに効率的な作業手順の検討が必要であると感じられた。

(2) 複合語と複合語以外の採用率

前述 3. 3 (3) (b)では、複合語情報と採用不採用の関係を調査した。調査の結果、複合語は 90%前後、複合語以外はおよそ 60%の採用率であった。この結果から、例えば、複合語以外では本調査で行ったような人手確認を行い、複合語については少し簡易な確認手順にするなどの検討が考えられる。

また、コストの面では今回のように辞書に入れる複合語の優先順位を下げるとコスト増につながる。その理由は、採用率が高い複合語の比率が下がることで、人手確認する語の採用率が下がるため、複合語と単語の優先順位を同じにする場合に比べて、多くの辞書候補データを確認する必要性が生ずるためである。

（３）複数訳語または複数見出し語を持つ用語の確認

前述 3. 3（3）(d)では、1つの見出し語に複数の訳語を持つものや、反対に1つの訳語に複数の見出し語を持つものについて再確認作業を行った。その結果、効率的に誤りを発見することができ、作業の有用性が確認できた。本調査では今回作成した対訳辞書候補内で再確認を行ったが、特許庁から貸与された対訳辞書データ内にも複数の訳語や見出し語を持つデータが存在するため今後の辞書作成では、既に作成済の対訳辞書候補を含めた確認作業も可能と考えられる。

（４）不適切な見出し語への対応

前述 3. 3（3）(a)では、人手確認結果について不採用理由別に件数を集計した。その結果、不採用理由「見出し語（中国語）自体が不適切」が全体の 47%を占める結果になった。見出し語が不適切になる主な要因は中国語の解析処理時の誤りが考えられるため、本調査では、平成 25 年度より高精度の中国語解析ツール¹⁴を使用した¹⁴が、さらなる高精度化が望まれる。しかし、中国語解析処理のさらなる精度向上は中国語及び自然言語処理の高度な専門的知識と経験が必要となるため長期的な課題となる。

¹⁴項番 3. 2（2）＜＜作業における工夫＞＞(e)中国語解析ツールの選定を参照されたい。

(5) 対訳コーパスの整備

人手確認用対訳辞書候補データで誤りと判定されたものの中には下記に示すような対訳コーパス自体の問題により誤った訳語候補が作成されたものが存在した（表 3.4-1）。

表 3.4-1 対訳コーパスの問題例

見出し語	訳語	対訳コーパス (上段：中国語、下段：日本語)
太阳能电池模块	太陽電池モジュ	因此, 耐热性作为太阳能电池模块的填充材料层所具有的特性是必不可少的。
		このため、耐熱性は、太陽電池モジュ-ルの充填材層の有する特性として必要不可欠のものである。
电磁场	電気場	该专利提出, 可以通过修改一个或多个特性信号参数来减小有害影响, 这些特性信号参数包括以下至少一项: 生命系统所暴露于的环境时变的电、磁或电磁场。
		この特許は、生体系が露出される周囲の時間的に変動する <u>電気場、磁気場、又は電磁気場</u> のうちの少なくとも1つを含む1以上の多くの固有信号パラメータを修正することによって有害な影響を低減することができると提言している。

上の例は対訳コーパスの日本語の“太陽電池モジュ”に後続する長音記号が日本語の長音でなくマイナス記号であるため訳語が正しく作成できなかった例である。下の例は中国語コーパスの下線部分の並列表現が省略された表現であるのに対し、日本語コーパスの下線部分の並列表現は省略されていない表現であることで正しい対訳候補が作成できなかった例である。このような辞書候補作成に適さない表現を収集・識別して、対象外とする対訳コーパスが整備できれば、さらに辞書候補作成の精度が向上すると考えられる。

（６）人手確認作業の支援情報

人手確認用対訳辞書候補データの解析では訳語確認者に支援情報を提供した。支援情報は２つあり、１つは人手確認用対訳辞書候補データそのものに見出し語と訳語を同時に含む対訳コーパスのデータを例文として提供した。もう１つは、見出し語と訳語を検索できる支援画面¹⁵を提供した。支援画面について、その利用状況を確認し、人手確認作業の効率化に有効であったかを確認した。利用状況を調べたところ、のべ６名の訳語確認者からの利用があり、利用回数は約 5,000 回であった。利用回数が多くなかった理由を訳語確認者に確認したところ、人手確認作業の期間が限られていたため、積極的な利用に至らなかったと回答を得た。また、訳語確認者による訳語判断の際のコメントを見ると英語の訳を参考にしているものが見られた。このような点を考慮し、より適切な参考情報の検討も必要であると考えられる。

（７）辞書データの機械翻訳システムでの利用方法

本調査で作成した辞書データは、ルールベース機械翻訳システムのユーザ辞書へ追加することで機械翻訳結果の品質向上に役立つと考えられる。しかしながら、ルールベース機械翻訳システムは、システムが備える基本用語辞書、専門用語辞書及びこれまで特許庁が作成してきた辞書データと連動して訳語が決められるため、本調査で作成した辞書データの一部について、対訳データとして誤りではないものの他の辞書データとの関係で機械翻訳結果に悪影響（副作用）を及ぼす可能性があるデータも含まれている。

例えば本調査で作成した辞書データ“光层（光層）”は、下記に示す文において有効に機能する（表 3.4-2）。

表 3.4-2 “光层”が正しく翻訳され则认为られる例

此时, 如图 11 所示, 照射模式 210 包括两个光层: 第一光层 212 和第二光层 214。
このとき、照射パターン 210 は、図 11 に示すように、第 1 の光層 212 と、第 2 の光層 214 との 2 つの光層からなる。

一方、表 3.4-3 のような文を機械翻訳する際、辞書データに“发光层（発光層）”が存在すれば最長一致で問題なく正しい訳語が選択され则认为られる。しかし、仮に辞書データに見出し語“发光层”が存在せず、見出し語“光层”が存在した場合、機械翻訳の中国語解析処理で“发”，“光层”と単語分割され一文字に区切られた“发”が 1 文字の語と認識される可能性が高い（表 3.4-3）。なお、“发”は動詞になり得る語なので影響が大きい、ここでは単純な未知語とみなされたと仮定する。

¹⁵項番 3. 2 （３）図 3.2-16 を参照されたい。

表 3.4-3 “光层”の副作用のハンドシミュレーション

可将第三发光层邻接于第一发光层。
“发光层（発光層）”と“光层”がある場合の形態素解析の想定 可将第三 发光层 邻接于 第一 发光层 。
“光层”だけが存在する場合の形態素解析の想定 可将第三 发 光层 邻接于 第一 发 光层 。

このような副作用の可能性のある辞書データは、複数の見出し語を組み合わせる際の曖昧性と機械翻訳エンジンの仕様による。このような副作用は辞書データ単独では判断できないため、本調査の辞書データ作成方法ではこのような副作用の有無は考慮していない。

この課題に対する対応案は、副作用の可能性のある用語を機械翻訳システムのユーザ辞書に登録しないことが考えられる。但し、辞書データ1つ1つについて副作用の有無を確認するのは現実的でないため、出現頻度の高い見出し語に限定して実際に使用する機械翻訳システムでの副作用の有無を確認するなどの方法で副作用の可能性のある用語を選定することが考えられる。

一方、先の例の様に、この削除により未知語が発生する可能性もある。最終的に機械翻訳結果に未知語が残ってしまった場合、機械翻訳の後処理として未知語となった部分と別に用意した対訳辞書データの見出し語を突き合わせて辞書データの訳語で置換する（かぶせ処理）ことが考えられる（表 3.4-3）が、この方法での対応は限定的である。

表 3.4-3 かぶせ処理のイメージ

原文	此时, 如图 11 所示, 照射模式 210 包括两个光层: 第一光层 212 和第二光层 214。
機械翻訳結果 (下線が未知語)	このとき、照射パターン 210 は、図 11 に示すように、第 1 の光层 212 と、第 2 の光层 214 との 2 つの光层からなる。
かぶせ処理後	このとき、照射パターン 210 は、図 11 に示すように、第 1 の光層 212 と、第 2 の光層 214 との 2 つの光層からなる。

4. 対訳辞書データの検証及び翻訳品質評価手順に関する調査

4. 1 評価の目的と概要

対訳辞書について、その効果を適切に評価して改善を図っていく必要がある。特許庁は機械翻訳の品質を評価するため、平成 25 年度に「特許文献機械翻訳の品質評価手順」を作成した。この評価手順は、英語から日本語への翻訳事例をベースにしており、手順書を使用していく中で複数の言語の事例を追加する等、適宜改定していくことを想定している。

本調査では、この品質評価手順と、特許庁が作成した中国語から日本語の機械翻訳結果の評価に用いる翻訳品質評価用テストセット 150 文を複数の方法で機械翻訳して評価者が翻訳結果の相対評価を行い、辞書データによる翻訳品質向上の効果を把握すること、対訳辞書の効果的な使い方を探ること及び中国語から日本語への機械翻訳の課題と知見を品質評価手順に加えることを目的とする。

翻訳品質評価用テストセット 150 文を機械翻訳する複数の方法として、分野別の中日辞書で中国語に対する訳語（日本語）を一つ選定する際、下記の 2 つの出現頻度に基づいて作成したそれぞれのユーザ辞書による機械翻訳結果を比較した。

- ① 中日対訳コーパスにおける文単位の出現頻度による選定
- ② 中日対訳コーパスにおける文献単位の出現頻度による選定

この二つの方法を選定した理由は、過去 2 年に渡って作成した約 200 万件の中日辞書から、機械翻訳のためのユーザ辞書を作成したが、文単位の出現頻度による訳語の選定を行ったところ、時々、期待されない訳語が採用されたためである。この原因は少数のパテントファミリーに期待されない訳語が頻繁に出現したため、一般的な訳語より出現頻度が大きくなったためである。このような少数のパテントファミリーの期待されない訳語の影響を軽減するため、1 つの文献に何回出てきても 1 回とカウントする文献単位の出現頻度を利用することが考えられた。そこで、この二つの方法で作成した機械翻訳辞書による翻訳結果を比較してどちらの方法が望ましいか検討する。

なお、実際にこの二つの辞書による比較を行ったところ機械翻訳結果に違いがあったのは 150 文中 61 文であった、違いが無い 89 文については、ユーザ辞書ありと無しの結果を比較・分析して中国語から日本語への機械翻訳の課題と知見を得るための素材とした。

4. 2 評価対象文の作成環境

(1) 評価用テストセット

中日機械翻訳の品質を評価するために、評価用テストセットとして特許庁の翻訳品質評価用評価対象文 150 文（テストセット）を使用した。この評価用テストセットは図 4.2-1 のような Excel ファイルである。A 列からセクション、IPC、文献番号、出願番号、通し番号、評価対象文（原文、参考人手翻訳）で構成される。なお、H 列の CEMP は IPC によって C：化学、E：電気、M：機械、P：物理（情報）の分野を付与した記号である。

各行は中国と日本のパテントファミリーの特許明細書から抽出された対訳文である。各文献から 3～4 文程度が採用されている。

図 4.2-1 特許庁の品質評価用評価対象文（全行に番号を補う等、整形済み）

	A	B	C	D	E	F	G	H
	セクション	IPC	文献番号	出願番号	通し番号	評価対象文(原文)	参考人手翻訳	CEMP
1								
2								
3	A	A61P 17/16	CN-A-101125155	200710049762.0	1	本発明用途之一的化粧品可以是包括软膏剂、溶液、乳液、乳膏、乳液、化妆水、洗剂、凝胶、美容液、粉底霜、填充物、面膜、口红、粘剂、淋溶液等各种皮肤外用剂在内的任何一种形式的化妆品。	当発明の用途の一つめである化粧品は、軟膏剤、溶液、クリーム、乳液、化粧水、洗剤、ゲル、美容液、ファンデーション、スタッフィング、フェイシャルパック、口紅、スティッキー、エージェント、ボディソープ等各種の皮膚外用剤を含むいかなる種類の形態の化粧品を含むことが出来る。	C04
4	A	A61P 17/16	CN-A-101125155	200710049762.0	2	本试验结果表明, 无论加与不加S9的试验, 自发对照组的每皿平均回变菌落数均在正常值范围内, 而阳性对照组诱发的每皿平均回变菌落数均为自发对照组二倍以上, 呈明显阳性反应。	当該試験結果により、S9を加えるか否かに拘わらず、自発対照群の各プレートの変異体コロニー平均数はいずれも正常値の範囲内にあり、陽性対照群が誘導した各プレートの変異体コロニー平均数は自発対照群の二倍以上あり、明らかに陽性反応を示した。	C04
5	A	A61P 17/16	CN-A-101125155	200710049762.0	3	针对现有技术的美白祛斑活性成分存在的不足, 本发明的目的旨在提供一种更为有效与安全的 Whitening 活性成分—灯盏花甲醇提取物, 以制备皮肤美白或祛斑化妆品、抑制酪氨酸酶活性或/和黑色素生物合成的药物组合物和抑制酪氨酸酶活性或/和黑色素生物合成的食品。	既存技術の美白・シミ取り活性成分の存在不足に対して、当発明の目的は更に効果的で安全な美白・シミ取り活性成分であるムカシヨモギメタノール抽出物を提供し、皮膚の美白或いはシミ取りの化粧品を調製し、チロシナーゼ活性及び/或いはメラニン生物合成による色素生成物を抑制し、チロシナーゼ活性及び/或いはメラニンの生物合成を抑制することである。	C04
6	A	A61P 17/16	CN-A-101125155	200710049762.0	4	就灯盏花而言, 其作为原料使用的部位不受特殊限制, 可以是茎、叶、花、果实等, 而且可以采集后立即提取, 也可以干燥后再提取, 优选将茎干燥后提取。	ムカシヨモギについて言えば、その原料として使用する部位は特別の制限を受けるものでなく、茎、葉、花、果実等であって、採取後すぐに抽出することが可能であり、乾燥後でも再度抽出することが可能であるが、茎を優先的に乾燥してから抽出を行う。	C04
7	A	A61K 35/78	CN-A-1686239	200510020655.6	5	故本研究将黄原胶和瓜儿豆胶进行复配, 将凝胶剂总用量、黄原胶与瓜儿豆胶的比例以及溶液温度这三个对本凝胶性能影响最大的因素作为考察对象, 采用凝胶强度测定仪分别测得不同条件所得凝胶的强度。	本研究ではキサンタンガムとグアーガムを配合し、ゲル製剤の容量、キサンタンガムとグアーガムの割合及びゲルの溶解温度といった3つが本ゲル製剤の性能に対して及ぼす最大の要素を考察の対象とし、ゲル強度測定装置を採用して異なる条件の下で得られるゲルの強度を測定した。	C12
8	A	A61K 35/78	CN-A-1686239	200510020655.6	6	原因可能是加入一定量的盐后, 盐的正负离子(Na ⁺ , Cl ⁻)把带相反电荷的-NH(CH ₃) ₃ 和-COO ⁻ 包裹起来, 削弱它们之间较强的静电作用力, 间接通过较弱的盐键以及氢键等加强凝胶强度。	その原因は一定量のナトリウムを加えた後、ナトリウムのプラス・マイナスイオン(Na ⁺ , Cl ⁻)が反対の電荷を帯びた-NH(CH ₃) ₃ と-COO ⁻ を覆い、それらの間の比較的強い静電作用力を弱め、間接的に比較的弱いナトリウムナトリウム結合及び水素結合等を通じてゲル強度を高めていることによる。	C12
9	A	A61K 35/78	CN-A-1686239	200510020655.6	7	末次给药后1h, 在每只大鼠右后足跖皮下注射蛋清0.1ml/只, 测量注入致炎剂后1、2、3h右后足跖体积, 减去致炎前大鼠右后足跖正常体积, 作为各鼠不同时间的足跖肿胀度。	最後の薬を与えてから1時間後、各ラットの右側後脚の皮膚に卵白0.1ml/匹を注射する。注射後炎症を起してから1、2、3時間後の右側後脚の体積を測定し、炎症を起す前のラットの右側後脚の正常な体積を差し引いて、各ラットの異なる時間における後脚の腫脹度を求める。	C12

(a) 評価対象文の分野

評価用テストセット 150 文の 4 分野の内訳は以下である。

表 4.2-1 評価用テストセットの分野別の文数

分野	文数
C：化学分野	60 文
E：電気分野	27 文
M：機械分野	41 文
P：物理分野	22 文

(b) 翻訳品質評価シート

評価作業を効率的に進めるために、1文1シートの翻訳品質評価シートを作成して評価作業を進めた。シートの番号と評価用テストセットの通し番号は一致している。

評価対象文に参考人手翻訳がつけられており、5名の評価者は評価対象文（原文）と参考人手翻訳の日本語訳を参考にして機械翻訳結果の品質を評価した。ただし、参考人手翻訳の日本語訳はパテントファミリーの明細書から採取された対訳であるため直訳でない場合がある。さらに文を詳しく見ると正しく翻訳されていない箇所もある。そのため、参考和訳は評価にあたって意味が正しくなるように修正した。表 4.2-2 に翻訳品質評価シートの例を示す。

表 4.2-2 翻訳品質評価シート 3

文番号	CN-A-101125155 200710049762.0	
分野	C04 A61P17/16	
中国語原文	针对现有技术的美白祛斑活性成分存在的不足，本发明的目的旨在提供一种更为有效与安全的美白祛斑活性成分-灯盏花甲醇提取物，以制备皮肤美白或祛斑化妆品、抑制酪氨酸酶活性或/和黑色素生物合成的药物组合物和抑制酪氨酸酶活性或/和黑色素生物合成的食品。	
参考和訳	□従来既存技術の美白・シミ取り活性成分にの存在する不足に鑑み対して、<u>本</u>発明の目的は、<u>更</u>に効果的で安全な美白・シミ取り活性成分であるムカシヨモギメタノール抽出物を提供し、 皮膚の美白或いはシミ取りの化粧品を調製し、 チロシナーゼ活性及び/或いはメラミン生物合成を抑制するによる医薬薬物組成物を抑制し、 チロシナーゼ活性及び/或いはメラミンの生物合成を抑制する食品の製造に用いることである。	
機械翻訳 1	機械翻訳 1	先行技術の美白シミ取りの活性成分存在する不足に対して、本発明の目的は提案がもっと有効にセキユリティの美白シミ取りの活性成分-油ランプとメタノール抽出物を使いを目指して、調製の皮膚の美白あるいはシミ除去化粧品、抑制のチロジンの酵素活性あるいは/とメラニン生合成の医薬組成物と抑制のチロジンの酵素活性あるいは/とメラニン生合成の食品で。
	評価点	1

機 械 翻 訳 2	評価根拠	・本センテンスのポイントとしては、①効果的で安全な美白・シミ取り活性成分であるムカシヨモギメタノール抽出物の提供、②ムカシヨモギメタノール抽出物の三つの用途（皮膚の美白或いはシミ取りの化粧品の製造、チロシナーゼ活性及び/或いはメラミン生物合成を抑制する医薬組成物の製造、チロシナーゼ活性及び/或いはメラミンの生物合成を抑制する食品の製造）を読み取る必要がある。 ・この訳文は文章の階層性が大きく崩れ、①も②も読み取れず、原文の意味を推察することは非常に困難である。 ・重要な用語として「灯盏花甲醇提取物」が「油ランプとメタノール抽出物」となってしまった。
	中日翻訳に特有の課題	・階層性のある文書の解説と翻訳 〔抑制（酪氨酸酶活性）或/和（黑色素生物合成）的藥物組合物〕和 〔抑制（酪氨酸酶活性）或/和（黑色素生物合成）的食品〕。
	機械翻訳 2	従来の技術の美白に対してぶちの活性成分存在する不足を取り除いて、本発明の目的はもっと有効で安全な美白を提供してぶちの活性成分-油ランプ還暦のアルコールエクストラクトを取り除きを目指して、皮膚美白を調合してあるいは、ぶちの化粧品を取り除く、チロシナーゼ活性あるいは/とメラニン色素の生物合成の医薬組成物がとチロシナーゼ活性あるいは/とメラニン色素の生物合成の食品を抑えるのを抑える。 【ユーザ辞書無し】
	評価点	1
	評価根拠	・本センテンスのポイントとしては、①効果的で安全な美白・シミ取り活性成分であるムカシヨモギメタノール抽出物の提供、②ムカシヨモギメタノール抽出物の三つの用途（皮膚の美白或いはシミ取りの化粧品の製造、チロシナーゼ活性及び/或いはメラミン生物合成を抑制する医薬組成物の製造、チロシナーゼ活性及び/或いはメラミンの生物合成を抑制する食品の製造）を読み取る必要がある。 ・この訳文は文章は全体が大きく崩れ、①も②もまったく読み取れず、原文の意味を推察不可能である。まったく意味不明となっている。
	中日翻訳に特有の課題	・階層性のある文書の解説と翻訳 〔抑制（酪氨酸酶活性）或/和（黑色素生物合成）的藥物組合物〕和 〔抑制（酪氨酸酶活性）或/和（黑色素生物合成）的食品〕。
1 と 2 の優劣		どちらも不可。
1 と 2 の違い		いずれも文が大きく崩れ、原文の意味を全く読み取ることができない。

(2) 評価用機械翻訳ソフト

辞書データによる翻訳品質向上の効果を把握するために、ルールベース機械翻訳(RBMT)方式の中日機械翻訳ソフトウェア A を利用した。本調査における機械翻訳結果は全てこの中日機械翻訳ソフトウェアにより作成した。各機械翻訳結果の作成方法は図 4.2-4 のフローを参照されたい。

選定した評価用機械翻訳ソフトウェアは系統分けされた分野別の辞書を備えている。図 4.2-2 に特許辞書・専門辞書の一覧を示す。

図 4.2-2 製品の特許用語辞書

系統	分野	収録語数	系統	分野	収録語数	系統	分野	収録語数
工学系統	建設	40,000語	工学系統	鉱業	27,000語	理学系統	環境	26,900語
	機械工学	275,600語		軽工業	13,800語	農学系統	農業	26,700語
	電気工学	68,700語	理学系統	医学	192,200語	経済学系統	会計	17,300語
	電子工学	53,000語		生命科学	81,100語	スポーツ系統	貿易	1,200語
	情報処理	257,500語		数学	24,200語		スポーツ	11,900語
	金属工学	49,400語		地質学・地理学	63,600語	ファッション系統	アパレル	4,000語
	化学	142,700語		物理学	36,500語	合計 1,460,000語		
	原子力	26,700語		力学	20,000語			

※特許用語は、基本辞書に100,000語追加収録。

(3) ユーザ辞書の作成

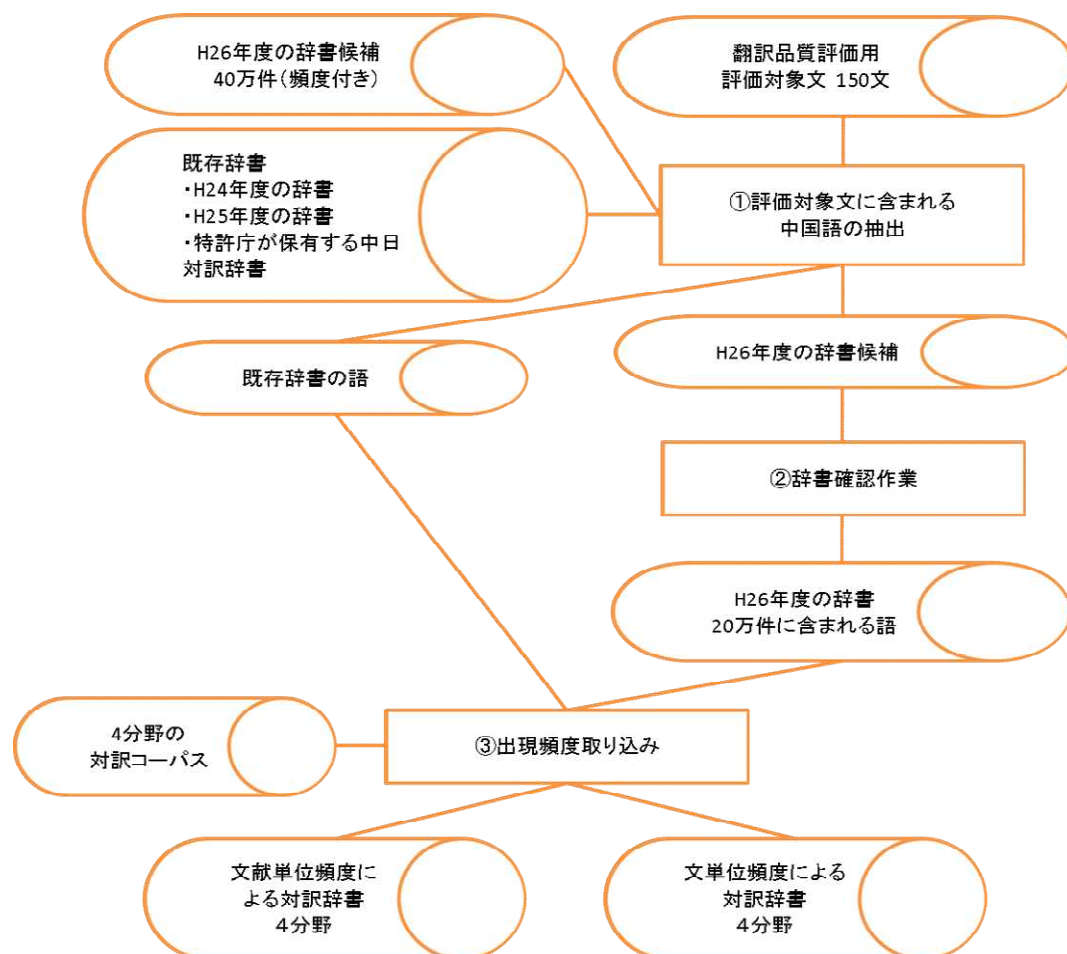
(a) ユーザ辞書の作成

分野別の辞書作成にあたって利用した2種類の出現頻度は、文単位の出現頻度と文献単位の出現頻度である。なお、両者の出現頻度の算出に用いた分野別の対訳コーパスは特許庁から貸与された対訳コーパス及び本調査で作成した対訳コーパスである。本報告書 3.3 (4) (b) 出現頻度を参照されたい。

翻訳品質評価の作業は今年度の辞書の人手確認と並行して行う必要がある。そのため、評価用テストセット 150 文の翻訳に利用される可能性のある語だけを網羅した辞書を作成した。機械翻訳結果の評価に使う辞書の入力は、平成 24 年度作成の辞書(約 100 万語)平成 25 年度の辞書(約 100 万語)、特許庁の追加辞書(8,497 語)、そして今年度の辞書(20 万語)である。その作成フロー(図 4.2-2)を示す。

なお、今年度の辞書(20 万語)については人手確認中に評価作業を行ったため、実際は辞書候補(約 40 万語)から、150 文の評価対象文に現れる中国語を抽出して最優先で辞書確認作業を行って必要な辞書を作成した。今年度の辞書作成では約 40 万語から人手確認の結果 26 万語を選定してそこから単語を優先して出現頻度上位 20 万語を採用した。そのため、実際には辞書に採用されなかった 6 万語に該当する語が分野別の辞書に含まれる可能性がある。

図 4.2-3 文単位と文献単位の出現頻度による対訳辞書の作成フロー



次に訳語の決定方法について説明する。入力する対訳辞書（約 220 万語）は中国語 1 語に対して日本語の訳語が N 語対応付けられている。翻訳ではこの N 語から 1 語を選ばなければならないが、複数の訳語候補の中から翻訳時に採用する語を決める情報は対訳コーパスにおける中国語と日本語の訳語が共に出現する頻度である。一方、今回採用した機械翻訳システムは出現頻度で訳語を選択する仕組みを持たない。そこで、出現頻度に基づいて中国語に対して日本語の訳語を 1 語だけ対応付けた辞書を作成した。図 4.2-3 のフローの最下段の対訳辞書である。

訳語の対応付けに用いる出現頻度は 2 種類ある。どちらも対訳コーパスの対訳文における見出し語（中国語）と訳語（日本語）の共起頻度だが、一方は文単位の出現頻度、もう一方は文献単位の出現頻度である。

このようにして分野別の対訳コーパスから 4 分野×2 種類の対訳辞書（表 4.2-3）を作成した。表 4.2-3 では文献単位の出現頻度による対訳辞書と文単位の出現頻度による辞書で見出し語（中国語）に対して訳語（日本語）が異なる行数を「文献単位と文単位の差」として示した。

表 4.2-3 文単位と文献単位の出現頻度による対訳辞書の差

No.	コード	分野	出現頻度の種類	品詞	行数	文献単位と文単位の差
1	C-D-N	化学	文献単位	名詞	2,147	81
2	C-S-N	化学	文単位	名詞	2,147	--
3	C-D-V	化学	文献単位	サ変	423	7
4	C-S-V	化学	文単位	サ変	423	--
5	E-D-N	電気	文献単位	名詞	2,147	81
6	E-S-N	電気	文単位	名詞	2,147	--
7	E-D-V	電気	文献単位	サ変	423	7
8	E-S-V	電気	文単位	サ変	423	--
9	M-D-N	機械	文献単位	名詞	2,147	84
10	M-S-N	機械	文単位	名詞	2,147	--
11	M-D-V	機械	文献単位	サ変	423	2
12	M-S-V	機械	文単位	サ変	423	--
13	P-D-N	物理	文献単位	名詞	2,147	100
14	P-S-N	物理	文単位	名詞	2,147	--
15	P-D-V	物理	文献単位	サ変	423	6
16	P-S-V	物理	文単位	サ変	423	--

(なお、化学分野と電気分野の文献単位と文単位の差が名詞 81 件、サ変 7 件で一致している。これは偶然の一致であり、差のある語は異なっている。)

今回作成する辞書は日本語の品詞が名詞及びサ変動詞になり得るものを対象とした。
C：化学分野の辞書を例にして文献単位頻度による対訳辞書と文単位頻度による対訳辞書の訳語の違いをサ変動詞辞書（表 4.2-4）と名詞辞書（表 4.2-5）で示す。

表 4.2-4 C：化学分野のサ変動詞の違い 7 件

No	中国語	日本語(文献単位)	日本語(文単位)
1	对照	対比する	照合する
2	导致	招来する	帰着する
3	暴露	露出する	曝露する
4	枢转	旋回する	枢動する
5	涉及	関連する	関与する
6	清洁	清浄する	クリーニングする
7	聚集	集結する	集約する

表 4.2-5 C：化学分野の名詞辞書の違い 81 件

No	中国語	日本語(文献単位)	日本語(文単位)	No	中国語	日本語(文献単位)	日本語(文単位)
1	主板	マザーボード	メインボード	51	电压源换流器	電圧源インバータ	電圧形変換器
2	交织器	インターリーバ	インタリーバ	52	电致薄膜	電極薄膜	電極フィルム
3	供给机构	供給機構	供給手段	53	电站	発電ステーション	電力ステーション
4	光阻层	フォトリジスト層	光遮蔽層	54	电还原	電解還元	電気還元
5	内腔	内部空洞	内部チャンバ	55	相反电荷	反対電荷	逆帯電
6	冲切	ダイカット	打抜	56	短切纤维	短繊維	チョップド繊維
7	冷态	冷態	コールド	57	空气喷出	空気噴出	エア噴出
8	减压蒸馏	減圧蒸留	減圧留去	58	管式反应器	管型反応器	管状反応器
9	分发	配給	分注	59	管道	パイプ	導管
10	分散片	分散錠	分散性錠剤	60	粒状多晶硅	粒状ポリシリコン	粒状多結晶シリコン
11	制品	製品	物品	61	粘连	ブロッキング	癒着
12	匹配	適合	マッチ	62	系統接口	システムインターフェース	システムインタフェース
13	压电晶片	圧電ウエハ	圧電ウェーハ	63	纤维带	繊維テープ	繊維帯
14	发电机组	発電機セット	発電機装置	64	组合加工	組み合わせ加工	組み合わせ処理
15	呼叫流	通話フロー	呼出フロー	65	色谱载体	クロマトグラフィー担体	クロマトグラフィー支持体
16	喉部	のど	スロート部	66	药物	薬物	医薬
17	喷出	噴出	吐出	67	调节系统	調節系	調節システム
18	回转阀	ロータリーバルブ	回転弁	68	货物	貨物	カーゴ
19	固体药剂	固形薬剤	固体薬剤	69	超导磁体	超伝導磁石	超伝導マグネット
20	基本流	エレメンタリストリーム	基本ストリーム	70	超细粉	超微粉	超微細粉
21	增强流	強化ストリーム	拡張ストリーム	71	轨道	軌道	レール
22	大部分	主要部分	実質的部分	72	轴转动	軸回転	軸回転
23	室壁	チャンバ壁	チャンバ壁	73	辐射单元	放射ユニット	放射素子
24	封套	封筒	ラッパー	74	过程	プロセス期間	プロセス
25	执行步骤	実行ステップ	実施工程	75	退火	アニール	焼鈍
26	抗皱性	防皺性	耐リジニング性	76	逻辑状态	論理状態	ロジック状態
27	折断	折断	断	77	量化	数量化	量子化
28	指纹图谱	フィンガープリンティング	指紋図譜	78	防粘连剂	ブロッキング防止剤	アンチブロッキング剤
29	挡水	水止め	遮水	79	除水	水除去	除水
30	挤压	押し出し	押し出し	80	集成技术	統合技術	集積技術
31	振荡	振盪	共振	81	预先设定	既設定	プリセット
32	控制机器	制御機器	制御機械				
33	推力	推進力	押し込み力				
34	方格	マス目	柵目				
35	明信片	葉書	ハガキ				
36	末端	最端	端末側終端				
37	杆体	ロッド本体	レバー体				
38	材料组	材料群	材料セット				
39	枢转轴	旋回軸	枢動軸				
40	样品	サンプル	試料				
41	模板	鋳型	テンプレート				
42	步进	ステップアップ	歩進				
43	步骤	工程	ステップ				
44	激振器	起振器	加振機				
45	灯壳	ランプケース	ランプハウス				
46	热压成型	熱プレス成形	熱間プレス成形				
47	热蒸汽	高温蒸気	熱スチーム				
48	现有	既存	先行				
49	生活垃圾	生活ゴミ	生活ごみ				
50	生物利用	生物利用	生体利用				

(b) ユーザ辞書のロード

評価に用いる機械翻訳文（評価対象文）を作成するため、表 4.2-3 の対訳辞書をユーザ辞書としてロードした。なお、RBMT はユーザ辞書に語が登録されていても使われない場合がある。特に文型を規定する動詞にこの現象が強く表れる。本調査で利用した機械翻訳ソフトのマニュアルにも「※ユーザ辞書に登録したけれども翻訳に反映されない単語について：翻訳ソフトは構文解析時の情報を最も優先するため、通常の翻訳方法では難しい特殊な単語に関しては、プログラムの内部で処理して適切な訳語を出力しています。」（製品名等一部修正）と明示している。

今回の辞書作成は日本語の品詞が名詞及びサ変動詞の 2 種類の辞書を作成・使用した。

名詞辞書

日本語の品詞が名詞である語は、中国語も名詞とみなして「中国語：名詞、日本語：一般名詞」としてユーザ辞書に登録した。表 4.2-6 に C：化学分野の例を挙げる。

表 4.2-6 ユーザ辞書 名詞辞書の例

	中国語	中国語品詞	日本語	日本語品詞
1	CCD 摄像头	名詞	C C D カメラ	一般名詞
2	DC 状態	名詞	D C 状態	一般名詞
3	HPLC 指纹图谱	名詞	HPLC フィンガープリント	一般名詞
4	ISO 系统	名詞	ISO システム	一般名詞

動詞辞書（サ変動詞）

日本語の品詞がサ変動詞語は、中国語も動詞とみなして「中国語：動詞、日本語：動詞」としてユーザ辞書に登録した。表 4.2-7 に C：化学分野の例を挙げる。

表 4.2-7 ユーザ辞書 動詞辞書の例

	中国語	中国語品詞	日本語	日本語品詞
1	一致	動詞	一致する	動詞
2	上下	動詞	上下する	動詞
3	不同	動詞	相違する	動詞
4	不溶	動詞	不溶化する	動詞

（４）機械翻訳の実行

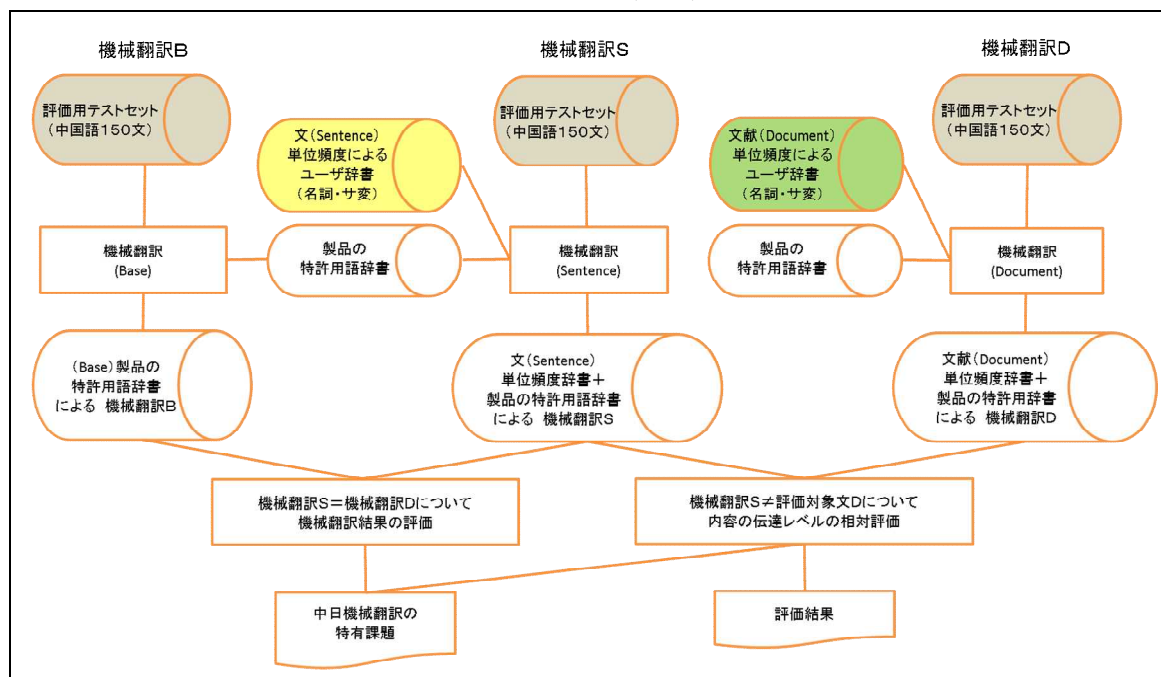
辞書データによる翻訳品質の改善を検証するために、同一の評価用テストセット（中国語 150 文）を 3 つの異なる条件で機械翻訳を実行した。

機械翻訳 B：機械翻訳ソフトに付属する専門辞書の「工学系統」、「理学系統」、「農業系統」及び「経済学系統」を選定し、評価対象文を機械翻訳した。（Bは 基礎(Base)を意味する）

機械翻訳 S：機械翻訳ソフトに付属する専門辞書と文単位の頻度に基づいて作成したユーザ辞書を利用して評価対象文を機械翻訳した。（Sは 文(Sentence)を意味する）

機械翻訳 D：機械翻訳ソフトに付属する専門辞書と文献単位の頻度に基づいて作成したユーザ辞書を利用して評価対象文を機械翻訳した。（Dは 文献(Document)を意味する）図 4.2-4 に機械翻訳の辞書との関係を図示した

図 4.2-4 評価対象文 B, S, D の機械翻訳



評価用テストセット 150 文を 3 種類の方法で中日機械翻訳して得た 3 種類の機械翻訳 B、機械翻訳 S 及び機械翻訳 D について、「機械翻訳 S」と「機械翻訳 D」に違いがある 61 文は特許文献機械翻訳の品質評価手順に従って内容の伝達レベルの評価を行い、どちらの方法による機械翻訳結果が優れているか調査して 4. 4 辞書の違いに関する評価結果にまとめた。

一方、「機械翻訳 S」と「機械翻訳 D」の評価対象文に違いが無い 89 文は、「機械翻訳 S」と「機械翻訳 B」に対して同様の評価を行い、4. 5 ユーザ辞書の有無に関する評価結果にまとめた。

また、61 文の評価結果と 89 文の評価結果を合わせて 4. 8 中日機械翻訳の特有の課題の知見を得るための素材とした。

4. 3 評価の実施

(1) 評価体制

機械翻訳の品質評価の体制は、評価者 5 名と中国語の文法の専門家 1 名（分析者）の体制で評価した。評価者 5 名は特許文献翻訳の専門家、C：化学、E：電気、M：機械そして P：物理（情報）からそれぞれの得意分野の文の評価を担当した。評価者は自身の専門分野の知識に基づいてテストセット文の評価シート（表 4.2-2）に判断の要素を記入するとともに評価点、優劣を記入した。

一方、中国語の文法の専門家である分析者 1 名は 評価シートの記入例・判断の要素の例示の作成等の事前準備、評価中は評価者へのフィードバック指導 そして 評価後は記入済

みの評価シートの分析と報告書の作成を担当した。

このような分担としたのは、①文の重要情報が正確に伝達されているかどうかは特許分野の専門家でなければ判断できないこと、②特許分野の専門家に機械翻訳の文法的な判断・分析が難しいこと、③複数人で評価を分担することで評価期間を短縮するためである。

(2) 評価基準（内容の伝達レベルの評価）

翻訳品質評価は「特許文献機械翻訳の品質評価手順 Ver.1.0 平成26年6月」の「3.1.1. 内容の伝達レベルの評価」に従った。内容の伝達レベルの評価は5名の評価者が分担して各人の担当する分野のテストセット1文と2種類の翻訳結果を評価手順により機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを、人手翻訳の内容に照らして、下記5段階の評価基準で主観的に評価した。詳細な基準は「特許文献機械翻訳の品質評価手順 Ver.1.0 平成26年6月」5～6ページを参照されたい。

- 5：すべての重要情報が正確に伝達されている。(100%)
- 4：ほとんどの重要情報が正確に伝達されている。(80～%)
- 3：半分以上の重要情報が正確に伝達されている。(50～%)
- 2：いくつかの重要情報が正確に伝達されている。(20～%)
- 1：文意が分からない、もしくは正確に伝達されている重要情報がほとんどない。
(～20%)

今回の内容伝達レベルの評価はテストセット150文の機械翻訳Sと機械翻訳Dで違いがあった61文について行った。評価手順は英日翻訳を想定しており、中日の評価が分かり難い。今回の評価対象文から分析者が各評価点の文を選び、例として挙げる

(a) 評価点5（100%）の例

評価点が5である文は存在しなかった。文の骨格構造の崩れが無くかつ重要な技術用語が適切に翻訳された事例が見つからなかったためである。

(b) 評価点4（80～%）の例

下記の機械翻訳文は、文の骨格構造の崩れがなく、ほとんどの重要情報が正確に伝達されている。しかし、重要な技術用語「癒着防止」が翻訳できていないため、評価点4と判断した。

表 4.3-1 評価点 4 の例 評価シート 21

原文	本发明所用的铁交联的透明质酸在预防粘方面有很明显的作用，其基本原理是基于其分子良好的生物相容性、物理(流变学)特性和独特的生物作用。
人手翻訳	本発明において使用される鉄架橋ヒアルロン酸は、癒着を防止する面で明らかな効果を有している。その基本原理は、その分子の良好な生物学的相性、物理的（レオロジー）特性と独自の生物的作用に基づいている。
機械翻訳	本発明使った鉄の架橋ヒアルロン酸は予防がくつつく方面にとっても明らかな作用があつて、その基本原理はその分子の良好な生体適合性、物理(レオロジー)特性と独特な生物作用に基づくのである。
評価点	4
評価根拠	・本センテンスが伝達しようとする情報は、①本発明において使用される鉄架橋ヒアルロン酸は、癒着を防止する面で明らかな効果を有する。②その基本原理は、その分子の良好な生体適合性、物理的（レオロジー）特性と独自の生物的作用に基づいている。 ・重要技術用語としては、「鉄架橋ヒアルロン酸」「癒着防止」がある。「癒着防止」が翻訳できていないものの、「予防がくつつく方面」からも概ね意味が推察できるため、①の意味も概ね伝達できると考えられる。 ・②の意味は問題なく伝達できる。

(c) 評価点 3（50～%）の例

下記の例文の場合、文前半の骨格構造が崩れなかったため、意味が推察できる。一方、文後半の骨格構造は崩れている。重要情報の一部が読み取りにくいため、評価点 3 と判断した。

表 4.3-2 評価点 3 の例 評価シート 9

原文	蜂王浆经科学试验证实是一种能促进增强人体机能，延长人体寿命并对机体的多种功能有调节作用，对多种疾病(包括肿瘤)有预防和辅助治疗作用的天然补品。
人手翻訳	同成分は、一種の人体の機能を促進し、人体の寿命を延ばすと共に、人体の多くの機能を調節する役割を有しており、多数の病気（腫瘍を含む）を予防し、治療を助ける天然のサプリメントである
機械翻訳	ロイヤル・ゼリーは科学を通じて実証するのが人体の機能を増強するのを促進することができるのを試験して、人体寿命を延長してそして機体の補助機能に対して調節作用があつて、多種の疾病(腫瘍を含む)に対して予防とアジュバント療法効果の天然の栄養食品がある。

評価点	3
評価根拠	・本センテンスの重要情報は、①ローヤルゼリーは天然栄養品である。②ローヤルゼリーの作用は科学的実験によって証明されている。③上記作用とは、(A) 一種の人体の機能を促進し、(B) 人体の寿命を延ばすと共に、(C) 人体の多くの機能を調節する役割を有しており、(D、E) 多数の病気（腫瘍を含む）に対し予防し、治療を助けるとの五つである。 ・本機械翻訳文において、「天然の栄養食品がある。」の「が」は適切ではないが、読み手は上下文から①の意味を概ね推察できると考えられる。 ・本機械翻訳文の「ロイヤル・ゼリーは科学を通じて実証するのが人体の機能を増強するのを促進することができるのを試験して、」では、人体機能増強のみが実験によって証明されているかにも読み取れる。文の構造が崩れている。

(d) 評価点 2 (20～%) の例

下記の機械翻訳文では、人手翻訳の「各ラットの右側後脚の皮膚に卵白 0.1ml/匹を注射する」に対応する箇所など、複数の部分の骨格構造が崩れている。そのため、正しく読み取れる情報が少なく、評価点 2 と判断した。

表 4.3.3 評価点 2 の例 評価シート 7

原文	末次给药后 1h, 在每只大鼠右后足跖皮下注射蛋清 0.1ml/只, 测量注入致炎剂后 1、2、3h 右后足跖体积, 减去致炎前大鼠右后足跖正常体积, 作为各鼠不同时间的足肿胀度。
人手翻訳	最後の薬を与えてから 1 時間後、各ラットの右側後脚の皮膚に卵白 0.1ml/匹を注射する。注射後炎症を起こしてから 1、2、3 時間後の右側後脚の体積を測定し、炎症を起こす前のラットの右側後脚の正常な体積を差し引いて、各ラットの異なる時間における後脚の膨張度を求める。
機械翻訳	末次投薬した後 1h, ハムスター右後ごとに十分に踏む皮下注射の卵白 0.1ml/匹、注入起炎剤を測定した後に 1、2、3h 右後十分に踏む体積、マイナスして炎を送る前にハムスター右後十分に踏む正常容積、各ネズミ異なる時間的十分に腫れる度として。
評価点	2

評価根拠	・本センテンスは重要情報として、①各ラットの足浮腫度を測定する方法について記載している。該方法は具体的に、②最終投薬後 1 時間後、③各ラットの右後ろ足に卵白 0.1ml/匹を皮下注射、④催炎剤を注射してから 1、2、3 時間後の右側後足の体積を測定し、炎症を起こす前のラットの右側後足の正常な体積を差し引いて、各ラットの異なる時間における足の浮腫度とする。 ・機械翻訳の「各ネズミ異なる時間的十分に腫れる度として」の部分は最重要の①の意味が伝達できるか疑問が残る。 ・「末次投薬した後 1h」は、記載上最終回投薬後 1 時間をかけて卵白 0.1ml/匹を注射すると誤解される余地はあるものの、常識から「0.1ml/匹」を 1 時間かけて注射するとは考えづらいので、読み手は「②最終投薬後 1 時間後」を推察できると思われる。 実験動物の「ラット」を「ハムスター」に誤訳しているのが良くない。 ③の注射部位の「右後ごとに十分に踏む」は「足」が翻訳されておらず、さらに「各ネズミ異なる時間的十分に腫れる度」も語のつながりが崩れているため、足浮腫度実験を連想できる可能性は低くなる。④「、注入起炎剤を測量した後に 1、2、3h 右後十分に踏む体積、マイナスして炎を送る前にハムスター右後十分に踏む正常容積」は文が崩れ、何をマイナスするのか読み取れない。 ・この翻訳文からはラットを用いた足浮腫度実験を連想しにくい。読み手は十分な情報が得られない。
------	---

(e) 評価点 1（～20%）の例

下記の機械翻訳文は文の骨格構造が崩れていて、重要な情報がほとんど伝達できなかったため、評価点 1 と判断した。

表 4.3-4 評価点 1 の例 評価シート 117

原文	干散货码头由于装卸的货物具有不同特性，另外船舶的类型也较多，再加上散货粉尘的影响使作业环境比较恶劣，对检测设备和技術有较苛刻的要求，因此目前 散货装船系统 自动化程度较低，基本上靠人工现场操作。
人手翻訳	ドライバルク貨物用埠頭 は、積み卸される貨物が異なる特性を持ち、また船舶のタイプが多く、バルク粉塵の影響と相まって作業環境が悪くなり、検出装置と技術に対する要求がより厳しいため、現在のバルク貨物用船積みシステムが自動化程度が低く、基本的に人に頼って現場で操作される。

機械翻訳	積み卸しする貨物が異なる特性があるため ルーズ貨物の埠頭 をして、その他に船舶のタイプもわりに多くて、ルーズ貨物の粉塵の影響をさらに加えて作業環境が比較的極めて悪くて、テストセットと技術に対して比較的 ひどい要求 があつて、そのため、現在ルーズ貨物船積みするシステムの自動化程度が比較的低くて、基本的に人力に頼る現場操作。
評価点	1
評価根拠	・「操作」が述語として判定されておらず、「現場」の後ろに「で」という格助詞も欠落 ・文の骨格構造が正しく翻訳されていないので、文の意味を理解しにくい ・「ドライバルク貨物用埠頭」が主語として判定されていない ・「ルーズ貨物の埠頭」という訳語が適切でない ・「ひどい要求」という訳語が適切でない ・「散货装船系统」という複合語が正しく認識されておらず、複合語の一部である「装船」が述語として認識されてしまう ・「バルク貨物用船積みシステム」が「ルーズ貨物船積みするシステム」と訳されているうえ主語として判定されておらず理解しにくい。さらにその後ろに「は」という格助詞が欠落している。

4. 4 辞書の違いに関する評価結果

(1) 翻訳の評価結果とその分析

(a) 評価点と評価対象文数

機械翻訳 S と機械翻訳 D に違いがあった 61 文の評価点と評価対象文数を表 4.4-1 に示す。評価点 5 は共に 0 文、評価点 4 は機械翻訳 D が多く、評価点 2 は機械翻訳 S が多いことから、機械翻訳 D の方が評価点が高く翻訳品質が良いが各評価点の文数の差は少ない。

表 4.4-1 評価点と評価対象文数

評価点	機械翻訳 S (文単位辞書)	機械翻訳 D (文献単位辞書)
5 (100%)	0 文	0 文
4 (80%～)	6 文	8 文
3 (50%～)	11 文	11 文
2 (20%～)	17 文	14 文
1 (～20%)	27 文	28 文

(b) 機械翻訳 S と機械翻訳 D の比較

機械翻訳 S と機械翻訳 D に違いがあった 61 文について、2 つの翻訳結果の評価点について、どちらが高いか評価点を比較した（表 4.4-2）。機械翻訳 D の評価点が高いのが 3 文、低いのが 1 文であった。機械翻訳 D の方が評価点が高いが、その文数の差は僅差である。

表 4.4-2 同じ原文（中国語）の翻訳結果の比較

評価点の比較	文数
機械翻訳 S の評価点＝機械翻訳 D の評価点	57 文
機械翻訳 S の評価点＜機械翻訳 D の評価点	3 文
機械翻訳 S の評価点＞機械翻訳 D の評価点	1 文

機械翻訳 S の評価点＜機械翻訳 D の評価点 の例

評価シート 69 の文の評価点は機械翻訳 S が 2 点で機械翻訳 D が 3 点で後者が優れている。この評価点の差は機械翻訳 S のユーザ辞書の「抗皺性（耐リジング性）」が機械翻訳 D のユーザ辞書の「抗皺性（防皺性）」に変わったためで、機械翻訳 D が優れている。原因はユーザ辞書の差である。（表 4.2-5 の 26 番を参照。）

表 4.4-3 翻訳品質評価シート 69

文番号	CN-A-1432678 02100213.4	
分野	C10 D03D1500	
中国語原文	例如，涤棉混纺织物是由涤纶和棉采用混纺的方式织成的，它一方面具有涤纶纤维强力高、 抗皱性 好的特性，同时，又具有棉的吸湿性好、穿着舒适的优点。	
参考和訳	例えば、コットン－ポリエステル混紡織物はポリエステル繊維とコットンを混紡する方法を用いて織上げたものであり、ポリエステル繊維の高強度と 防しわ性 が優れるという特性を有し、また綿の吸湿性が良く、着心地が良いというメリットも有する。	
評価 S	機械翻訳 S	例えば、洗う綿混紡織物はテリレンと綿の採用の混紡の形態から織ったので、それ一様相はポリエステル繊維があつて 耐リジング性 の良い特性が強く高くて、同時に、また綿の吸湿性の良くて、着心地が良い利点がある。
	評価点	2

評価 D	機械翻訳 D	例えば、洗う綿混紡織物はテリレンと綿の採用の混紡の形態から織ったので、それ一様相はポリエステル繊維があつて 防皺性 の良い特性が強く高くて、同時に、また綿の吸湿性の良くて、着心地が良い利点がある。
	評価点	3
優劣の判断のポイント		「 防皺性 」が翻訳できていることは高く評価できる。文の構造が崩れているものの、前の「ポリエステル繊維があつて」と組み合わせれば、「ポリエステル繊維に由来する防皺性に優れる」という意味が読み取れる。

翻訳品質評価シート 69

他の2文は評価シート112と113で共にP（物理）分野で機械翻訳Sのユーザ辞書の「質量（質量）」と機械翻訳Dのユーザ辞書の「質量（品質）」に関するもので、どちらも機械翻訳Dの用語選定が優れている。

表 4.4-4 翻訳品質評価シート 112

文番号	CN-A-1900710 200510035987.1	
分野	P01 G01N30/02	
中国語原文	(1) 以灵芝孢子及其脂溶性提取物灵芝孢子油建立起来的 HPLC 指纹图谱，代表着灵芝孢子及其脂溶性提取物灵芝孢子油大部分药理活性，能有效地表征灵芝孢子及其脂溶性提取物灵芝孢子油的 质量 。	
参考和訳	(1) レイシ胞子及びその脂溶性抽出物であるレイシ胞子油から作成された HPLC フィンガープリントは、レイシ胞子及びその脂溶性抽出物であるレイシ胞子油の大部分が薬理活性を示し、レイシ胞子及びその脂溶性抽出物であるレイシ胞子油の 品質 を効果的に <u>キャラクターゼーション</u> することができる。	
評価 S	機械翻訳 S	(1) 霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）作り上げる HPLC 指紋スペクトルで、代表霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）の実質的部分の薬理の活性、有効に特性評価の霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）の 質量 。
	評価点	3
評価 D	機械翻訳 D	(1) 霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）作り上げる HPLC 指紋スペクトルで、代表霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）の実質的部分の薬理の活性、有効に特性評価の霊芝胞子と脂溶性抽出物の霊芝胞子の油（脂）の 品質 。
	評価点	4

優劣の判断のポイント	・「品質」が正確に翻訳できている。
------------	-------------------

翻訳品質評価シート 112

機械翻訳Sの評価点>機械翻訳Dの評価点 の例

評価シート23の文の評価点は機械翻訳Sが2点で機械翻訳Dが1点である。これは、ユーザ辞書において、機械翻訳Sのユーザ辞書の「粘连（癒着）」が機械翻訳Dのユーザ辞書の「粘连（ブロッキング）」に変わってしまったため、機械翻訳Dの評価点が下がってしまった。原因はユーザ辞書の差である。4.2(3)の表4.2-5の61番を参照。

表 4.4-5 翻訳品質評価シート 23

文番号	CN-A-1473572 03133843.7	
分野	C12 A61K31/728	
中国語原文	且铁交联的透明质酸既是良好的隔离层，又是良好的给药体系，配合胶原抑止剂，能够从根本上解决手术后防粘连的问题。	
参考和訳 (調整有)	また、鉄架橋ヒアルロン酸は良好な隔離層でありながら、良好な投薬体系でもある。コラーゲン抑止剤を調合することで、根本的に手術後の癒着防止の問題を解決することが可能となる。	
評価 S	機械翻訳S	しかも鉄の架橋ヒアルロン酸良好な絶縁層、また良好な投与系で、衝合のコラーゲンの抑制物質、根本的に解決の手術の後方守備の癒着の問題。
	評価点	2
評価 D	機械翻訳D	しかも鉄の架橋ヒアルロン酸良好な絶縁層、また良好な投与系で、衝合のコラーゲンの抑制物質、根本的に解決の手術の後方守備のブロッキングの問題。
	評価点	1
	優劣の判断 のポイント	機械翻訳Sの方が「癒着」という重要技術用語を適切に翻訳できているため、内容の伝達度が高い。

翻訳品質評価シート 23

(2) 主観評価

ここまで、機械翻訳Sと機械翻訳Dの61文の評価点で優劣を評価してきた。その結果、文献単位の出現頻度に基づくユーザ辞書を使用した機械翻訳Dが優れていることがわかったが僅差である。そこで、判断をさらに精密にするため、機械翻訳Sと機械翻訳Dのどちら

らが良いか評価点では判断できなかった翻訳結果について評価者が主観的に良い方を判断して、その理由をコメントしたもの进行分析者が分析した。

評価者が表 4.4-1 と同じ原文（中国語）の翻訳結果の比較で「機械翻訳 S＝機械翻訳 D」と判定した 57 文について、どちらの機械翻訳の方が良いかを同じ評価者が判断した結果を表 4.4-6 に示す。

「機械翻訳 D がよい」という判断と「機械翻訳 S がよい」と判断された文数は共に 1 文で同数、さらに「どちらかといえば S がよい」と「どちらかといえば機械翻訳 D がよい」の文数も共に 2 文で同数であった。したがって、表 4.4-1 において「機械翻訳 S＝機械翻訳 D」と判定された結果の精度が高いことと、評価点による判定を変更するような結果が得られないことが分かった。

表 4.4-6 評価者による機械翻訳 S と機械翻訳 D の相対評価

判断分類	文数（全 57 文）
変わらない	51 文
機械翻訳 S がよい	1 文（評価シート 6）
機械翻訳 D がよい	1 文（評価シート 109）
どちらかといえば機械翻訳 S がよい	2 文（評価シート 20, 55）
どちらかといえば機械翻訳 D がよい	2 文（評価シート 8, 108）

(a) 機械翻訳 S が良い例

表 4.4-7 に機械翻訳 S が良い例を挙げる。これは機械翻訳 S のユーザ辞書に「相反电荷（逆帯電）」、機械翻訳 D のユーザ辞書に「相反电荷（反対電荷）」と異なる訳語が採用されたためである。辞書を参照することのできない評価者は判定理由で「帯びる」の有無をあげているが、これは採用した機械翻訳ソフトがユーザ辞書の用語の影響で原文の「帯」をどのように翻訳するかにより、辞書の訳語が優劣の直接的な原因ではない。

表 4.4-7 機械翻訳 S が良い例（評価シート 6）

原文	原因可能是加入一定量的盐后，盐的正负离子(Na ⁺ , Cl ⁻)把带相反电荷的-N ⁺ (CH ₃) ₃ 和-COO ⁻ 包覆起来，削弱它们之间较强的静电作用力，间接通过较弱的盐键以及氢键等加强凝胶强度。
参照和訳	その原因は一定量のナトリウムを加えた後、ナトリウムのプラス・マイナスイオン (Na ⁺ 、Cl ⁻) が反対の電荷を帯びた—N ⁺ (CH ₃) ₃ と—COO ⁻ を覆い、それらの間の比較的強い静電作用力を弱め、間接的に比較的弱いナトリウムナトリウム結合及び水素結合等を通じてゲル強度を高めていることによる。

機械翻訳 S (評価点 2)	理由おそらく参加の一定量の塩の後で、塩の正負イオン(Na ⁺ 、Cl ⁻)は 逆帯電 の-N+(CH ₃) ₃ と-COO ⁻ の被包を持ちを、弱化それら間のより強い静電気の作用力、間接的に比較的弱い塩結合と水素結合などの補強のゲル強度を通じて。
機械翻訳 D (評価点 2)	理由おそらく参加の一定量の塩の後で、塩の正負イオン(Na ⁺ 、Cl ⁻)は 反対電荷 の-N+(CH ₃) ₃ と-COO ⁻ の被包を持ちを、弱化それら間のより強い静電気の作用力、間接的に比較的弱い塩結合と水素結合などの補強のゲル強度を通じて。
どちらが良 いか	機械翻訳 S
理由	「 帯相反電荷的 」を翻訳する際、Sは「 逆帯電の 」、Dは「 反対電荷の 」と翻訳した。意味の理解に影響しないが、Dは「反対電荷を帯びた」の「 帯びる 」の意味を訳せていない。Sの「 逆帯電の 」がより優れた翻訳である。

(b) 機械翻訳Dが良い例

表 4.4-8 に機械翻訳Dが良い例を挙げる。本例は機械翻訳Sのユーザ辞書に「**反向引物**(後方プライマ)」、機械翻訳Dのユーザ辞書に「**反向引物**(逆方向プライマ)」が登録されている。この用語の違いで機械翻訳Dの方が良いと判定された。

表 4.4-8 機械翻訳Dが良い例 (評価シート 109)

原文	本発明所提供的的检测松材线虫的专用引物由正向引物和 反向引物 组成, 正向引物是 3' 末端序列为 3' -TCAGT 的 16-30 个核苷酸的寡核苷酸序列, 反向引物 是 3' 末端序列为 3' -CCAAG 的 16-30 个核苷酸的寡核苷酸序列; 探针是 3' 末端序列为 3' -GGTAGTACGCGCC 的 16-25 个核苷酸的寡核苷酸序列。
人手翻訳	本発明が提供するマツノザイセンチュウ検出専用プライマーは、センスプライマーと アンチセンスプライマー からなる。センスプライマーは、3' 末端配列が 3' -TCAGT の 16-30 個のヌクレオチドからなるオリゴヌクレオチド配列である。 アンチセンスプライマー は、3' 末端配列が 3' -CCAAG の 16-30 個のヌクレオチドからなるオリゴヌクレオチド配列である。プローブは 3' 末端配列が 3' -GGTAGTACGCGCC の 16-25 個のヌクレオチドからなるオリゴヌクレオチド配列である。

機械翻訳 S (評価点 4)	本発明所の提案の検出のマツ材線虫の専用プライマは順方向プライマと 後方プライマ の組成から、順方向プライマは 3' 末端配列が 3' -TCAGT の 16-30 ヌクレオチドのオリゴヌクレオチド配列なると、 後方プライマ は 3' 末端配列が 3' -CCAAG の 16-30 ヌクレオチドのオリゴヌクレオチド配列なるとである；プローブは 3' 末端配列が 3' -GGTAGTACGCGCC の 16-25 ヌクレオチドのオリゴヌクレオチド配列なるとである。
機械翻訳 D (評価点 4)	本発明所の提案の検出のマツ材線虫の専用プライマは順方向プライマと 逆方向プライマ の組成から、順方向プライマは 3' 末端配列が 3' -TCAGT の 16-30 ヌクレオチドのオリゴヌクレオチド配列なると、 逆方向プライマ は 3' 末端配列が 3' -CCAAG の 16-30 ヌクレオチドのオリゴヌクレオチド配列なるとである；プローブは 3' 末端配列が 3' -GGTAGTACGCGCC の 16-25 ヌクレオチドのオリゴヌクレオチド配列なるとである。
どちらが良 いか	機械翻訳 D
理由	「 逆方向プライマ 」という技術用語が適切に翻訳されているため。

(3) 優劣のポイントとなった用語

「機械翻訳 S」と「機械翻訳 D」の優劣の判断の理由の内、用語の選定によるものを表にした。○印は望ましい訳語、×印は良くない訳語である。機械翻訳 D が望ましい訳語が 3 語、機械翻訳 S が望ましい訳語が 1 語であった。このことから僅差であるが、文献単位の出現頻度に基づいて訳語を選定したユーザ辞書の方が妥当な語を選定していることが分かった。

表 4.4-9 優劣のポイントとなった用語

No	シート番号と分野	中国語	機械翻訳 D (文献単位)	機械翻訳 S (文単位)
1	69C	抗皺性	○防皺性	×耐リジグ性
2	112P, 113P ※	質量	○品質	×質量
3	20C, 22C, 23C ※※	粘連	×ブロッキング	○癒着
4	109P	反向引物	○逆方向プライマ	×後方プライマ

※112P,113P は同一文献 CN-A-1900710 から抽出されたテストセットである。

※※20C,22C,23C は同一文献 CN-A-1473572 から抽出されたテストセットである。

(4) 二つの辞書の優劣 (結果)

翻訳品質評価用テストセット 150 文を機械翻訳する複数の方法として、分野別の中日辞

書で中国語に対する訳語（日本語）を一つ選定する際、下記の2つの方法で選定したユーザ辞書による機械翻訳結果を比較した。

①中日対訳コーパスにおける文単位の出現頻度による選定（機械翻訳S）

②中日対訳コーパスにおける文献単位の出現頻度による選定（機械翻訳D）

辞書の違いは分野毎に異なるがその違いは見出し語（中国語 2,147 語）に対して 81～100 語（約 4.0%）であった。また、機械翻訳結果に違いがあったのは、テストセット 150 文の内、61 文であった。

機械翻訳 S と機械翻訳 D を「特許文献機械翻訳の品質評価手順 Ver.1.0 平成 26 年 6 月」の「3.1.1.内容の伝達レベルの評価」に基づいて評価した結果、中日対訳コーパスにおける文献単位の出現頻度による選定によって作成した機械翻訳 D の方が翻訳品質が優れていた。

4. 5 ユーザ辞書の有無に関する評価結果

今回の調査は2種類のユーザ辞書による機械翻訳 S と機械翻訳 D の比較評価が主目的であるが、2種類のユーザ辞書の違いは僅かであるため、機械翻訳 S と機械翻訳 D で全く同じ翻訳文になるテスト文が 89 文存在した。この 89 文について、ユーザ辞書の有無による違いを調べるため、ユーザ辞書を利用しない機械翻訳 B を作成して、ユーザ辞書を利用した機械翻訳 S と比較評価した。

（1）評価点と評価対象文数

機械翻訳 S と機械翻訳 B に違いがあった 89 文の評価点と評価対象文数を表 4.5-1 に示す。評価点 5 は共に 0 文、評価点 4 は機械翻訳 S が多く、評価点 2 は機械翻訳 S が多い。一方、機械翻訳 B は評価点 1 が圧倒的に多いことから、機械翻訳 S の方の評価点が高く翻訳品質が良いことがわかる。

試みに「評価点×文数」の総和を得点として計算すると、機械翻訳 S の得点は 196 点、機械翻訳 B の得点は 168 点となる。

表 4.5-1 評価点と評価対象文数

評価点	機械翻訳 S (文／文献単位共通)	機械翻訳 B (ユーザ辞書無し)
5 (100%)	0 文	0 文
4 (80%～)	13 文	7 文
3 (50%～)	21 文	20 文
2 (20%～)	26 文	18 文
1 (～20%)	29 文	44 文

(2) 機械翻訳Sと機械翻訳Bの比較

機械翻訳Sと機械翻訳Bを比較した 89 文について、2つの翻訳結果の評価点について、どちらが高いか評価点を比較した。機械翻訳Sの評価点が高いのが 35 文、低いのが 13 文であった。したがって機械翻訳Sの方の評価点が高いが、低くなる例も存在する。

表 4.5-2 同じ原文（中国語）の翻訳結果の比較

評価点の比較	文数
機械翻訳Sの評価点＝機械翻訳Bの評価点	41 文
機械翻訳Sの評価点＜機械翻訳Bの評価点	13 文
機械翻訳Sの評価点＞機械翻訳Bの評価点	35 文

(a) 機械翻訳Sが良い例

表 4.5-3 に機械翻訳Sが良い例を挙げる。ユーザ辞書の無い機械翻訳Bでは、「**纤维可以是牵伸丝也可以是弹力丝**（繊維は延伸加工糸あるいはストレッチヤーンであって）」という文の構造が崩れており、「**纖維は糸に纖維束を引き伸ばしてもであるのがストレッチシルクであるなことができるで**」と翻訳された。一方、ユーザ辞書を使った機械翻訳Sの場合、「**纖維は糸に纖維束を引き伸ばしてもであるのがストレッチシルクであるなことができるで**」と翻訳され、文前半の構造がよくなった。その他に、「**截面（断面）**」と「**单丝细度（単纖維の繊度）**」の訳語の質も良くなった。

表 4.5-3 機械翻訳Sが良い例（評価シート 70）

原文	所述涤纶纱的 截面 可以是圆形或三叶形、五叶形异形丝等其它形状， 纤维可以是牵伸丝也可以是弹力丝 ，可以加网络也可以空变，其规格为 56dtex～222dtex， 单丝细度 在 0.7dtex～2.0dtex。
参照和訳	上記ポリプロピレン繊維の 断面 は円形あるいは三つ葉形、五葉形等異形繊維および他の形であることができ、 纖維は延伸加工糸あるいはストレッチヤーンであって 空気加工も空気変形加工も施すことができ、その規格は 56dtex～222dtex であり、 単纖維の繊度 は 0.7dtex～2.0dtex である。
機械翻訳S (評価点 4)	前記のテリレンの紗の 断面 は円形あるいは三葉形、 5 葉形 の異形の糸などのその他の形状であるなことができるで、 纖維は牽伸の糸もであるのが弾力糸であるなことができるで 、ネットワークを加えてもヌルの引数のことができることができるで、その規格は 56dtex～222dte である x， 単糸繊度 は 0.7dtex～2.0dte にある x。

機械翻訳 B (評価点 2)	前記のテリレンの紗の切断面は円形あるいは3葉形、五葉形の異形の糸などのその他の形であることができるで、繊維は糸に繊維束を引き伸ばしてもであるのがストレッチシルクであることができるで、ネットワークを加えてもヌルの引数のことができることができ、その規格は 56dtex～222dte である x, 単繊維の粒度は 0.7dtex～2.0dte にある x。
どちらが良 いか	機械翻訳 S
理由	・崩れた文構造がよくなる ・技術用語がよく翻訳されているため、内容の伝達度が高い

(b) 機械翻訳 B が良い例

表 4.5-4 に機械翻訳 B が良い例を挙げる。機械翻訳 S の場合、述語「**採用 (使用する)**」と「**加工 (加工する)**」は述語として認識できず、名詞として翻訳されているが、機械翻訳 B では共に述語として正しく認識された。そのため、「**採用常规设备 (通常の設定が使用されている)**」と「**加工竹木复合结构长材 (竹木複合構造長材を加工する)**」の翻訳結果は共に機械翻訳 B の方が正しい。

表 4.5-4 機械翻訳 B が良い例 (評価シート 24)

原文	按照本发明的加工方法, 採用常规设备 替代一台套价格昂贵的进口连续拼接、平压专用设备 加工竹木复合结构长材 , 可以大大降低竹木复合结构长材的价格, 使其有条件替代大量的建筑用材。
人手翻訳	本発明における加工方法においては、価格が高い輸入設備に代わって 通常の設定が使用されている 。連続加工、平圧専用設備により 竹木複合構造長材を加工する ことで、徳木複合構造長材の価格を大幅にカットすることができる。その結果、条件付きで大量の建築資材として使用することが可能となっている。
機械翻訳 S (評価点 2)	本発明の加工方法によって、 採用の常法の装置 の代替 1 台高価格をかぶせる入口の連続の添接は、平らに専用設備 加工の竹木の複合構造長い材料 をおさえて、とても大きい低下の竹木の複合構造長い材料の価格のことができ、それ十分な状況がある代替の大量の建築用材。
機械翻訳 B (評価点 4)	本発明の処理方法によって、 常規の装置を採用して 1 台が価格の高価な輸入の連続の接合をかぶせて、平らに専用設備に 竹木の複合構造長い材料を加工する ようにおさえるのに取って代わって、大いに竹の木の複合構造の長い材料の価格を低減することができて、それは十分な状況があって大量の建物の用材に取って代わる。

どちらが良 いか	機械翻訳 B
理由	・ 述語の認識が、より正しい ・ 文構造が比較的正しく分析される

4. 6 複合語の追加による評価結果

今年度、本事業の辞書作成では単語と複合語を識別した。本調査における複合語であるか否かの情報は、見出し語（中国語）が対訳辞書データの中の複数の見出し語のみによって分解できる、かつその訳語が当該複数の見出し語の訳語のみによって分解できる場合をいう。例えば、対訳辞書データに、「見出し語：訳語」の対として「形成：形成」、「区域：領域」という語対がある場合、それらの語対を順に組み合わせた「形成区域：形成領域」は複合語であるとする。但し「形成区域：領域形成」というような語順の異なる語対は複合語でない。複合語の定義は本報告書 3. 2 章の（4）対訳辞書データの作成を参照されたい。

本調査における複合語と単語を区別する動機は、複合語は単語の名詞連鎖として翻訳できるので、辞書登録の必要が無いのではないかと、または 辞書を作成する場合は単語を優先して辞書登録すべきでないかという仮説による。この仮説を検証することは本調査の仕様の外であるが、今年度以降の辞書作成でも重要な要素であるため、少数の文であるが実験・評価した。

実験の対象文の選定はプログラムでテストセット 150 文から複合語を含む文と複合語を見つけ出し、分析者がその中から適当と思われる文と複合語を選定した。次に分析者はユーザ辞書にその複合語がある場合の機械翻訳 F（複合語）とユーザ辞書からその複合語を削除した場合の機械翻訳 T（単語）を作成して比較した。以下に典型的な例と複雑な例を解説する。

複合語の辞書登録の翻訳結果への影響について少数の事例で分かったことは下記の 2 点である。

- ① 単純な名詞連鎖は翻訳結果の助詞「の」の有無程度で複合語の辞書登録による良い／悪い影響は共に少ない。（典型的な例）
- ② 一方、中国語の形態素解析や述語動詞の認識に関係する場合は影響が大きい、その影響の予測は難しい。（複雑な例）

（1）複合語の追加により、翻訳が良くなる例（典型的な例）：

名詞の連続である複合語の追加により、名詞の間の助詞「の」をなくすることができる。一方、意味的には「の」の有無で翻訳品質は変化しない。

名詞連鎖の複合語の場合はこのような例が多いと思われる。ユーザ辞書に複合語がある

場合の機械翻訳結果を機械翻訳 F、無い場合を機械翻訳 T とする。

表 4.6-1 複合語の追加により、翻訳が良くなる例 シート 133

原文	本发明为完整、准确评价灵芝孢子及其脂溶性提取物灵芝孢子油的质量提供了新的对照标准, 将为提高 灵芝孢子 及其脂溶性提取物灵芝孢子油制剂的质量及疗效做出贡献。
人手翻訳	本特許はレイシ孢子及びその脂溶性抽出物であるレイシ孢子油の質量を全体的、正確に評価する上で、新しい対照基準を提供するものである。これは レイシ孢子 及びその脂溶性抽出物であるレイシ孢子油の質量及び治療効果を高めるのに役立つものとなる。
複合語	灵芝孢子：靈芝孢子
機械翻訳 F	本発明は完全で、正確な評価の靈芝孢子と脂溶性抽出物の靈芝孢子的油（脂）の質量の提案新しい対照標準で、 靈芝孢子 と脂溶性抽出物の靈芝孢子的油処方物の質量と効能を高めるために貢献する。
機械翻訳 T	本発明は完全で、正確な評価の靈芝の孢子和脂溶性抽出物の靈芝の孢子的油（脂）の質量の提案新しい対照標準で、 靈芝の孢子 と脂溶性抽出物の靈芝の孢子的油処方物の質量と効能を高めるために貢献する。

なお、4. 8（3）の CJ-9 のように複合語の境界の認識を誤る場合には、複合語の登録が有効と考えられる。

（2）複合語の追加により、述語の認識ができなくなる複雑な例：

本例は動詞になり得る単語を含む複合語の追加により、述語の認識が不可能になってしまう例 と 長い単位の複合語を登録することで、単語の誤分割を防ぐ例 が複合した例である。

この例文の場合、ユーザ辞書の二つの用語「**材料组成（材料組成）**」（複合語）と「**材料组（材料群）**」（単語）が関係して二つの側面を持つ。ユーザ辞書に先の 2 語がある場合の機械翻訳結果を機械翻訳 F、単語だけを登録した場合を機械翻訳 T とする。

機械翻訳 F は「**材料组成（材料組成）**」（複合語）が利用されており、文法的には 中国語の述語動詞「**组成（構成する）**」が正しく翻訳されない課題がある。しかし、文全体を見ると理解可能な日本語が生成されている。大きな文法的な誤りがあるが翻訳結果は悪くないという 複雑な例である。

一方、機械翻訳 T はユーザ辞書に「**材料组（材料群）**」（単語）だけが存在する。この場合、中国語の述語動詞「**组成（構成する）**」が「**组**」と「**成**」に誤って分割されてしまい、「**材料组（材料群）**」と「**成（なつて）**」と翻訳された。こちらも中国語の述語動詞が「**组成**」から「**成**」に変わったがそれなりに結果は良くなっている。このように、ユーザ辞書内の

複合語の影響はユーザ辞書内の他の語やメーカーの専門用語辞書の用語と影響し合うため予測が難しい。

表 4.6-2 複合語の追加により、述語の認識ができなくなる複雑な例 シート 140

原文	天线辐射单元印刷在 PCB 板上，该 PCB 板的结构分为三层如图 2 所示，上、下两层分为金属表层 3 和金属表层 5，中间为介质层 4，介质层由微波损耗小的聚四氟乙烯或者环氧树脂等材料组成，相对介电常数为 4.4。
人手翻訳	アンテナ放射エレメントを PCB 基板上にプリントし、この PCB 基板の構造は図 2 に示す通り 3 層に分かれている。上、下の 2 層は金属表層 3 と金属表層 5 に分かれ、中間層は誘電体層 4 である。誘電体層はマイクロ波損失の少ないフロン、或いはエポキシ樹脂等の材料で構成され、比誘電率は 4.4 である。
単語用語	材料組：材料群
複合語	材料 組成：材料 組成
機械翻訳 F 複合語（材料 組成）あり	アンテナの放射ユニットの印刷は PCB オンボードにあって、この PCB 板の構造は図 2 のように三重層に分けて、上下二層は金属表層 3 と金属表層 5 に分けて、中間は誘電体層 4 で、誘電体層はマイクロ波損失の小さいポリテトラフルオロエチレンあるいはエポキシ樹脂などの材料組成から、比誘電率は 4.4 である。
機械翻訳 T 単語用語の み	アンテナの放射ユニットの印刷は PCB オンボードにあって、この PCB 板の構造は図 2 のように三重層に分けて、上下二層は金属表層 3 と金属表層 5 に分けて、中間は誘電体層 4 で、誘電体層はマイクロ波損失の小さいポリテトラフルオロエチレンあるいはエポキシ樹脂などの材料群からな って、比誘電率は 4.4 である。

4. 7 評価手順の課題

今回の調査では「特許文献機械翻訳の品質評価手順 Ver.1.0 平成26年6月」の「3.1.1. 内容の伝達レベルの評価」を中日機械翻訳結果の評価に用いたが、この評価作業で判明した評価手順そのものの特徴と課題を整理した。

(1) 評価手順の特徴と課題

今回の品質評価手順を使用した評価の長所は、①評価結果の信頼性が良いこと、②機械翻訳結果が原文の内容をどの程度伝えているか明確であることである。一方、短所は、③評価そのものの難しさ、④評価の再実行の難しさ、⑤誤りの識別の難しさであった。

①評価結果の信頼性が良いことは従来の機械翻訳の品質評価値が5:all meaning、4:most meaning、3:much meaning、2:little meaning、1:noneのように感覚的に決められていたのに対して、内容の伝達レベルが5:100%、4:80%～、3:50%～、2:20%～、1:～20%のように数値で目安が示された上に、英日翻訳であるが例が示されていることに依る。今回、同一の評価点の機械翻訳Sと機械翻訳Dに限定して評価者が主観による相対評価をした結果、機械翻訳Sと機械翻訳Dで差が無いという結果（表4.4-6）が得られたことは、この品質評価手順による評価点が安定していて信頼できることを示している。

②機械翻訳結果が原文の内容をどの程度伝えているかについては、今回の英日翻訳の例がある程度役に立ったが、本調査の中日翻訳の評価点の例を生かしてさらに明確になることを期待する。

次に短所について説明する。③評価そのものの難しさの原因は今回の中日機械翻訳結果の品質の低さである。評価者は最初に、原文の単語がどのように翻訳され、どのように並べられているかを把握するが、機械翻訳結果を読んでも理解できない文の場合には非常に時間と集中力を要する作業となる。また、今回評価者は特許翻訳の専門家5名の文法知識に差があり、評価基準の理解及び指導に時間がかかる人がいた。また、文法知識の少ない人は誤りの指摘に主語・述語・目的語等の用語を使わずにコメントするため、とりまとめの分析者が誤りの分類に苦勞する場面があった。

④評価の再実行の難しさは別の機械翻訳ソフトの評価にも今回と同様の時間とコストが必要となることである。これは本調査に限らず、人手による機械翻訳結果の評価に共通する課題である。

⑤誤りの識別の難しさは中日機械翻訳の誤りは複数の誤りが複合して現れることが多く、分析者にとってもこれらを個別の誤りに分離することが難しかったことである。分析者は自身の機械翻訳プロセスの知識及び中国語と日本語の文法知識をもとに、機械翻訳の過程でどのような現象が起きたかを想像して誤りの整理・分類や評価者が挙げなかった課題の抽出等を行った。今回の調査で中日の翻訳言語別任意項目の抽出ができたので、誤りの識別の難しさが軽減できることを期待する。

(2) テストセットの課題

品質評価に用いたテストセットの課題を挙げる。まず、各テスト文に付けられた参考人手翻訳が対訳文として正しくない場合がある。本テストセットはパテントファミリーの特許明細書から作成した対訳コーパスから抽出したと思われる。この抽出は文献の明細書と辞書による単語対応の判定に基づくが、単体の文として取り出した場合、周辺の文脈から独立してしまうため、対訳文として正しくない場合がある。例えば、表 4.3-2 評価シート 9 の人手翻訳「同成分」と機械翻訳「ロイヤル・ゼリー」がある。つまり、この参考人手翻訳に基づいて機械翻訳結果を評価することは評価精度の低下につながるので、現状では中国語原文を理解することが必須であり、特許文献の中日翻訳の知識無しに、機械翻訳結果の品質を精度よく評価することは難しい。今回、一部のテスト文について評価時に気付いた参考人手翻訳の誤りを記録したので、この課題が軽減されることを期待する。

次に、テスト文の選定方法が課題である。今回のテストセットは一つのパテントファミリーから複数文を選定している（図 4.2-1）。そのため、ある特定の技術用語が複数のテスト文の評価のキーポイントとなる場合がある。辞書作成において文献単位の出現頻度に基づいた用語選定を行って特定の文献の影響を排除しているのに、テストセットが特定の文献に偏るのは好ましくない。

4. 8 中日機械翻訳に特有の課題

(1) 課題の抽出対象

今回の評価用テストセット 150 文の機械翻訳 B、機械翻訳 S 及び機械翻訳 D について、評価者が表 4.2-1 の品質評価シートを作成した。その 150 枚の品質評価シートを中国語の文法の専門家 1 名（分析者）が熟読して中日翻訳特有の課題を抽出した。

(2) 課題の指摘件数

品質評価シートで課題の指摘があった翻訳文は、評価対象の機械翻訳 B、機械翻訳 S 及び機械翻訳 D いずれも約 90% であった。この割合を表 4.8-1 に示す。次に課題が記入されなかった翻訳文の評価点の表 4.8-2 を示す。この表 4.8-2 を見ると評価点 3 又は 4 の文より、翻訳がよく理解できない評価点 1 又は 2 の文の方が若干多い。この理由であるが、技術分野の翻訳の専門家である評価者が翻訳品質の低い機械翻訳文から機械翻訳の課題を洗い出すことが難しいと考えられる。そのため、中国語の文法の専門家である分析者が各評価シートを熟読したうえで評価者があげた課題の分類や評価者があげなかった課題の抽出を行ってテストセット 150 文全てを分析の対象とした。

表 4.8-1 品質評価シートの課題ありの指摘件数

課題あり・無しの判断	機械翻訳 S	機械翻訳 D	機械翻訳 B
課題あり	139 文(92.7%)	57 文(93.4%)	78 文(87.6%)
課題無し	11 文(7.3%)	4 文(6.6%)	11 文(12.4%)
総文数	150 文(100%)	61 文(100%)	89 文(100%)

表 4.8-2 課題が記入されなかった翻訳文の評価

評価点	機械翻訳 S	機械翻訳 D	機械翻訳 B
1 又は 2	7 文(63.6%)	2 文(50%)	9 文(81.8%)
3 又は 4	4 文(36.4%)	2 文(50%)	2 文(19.2%)
合計	11 文(100%)	4 文(100%)	11 文(100%)

(3) 中日の翻訳言語別任意項目

評価者が作成した品質評価シートからの課題の抽出にあたって、分析者が分類した課題毎の件数の算出を試みたが、中日機械翻訳の誤りは複数の誤りが複合して現れることが多く、しかも機械翻訳の品質が低い文では特に複合した誤りを個別の誤りに分離することが難しかった。そのため、誤りの出現回数を集計することはあきらめ、分析者の知識と主観により、目立つ順に誤りを整理・分類して、表 4.8-3 の中日の翻訳言語別任意項目リストを作成した。本リストは特許文献機械翻訳の品質評価手順の 4.1.3.【翻訳言語別任意項目】の中日版の素材となることを意識して作成した。

表 4.8-3 中日の翻訳言語別任意項目リスト

CJ-1.中国語の文型が対応する日本語文型に翻訳されていない誤り
CJ-2.中国の特許文でよく用いられるフレーズが日本語に正しく翻訳されない誤り
CJ-3.分離前置詞（「在～两侧」で「～両側に」）の翻訳誤り
CJ-4.辞書の不足による誤り
CJ-5.特許明細書の見出しに使われる語が文中でも使われる誤り
CJ-6.動詞が連続して現れる場合の翻訳の誤り
CJ-7.中国語文の述語が述語として認識されない誤り
CJ-8.単語分割の誤り
CJ-9.中国語文の複合語の境界が正しく判定されない誤り
CJ-10.中国語文の名詞句の境界が正しく認識されない誤り
CJ-11.中国語の複合語に含まれる動詞が述語として翻訳されてしまう誤り
CJ-12.目的語が目的語として認識されない誤り
CJ-13.主語が主語として認識されない誤り
CJ-14.目的語が前置されている場合の翻訳誤り
CJ-15.主語でないものが主語として認識されてしまう誤り
CJ-16.中国語の特有の表現を日本語に直訳する不自然さ

以下に中日の翻訳言語別任意項目を例で説明する。

CJ-1 中国語の文型が対応する日本語文型に翻訳されていない誤り

評価者が最も頻繁に挙げた課題は、中国語の文型が対応する日本語の文型に正しく翻訳されていないことである。例えば、表 4.8-4 で評価者が「包括 A, B, C, D, E 在内的」というパターンは日本語の「A, B, C, D, E を含む」に翻訳するべきである。指摘箇所を明示するため、この例のみ、人手翻訳と機械翻訳 S に【始まり】を補った。

表 4.8-4 中国語の文型が対応する日本語文型に翻訳されていない誤り（評価シート 1）

原文	本发明用途之一的化妆品可以是 包括 软膏剂、溶液、乳霜、乳液、化妆水、洗剂、凝胶、美容液、粉底霜、填充物、面膜、口红、粘粘剂、淋浴液等各种皮肤外用剂 在内的 任何一种形式的化妆品。
人手翻訳	当発明の用途の一つめである化粧品は、【始まり】軟膏剤、溶液、クリーム、乳液、化粧水、洗剤、ゲル、美容液、ファンデーション、スタッフイング、フェイシャルパック、口紅、スティッキー・エージェント、ボディソープ等各種の皮膚外用剤 を含む いかなる種類の形態の化粧品を含むことが出来る。

機械翻訳 S	本発明の用途の中の一つの化粧品は【始まり】軟膏剤、溶液、クリーム、エマルジョン、化粧水、ローション、ゲル類、美容液、ファンデーション、詰め物、フェイスマスク、リップスティック、が接着剤にくっついてを、浴液などの各種の皮膚外用剤をかける含めるいかなるフォームの化粧品であるなことができるである。
挙げられた課題	・「包括 A, B, C, D, E 在内的」を「A, B, C, D, E を含む」に翻訳する必要がある。このパターンは特許明細書において頻繁に出てくるものである。 ・「可以是～」を「～とすることができる」に翻訳する必要がある。 ・中国語単語分割の誤り

CJ-2 中国の特許文でよく用いられるフレーズが日本語に正しく翻訳されない誤り

中国の特許文献でよく用いられるフレーズが、意味的に対応する日本語パターンに正しく翻訳できていない場合がある。例えば、下記の例では、中国語の「本发明的目的是（在于）提供一种～」は特許文でよく用いられるフレーズであり、日本語の「本発明は～を提供することを目的とする」と翻訳されるべきだが、正しく翻訳されていない。

表 4.8-5 中国語特許文でよく用いられるフレーズが日本語に正しく翻訳されていない誤り
(評価シート 94)

原文	本发明的目的是提供一种汽轮机数字电液调节系统，该调节系统将计算机技术、现代控制理论和计算机网络通讯技术结合在一起，集测量、控制、故障应急处理、分析诊断、操作指导、优化运行以及生产管理于一体，有效地满足了电力企业对汽轮机综合控制的要求。
人手翻訳	本発明は、蒸気タービンデジタル電気油圧制御システムを提供することを目的とし、この制御システムがコンピュータ技術、現代制御理論及びコンピュータネットワーク通信技術を一緒に合わせて、測定、制御、故障応急処理、分析診断、操作指導、最適化運転及び生産管理を一体に集め、電力企業が蒸気タービンを総合制御する要求を有効的に満たす。
機械翻訳 D	本発明の目的は提案蒸気タービンのデジタルの電気油圧の調節システムで、この調節システムはコンピュータ技術、現代制御理論とコンピュータネットワークの通信技術を互いに結合して、一体で測定、制御、故障応急処理、分析診断、操作案内、最適化運行と生産管理に集まって、有効に満足電力企業蒸気タービンの統合制御に対する要求。

挙げられた課題	・ 述語の判定の誤り ・ 助詞の利用の誤り ・ 一般的には、「本発明的目的是（在于）提供一种～」は「本発明は～を提供することを目的とする。」に翻訳する必要がある。このパターンは特許明細書において頻繁に出てくるものである。
---------	--

CJ-3. 分離前置詞（「在～两侧」で「～両側に」）の翻訳誤り

この課題は、特許庁の仕様書で例示されているものである。今回の評価でも、評価者から課題として挙げられた。例では中国語の「在～两侧」というパターンが日本語の「～両側に」と正しく翻訳されなかった。

表 4.8-6 分離前置詞（「在～两侧」で「～両側に」）の翻訳誤り（評価シート 51）

原文	本发明进一步的技术方案是：构成消毒液喷射单元、热水喷射单元及无菌水喷射单元的喷头分别设置在瓶盖通道的上、下 两侧 。
人手翻訳	本発明の一步進んだ技術は：消毒液噴射口ユニット、お湯噴射口ユニットおよび無菌水噴射口ユニットの噴射ノズルはそれぞれキャップ（消毒）ラインの上、下 両側に 設けてある。
機械翻訳 S	本発明更にの技術方案は：構成の消毒液はパンチユニット、温水を噴き出してパンチユニットと無菌水を噴き出してパンチのユニットの吐出ヘッドの個別設定がキャップの通路のの上で、下で 両側を 噴き出す。
挙げられた課題	・ 中国語文の複合語の境界の判定の誤り ・ 中国語複合語に含まれる動詞を述語と認識している ・ 中国語文の述語の判定の誤り ・ 「将～设置在～两侧」は「～を～に設置する」に、「设在～两侧的」は「～に設置された～」に、「～设置在～」は「～が～に設置される」に翻訳する必要がある。このパターンは特許明細書において頻繁に出てくるものである。

CJ-4. 辞書の語彙不足による誤り

必要な中日対訳用語が辞書に登録されていない課題が多数あった。下記の例では、「**初镀膜**（初期成膜）」を辞書登録する必要がある。

表 4.5-7 辞書の語彙不足による誤り（評価シート 138）

原文	制程热处理测试实验中依序以 初镀膜 (As-deposited)、银胶固化 (curing)、退火 (anneal)、老化处理 (aging)、回流 (reflow) 等程序进行，且热处理测试实验的 操作温度 范围介于 150～350℃ 之间。
----	--

人手翻訳	プロセス熱処理テスト実験中、順次に 初期成膜 (As-deposited)、銀接着剤硬化(curing)、アニール(anneal)、老化処理(aging)、逆流(reflow)等の工程を行い、且つ熱処理テスト実験の 操作温度 範囲が150～350℃の間にある。
機械翻訳S	制作プロセスの熱処理の試験的実験の中で順番に 初めのメッキ膜 (As-deposited)で、銀ペーストの硬化(curing)、アニール(anneal)、エージング処理(aging)、環流(reflow)などプログラムは行って、しかも、熱処理の試験的実験の 動作温度 範囲は150～350℃間に介在する。
挙げられた課題	・「初镀膜：初期成膜」を辞書に登録する必要あり。 ・「操作温度：操作温度」を辞書に登録する必要あり。

CJ-5. 特許明細書の見出しに使われる語が文中でも使われる誤り

中国特許明細書の見出しの【实施方式】を日本語の【実施方法】と翻訳するためにユーザ辞書に登録されている語が、文中の訳語として使われると誤った翻訳になってしまう。

表 4.8-8 特許明細書の見出しに使われる語が文中でも使われる誤り（評価シート 62）

原文	本发明具体的 实施方式 是将催化剂均匀的分散在离子液体中，加入氧化剂，醛和醇在适当的温度和 pH 值条件下反应生成羧酸酯。
人手翻訳	本発明の具体的 実施方法 については次の通りである。触媒剤をイオン液体の中にまんべんなく分散させ、酸化剤を加える。アルデヒドとアルコールを適切な温度と pH 値の条件の下で反応させてカルボン酸樹脂を生産する。
機械翻訳S	本発明の具体的 実施形態 は触媒のむらがない非集中をイオン液の中で、参加の酸化剤、アルデヒドとアルコール適切な温度と水素イオン指数の条件の下で反応生成のカルボン酸エステル。
挙げられた課題	・「 实施方式 」は「 実施形態 」が特許明細書における慣用表現であるがこれが文中で使われると正しくない場合がある。

CJ-6. 動詞が連続して現れる場合の翻訳の誤り

中国語の動詞が連続して現れると、翻訳が誤ってしまう場合がある。下記の例では、「**选择**（選択する）」と「**采用**（採用する）」が連続して現れている。「を採用することを選択する」と翻訳されるべきだが、正しく翻訳できなかった。

表 4.8-9 動詞が連続して現れる場合の翻訳の誤り（評価シート 10）

原文	水解过程采用一种以上蛋白酶时,可根据不同蛋白酶适宜水解的PH值和温度, 选择采用 同时加入或依次加入进行水解。
----	--

人手翻訳	加水分解には、一種類以上のプロテアーゼを使用する場合、異なるプロテアーゼの加水分解に適した PH 値と温度に基づいて、同時に添加または順番に添加して加水分解を行うことを 採用することを選択できる 。【人手翻訳を一部修正】
機械翻訳 S	加水分解プロセスの採用の一種類以上のプロテアーゼの時に、しかし、異なるプロテアーゼの好適の加水分解の pH 値と温度によって、 選択の採用 の同時添加はあるいは、順次参加は加水分解を行う。
挙げられた課題	・「 选择采用同时加入或依次加入进行水解。 」は「同時添加または順次添加して加水分解を行うことを 採用することを選択できる 。」と翻訳するべき。ここで「 选择 」、「 采用 」と二つの動詞が続いている。

CJ-7. 中国語文の述語が述語として認識されていない誤り

中国語文の述語が述語として認識されない誤りがある。下記の例文では、中国語「**延伸**（延ばし）」も「**具备**（備える）」も述語であるが、機械翻訳ソフトが述語と認識しなかったので、訳文では、それぞれ「延在」「備え付け」と翻訳された。

表 4.8-10 中国語文の述語が述語として認識されていない誤り（評価シート 142）

原文	经数值仿真计算和天线测试，本例角反射天线的工作频段可以从 1.25GHz 延伸到 2.02GHz， 具备 超宽带工作性能。
人手翻訳	数値シミュレーション計算とアンテナの測定試験の結果、本例のコーナアンテナの作業周波数帯は、1.25GHz から 2.02GHz まで 延ばし 、スーパーブロードバンドの動作性能を 備える ことが可能になった。
機械翻訳 S	数値のシミュレーション計算とアンテナテストを通じて、本例の角反射のアンテナの作業の周波数帯域は 1.25GHz 延在 のから 2.02GHz まで(に)ことができて、 備え付け の超広帯域の動作性能。
挙げられた課題	・ 述語の判定の誤り ・ 中国語文の後半の構造を正しく取り扱うのが難しい

CJ-8. 単語分割の誤り

中国語の単語分割の誤りが機械翻訳の誤りの原因となる場合がある。下記の例では、「**后防**」が「後方守備」と翻訳されているが、「**后**（後）」と「**防**（防止）」に分けて翻訳されるべきである。なお、「**后防**」はユーザ辞書には無いので、採用した機械翻訳ソフトの辞書による誤りと思われる。

表 4.8-11 単語分割の誤り（評価シート 20）

原文	本发明属于一种新型的防粘连剂，特别是用于手术后 后防 粘连的带药加铁离子的透明质酸凝胶，在防止外科手术组织粘连领域具有非常广泛的应用。
人手翻訳	本発明は、一種の新しい癒着防止剤であり、特に手術 後 の癒着 防止 に用いられる薬を付した鉄イオンのヒアルロン酸ゲルの調製方法とその用途に用いられ、外科手術後の組織の癒着を防止する分野において、広く使用されている。
機械翻訳 S	本発明は新型のアンチブロッキング剤に属して、特に手術の 後方守備 の癒着の付いている薬の鉄イオンのヒアルロン酸ゲルに用いて、防止の外科手術の後で組織癒着の分野はとても広範な施用がある。
挙げられた課題	・「手術 后防 粘连」の「 后防 」が一つの用語として誤って認識され、「 後方守備 」になった。

CJ-9. 中国語文の複合語の境界が正しく判定されない誤り

中国語文の複合語の境界が正しく認識されない誤りが多数存在している。下記の例文では、複合語「**外卡套钎杆(ドリルロッドカバーシャンク)**」という複合語の訳語が辞書に無かったため、複合語の先頭の「外」が別の語として「外に」と翻訳されてしまい「**外にフェルールのたがね竿尻**」と翻訳された。本来は「**外卡套钎杆**」を辞書登録して、一つの複合語として扱うべきである。

表 4.8-12 中国語文の複合語の境界が正しく判定されない誤り（評価シート 80）

原文	所述的采用高温挤压固定在新型外卡套钎杆尾中空六角钢的内接圆 22mm 处，为采用高温挤压固定在新型 外卡套钎杆 尾 108mm 后 32mm 宽的中空六角钢的内接圆 22mm 处。
人手翻訳	前記高温押出は新型ドリルロッドカバーのシャンク中空六角鋼の内接円 22mm の部分に固定し、高温押出を新型 ドリルロッドカバーシャンク 後 108mm、幅 32mm の中空六角鋼の内接円 22mm の所に固定する。
機械翻訳 S	前記した採用の高温の押出しは新型である外フェルールのたがねの竿尻の空洞の六角鋼の内接円 22mm 所で固定して、採用の高温の押出しのために新型の 外にフェルールのたがね竿尻 108mm 後 32mm の広い空洞六角鋼の内接円 22mm 所で固定する。
挙げられた課題	・ 中国語文の複合語の境界の判定の誤り ・ 中国語文の述語の判定の誤り

CJ-10. 中国語文の名詞句の境界が正しく認識されない誤り

複合語でない名詞句の境界が正しく判定できない誤りがある。下記の例文では、「**其数量和间距**（その数量と間隔）」という名詞句が一つのかたまりとして認識できなかった。

表 4.8-13 中国語文の名詞句の境界が正しく認識されない誤り（評価シート 84）

原文	上述钢支架沿下定位板环向设置， 其数量和间距 按照预埋地脚螺栓型号、间距、上下定位板系统的重量、平面外刚度、混凝土浇筑方式和水平推力大小等由施工设计综合确定，一般为均匀布置。
人手翻訳	前記鋼製ブラケットを下部位置決め板に沿って円周方向に設置し、 その数量と間隔 を埋め込みアンカーボルトのモデル、間隔、上下部位置決め板システムの重量、平面外剛性、コンクリートの注入方式と水平推力の大きさ等によって工事設計で総合的に確定し、一般的には均一に配置する。
機械翻訳 S	上述の鋼製ブラケットは下位置を定める板リングに沿って設置に向って、 その数量はピッチ とあらかじめ上から下まで位置を定める板システムの重量、平面の外でアンカーボルトモデル、ピッチ、剛性、コンクリート注入の形態と水平推力の大きさなどを埋めるによって施工設計の総合の決定から、普通はむらがない配置である。
挙げられた課題	・ 名詞句の境界の誤り ・ 複合語の境界の誤り ・ 述語の判定の誤り ・ 中国語文の構造を正しく取り扱うのが、難しい

CJ-11. 中国語の複合語に含まれる動詞が述語として翻訳されてしまう誤り

中国語の複合語に含まれる動詞が文全体の述語として認識されて翻訳される誤りがある。下記の文では、複合語「**排砂系统**（砂を排出するシステム）」に含まれる動詞「**排**（排出）」が述語動詞「**並ぶ**」として訳された。

表 4.8-14 中国語の複合語に含まれる動詞が述語として翻訳されてしまう誤り

（評価シート 29）

原文	所述的工业机器人砂型铣削加工方法中的砂型铣削，在砂型铣削过程中，工作台的绕轴转动以及机器人端部增加 排砂系统 能够将铣削过程中产生的废砂及时清理。
人手翻訳	前述のロボットを用いた砂型フライス加工方法中のフライスにおいて、砂型のフライスのプロセスで作業プラットフォームの軸回転及びロボットの先端部に 砂を排出するシステム により、フライスの過程において発生する廃棄砂を速やかに処理することができる。

機械翻訳 S	前記した産業用ロボットの砂型のフライス加工方法の中の砂型のフライス、砂型のフライスで過程で、作業台の軸周りの回転とロボットの端部の増加はフライスを砂のシステム過程で生成の廃砂タイムリーなクリーニングに並ぶ。
挙げられた課題	・中国語文の複合語の境界の判定の誤り ・中国語複合語に含まれる動詞は述語と認識してしまう ・中国語文の述語の判定の誤り ・「所述」を「前記」に翻訳する必要がある。このパターンは特許明細書、特に特許請求の範囲において頻繁に出てくるものである

CJ-12. 目的語が目的語として認識されない誤り

中国語の場合、目的語は述語動詞の後ろに置くのが一般的である。一方、日本語は目的語を述語動詞の前に置く必要がある。そのため、機械翻訳システムが述語を正しく認識しないと日本語の順序が正しく生成されないこととなる。この機械翻訳文では、「**以上の 76g 亚硝基丙二酸二乙酯(0.4mol)** (以上の 76g ニトロソ基マロン酸ジエチル (0.4mol))」という目的語の位置が誤っており、目的語として認識されていない。本件は CJ-7,CJ-11 と同時に現れることが多い。

表 4.8-15 目的語が目的語として認識されない誤り (評価シート 67)

原文	醋酸钠、醋酸的缓冲溶液(PH=4-5)中 加入以上的 76g 亚硝基丙二酸二乙酯(0.4mol) 在 30-40℃下, 通入 5-6 伏的直流电压进行电还原, 点板, 当原料点消失反应到达终点, 经萃取得到氨基丙二酸二乙酯, 其纯度≥99%。
人手翻訳	酢酸ナトリウム、酢酸の緩衝溶液(4~5)中に、 以上の 76g ニトロソ基マロン酸ジエチル (0.4mol) を加える。30~40℃で 5~6 ボルトの直流電圧をかけて電気還元を行い、薄層クロマトグラフィーにかける。原料ポイントが消失し反応が最終ポイントに達する。アミノマロン酸ジエチルを抽出して得る。その純度は≥99%となる。
機械翻訳 S	酢酸ナトリウム、酢酸の緩衝液(PH=4-5)の中で 参加以上の 76g ニトロソマロン酸ジエチル(0.4mol) は 30-40℃の下で、通入 5-6 ボルトの直流電圧は電気還元を行って、点板、原料が消失の反応を注文して時に終点に到達する、抽出するを通じてアンモニア系のマロン酸ジエチルを得て、その純度≥99%。
挙げられた課題	「 加入 」は「参加」でなく、「 添加 」に翻訳し、化学物質名の後ろに置くべきである。

CJ-13 主語が主語として認識されない誤り

中国語の主語となる名詞句が正しく判定されない場合がある。例えば、下記の翻訳文では、「**产品更新换代的速度（製品の更新、リニューアルのペース）**」が主語である。日本語に翻訳する際に、主語の後ろに「が」という助詞が必要である。しかし、機械翻訳システムは主語の判定に失敗して翻訳文では「が」が欠落した。

表 4.8-16 主語が主語として認識されない誤り（評価シート 28）

原文	随着市场全球化以及竞争的不断加剧， 产品更新换代的速度 不断加快，单件、小批量铸件的需求越来越多，单件、小批量尤其是大型铸件的砂型制造工艺生产要求生产周期短，制造手段更加具有柔性。
人手翻訳	市場のグローバル化及び競争の絶え間ない激化に伴い、 製品のリニューアルのペース が絶え間なく加速化し、単品、小ロットの铸件に対する需要がますます増加している。単品、小ロット、特に大型の铸件の砂型の製造プロセスは、生産サイクルが短くなり、製造手段がより柔軟になることを必要とする。【人手翻訳を一部修正】
機械翻訳 S	市場のグローバル化と競争の絶え間ない助長に従って、 モデルチェンジの速度 絶え間ない早め、ワンピース、小ロットの铸件の需要はだんだん多くなって、ワンピース、小ロットは特に大型铸件の砂型の製造プロセスの生産要求の生産周期が短くて、製造手段はいっそう可撓性がある。
挙げられた課題	・ 述語の判定の誤り ・ 主題語の判定の誤り ・ 格助詞「が」の欠落（中国語の主語が主語として認識されない誤り） ・ 格助詞「は」の誤用

CJ-14. 目的語が前置されている場合の翻訳誤り

この課題は、特許庁の仕様で例示されている。中国語文は通常 SVO（主語・動詞・目的語）の語順であるが、目的語を強調する目的で SOV と語順が変わる場合がある。この目的語の前置による強調が今回のテストセットにも含まれている。下記の文では、「**塞锥（テーパプラグ）**」は述語動詞「**设计（設計する）**」の目的語であり、それを強調するために述語動詞の前に置かれている。翻訳文では、「きりを詰め込む設計は“瓦”形の曲面で」と誤って翻訳された。

表 4.8-17 目的語が前置されている場合の翻訳誤り（評価シート 87）

原文	本发明的一种塞式喷管，由多个内喷管、塞锥组成，其内喷管为轴对称结构， 塞锥设计为“瓦”状曲面 ，内喷管可分为收敛段、喉部、扩张段，内喷管和塞锥为一体设计。
----	--

人手翻訳	本発明のプラグノズルは、複数のインナーノズル、テーパプラグからなり、そのインナーノズルが軸対称構造であり、 テーパプラグが「瓦」状曲面に設計され 、インナーノズルが収束段、スロート部、拡張段に分けられることができ、インナーノズルとテーパプラグとが一体設計である。
機械翻訳 S	本発明のプラグノズル、多数の内ですパウトの、からきりの組成を詰め込んで、その内ですパウトは軸対称構造で、 きりを詰め込む設計は“瓦”形の曲面で 、内スパウトは収束段、スロート部、拡張段に分けることができ、内スパウトがときりを詰め込むのは一体の設計である。
挙げられた課題	・ 複合語や名詞句の境界の誤り ・ 主題語や述語の判定の誤り ・ 中国語文構造を正しく取り扱うのが、難しい ・ 一般的には、「多个～」は、「複数の～」に翻訳する必要がある。このパターンは特許明細書において頻繁に出てくるものである。 ・ 一般的には、「由～组成」は、「～からなる、～で構成される」に翻訳する必要がある。このパターンは特許明細書において頻繁に出てくるものである。

CJ-15 主語でないものが主語として認識されてしまう誤り

主語でないものが主語として認識されてしまう誤りがある。「**码头（埠頭）**」は主語でなく、場所を表すものであるが、機械翻訳 S では「**埠頭**」が主語として翻訳されたため、格助詞「が」が付けられた。

表 4.8-18 主語でないものが主語として認識されてしまう誤り（評価シート 116）

原文	本发明涉及散货 码头 装船过程中自动化检测装置和方法，检测装船系统和船之间的位置关系及物料添加情况，为自动化装船系统提供检测数据。
人手翻訳	本発明は、バルク貨物が 埠頭 で船積みされる過程における自動化検出装置および方法に関し、船積みシステムと船舶との位置関係及び貨物の添加状況を検出して、自動化船積みシステムに検出データを提供する。
機械翻訳 S	本発明はルーズ貨物の 埠頭が 船積みするのに過程でオートメーション化の検出装置と方法関連して、検出の船積みするシステムと船の間の位置関係と物の付加の進行度合、自動化船積みするシステム提供の検出データのため。
挙げられた課題	・ 文の構造が正しくないので、文の意味が理解しにくい ・ 助詞の判定の誤り ・ 複合語の境界の判定の誤り ・ 主語の判定の誤り

CJ-16. 中国語の表現を日本語に直訳する不自然さ

中国語の表現を日本語に直訳すると、不自然になる場合がある。下記の文では、「不能达到」は「実現できない」と翻訳するのが自然だが、機械翻訳Sでは、「達することはできない」と訳され、不自然さが感じられる。今回のユーザ辞書に「达到」を含む語は無いので、この翻訳は機械翻訳システムの辞書に由来する。

表 4.8-19 中国語の表現を日本語に直訳する不自然さ（評価シート 92）

原文	我国城市生活垃圾的收集、储运和处理问题越来越受到人们的重视，城市生活垃圾资源化利用技术和途径越来越多，目前的处理主要有填埋、堆肥、焚烧等方式，而垃圾焚烧发电处理有着其他方式不能达到的资源化、减量化和无害化。
人手翻訳	我が国の都市家庭ゴミの回収、保管、輸送および処理問題は、徐々に人々の関心を集め、都市家庭ゴミの資源化利用の技術および方法は徐々に多様化し、現在主要な処理方法は埋設、堆肥、焼却等であり、ごみ焼却発電処理はその他の方法では実現できないゴミの資源化、減量化および無害化を実現する。
機械翻訳 S	我が国の都市生活ゴミの収集、貯蔵輸送と処理問題はだんだん人々の重視を受けて、都市生活ゴミの資源化利用技術と径路はだんだん多くなって、現在の処理は主に穴埋め、堆肥、焼却など形態があって、ゴミ焼却発電の処理はその他の形態達することはできない資源化、減量化と無害化を持っている。
挙げられた課題	・ 主題語の判定の誤り ・ 助詞の利用の誤り ・ 「達することはできない」は中国語の特有の表現で、日本語に直訳すると、不自然になる

(4) 特許庁の仕様書で例示された課題

特許庁の仕様書で例示された課題と本調査で判明した中日の翻訳言語別任意項目の関係を整理する。特許庁の仕様書で例示された課題は以下の 5 つである。

表 4.8-20 特許庁の仕様書で例示された課題

- | |
|---|
| <ul style="list-style-type: none">(1) 分離前置詞（「在～上」で「～の上に、～に関して」、等）の翻訳誤り(2) 4 文字熟語の翻訳誤り(3) 目的語が前置されている場合（目的語の強調のために、SOV とされる場合がある）の翻訳誤り(4) 表記の一部を重ねて表現する重ね方に対処できていない。（例：「散歩（散歩する）」→「散散歩」（ちょっと散歩する））(5) 外来語の翻訳誤り |
|---|

そのうち、(1)と(3)については今回の評価で「CJ-3」と「CJ-14」の課題として顕在化した。一方、(2)と(4)は、特許文献であまり出てこない表現であり、今回の評価で事例が出現しなかった。(5)は特許文献に多く現れると考えられるが、残念ながら今回のテストセットに見つけることができなかったので中日の翻訳言語別任意項目の課題として挙げるができなかった。

(5) 用語辞書の副作用に関する分析

ここまで評価用テストセット 150 文の機械翻訳 B、機械翻訳 S 及び機械翻訳 D について、中国語の文法の専門家 1 名（分析者）が熟読して中日翻訳特有の課題を抽出した。この中で、ユーザ辞書の使用が翻訳品質を改善しない場合があった。このような場合は複数あると予想されるが、テストセットで判明した副作用を一つ挙げる。

(a) 中国語の連続して現れる動詞を用語として採用

中国語の連続して現れる動詞を「反応生成：反応生成」としてユーザ辞書に登録した。その結果、翻訳結果が「反応生成」となってしまったが、ユーザ辞書無しの翻訳結果 B はそれぞれを正しく動詞と認識して「反応して～を形成する」と翻訳することができた。

中国語の連続して現れる動詞を名詞として辞書登録した方が良いかどうかは検討が必要である。

表 4.8-21 用語辞書の副作用の例 （評価シート 62）

原文	本発明具体的实施方式是将催化剂均匀的分散在离子液体中，加入氧化剂，醛和醇在适当的温度和 pH 值条件下 反应生成 羧酸酯。
人手翻訳	本発明の具体的実施形態については次の通りである。触媒をイオン性液体中に均一に分散させ、酸化剤を加え、アルデヒドとアルコールを適切な温度と pH 値の条件の下で 反応させて カルボン酸エステルを 生成する 。
翻訳結果 S	本発明の具体的実施形態は触媒のむらがない非集中をイオン液の中で、参加の酸化剤、アルデヒドとアルコール適切な温度と水素イオン指数の条件の下で 反応生成 のカルボン酸エステル。
翻訳結果 B	本発明の具体的な実施方法は触媒のむらがない分散をイオン液体中で、酸化剤に参加して、アルデヒドとアルコールは適切な温度と水素イオン指数条件下で 反応して カルボン酸塩を 形成する 。
副作用の原因 となった用語	反応生成：反応生成

添付資料

- 添付資料 3. 1 対応中国・日本公開特許公報番号リストレイアウト説明
- 添付資料 3. 2 対応中国公開特許公報・和文抄録番号リストレイアウト説明
- 添付資料 3. 3 対訳コーパスレイアウト説明
- 添付資料 3. 4 人手確認用対訳辞書候補データレイアウト説明
- 添付資料 3. 5 人手確認により対訳辞書から除外したデータレイアウト説明
- 添付資料 3. 6 対訳辞書データレイアウト説明
- 添付資料 4. 1 機械翻訳結果評価シート（サンプル）

添付資料 3. 1 対応中国・日本公開特許公報番号リストレイアウト説明

1. ファイル名

h26fy_family_numlist.tab.gz

(補足) ファイルは **gunzip** で圧縮されている。

2. ファイル形式

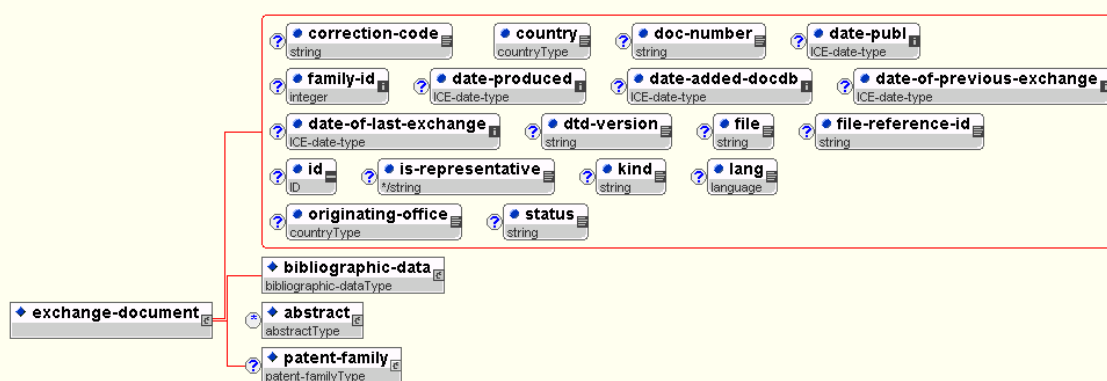
対応中国・日本公開特許公報番号リストは、Unix(Linux)で作成したテキストデータである。各行の項目はタブで区切られている。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	ascii
3	レコード区切り文字	0x0a

3. レコードフォーマット

本データは DOCDB のバックファイルから作成した。具体的には日本データの **kind='A'** と中国データの **kind='A'** のデータからファミリーID が共通するデータを抽出した。設定される番号等の形式は下記の DOCDB の基準に準ずる。また、IPC は日本データの **ipcr** の先頭のデータを **ipc** 使用した。

7 ELEMENT <exch:exchange-document>



<http://www.epo.org/patents/patent-information/raw-data/manuals.html> より

対応中国・日本公開特許公報番号リストの各レコードは下記表に示す項目から構成されている。各項目は区切り文字（タブコード(0x09)）で区切られている。

項番	項目名 DOCDB 名	説明
1	国コード country	C ‘CN’固定
2	公報種別 kind	C ‘A’固定
3	公開番号 doc-number	中国特許の公開番号が設定される。
4	公開日 date-publ	中国特許の公開日が設定される。
5	ファミリーID family-id	日本のデータと中国のデータを結びつけるキー。
6	国コード country	C ‘JP’固定
7	公報種別 kind	C ‘A’固定
8	公開番号 doc-number	日本の特許の公開番号が設定される。
9	公開日 date-publ	日本の特許の公開日が設定される。
10	国際分類 (IPC) ipc or ipcr	日本の文献の先頭 1 つ目の ipc または ipcr が設定される。 記事の優先順位は ①ipc ②ipcr とした。

4. レコードサンプル

CN	A	1816909 20060809	33562464	JP	A	2005051217	20050224	H01L	21/683
CN	A	1579344 20050216	34395279	JP	A	2005052652	20050303	A61B	17/88
CN	A	1580634 20050216	34368610	JP	A	2005061688	20050310	F23D	14/06
CN	A	1603645 20050406	34373274	JP	A	2005106105	20050421	F16C	33/14

添付資料 3. 2 対応中国公開特許公報・和文抄録番号リストレイアウト説明

1. ファイル名

h26fy_wabun_syouroku_docnum_list.txt.gz

(補足) ファイルは **gnuzip** で圧縮されている。

2. ファイル形式

対応中国・日本公開特許公報番号リストは、Unix(Linux)で作成したテキストデータである。各行の項目はタブで区切られている。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	ascii
3	レコード区切り文字	0x0a

3. レコードフォーマット

対応中国・日本公開特許公報番号リストの各レコードは下記表に示す項目から構成されている。各項目は区切り文字（タブコード(0x09)）で区切られている。

項番	項目名	説明
1	国コード	C 'CN'固定
2	公報種別	C 'A'固定
3	公開番号	和文抄録に記録されている中国特許の公開番号が設定される。
4	公開日	和文抄録に記録されている中国特許の公開日が設定される。
5	予備 1	空値
6	予備 2	空値
7	予備 3	空値
8	予備 4	空値
9	予備 5	空値
10	国際分類 (IPC)	和文抄録に記録されている IPC 1 つ目が設定される。

4. レコードサンプル

CN	A	101619502	20100106	D01D5/06
CN	A	101619503	20100106	D01D5/098
CN	A	101619504	20100106	D01D5/10
CN	A	101619505	20100106	D01D5/10

添付資料 3. 3 対訳コーパスレイアウト説明

1. ファイル構成

中日対訳コーパスは発明の名称、要約、請求項、明細書から作成したもので構成される。
各中日対訳コーパスのファイル名は

h26fy_jpo_cj_corpus_tit.txt.gz

h26fy_jpo_cj_corpus_abs.txt.gz

h26fy_jpo_cj_corpus_clm.txt.gz

h26fy_jpo_cj_corpus_des_xx.gz (xx は'01'から'03')

である。

(補足) ファイルは **gunzip** で圧縮されている。

2. ファイル形式

中日対訳コーパスは、区切り文字 (c' ||| ') で区切られた以下に示す項目が設定される。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	UTF-8
3	レコード区切り文字	0x0a

項番	項目名	説明
1	SSR	文対応スコア。値が大きいほど文と文の対応付け精度が高いこと意味する。
2	DID	どの文献対のどの部分から作成したコーパスかを示す。 項目は作成元のデータにより以下の2つ形式で格納する。 【形式1】 CNAXXX...XXX_JPAYYY...YYY_ZZZ の形式で格納されている。 CN 国コード（中国） A 種別（公開） XXX...XXX は中国特許公開番号 JP 国コード（日本）又は WO（再公表の場合） A 種別（公開） YYY...YYY は（日本特許公開番号、再公表番号または国際公開番号） ZZZ は公報のどの部分から作成した対訳コーパスであるかを示す。(tit:発明の名称、abs:要約、clm:請求項、des:明細書) (例)CNA101336677_JPA2009011234_des 意味：中国公開特許公報の公開番号 101336677 と日本公開特許公報の公開番号 2009011234 の明細書から作成した対訳コーパス 【形式2】 中国公開特許公報と和文抄録データから作成した場合以下の形式で格納する。 CNAXXX...XXX_YYYY YYYY _ZZZ CN 国コード（中国） A 種別（公開） XXX...XXX は中国特許公開番号 YYYY YYYY は c'syouroku'固定 ZZZ は公報のどの部分から作成した対訳コーパスであるかを示す。(tit:発明の名称、abs:要約、clm:請求項、des:明細書)
3	SID	DID における文対応の順番である。

4	AR	文書(DID)レベルでの文対応スコア。値が大きいほど文書(DID)レベルでの対応付け精度が高いことを意味する。
5	JL	文書(DID)に含まれる日本語文の数。
6	CL	文書(DID)に含まれる中国語文の数。
7	R	JL と CL の比率。 (例) JL= 58, CL= 60 のとき、R=0.9666666666666667
8	NM	文対応に含まれ文の数。n・m の形式で格納されている。n は日本語文数、m は中国語文数を示す。
9	IPC	特許分類 (IPC) (先頭 1 つ目)
10	分野情報	分野を表すコードが格納されている。 C00:化学, E00:電気 ,M00:機械, P00:物理 コードは IPC に基づいて作成される。
11	日本語文	対応する文の日本語。複数文があるときには、区切り記号'///' 'で区切られている。
12	中国語文	対応する文の中国語。複数文があるときには、区切り記号'///' 'で区切られている。

0.13471383533333 ||| CNA101336677_JPA22009011234_des.txt ||| 1 ||| 0.350255965583756 ||| 58 ||| 60 ||| 0.9666666666666667 ||| 1-1 ||| A23K 1/18 ||| C00 ||| 本発明は、ペットにおやつとして与えるペット用スナックと、その製造方法に関する。||| 本发明涉及一种作为给予宠物零食的宠物用零食及其制备方法。

添付資料 3. 4 人手確認用対訳辞書候補データレイアウト説明

1. ファイル名

対訳辞書候補データは以下のファイル名で作成される。

cj_candidate_data_h26fy.gz

2. ファイル形式

中日対訳辞書は、下記表に示すファイル形式を下記表に示す。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	UTF-8
3	レコード区切り文字	LF(0x0a)

3. レコードフォーマット

対訳辞書候補データの各レコードは、区切り文字 0x09（タブコード）で区切られた以下の項目から構成される。

項番	項目名	説 明
1	見出し語	中国語が設定される。
2	訳語	日本語が設定される。
3	分野情報	C"C00"(化学) or C"E00"(電気) or C"M00"(機械) or C"P00"(物理)が設定される。
4	複合語フラグ	値が 1 のとき単語、それ以外の場合は複合語
5	出現頻度 (全分野)	全分野の中国語と日本語が同時に出現する頻度（正規化）が設定される。

4. サンプル(項目ごと改行表示)

相互作用 相互作用 C00 1 0.03183123205592311

添付資料 3. 5 人手確認により対訳辞書から除外したデータレイアウト説明

1. ファイル名

対訳辞書候補データは以下のファイル名で作成される。

jyogai_data.gz

2. ファイル形式

除外データは、下記表に示すファイル形式を下記表に示す。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	UTF-8
3	レコード区切り文字	CR+LF(0x0a)

3. レコードフォーマット

項番	項目名	説 明
1	除外理由	削除理由として以下の記号が設定される。 C”D1”：見出し語自体が不適切 C”D2”：訳語自体が不適切 C”D3”：見出し語と訳語が対応していない。 C”D4”：辞書に登録する用語として不適切
2	見出し語	中国語が設定される。
3	訳語	日本語が設定される。
4	分野情報	C”C00”(化学) or C”E00”(電気) or C”M00”(機械) or C”P00”(物理)が設定される。
5	複合語フラグ	値が 1 のとき単語、それ以外の場合は複合語
6	出現頻度 (全分野)	全分野の中国語と日本語が同時に出現する頻度が設定される。

4. サンプル(項目ごと改行表示)

D3
描述
範囲記載
E00
1
83

添付資料 3. 6 中日対訳辞書データレイアウト説明

1. ファイル名

JPO-CJ-DICT-26fy.utx.gz

2. ファイル形式

中日対訳辞書は、下記表に示すファイル形式で作成されている。

項番	項目	内容
1	形式	テキストファイル形式
2	文字エンコーディング	UTF-8
3	レコード区切り文字	CR+LF(0x0d 0x0a)

3. データフォーマット

データフォーマットは UTX1.11 仕様に準拠する。

4. レコードフォーマット

データは、ヘッダ部と本文から構成される。ヘッダ部はさらに辞書情報と列定義から構成される。以下に各レコードのレイアウトを説明する。

4. 1 辞書情報

辞書情報は先頭が C'# で始まり、区切り文字 C'; (半角セミコロン) で区切られた以下の項目が設定される。

項番	項目名	説 明
1	バージョン	C 'UTX 1.11' 固定
2	言語	C 'zh/ja' 固定
3	作成日付	作成日が設定される。(形式は YYYY-MM-DD)
4	作成者	C 'JPO' 固定

4. 2 列定義

列定義は先頭が C'# で始まり、区切り文字（タブコード(0x09)）で区切られた以下の項目が設定される。

項番	項目名	説 明
1	見出し語（中国語）	C'src' 固定
2	訳語（日本語）	C'tgt' 固定
3	品詞（中国語）	C'src:pos' 固定
4	品詞（日本語）	C'tgt:pos' 固定
5	複合語フラグ	C'composit-f'
6	文献単位頻度（全分野）中国語	C'freq-d-all-src' 固定
7	文献単位頻度（全分野）日本語	C'freq-d-all-tgt' 固定
8	文献単位頻度（全分野）中国語・日本語	C'freq-d-all-src-tgt' 固定
9	文献単位頻度（化学分野）中国語	C'freq-d-c-src' 固定
10	文献単位頻度（化学分野）日本語	C'freq-d-c-tgt' 固定
11	文献単位頻度（化学分野）中国語・日本語	C'freq-d-c-src-tgt' 固定
12	文献単位頻度（電気分野）中国語	C'freq-d-e-src' 固定
13	文献単位頻度（電気分野）日本語	C'freq-d-e-tgt' 固定
14	文献単位頻度（電気分野）中国語・日本語	C'freq-d-e-src-tgt' 固定
15	文献単位頻度（機械分野）中国語	C'freq-d-m-src' 固定
16	文献単位頻度（機械分野）日本語	C'freq-d-m-tgt' 固定
17	文献単位頻度（機械分野）中国語・日本語	C'freq-d-m-src-tgt' 固定
18	文献単位頻度（物理分野）中国語	C'freq-d-p-src' 固定
19	文献単位頻度（物理分野）日本語	C'freq-d-p-tgt' 固定
20	文献単位頻度（物理分野）中国語・日本語	C'freq-d-p-src-tgt' 固定
21	文単位頻度（全分野）中国語	C'freq-s-all-src' 固定
22	文単位頻度（全分野）日本語	C'freq-s-all-tgt' 固定
23	文単位頻度（全分野）中国語・日本語	C'freq-s-all-src-tgt' 固定
24	文単位頻度（化学分野）中国語	C'freq-s-c-src' 固定
25	文単位頻度（化学分野）日本語	C'freq-s-c-tgt' 固定
26	文単位頻度（化学分野）中国語・日本語	C'freq-s-c-src-tgt' 固定
27	文単位頻度（電気分野）中国語	C'freq-s-e-src' 固定
28	文単位頻度（電気分野）日本語	C'freq-s-e-tgt' 固定
29	文単位頻度（電気分野）中国語・日本語	C'freq-s-e-src-tgt' 固定
30	文単位頻度（機械分野）中国語	C'freq-s-m-src' 固定

31	文単位頻度（機械分野）日本語	C'freq-s-m-tgt' 固定
32	文単位頻度（機械分野）中国語・日本語	C'freq-s-m-src-tgt' 固定
33	文単位頻度（物理分野）中国語	C'freq-s-p-src' 固定
34	文単位頻度（物理分野）日本語	C'freq-s-p-tgt' 固定
35	文単位頻度（物理分野）中国語・日本語	C'freq-s-p-src-tgt' 固定

4. 3 本文

本文は、区切り文字（タブコード(0x09)）で区切られた以下の項目が設定される。

項番	項目名	説 明
1	見出し語（中国語）	
2	訳語（日本語）	
3	品詞（中国語）	C'noun'(名詞)又は C'verb'が設定される。
4	品詞（日本語）	同上
5	複合語フラグ	値が 1 のとき単語、それ以外の場合は複合語
6	文献単位頻度（全分野）中国語	
7	文献単位頻度（全分野）日本語	
8	文献単位頻度（全分野）中国語・日本語	
9	文献単位頻度（化学分野）中国語	
10	文献単位頻度（化学分野）日本語	
11	文献単位頻度（化学分野）中国語・日本語	
12	文献単位頻度（電気分野）中国語	
13	文献単位頻度（電気分野）日本語	
14	文献単位頻度（電気分野）中国語・日本語	
15	文献単位頻度（機械分野）中国語	
16	文献単位頻度（機械分野）日本語	
17	文献単位頻度（機械分野）中国語・日本語	
18	文献単位頻度（物理分野）中国語	
19	文献単位頻度（物理分野）日本語	
20	文献単位頻度（物理分野）中国語・日本語	
21	文単位頻度（全分野）中国語	
22	文単位頻度（全分野）日本語	
23	文単位頻度（全分野）中国語・日本語	
24	文単位頻度（化学分野）中国語	
25	文単位頻度（化学分野）日本語	
26	文単位頻度（化学分野）中国語・日本語	
27	文単位頻度（電気分野）中国語	
28	文単位頻度（電気分野）日本語	
29	文単位頻度（電気分野）中国語・日本語	

30	文単位頻度（機械分野）中国語	
31	文単位頻度（機械分野）日本語	
32	文単位頻度（機械分野）中国語・日本語	
33	文単位頻度（物理分野）中国語	
34	文単位頻度（物理分野）日本語	
35	文単位頻度（物理分野）中国語・日本語	

5. サンプル(項目ごと改行表示)

相互作用

相互作用

noun

noun

1

0.034330807188308846

0.03995259530194176

0.03183123205592311

0.05220714424727092

0.054592098512057464

0.04970188071690168

0.020653030559708914

0.03072911679568673

0.018215030161968668

0.013903684510759377

0.01490711404045333

0.01193806228135889

0.03212589073634204

0.04746634996041172

0.02879057798891528

0.0014516355764857527

0.0017100606193756478

0.001368413623323029

0.00263550536904415

0.0027933517065687296

0.002518495165508851

0.0006366279043393994

0.0008485995935794417

0.0005935984938669482

0.0006885806144583298

0.000791212441882754

0.0005870540528677069

0.0010177616976403414

0.0016026152818019823

0.0009462472425763947

添付資料 4. 1 機械翻訳結果評価シート（サンプル）

文番号		CN-A-1624473
分野		C04 A61P17/16
中国語原文		目前，商品化的多糖衍生物固定相是采用物理涂敷的方法将多糖类衍生物固定在色谱基质上，因此存在手性选择剂流失的问题，在实际应用中流动相的组分受到了较大限制
参考和訳		現在、商品化された多糖体誘導物固定相は、物理的にコーティングする方法を採用し、多糖類誘導物がクロマトグラフィーマトリックス上に固定される。このため、キラル選択剤が流出するといった問題が生じる。実際の応用において、流動相の成分が大きな制限を受けることになる。
コメント		原文は1文だが参考和訳は2文に分割されている。
機械翻訳1	機械翻訳	現在、商品化の多糖類の派生物が様子に固定するのは物理を採用してつける方法を塗って多い炭水化物の派生物をクロマトグラフィーマトリックス上に固定するので、そのため、カイラリティ選択剤流失する問題が存在して、実用上移動する様子の成分は比較的大きい制限を受けた。
	評点	1
	評価根拠	・ 文前半の骨格構造が崩れている ・ 「多糖体誘導物固定相」という複合用語の境界が正しく判定されていない ・ 複合用語の一部である「固定」は述語として認識されてしまう ・ 「物理的にコーティングする方法」という名詞句の境界が正しく判定されていない ・ 目的語「方法」の修飾節の一部である「涂敷（コーティング）」は述語として認識されてしまう ・ 複合用語「流動相」の一部である「流動」は述語として認識されてしまう ・ 主語「商品化の多糖類の派生物の定常期」と意味不明の訳語となる ・ 主語「カイラリティ選択剤」の後ろに助詞「が」が欠落 「制限を受けた」という時制が間違っている
	中日翻訳に特有の課題	・ 主語の後ろの格助詞「が」が欠落する ・ 時制の対応が難しい

機 械 翻 訳 2	機械翻訳	現在、商品化の多糖類の派生物の定常期は物理被覆の方法を採用して多糖類派生物をクロマトグラフィマトリックス上に固定するので、そのため、カイラリティ選択剤流失する問題が存在して、実用上流動相の成分は比較的大きい制限を受けた。
	評点	3
	評価根拠	・ 文の骨格構造は大体あっている ・ 主語「商品化の多糖類の派生物の定常期」と意味不明の訳語となる ・ 主語「カイラリティ選択剤」の後ろに助詞「が」が欠落 ・ 「制限を受けた」という時制が間違っている
	中日翻訳に特有の課題	・ 主語の後ろの格助詞「が」は欠落 ・ 時制の対応が難しい
1 と 2 の 優 劣	【機械翻訳 1と2どちらが良いかの 記入してください。】	機械翻訳2の訳質がよい
1 と 2 の 違 い	【優劣の判断のポイントを記述してください。】	・ 翻訳1の骨格構造は崩れてしまうが、翻訳2のほうが大体あっている ・ 翻訳1では、「多糖体誘導物固定相」という複合用語の境界が正しく判定されていないのに対して、翻訳2ではそれがうまく判断できた ・ 翻訳1では、「物理的にコーティングする方法」という名詞句の境界が正しく判定されていないが、翻訳2ではそれがうまく判断できた