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

調査報告書

平成２６年２月

一般財団法人　日本特許情報機構

（空白）

目次

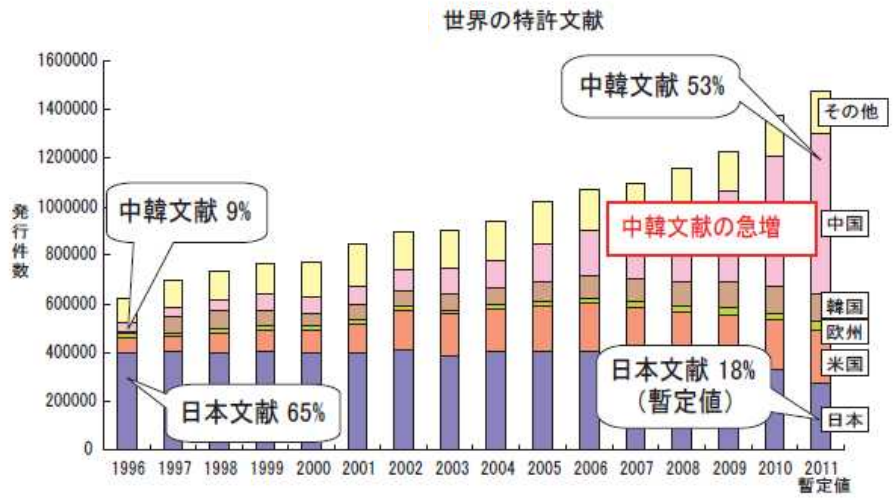
（空白）

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、我が国審査官およ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。

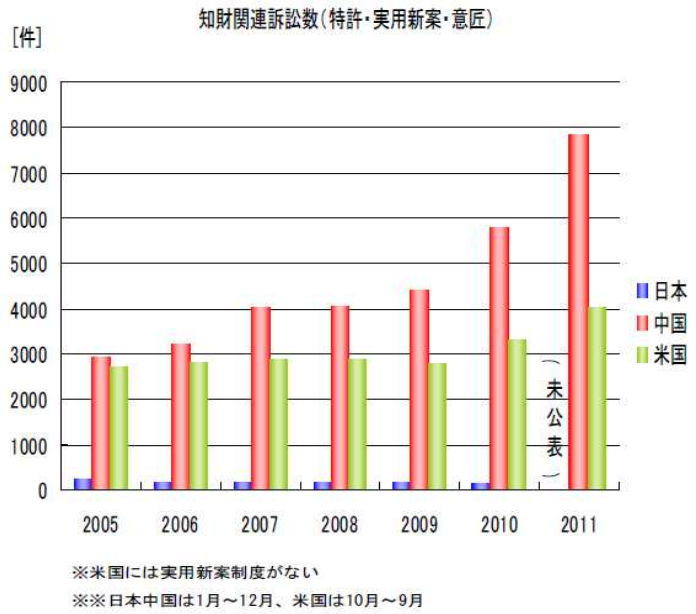
また、中国の公開特許公報及び実用新案公報などの特許関連文献の年間発行数は既に米国特許公報、米国公開特許公報、欧州公開特許公報を上回っており、今後さらに増加するものと見込まれている。



（出典）特許行政年次報告書 2013 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。

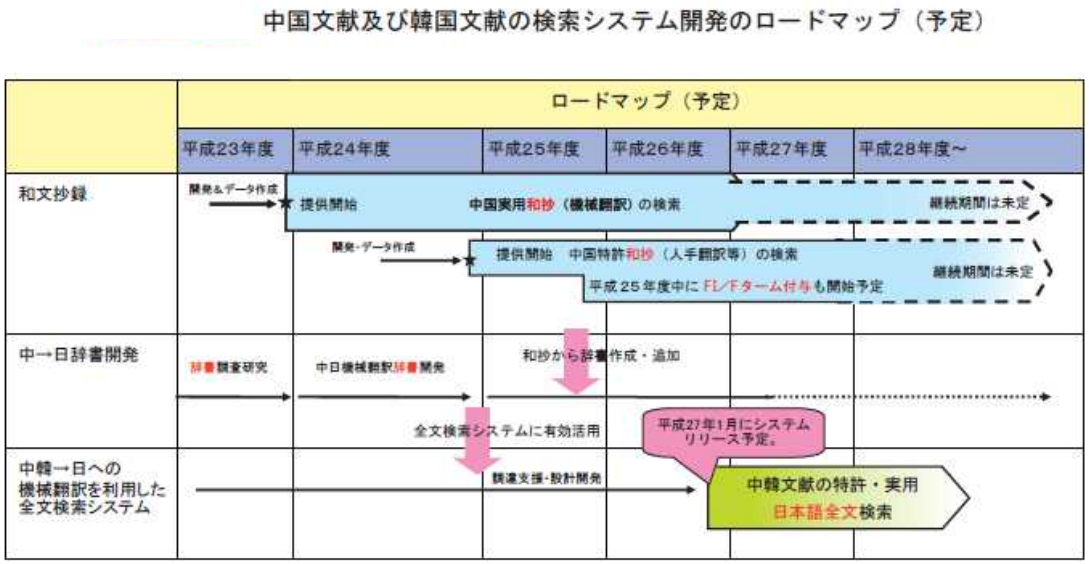
出願件数の増加を背景に、中国での知財民事訴訟件数は米国を越え、なおも増加傾向にある。



（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、特許庁は中国語の特許関連文献の和抄の提供を開始し、さらに日本語により調査・理解を可能とするインフラとして中韓文献翻訳・検索システムのリリースが今後予定されている。



（出典）特許行政年次報告書 2013 年版　特許庁

そうしたなか、特許庁は平成 23 年度には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う「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る。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 年度には「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」を実施し、2005 年～2009 年の中日パテントファミリーから中日対訳コーパスと 100 万語の中日機械翻訳辞書を作成した。

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、従来は使われていなかった新たな技術用語・専門用語（新語）のが増大していると考えられる。さらに、日本語と同様に中国語にも同義語が存在する。これら新語および同義語の発生状況や規模、傾向等については未知の部分がある。

本調査では、2010 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新語 100 万語の中日機械翻訳辞書を作成するとともに、今後、辞書データを適切にアップデートさせていくために、新語についての調査を行った。

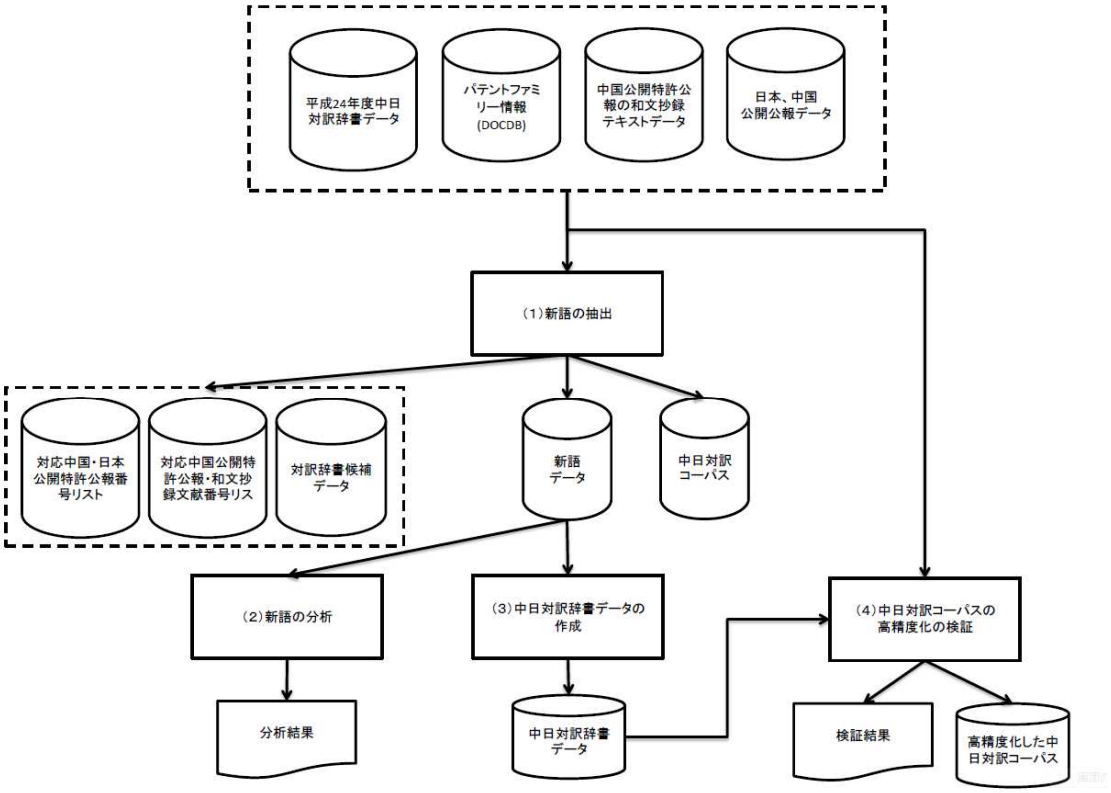
２．調査概要

本調査は、平成 24 年度に特許庁が実施した「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」で中国特許公開明細書（2005～2009 年）から作成した中日辞書 100 万語を基礎として、中国特許公開明細書（2010～2012 年）と中国公開特許公報の和文抄録（約 10 万件）から中日対訳辞書データと高精度化した中日対訳コーパスを作成した。

具体的には 2010 年～2012 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・公表特許公報および再公表特許の対応中国・日本公開特許公報番号リストを作成し、そのパテントファミリーと特許庁から提供された中国公開特許公報の和文抄録と対応する中国特許公報の要約から中国語と日本語の対訳コーパス（以下 中日対訳コーパス）を作成し、中日対訳コーパスから対訳辞書候補データを作成し、対訳辞書候補データを人手により確認し、100 万語の新語データを抽出した。

そして抽出した新語データについて、新語の状況や規模、傾向などを分析した。
また、中日対訳コーパスついて高精度化を行い、どの程度高精度化が図れたのか検証を行った。図 2.1 に調査概要図を示す。

図 2-1 調査概要図



２．１　調査実施体制及びスケジュール

前述調査概要図に示すように、本調査は複数の工程（新語の抽出、新語の分析、中日対訳辞書データの作成、中日対訳コーパスの高精度化の検証）から構成されている。

これらの工程が円滑に連携して調査を実施可能とするため、下記表に示すチーム体制にて調査を実施した。各チームは表 2.1-1 に示すように役割を明確にし、各工程が円滑に連携するようにした。

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	「中日機械翻訳用辞書データ」及び「中日対訳コーパス」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、校閲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）
辞書校閲・翻訳チーム	校閲者と翻訳者のワークフローを管理。 辞書作成チームが作成した人手確認用中日対訳辞書候補データの校閲作業。	管理者 1 名、校閲担当者 10 名
調査・分析・検証チーム	新語データの分析及び、中日対訳コーパスの高度化の検証作業。	3 名（うち 1 名は中国語を母国語とする者）
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、報告書を作成する。	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 25 年 10 月から平成 26 年 2 月の期間で実施した。実施スケジュールを図 2.1-1 に示す。

図 2.1-1 調査実施実施スケジュール

項番	作業項目	担当	開始	終了	平成２５年１０月					平成２５年１１月				平成２５年１２月					平成２６年１月				平成２６年２月				
1	新語の抽出	辞書作成チーム 辞書校閲・翻訳チーム	10月7日	1月26日	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●						
2	新語の分析	調査・分析・検証チーム	12月2日	2月16日										●	●	●	●	●	●	●	●	●	●	●			
3	中日対訳辞書データの作成	辞書作成チーム	11月18日	1月31日								●	●	●	●	●	●	●	●	●	●	●					
4	中日対訳コーパスの高精度化の検証	調査・分析・検証チーム アドバイザー	12月2日	2月23日										●	●	●	●	●	●	●	●	●	●	●	●		
5	報告書作成	報告書作成チーム アドバイザー	12月2日	2月27日										●	●	●	●	●	●	●	●	●	●	●	●	●	●

２．２　調査結果

（１）新語の抽出

2010 年～2012 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、6,500 万件の中日対訳コーパス（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。中日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、平成 24 年度調査において作成した 100 万語の中日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(219,196 件)(図 2.2.-1)、対応中国公開特許公報・和文抄録番号リスト(128,742 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うるものを抽出し、抽出した日本語に対応する中国語を抽出し、760 万件の対訳辞書候補データを作成した。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除き、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と判断した 100 万語を抽出し新語データとした。

図 2.2-1 新語抽出対象一覧[1]（抜粋）

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

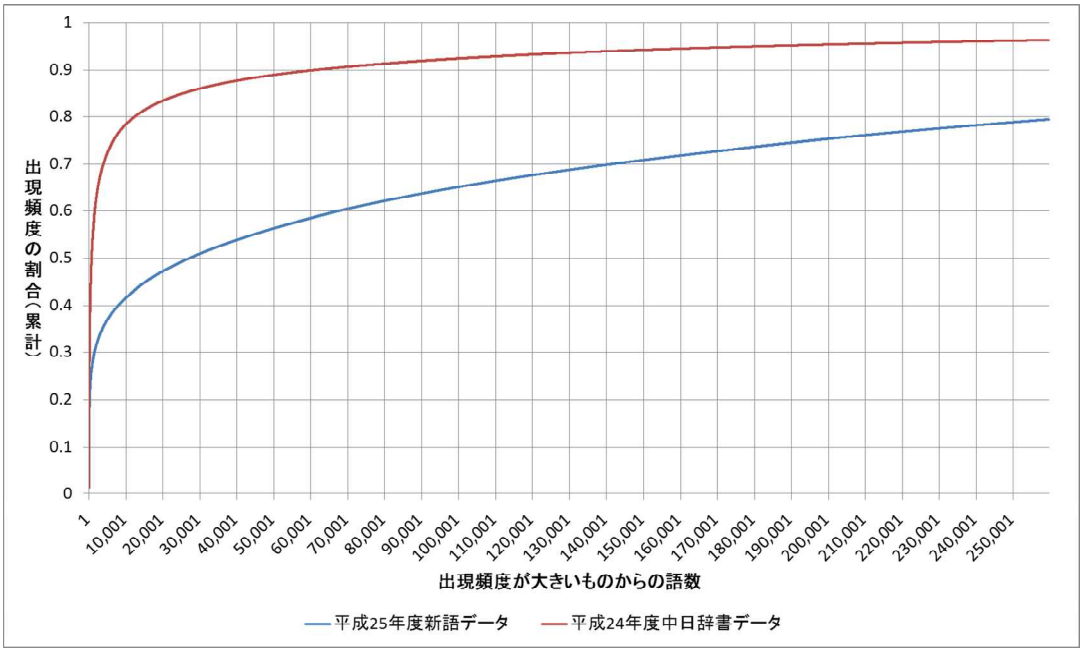
（２）新語の分析

前述（１）「新語の抽出」で作成した新語データは、出現頻度が大きい新語データほど、機械翻訳の品質向上に大きく資する重要なデータであるため、新語データの状況や規模、傾向等を分析する際には、新語データの出現頻度が有用な指標となる。

新語の分析は、全体の分布の分析として、新語データにおける出現頻度をすべて足しあわせた全体合計値を集計した上で、新語データにおける出現頻度を頻度が高い順に足し合わせていったときの合計値が全体合計値の 0.5 倍の値となるときの出現頻度、全体合計値の 0.8 倍の値となるときの出現頻度を求めた。調査の結果、出現頻度上位から 27,192 位(出現頻度 128)までの累計が総出現頻度の 0.5 倍となり、出現頻度上位から 269,977 位(出現頻度 22)までの累計が総出現頻度の 0.8 倍となった。同様の分析を平成 24 年度中日対訳辞書データについて行った結果、出現頻度上位から 696 位(出現頻度 23,589)までの累計が総出現頻度の 0.5 倍となり、出現頻度上位から 12,360 位(出現頻度 728)までの累計が総出現頻度の 0.8 倍に達する結果となった。また、新語データを出現頻度が 1,000 以上、999～500、499～250、249～100、99～50、49～25、24～10、9～5、4、3、2、1 回の範囲に区分し、各範囲に属する語数を調査した。調査の結果、出現頻度 9～5 の語数が最も多く 41 万語となった。これら全体の分布の調査結果を分析した結果、新語データは少数の高頻度語と多数の低頻度語から構成される傾向が見られた。

[1] 一覧の項目は、左から国コード（中国)、公報種別（中国)、公開番号（中国)、公開日（中国)、中国公報と日本公報を対応付けるコード、国コード（日本)、公報種別（日本)、公開番号（日本)、公開日（日本)、国際特許分類となる。

図 2.2-2 新語データ出現頻度による分析（出現頻度の割合（累計））



次に、新語データの出現頻度上位 100 語について、技術分野、初出年、文字数、見出し語重複の観点から調査を行った。技術分野の調査結果からは、化学と電気分野に頻出する用語が多い傾向が見られた。初出年の調査からは、ある年から出現頻度が上昇するような新語の存在は確認出来ず、2010 年から 2012 年まで一定の出現頻度を持つ比較的一般的な用語が多く含まれていることが分かった。見出し語重複の調査からは、新語の中国語のうちおよそ 15%の用語が平成 24 年度中日対訳辞書データと重複した。この重複した中国語の用語は平成 24 年度とは異なる日本語の用語を抽出したことになる。この異なる日本語は技術分野による訳し分け、異表記・類義語など検索及び日中辞書への利用が考えられる。

定期的に新語収集を行う必要性の観点から、平成 24 年度中日対訳辞書データの 2005～2009 年の中日対訳コーパスにおける出現頻度と 2010～2012 年の対訳コーパスにおける出現頻度を調査した。調査の結果、2005～2009 年の中日対訳コーパスにおける出現頻度に比べ、2010～2012 年の対訳コーパスにおける出現頻度が 1/5 以下に減少または出現頻度ゼロになってしまう用語が全体の 50.4%存在した。これらの用語は用語を採取した年範囲より新しい 3 年間の中日対訳コーパスにおいてほとんど使われなかった用語である。このような年により出現頻度が著しく増減する語であっても出現頻度が多い年代の文献の翻訳に必要な語であり、辞書化する必要がある。

そして、このようなある年代に高頻度で現れる用語は今後も発生し得るため、新しい中日対訳コーパスがある程度蓄積される度に、定期的に新語の抽出を行うことが望ましい。

（３）中日対訳辞書データの作成

前述（１）新語の抽出で作成した新語について、「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))、出現頻度情報から成る中日対訳辞書データを作成した。

出現頻度情報は、化学分野、電気分野、機械分野、物理分野および全分野について中日対訳コーパスに基づく、中国語単独で出現する頻度、日本語単独で出現する頻度、中国語と日本語に同時に出現する頻度を作成した（図 2.2-3)。また、中日対訳辞書データは UTX 形式(UTX1.11)で作成した。

図 2.2-3 中日対訳辞書データ[2]（イメージ)

组合物	組成物	noun	noun	176846	180073	170343 …
端部	端部	noun	noun	157989	190932	144120 …
控制部	制御部	noun	noun	114633	133468	110810 …
反应混合物	反応混合物	noun	noun	111709	108845	105387 …

（４）中日対訳コーパスの高精度化の検証

前述（１）新語の抽出において作成された中日対訳コーパスは中日自動文アライメントツールを用いて作成した。自動文アライメントツールは「対応する文を推定する」ものであり、自動文アライメントによって作成された中日対訳コーパスは意味内容が対応しない対訳文も含む可能性がある。中日対訳コーパスが正しい対訳文対をどれくらい含むかは、

・自動文アライメントの分析対象となった「対訳関係にある２つの文書の文章」の内容
・自動文アライメントツールのアルゴリズム
・中日自動文アライメントツールが学習する対訳辞書データの品質や数量

に依存する。

本調査では、中日自動文アライメントツールが学習する辞書を語数により３種類用意し、それぞれの学習用辞書で作成した中日対訳コーパスの対訳文としての正解率（精度）がどのように変化するか検証した。表 2.2-1 に使用した検証データを示す。

[2] 中日対訳辞書データの項目は左から、見出し語（中国語)、訳語（日本語)、品詞（見出し語)、品詞（訳語)、中国語の出現頻度、日本語の出現頻度、中日の出現頻度となり、以降、分野別の出現頻度が続く。

表 2.2-1 検証データ及び精度

検証データ	学習辞書の語数＊	検証データ件数	正解率 90%の件数（推定値）
①基準となる中日対訳コーパス	156,845	65,803,582	1,760 万件
②新語の抽出おいて作成された中日対訳コーパス	973,865	65,839,814	3,943 万件
③高精度化した中日対訳コーパス	1,704,744[3]	65,862,370	3,645 万件

＊学習辞書の語数は中国語のユニークな語数である。

表 2.2-1 の各検証データは、

①基準となる中日対訳コーパスは中日自動文アライメントツール付属の辞書を学習したツールで作成した中日対訳コーパス、

②新語の抽出において作成された中日対訳コーパスは、さらに平成 24 年度中日対訳辞書データを学習したツールで作成した中日対訳コーパス、

③高精度化した中日対訳コーパスは、さらに本調査で作成した中日対訳辞書データを学習したツールで作成した中日対訳コーパス

となる。

検証は、これら各検証データについて、中日自動文アライメントツールが出力する部の対応度合いを示す文対応スコアを使い、所定の文対応スコア値を持つ中日対訳文 30 件を人手によりチェックし、その結果から各検証データの正解率とその正解率で得られる対訳文の文数を推定した。

検証の結果、①基準となる中日対訳コーパスは正解率 90%の中日対訳コーパスの件数が 1,760 万件（推定値）となった。②新語の抽出において作成された中日対訳コーパスは３つのうち最も高精度となり、正解率 90%の中日対訳コーパスの件数が 3,943 万件（推定値）となった。③高精度化した中日対訳コーパスは、正解率 90%の中日対訳コーパスの件数が 3,645 万件（推定値）となった。これらの結果から、中日自動文アライメントツールに対して中日対訳辞書データを学習することで中日対訳コーパスが高精度化することが確認できた。また、検証結果は、③の辞書数を最も多く学習したときの精度が二番目の結果となり、辞書数の増加による頭打ち傾向が見られた。

[3] 高精度化した中日対訳コーパスで使用した中日対訳辞書データは、調査工程の都合から最終的な人手確認前の辞書データを使用した。

３．新語の抽出

2010 年～2012 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、2010 年～2012 年発行分の中国公開公報における 100 万語の新語データを抽出した。

また、新語データの抽出過程において、219,196 件の対応中国・日本公開特許公報番号リスト、128,742 件の対応中国公開特許公報・和文抄録番号リスト、6,500 万件の中日対訳コーパス、760 万件の対訳辞書候補データを作成した。

３．１　抽出環境

（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特に、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する必要がある。そして中日対訳コーパスを高精度化するため、中日対訳辞書データを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 24 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。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 （ http://alaginrc.nict.go.jp/cnp/index.html ）
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

新語の抽出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[4]×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　抽出手法

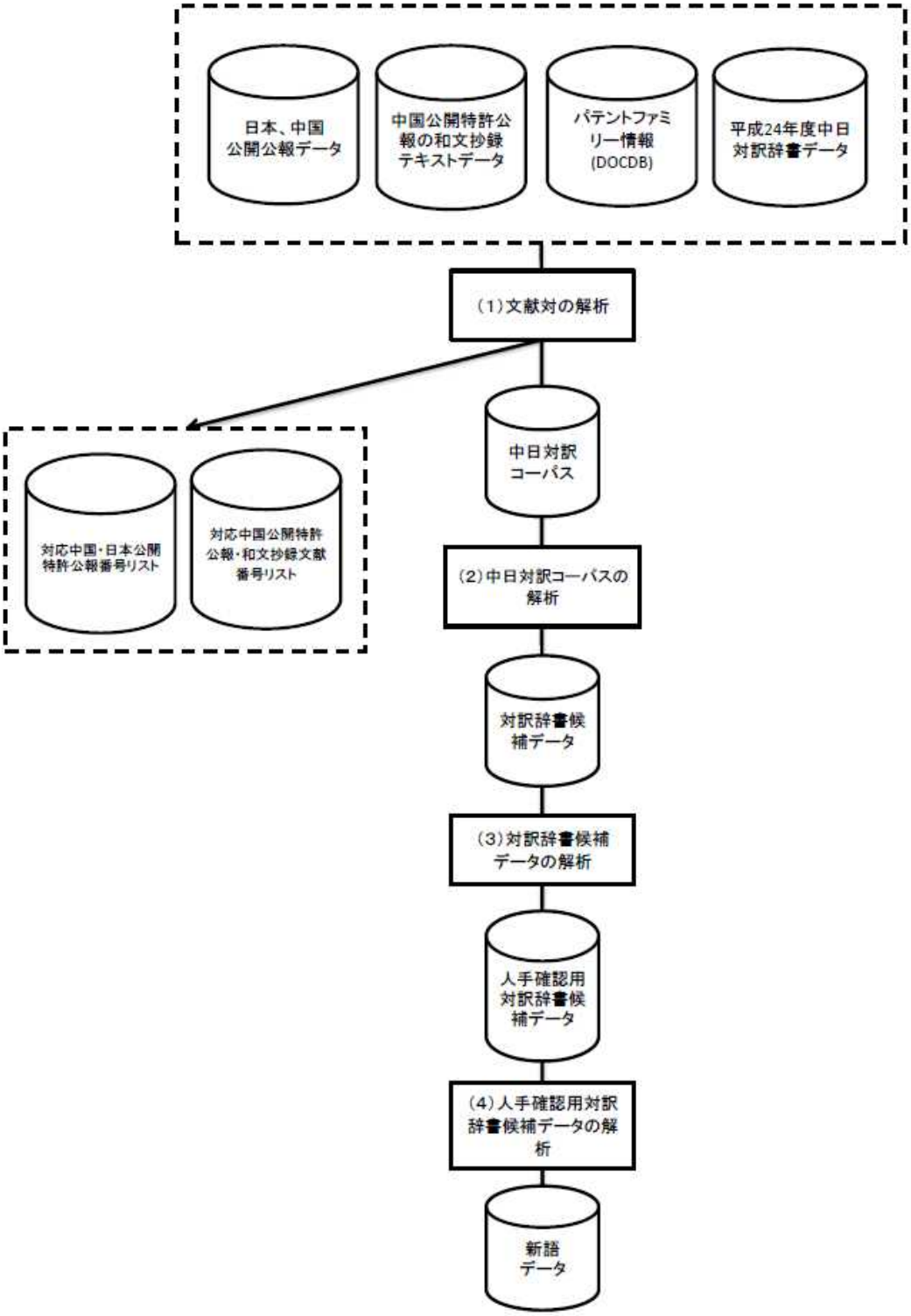
新語データは図 3.2-1 新語抽出概要図に示す手順で作成した。手順は下記 （1）から（4）の処理で構成される。

（1）文献対の解析（中日対訳コーパス作成）
（2）中日対訳コーパスの解析（対訳辞書候補データの作成）
（3）対訳辞書候補データの解析（不要語の除去）
（4）人手確認用対訳辞書候補データの解析（新語データの抽出）

（1）文献対の解析では技術内容が対応する中国公開特許公報と日本公開公報等から、中日対訳コーパスを作成した。（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から、中日対訳辞書データとなりうる中日対訳辞書候補データを作成した。（3）対訳辞書候補データの解析では（2）で作成した中日対訳辞書候補データから明らかに誤りと思われる用語を除去した。（4）人手確認用対訳辞書候補データの解析では技術知識を有する者が、（3）で取り除いた後のデータのチェック（校閲）を行い、新語データを作成した。

[4] 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。

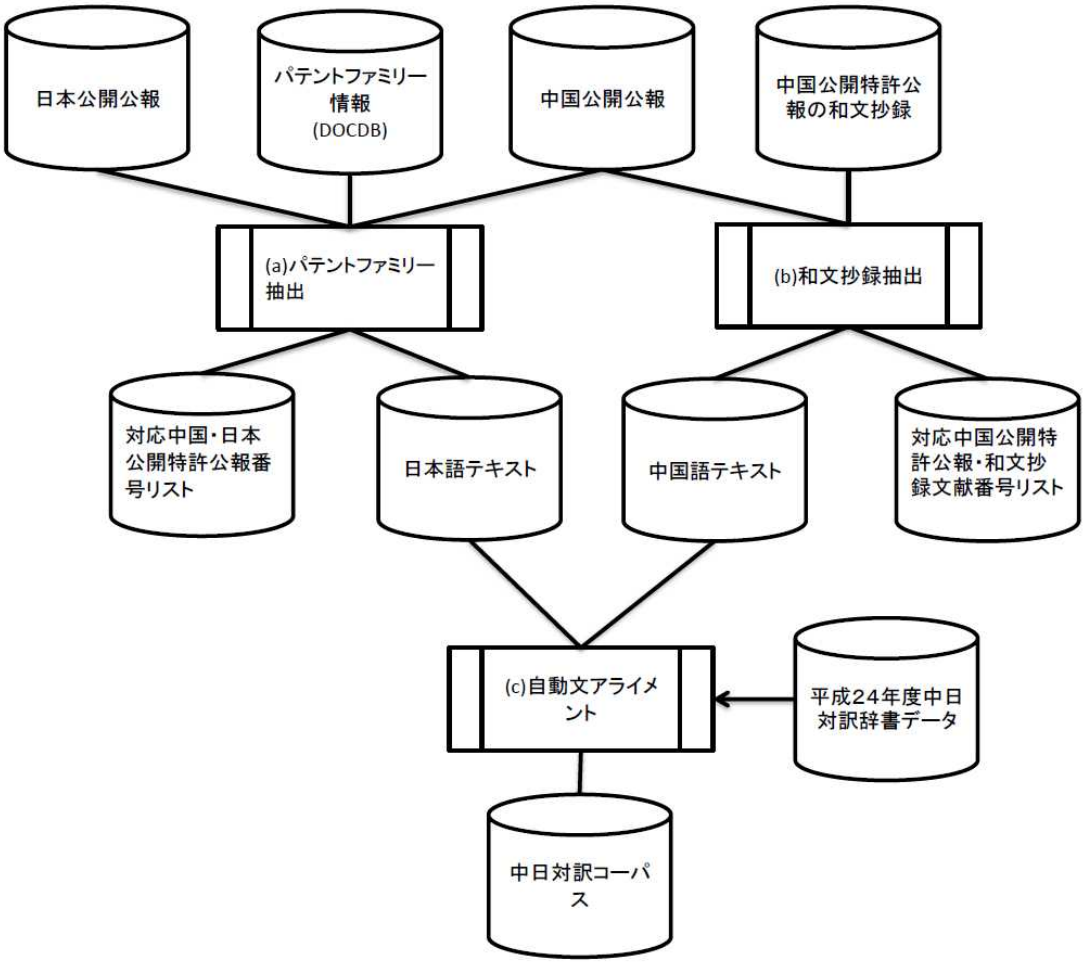
図 3.2-1 新語抽出概要図



（１）文献対の解析（中日対訳コーパス作成）

文献対の解析では、技術内容が対応する中国公開特許公報と日本公開公報や、中国公開特許公報の和文抄録と中国公開特許公報の要約の解析を行い、中日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

中日対訳コーパスは、技術内容が対応する中国公開特許公報と日本公開公報の文献対から作成した。本調査では、DOCDB[5]に蓄積されているパテントファミリー情報を利用し、この文

[5] EPO(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

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 対応中国・日本公開特許公報番号リスト）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公報データ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の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。本調査の中日自動文アライメントツールは 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）

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

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

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

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。

(b)和文抄録抽出

本調査では、前述(a)のパテント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、このリストを基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)で抽出した日本語テキストと中国語テキストから、中日自動文アライメントツールを使い、中日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 24 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として論文[6]が知られている。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）を基に中国語と日本語文の対応づけを行う。

新語の抽出では、この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、平成 24 年度調査で作成した 100 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。

技術的課題

文献対の解析における技術的課題を以下に示す。

(a)適切な中日自動文アライメントツールの選定

<<課題>>

文献対の解析における最も重要な要素は中日自動文アライメントツールの精度である。自動文アライメントツールの精度は、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対訳辞書の品質にも大きく影響する。そのため、適切な中日自動文アライメントツールの選定が課題となる。また本調査では前述の手法で述べたように、平成 24 年度調査で作成した 100 万語の中日対訳辞書データを追加する必要があるので、辞書追加が可能なツールの選択が必要となる。

<<対応>>

本調査では、独立行政法人 情報通信研究機構(NICT)[7]が開発した中日自動文アライメント

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

[7] http://www.nict.go.jp/

ツールを使用した。このアライメントツールは平成 24 年度「中国特許文献の機械翻訳のための辞書データ整備に関する調査」での使用実績、ツール自体の精度が高いこと、同一のツールを使用することによる一貫性や、辞書追加が可能なツールであることを理由に使用した。

(b) 公報の構造を考慮したデータ処理

<<課題>>

中日自動文アライメントツールは、基本的に中国語文と日本語文を上から順番に対応付を行う。そのため、中国語文献と日本語文献の文の並び順がずれていると作成される対訳文コーパスの品質が低下する課題が存在する。

<<対応>>

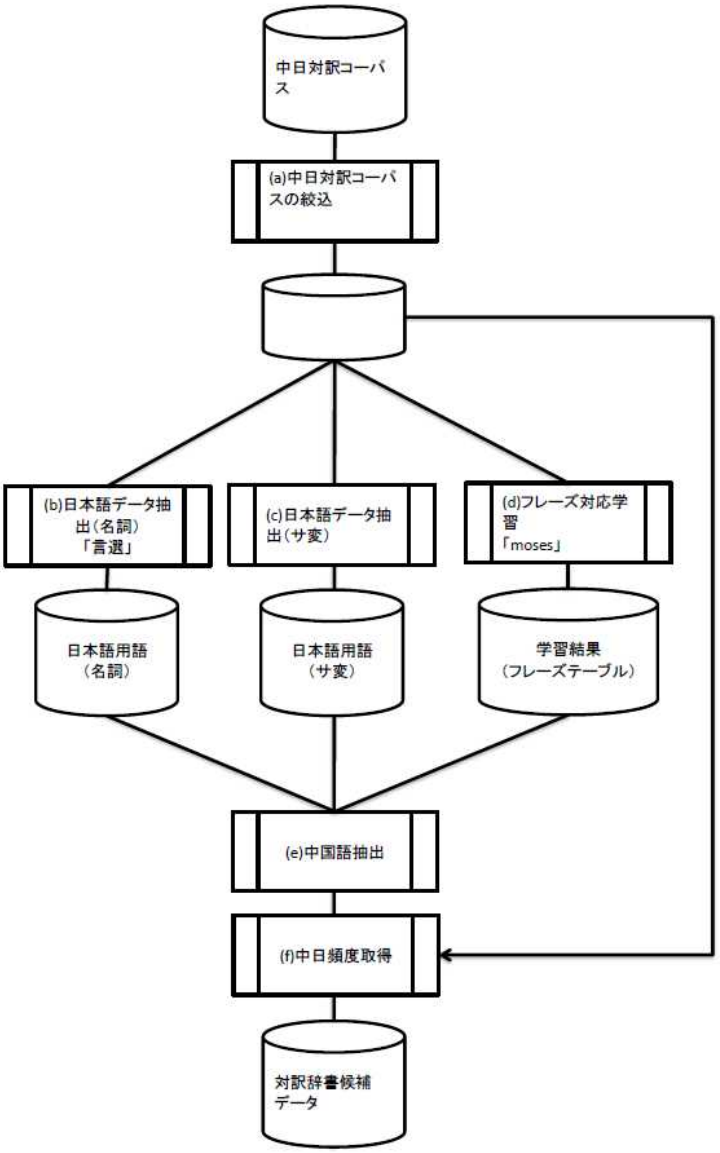
中国語公報と日本語公報の明細書との間にはパラグラフの並びがずれている部分がある。そのため、本調査では、事前に日本語と中国語文献の並び順をそろえ、文アライメント処理を行った。具体的には中国公報の“図面の簡単な説明”に相当する部分が“課題”と“実施例”の間に記載されることが多いので、これを日本の公報の並び順に合わせた。

（２）中日対訳コーパスの解析（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中日対訳コーパスを解析し、中日対訳辞書データとなりうるデータ（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、出現頻度値の大きい順に並び替えし、対訳辞書候補データを作成した。

図 3.2-6 中日対訳コーパスの解析概要図



(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中日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文対応度合いがよ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文で対応するものかつ、文の対応度合いを示す文対応スコアが一定値以上のものを抽出した。本調査では、この文対応スコアの閾値を 0.08 とした。この閾値は Japio が中日対訳コーパス作成の経験から得た値である。

次に、中日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。これら詳細は、後述「技術的課題」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う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 24 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”なども抽出対象とした。

図 3.2-7 サ変動詞抽出例

```
・・・・
インサート　　名詞　　一般
を　　助詞　　一般
嵌合　　名詞　　一般
する　　動詞　　自立
こと　　名詞　　一般
による　助詞　　連語
・・・・
```

(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 中国語抽出例にあるように、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」を使用し作成した。なお、同ツールは平成 24 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）

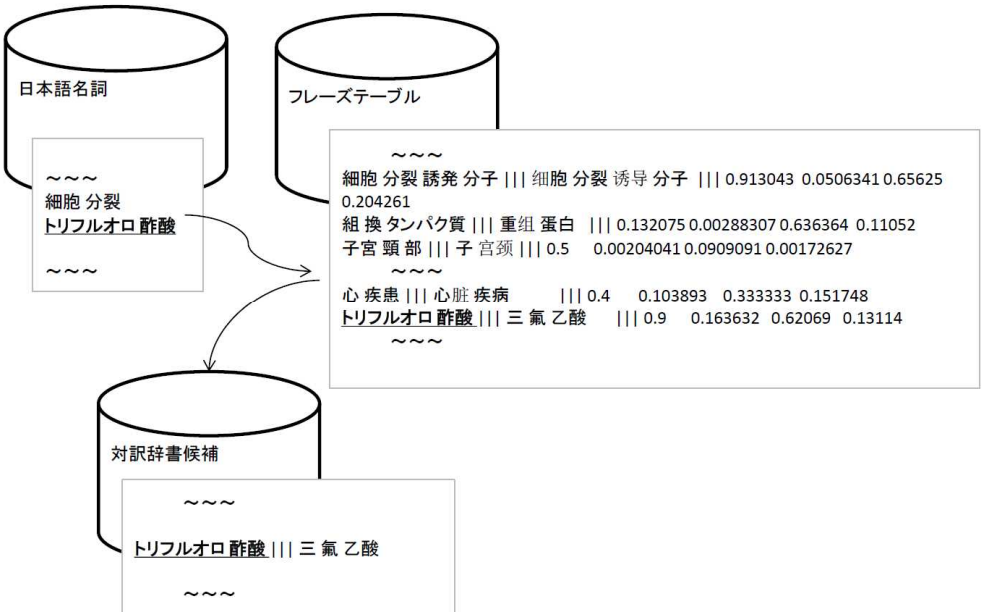


図 3.2-8 は日本語名詞“トリフルオロ酢酸”に対応する中国語“三氟乙酸”の抽出例である。抽出は、図の日本語名詞内の各日本語用語について、同一の用語がフレーズテーブルに含まれているかを検索する。検索の結果、フレーズテーブルに含まれている場合、その内容を対訳辞書候補として出力する。

(f)中日頻度取得

本調査の新語データは中日対訳コーパス中に頻出する順に 100 万語作成する必要がある。そこで、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、前述(a)で絞り込んだ中日対訳コーパス内の出現頻度情報を付加し、出現頻度値の大きい順に並び替えを行った。

技術的課題

中日対訳コーパスの解析における課題を以下に示す。

(a)対応していない括弧書き表現

<<課題>>

前述（１）文献対の解析で作成した中日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。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。例を、図 3.2-9 に示す。

図 3.2-9 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（ｆｒｏｎｔ−ｐｒｏｊｅｃｔｉｏｎ−ｔｙｐｅ）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 3.2-9 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よく見られる表現である。

<<対応>>

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 3.2-10 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 3.2-10 対応していない括弧書き表現（対応例）

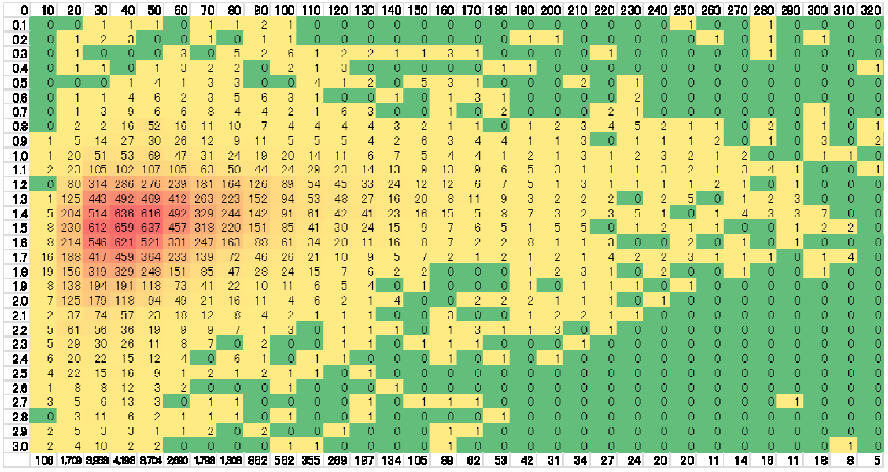
図４は、本発明に基づく反射投影型表示装置１０Ｃを示している。 图 4 表示根据本发明的一个前投射型显示器 10C。

(b)文の長さによるアライメントの選択

<<課題>>

Japio で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 3.2-10 は、その調査結果の一部である。図は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 3.2-10 中国語文と日本語文の文字数比率調査結果(Japio)



<<対応>>

本調査では、課題で説明した Japio 独自調査結果を利用し、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。

(c) 日本語専門用語抽出ツールの選定とチューニング

<<課題>>

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、図 3.2-11 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課題が存在する。

<<対応>>

Japio 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、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を向上した。(図 3.2-12)

図 3.2-11 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　サ変・スル　基本形
くわえ　クワエ　くわえる　動詞-自立　一段　連用形
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・ースル　基本形
```

図 3.2-12 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
```

(d)中国語解析ツールの選定

<<課題>>

前述(d)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、および中国語文を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を解析するツール（形態素解析ツール）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。

<<対応>>

Japio 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。精度の比較は中国語特許文書を複数のツール(NICT ツール、ICTCLAS(中国科学院)、中国語 Mecab (奈良先端大)、中国語 Chasen (Japio が開発中)) の結果を中国語ネイティブが評価・比較した。また、処理の安定性の観点で大量の特許データを問題なく解析できるか等の評価も行った。これらの比較検討結果から、最も適切と思われる中国語解析ツールは NICT が開発した中国語解析ツールと判断し、これを使用した。

(e)用語の出現頻度データの効率的取得

<<課題>>

中日対訳コーパスの解析では、作成した対訳辞書候補データに対して、その中国語と日本語の対についての出現頻度を取得し、出現頻度値の大きい順に並び替えを行う必要がある（図 3.2-6 中日対訳コーパスの解析概要図の(f)参照)。この出現頻度を得るためには、中日対訳コーパスを高速に検索する必要がある。

<<対応>>

本事業においては安価かつ効率的な処理を実現する観点から、オープンソース Apache solr のフルテキスト検索ツールを用いて、効率的な出現頻度集計を行った。

なお、このツールは平成 24 年度「中国特許文献の機械翻訳のための辞書データ整備に関する調査」で使用実績のあるものである。

（３）対訳辞書候補データの解析（不要語の除去）

対訳辞書候補データの解析では、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除去した。この除去により無駄な人手確認作業を抑止し、効率的な辞書作成が可能となる。明らかに解析誤りと思われる用語として、(a)単語の切り方を誤っており日本語として意味をなさない語、(b)接頭語の登録が行われていれば適切に翻訳されうる語、(c)中国語の品詞情報が名詞以外の語、(d)中国語用語の先頭・末尾の機能語[8]を除去の対象とした。なお、本調査仕様書記載の例示 “汎用的な形容詞と名詞の単純な組み合わせが一語として抽出される語“は、対訳辞書候補データ内に確認することが出来なかった。また、”単純な接続語で結ばれた２つの名詞が一語として抽出される語“は、単語の区切り方の誤りに原因があるため、(a)の事象の一部として対応を行った。

(a) 単語の切り方を誤っており日本語として意味をなさない語

本調査仕様書では（“て効果”、“や開口部”）などを解析誤りとして除去することが例示されている。このような解析誤りは日本語解析ツールが本来“や”の部分の品詞を助詞と解析しなければならない所、名詞と誤解析し、後続する名詞“開口部”と連接して１つの用語と解釈されため発生したものである。このような用語について不要語の除去をおこなった。詳細については後述の技術的課題(b)解析誤りの原因の究明と対策で説明する。

(b) 接頭語の登録が行われていれば適切に翻訳されうる語

本調査の仕様で例示されている“前記開口部”、“上記間隔”、“該送風ファン”とった表現は特許文書で頻出する表現である。これらの用語を日本語解析ツールで解析すると、図3.2-13 の様な結果になる。

図 3.2-13 日本語解析結果（前記・上記・該）

前記(名詞) /// 開口(名詞) /// 部(名詞)
上記(名詞)/// 間隔(名詞)
該(未知語)/// 送風(名詞) // /ファン(名詞)

上記解析結果のように日本語解析ツールは、不要な“前記”、“上記”、“該”を名詞又は未知語と解析する。これら解析結果から日本語用語を抽出すると、名詞の連続を１つの用語として抽出するために“前記”を含む“前記開口部”が出力される。
なお、本調査では例示された表現に加え“各・・・”、“当該・・・”も除去対象とした。

[8] 助詞や助動詞といった、主に文の構成に関わる要素のこと。

技術的課題

対訳辞書候補データの解析における技術的課題を以下に示す。

(a)日本語不要語パターンによる除去

<<課題>>

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することは無駄な人手確認作業を抑止するために重要である。本調査仕様書上では（“て効果”、“や開口部”）などが解析誤りとして例示されている。Japio は他にも“かつ内部電極”、“からカラーフィルタ”といった解析誤りが存在することを把握している。

<<対応>>

このような解析誤りに関する課題を Japio は日英や中日対訳辞書作成事業を通じたノウハウとして不要な用語の事例などを蓄積している。本事業ではこのノウハウを活用して対応した。

(b)解析誤りの原因の究明と対策

<<課題>>

Japio は日本語解析ツールの解析誤りの原因を特定しその対応を行なっている。

例えば、前述記述的課題(a)の解析誤り例“や開口部”の場合、一般的な日本語解析ツールでは図 3.2-15 のように、本来、助詞として解析されるべき“や”を名詞として解析誤りをしたため“や”が名詞の一部と判断されたものである。なお、日本語解析ツールは“や”を常に誤るわけでなく図 3.2-14 に示すように通常は正しく解析を行う。解析誤りの原因を調査したところ、この誤りは日本語解析ツールが特許文で頻出する “スリット A”の“A”のような表現に対応していないのが原因として判明した。

図 3.2-14 日本語解析ツール解析結果（助詞“や”が正しく解析される例）

スリットや開口部
↓
スリット(名詞) /// **や(助詞)** /// 開口(名詞) /// 部(名詞)

(///は単語の区切りを示す記号を表し、各単語の後ろに括弧書きで名詞を表している。)

<<対応>>

本調査では、このような誤解析に対応するため、日本語解析ツールの設定をカスタマイズすることで対応した。カスタマイズは、日本語解析ツールが備える単語間の品詞組み合わせのルールに、記号と助詞の組み合わせを追加する対応を行った。

図 3.2-15 日本語解析ツール析結果（助詞“や”が誤って解析される例）

スリットＡや開口部
↓
スリット(名詞) /// Ａ(記号) /// **や(名詞)** /// 開口(名詞) /// 部(名詞)

(c)中国語の機能語による除去

<<課題>>

対訳辞書候補データ内には日本語のみならず中国語の誤りも存在し、対訳辞書候補データの精度を下げる要因となる。図 3.2-16 のその例を示す。

図 3.2-16 対訳辞書候補の中国語の誤り例

日本語	中国語
異物除去チェック処理	在进行异物除去检验处理
液空間速度	在液体空间速度
渦電流損	的涡流损耗
右側副クランク室内	了右辅助曲柄室内

<<対応>>

図 3.2-16 の下線部分は、中国語の名詞とならない部分である。本調査ではこのような部分を除去するために、Japio が保有している「中国語用語（名詞）に含むべきでない語のリスト」を使用し、対訳辞書候補の中国語の先頭・末尾から不要な部分の除去を行った。

(d)中国語の品詞情報による不要語の除去

<<課題>>

対訳辞書候補データには日本語の品詞が名詞であるにも関わらず、中国語の品詞が名詞とならないものが存在する。この品詞不整合は対訳辞書候補データの精度を下げる要因となる。

<<対応>>

前述(c)中国語の機能語による除去は、Japio が保有している「中国語用語（名詞）に含むべきではない語のリスト」を使い、誤りのパターンを発見し、不要語の除去を行うもので

ある。対訳辞書候補データは大量であるため、この方法による誤りパターンの発見には限界がある。

そこで、本調査では前述(c)の除去に加え、対訳辞書候補データの中国語を中国語解析ツールで解析し得られた品詞情報を参考に、名詞以外の中国語を除去した。除去例を以下に示す。

図 3.2-17 中国語の品詞情報による不要語除去例１

但是,这些**中小型车辆多**不搭载压缩空气罐。 しかし、これらの**中小型車両**は、圧縮空気タンクを搭載しないことが多い。

図 3.2-17 は、日本語“中小型車両”に対する中国語が“中小型车辆”となるべきところ、“中小型车辆**多**“となり、中国語”多“が余分に抽出された例である。

この中国語“中小型车辆多“を解析すると“中小型”,”车辆”,“多“と単語分割され、末尾の単語”多”は品詞が「数詞」と解析された。この解析結果のように、中国語の末尾が１文字であることは中国語の区切りが誤っている可能性があり、かつ末尾の品詞が副詞であるということも誤りの可能性があると考えられる。このように中国語の解析結果を手掛かりに誤りパターンを発見することで不要語の除去を行った。

図 3.2-18 中国語の品詞情報による不要語除去例２

动作数据库 420 含有**上下文动**作。 アクションデータベース 420 は、**文脈アクション**を含んでいる。

図 3.2-18 の例は、日本語“文脈アクション”に対する中国語が“上下文动动**作**”となるべきところ、 “上下文动作”となり、中国語”作“が欠落した例である。

この例も前述の例と同様に中国語を解析すると、”上下文”,”动”(形容詞)となり、このような解析結果パターンを除去対象とすることで対応を行った。

（４）人手確認用対訳辞書候補データの解析（新語データの抽出）

対訳辞書候補データの解析で作成された人手確認用対訳辞書候補データについて、平成 24 年度調査において作成した 100 万語の中日対訳辞書データと突き合わせ、重複するデータを排除した上で、人手による確認（以下、校閲と記載）を行った。校閲作業は、人手確認用対訳辞書候補データの日本語、中国語の出現頻度の多い順に行った。

技術的課題

人手確認用対訳辞書候補データの解析における課題を以下に示す。

(a)校閲者による作業内容の正確な理解

<<課題>>

中日対訳辞書データの最終的な精度は、作業の趣旨や校閲の基準を正確に理解しているかどうかに依存する。このため、各校閲者に作業内容をいかに正確に理解させるかが課題として存在する。

<<対応>>

校閲作業の実施にあたり、各校閲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、専用の校閲手順書を作成し各校閲者に作業内容を正確に理解させた上で校閲作業を実施した。

(b)支援情報の提供

<<課題>>

中日対訳辞書データ校閲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、校閲者に利用し易い方法で提供するよう工夫する必要がある。

<<対応>>

人手確認用対訳辞書候補データを表計算ソフトウェアの形式に変換後、校閲者に提供した。(図 3.2-19)校閲作業の効率化のため、中日対訳コーパスの対訳文を支援情報として掲載し、日本語と中国語部分を強調表示した。こうすることで、校閲者が手間なく支援情報を活用できた。

図 3.2-19 人手確認用対訳辞書候補データ（イメージ）

チェック	品詞	日本語	中国語	日本語例文	中国語例文
		入射角調整ステップ	**入射角调整阶梯**	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のレーザ光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也变小。
		入口側端縁	**入口侧端缘**	例えば、図8に示すインペラーブレード416は、**入口側端縁**416xに複数の凸部416bと複数の凹部416aとを有している。	例如,图8所表示的叶轮叶片416,在**入口侧端缘**416x具有多个凸部416b和多个凹部416a。

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校閲者に提供した。（図 3.2-20）

図 3.2-20 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 入射角調整ステップ　除外する日本語　（例）エンジン
中国語　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。

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の光出射面として構成したものであり、この光出射面を光偏向装置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,使得该光射出面朝对着光偏转装置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のレーザ光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也变小。
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に対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,使其与两者的熔敷面R对向。

３．３　抽出結果

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（中日対訳コーパス作成）

(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、2010 年～2012 年まで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の番号を取得し、219,196 件の番号リスト（対応中国・日本公開特許公報番号リスト[9]）を作成した。この番号リストの日本の文献番号には公表公報及び再公表も含んでおり、それぞれ 89,428 件、26,632 件であった。分野別の件数を表 3.3-1 に示す。

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	72,337（ 33%）
電 気	64,863（ 30%）
機 械	38,084（ 17%）
物 理	43,912（ 20%）
合 計	219,196（100%）

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とを解析対象とし、128,742 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[10]）を作成した。分野別の件数を表 3.3-2 に示す。

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	82,650
電 気	2,502
機 械	49,908
物 理	91
合 計	128,742

[9] 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のこと。

[10] 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のこと。

上記表で電気、物理分野が少ないのは、本調査実施時点の和文抄録中に IPC の G セクション（物理学）、H セクション（電気）に該当するデータが存在しなかったためである。なお、調査実施時点で和文抄録作成予定の G セクション、H セクションはそれぞれ 37,828 件、42,373 件であった。

(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及び和文抄録からテキストデータを発明の名称、要約、請求項、明細書に分けて抽出し、中日自動文アライメントツールを使い、6,500 万件の中日対訳コーパスを作成した。表 3.3-3 に作成件数を示す。

表 3.3-3 中日対訳コーパス（平成 24 年度中日対訳辞書データ学習）作成件数

作成箇所	件数（文対数）
発明の名称	347,960
要 約	936,824
請求項	3,435,270
明細書	61,119,760
合 計	65,839,814

この中日対訳コーパス作成に使用した中日自動文アライメントツールは、ツールが備える標準の辞書に、平成 24 年度調査で作成した中日対訳辞書データを加え、学習したものを使用した。中日自動文アライメントツールの辞書学習や中日対訳コーパスの詳細については、後述の項番６中日対訳コーパスの高精度化の検証で説明する。

（２）中日対訳コーパスの解析（対訳辞書候補データの作成）

(a)中日対訳コーパス（絞り込み後）

前述（１）で作成した中日対訳コーパス（平成 24 年度中日対訳辞書データ学習）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、文の対応がよくないものが存在する。

図 3.3-1 中日対訳コーパスの文の対応がよくない例１

在一些实施例中，设备 110 的总体形状可以与所示出的不同。 色が違っても良い。 /// 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 3.1-1 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[11]。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よくない例となる。

図 3.3-2 中日対訳コーパスの文の対応がよくない例２

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ＲＩＦの薬物動態パラメーターは、ＴＩＭの同時投与に影響を受けなかった（表３）。

図 3.3-2 は、中国語文と日本語文は１文対１文で対応しているが、中国語文の下線部分に対応する日本語文がなく、文の対応がよくない例となる。

このようなデータは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[12]を行った。絞込後の件数を表 3.3-4 に示す。

表 3.3-4 中日対訳コーパス（絞り込み後）

分 野	件数
化 学	13,226,297
電 気	11,530,306
機 械	5,272,498
物 理	8,050,782
合 計	38,079,883

(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うる日本語の名詞及びサ変動詞を抽出した。抽出件数を表 3.3-5 に示す。

11 図3.3-1の日本語文中の /// は中日自動文アライメントツールが出力した分区切り記号である。

12 絞り込みの詳細は、項番 3.2 抽出手法の(2)中日対訳コーパスの解析を参照のこと。

表 3.3-5 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	3,144,583 語	4,160 語
電 気	1,952,478 語	3,595 語
機 械	1,228,060 語	3,643 語
物 理	1,735,303 語	3,834 語
合計 （分野間重複なし）	6,672,759 語	5,039 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、766 万件の対訳辞書候補データ[13]を作成した。作成件数を表 3.3-6 に示す。

表 3.3-6 対訳辞書候補データ作成件数

分 野	対訳辞書候補数
化 学	3,090,421 語
電 気	2,164,179 語
機 械	1,346,027 語
物 理	1,931,701 語
合計(分野間重複なし)	7,660,638 語

（３）対訳辞書候補データの解析（不要語の除去）

(a)人手確認用対訳辞書候補データ

前述（２）(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、人手確認用対訳辞書候補データを作成した。その結果、誤っている用語およそ 237 万件を除去し 529 万語の人手確認用対訳辞書候補データを作成した。作成件数を表 3.3-7 に示す。

[13] 詳細は、添付資料 3.3 対訳辞書候補データレイアウト説明を参照のこと。

表 3.3-7 人手確認用対訳辞書候補データ件数

分 野	対訳辞書候補数
化 学	2,217,285 語
電 気	1,252,667 語
機 械	824,964 語
物 理	998,973 語
合計(分野間重複なし)	5,293,889 語

（４）人手確認用対訳辞書候補データの解析（新語データの抽出）

(a)平成 24 年度調査中日対訳辞書データとの重複排除

前述（３）で作成した人手確認用対訳辞書候補データには平成 24 年度調査で作成した中日対訳辞書データと同一のデータが含まれている。重複した辞書データを避けるため、平成 24 年度調査で作成した中日対訳辞書データと突き合せ、中国語、日本語の対が重複するデータを排除した。その結果 43 万語が重複し、排除後の件数は 4,861,954 語となった。 分野毎の内訳を表 3.3-8 に示す。

表 3.3-8 平成 24 年度中日対訳辞書データ重複排除後件数

分 野	対訳辞書候補数
化 学	1,997,488 語
電 気	1,157,653 語
機 械	752,369 語
物 理	954,444 語
合計(分野間重複なし)	4,861,954 語

(b)人手確認結果

前述(a)で重複排除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 140 万件の人手確認を行い 100 万件が正しいデータと判断された。

４．新語の分析

前述項番３「新語の抽出」で作成した新語データについて、新語の状況、規模を調査・分析した。調査は新語データに付されている出現頻度情報による集計を行った。調査の結果、出現頻度上位から 27,192 位(出現頻度 128)までの累計が総出現頻度の 0.5 倍となり、出現頻度上位から 269,977 位(出現頻度 22)までの累計が総出現頻度の 0.8 倍に達する結果となった。また、出現頻度毎の件数を集計したところ、出現頻度 9～5 の語数が最も多く 41 万語となった。これらの結果を分析した結果、新語データは少数の高頻度語と多数の低頻度語から構成される傾向が見られた。さらに同様の分析を平成 24 年度中日対訳辞書データについても行った結果、出現頻度上位から 696 位(出現頻度 23,589)までの累計が総出現頻度の 0.5 倍となり、出現頻度上位から 12,360 位(出現頻度 728)までの累計が総出現頻度の 0.8 倍に達する結果となった。

次に、新語データの出現頻度上位 100 語について、技術分野、初出年、文字数、見出し語重複の観点から調査を行った。技術分野の調査結果からは、化学と電気分野に頻出する用語が多い傾向が見られた。初出年の調査からは、ある年から出現頻度が上昇するような新語の存在は確認出来ず、2010 年から 2012 年まで一定の出現頻度を持つ比較的一般的な用語が多く含まれていることが分かった。見出し語重複の調査からは、新語の中国語のうちおよそ 15%の用語が平成 24 年度中日対訳辞書データと重複した。この重複した中国語の用語は平成 24 年度とは異なる日本語の用語を抽出したことになる。この異なる日本語は技術分野による訳し分け、異表記・類義語など検索及び日中辞書への利用が考えられる。

以下に分析内容と結果を報告する。

４．１　分析の環境（分析用データ）

新語の分析では、本調査項番３「新語の抽出」で作成した新語データ 100 万語に加え、平成 24 年度調査で作成した中日対訳辞書データ 100 万語を分析用データとした。これら分析用データには出現頻度情報が付されている。この出現頻度情報は新語の作成元である中日対訳コーパス中の出現頻度である[14]。これら分析用データに対し、後述項番４．２で説明する分析手法による調査・分析を行った。

４．２　分析の手法

前述項番４．２で説明した分析用データには、中日対訳コーパスにおける出現頻度が付されている。この出現頻度が大きい新語データほど、機械翻訳の品質向上に資する重要なデータあると考えられる。このため、新語データの状況や規模、傾向を分析する際には新語

[14] 新語データは 2010 年から 2012 年までの中日対訳コーパスにおける出現頻度、平成 24 年度中日対訳辞書データは 2005 年から 2009 年の中日対訳コーパスにおける出現頻度となる。

データの出現頻度が有用な指標になる。
そこで、分析はこの考えに基づき、以下の（１）から（６）を行った。

（１）全体の分布の分析１
（２）全体の分布の分析２
（３）上位の新語データの傾向
（４）初出新語データ件数の推移
（５）平成 24 年度辞書データとの中国語／日本語見出し重複の割合
（６）平成 24 年度辞書データの、中日対訳コーパス（2010-2012）における出現頻度のカウント

（１）全体の分布の分析１

全体の分布の分析１では、新語データにおける出現頻度をすべて足し合わせた全体合計値を集計した上で、新語データにおける出現頻度を出現頻度が高い順に足し合わせていったときの合計値が、全体合計値の 0.5 倍の値となるときの出現頻度、全体合計値の 0.8 倍の値となるときの出現頻度をそれぞれ求めた。さらに、平成 24 年度調査で作成した中日対訳辞書データについても同様の処理を行い、両者を比較・分析した。

（２）全体の分布の分析２

全体の分布の分析２では、任意の出現頻度と、それに該当する新語数とを把握するため、新語に付されている出現頻度を 1,000 以上、999～500、499～250、249～100、99～50、49～25、24～10、9～5、4、3、2、1 回の範囲に区分し、各範囲に属する新語の語数を集計した。

また、全体の分析の分析１と同様、平成 24 年度辞書データについても同様の処理を行い、両者を比較・分析した。

（３）上位の新語データの傾向

出現頻度上位の新語 100 語について、以下の観点にて集計し、その傾向を分析した。

(a) 技術分野（化学、電気、機械、物理）
(b) 初出年
(c) 文字数分布（中国語、日本語）
(d) 平成 24 年度辞書データとの中国語／日本語見出し重複

さらに、平成 24 年度辞書データの出現頻度上位 100 語についても同様の集計を行い、両者を比較・分析した。

（４）初出新語データ件数の推移

新語データ全件について、2010 年、2011 年、2012 年にそれぞれ初出のものの件数を調査し、各年における中国公開特許公報の発行件数との関係推移について分析した。

また、この件数調査では、本調査にて作成する中日対訳コーパスが基準となることから、全ての語が 2010 年～2012 年のいずれかの年に初出と判断されるため、「従来は使われていなかった新たな技術用語・専門用語」という意味での「新語」であるか否かの分析も別途必要と考えた。具体的には、平成 24 度調査の成果物「対訳辞書候補データ」利用し、このデータ中に存在する語（すなわち、従来から使われていた語）と、存在しない語（新たな技術用語・専門用語とみなせる語）を区別した分析を行った。

（５）平成２４年度辞書データとの中国語／日本語見出し重複の割合

本調査の新語には、中国語または日本語見出しのいずれか一方だけが平成 24 年度辞書データと重複している場合がある。以下、このような中国語と日本語が１対 N で対応する場合、およびN対１で対応する場合、N語の別の訳語があるということで別訳語という。こうした別訳語が全体に占める割合を調査・分析した。

また、こうした別訳語は平成 24 年度辞書データ中にも存在していると考えられるため、平成 24 年度辞書データにおける別訳語の割合についても調査し、本年度データとの比較を行った。

（６）平成２４年度辞書データの、中日対訳コーパス（2010-2012）における出現頻度のカウント

平成 24 年度辞書データ 100 万語について、本調査にて作成した中日対訳コーパス（2010 年～2012 年）における出現頻度をカウントし、平成 24 年度の中日対訳コーパス（2005 年～2009 年）における出現頻度と対比した。対比により、平成 24 年度の辞書データのうち、年度を問わず定期的に使用される語、出現頻度が上昇した語、下降した語などを分析した。

４．２　分析結果

（１）全体の分布の分析１

本調査で作成した新語データ 100 万語を出現頻度の大きい順にしたファイルについて、全ての出現頻度を合計すると 36,400,270 となる。このファイルの出現頻度１位から 27,192 位までの用語の出現頻度の合計が総出現頻度の 0.5 倍となり、269,977 位までの出現頻度の合計が総出現頻度の 0.8 倍となった。一方、平成 24 年度中日対訳辞書データを同様に集計すると出現頻度の合計が 129,336,825 となり、0.5 倍が 696 位、0.8 倍が 12,360 位となった。

出現頻度の累計値と語数の関係を表 4.2-1、図 4.2-1 に示す。

表 4.2-1 出現頻度の割合（累計）の比較

総出現頻度に対する割合	新語データ（上段：出現頻度、下段：用語順位）	平成 24 年度中日対訳辞書データ（上段：出現頻度、下段：用語順位）
0.5 倍	128 27,192 位	23,589 696 位
0.8 倍	22 269,977 位	728 12,360 位

図 4.2-1 出現頻度の割合（累計）

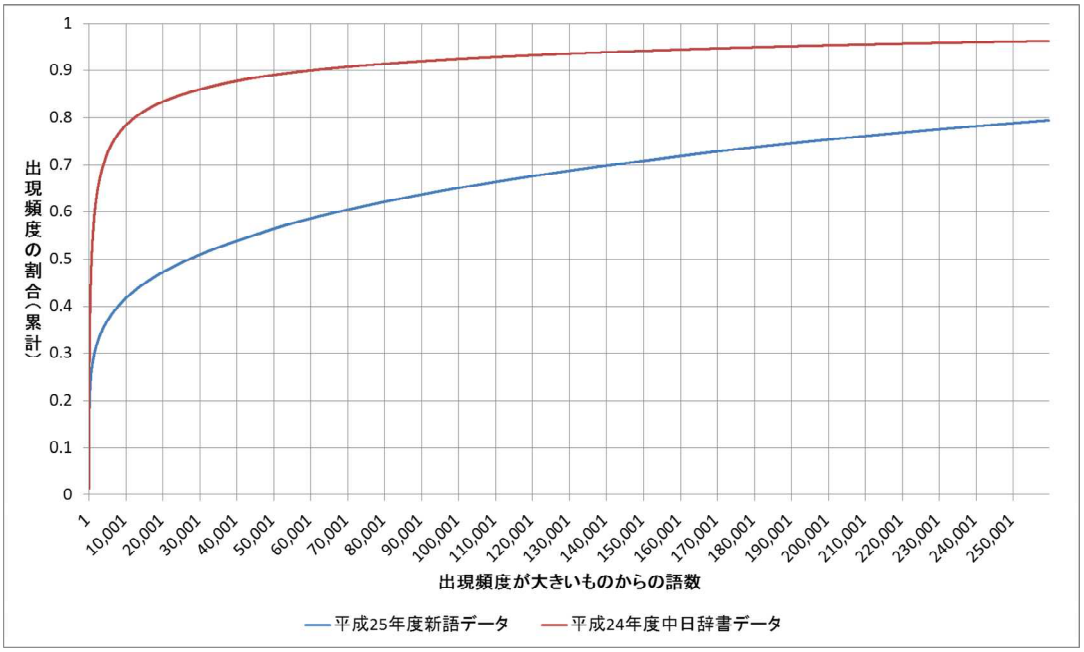


図 4.2-1 は用語を出現頻度の大きい順に並べて、先頭の語からその出現頻度を積算して全用語の出現頻度の合計で除した出現頻度の割合（累計）をグラフ化したものである。図 4.2-1 の結果から新語データと平成 24 年度中日対訳辞書データの傾向を比較すると、平成 24 年度中日対訳辞書データは総出現頻度に対する割合が 0.8 まで急激にグラフが上昇し、0.8 以降傾きが緩やかになる傾向が見られた。一方、新語データは総出現頻度に対する割合が 0.5 付近までは急な上昇傾向が見られ、その後緩やかに上昇し続ける傾向が見られた。

これらの結果を比較すると、平成 24 年度中日対訳辞書データは新語に比べ出現頻度の大きな用語の数が多い傾向にあると分かった。後述（２）全体の分布の分析２では、出現頻度範囲毎に語数集計を行うことで、さらに詳細な傾向の比較・分析を行った。

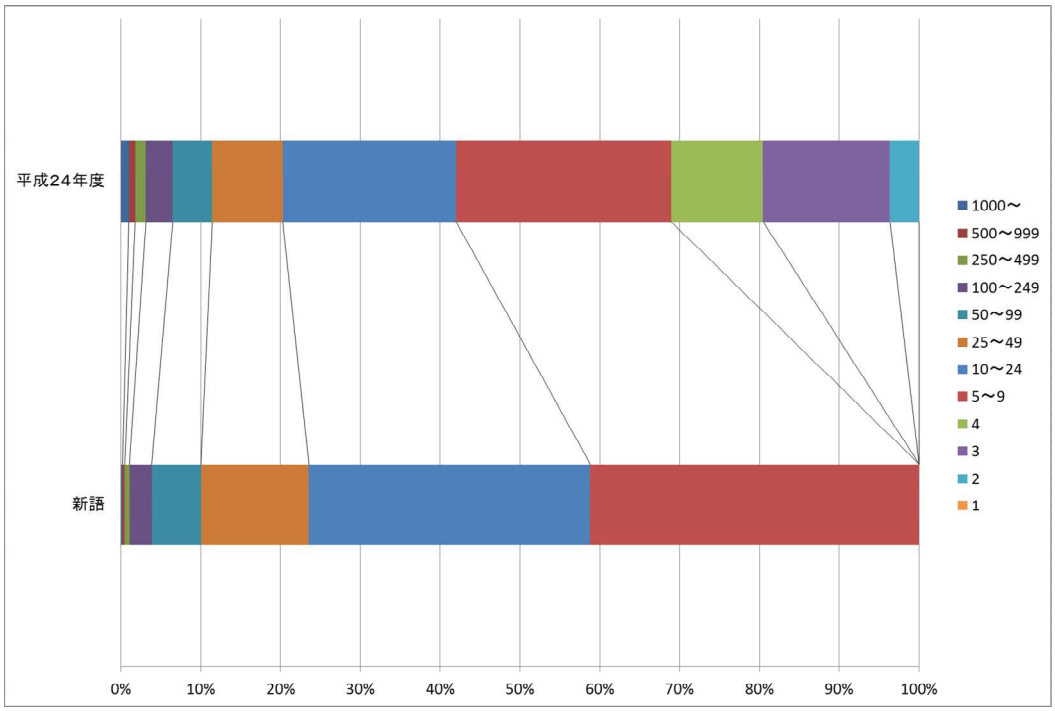
（２）全体の分布の分析２

新語データを出現頻度が 1,000 以上、999～500、499～250、249～100、99～50、49～25、24～10、9～5、4、3、2、1 回の範囲に区分し、各範囲に属する語数を調査した。さらに同様の集計を平成 24 年度中日対訳辞書データについても行った。調査結果を表 4.2-2、図 4.2-2 に示す。

表 4.2-2 出現頻度範囲毎の語数

出現頻度範囲	新語データ	平成 24 年度辞書データ	出現頻度範囲	新語データ	平成 24 年度辞書データ
1,000 以上	1,960	10,118	24～10	352,462	217,267
999～500	2,605	7,734	9～5	412,130	269,085
499～250	6,224	13,430	4	0	115,131
249～100	28,050	33,609	3	0	159,076
99～50	61,659	49,450	2	0	36,610
49～25	134,910	88,490	1	0	0

図 4.2-2 出現頻度範囲毎の語数構成比



集計結果から新語データは、語数が多い順から、出現頻度 9～5 が 41 万語、出現頻度 24～10 が 35 万語、49～25 が 13 万語となった。一方、平成 24 年度中日対訳辞書データは、語数が多い順から、出現頻度 9～5 が 27 万語、出現頻度 24～10 が 21 万語、出現頻度 4 が 11 万語となった。

新語データと平成 24 年度中日対訳辞書データの集計結果を比較すると、新語データは出現頻度 25～9 の範囲で平成 24 年度より語数が多くなり、出現頻度上位と出現頻度下位の語数が少なくなる傾向が見られた。

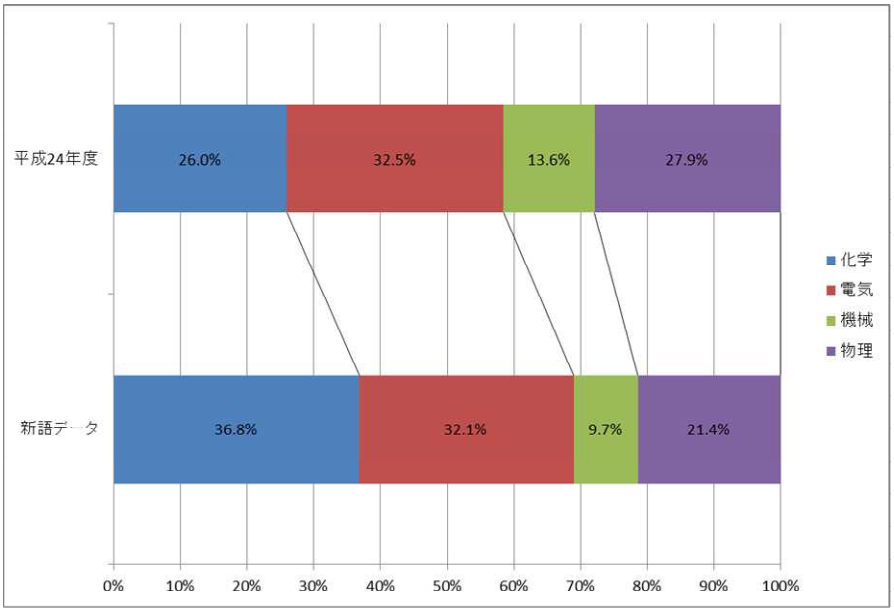
（３）上位の新語データの傾向

上位の新語データの傾向では、本調査で抽出した新語データの出現頻度上位 100 語と平成 24 年度中日対訳辞書データ出現頻度上位 100 語を調査対象とし、(a)技術分野、(b)初出年、(c)文字数分布、 (d)平成 24 年度辞書データとの中国語/日本語見出し重複の観点による分析を行った。

(a)技術分野（化学、電気、機械、物理）

技術分野別の傾向を把握するため、分野毎の出現頻度分布を調査した。分野毎の出現頻度は、上位 100 語について、前述の技術分野別の中日対訳コーパス(表 3.3-4)における出現頻度を調査した。また、同様の調査を平成 24 年度中日対訳辞書データについても行った。調査結果を図 4.2-3 に示す。

図 4.2-3 分野別出現頻度構成比（新語上位 100 語）



調査の結果、新語データは出現頻度の多い分野から、化学、電気、物理、機械の順番となり、出現頻度の多い化学と、電気分野の出現頻度で全体の 70%弱を占める傾向が見られた。平成 24 年度中日対訳辞書データは出現頻度の多い分野から、電気、化学、物理、機械の順番となり、出現頻度の多い電気と化学分野の出現頻度で全体の 60%弱を占める傾向が見られた。この電気分野と化学分野が多い割合を占める傾向は、新語作成元の公開公報[15]の分野の構成比が化学 33%、電気 30%、機械 17%、物理 20%であり、この傾向が影響したと考えられる。

(b)初出年

本調査で作成した新語は、2010 年から 2012 年までに公開された中国公開公報と内容がほぼ対応する日本公開公報のデータから作成したものである。新語データを中国公開公報の公開年毎に出現頻度を集計することで、ある年から出現した用語（初出年）や、ある年以降出現頻度が上昇した用語、逆にある年以降出現頻度が減少した用語など用語の使用傾向を調査した。調査は、新語データ頻度上位 100 語について中国公開特許の公開年別の出現頻度分布を集計した。調査結果を図 4.2-4 に示す。

図 4.2-4 中国公開年別出現頻度(新語出現頻度上位 100 語)

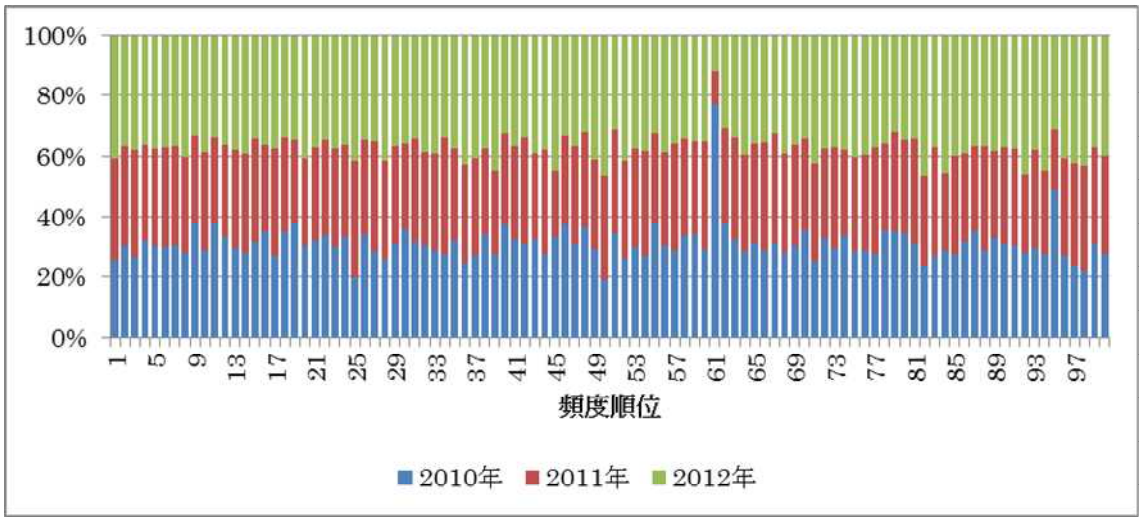


図 4.2-4 は出現頻度上位 100 語の年別出現頻度推移を比較しやすくするため、出現頻度の構成比でグラフ化した。調査結果から、出現頻度上位 100 語においては 2010 年、2011 年、2012 年各年全てが一定の出現頻度をもつ結果となった。また、各年の出現頻度の割合は、一部を除き大きな差がない結果となった。2010 年から 2012 年までの出現頻度推移を見ると、2010 年から 2012 年にかけて件数が増加している語が 100 語中、53 語見つかった。但しそれらの語の各年の出現頻度値に大きな差はなく、誤差範囲の差であった。これらの傾向か

[15] 前述表 3.3-1 対応特許中国日本公開特許公報番号リストの分野別件数を参照のこと。

ら、新語上位 100 語においては、ある年から新たに出現する意味での「新語」は存在しないことが分かった。

全体の傾向は以上であるが、調査結果中、全体の傾向と異なる用語が 1 語存在した。”基准”（中国語)、”参照”（日本語）は 2010 年から 2012 年までの出現頻度が 23,540、3,362、 3,574 となり 2010 年の出現頻度が極端に大きな用語であった。

また、本調査では、2010 年より前の年に出現した用語か否かの判別するために、平成 24 年度調査の成果物である対訳辞書候補の中に出現頻度上位 100 語が存在するかの調査を行った。調査は新語データの各語について、平成 24 年度対訳辞書候補の中国語と日本語を部分一致検索で、存在の確認を行った。部分一致検索は例えば、“溶液”（中国語)、“溶液”（日本語）を検索する場合、これら中国語と日本語用語を含む平成 24 年度対訳辞書候補の“溶液混合”（中国語)、“溶液ブレンド”（日本語）をヒットさせる方式である。調査の結果、出現頻度上位 100 語全てが平成 24 年度調査の対訳辞書候補に含まれており、新語データ出現頻度上位 100 語は 2010 年以降に新たに出現した用語でないことが分かった。

(c)文字数分布（中国語、日本語）

出現頻度上位 100 語について中国語、日本語の文字数を調査した。調査結果を表 4.2-3 に示す。

表 4.2-3 出現頻度上位 100 語文字数

	中国語			日本語		
	最小	最大	平均	最小	最大	平均
新語データ	2	5	2.28	2	6	3.09
平成 24 年度中日対訳辞書データ	1	4	2.13	1	4	2.28

調査の結果、新語の中国語文字数は 2 文字から 5 文字で構成され、その平均は 2.28 文字であった。日本語の文字数は 2 文字から 6 文字で構成され、平均 3.13 文字であった。日本語と中国語の文字数を比べると日本語の方が構成する文字が多い傾向が見られた。これは例えば、中国語「钟信号」に対して日本語「クロック信号」のカタカナ表記や、中国語「弯曲」に対する日本語「湾曲する」のようなサ変動詞の末尾「する」などが、その要因と考えられる。また、平成 24 年度についても同様の傾向が見られた。

(d)平成 24 年度辞書データとの中国語／日本語見出し重複

本調査で作成した新語データには中国語または日本語見出しのいずれか一方だけが平成 24 年度中日対訳辞書データと重複する場合がある。これは、ある中国語に対する日本語の

訳語が複数存在し、同様に、ある日本語に対する中国語の訳語が複数存在するためである。このような用語がどの程度存在するか調査を行った。調査は(1) 出現頻度上位 100 語の新語データの中国語が同一な平成 24 年度中日対訳辞書データ、(2)同新語データの日本語と同一な平成 24 年度中日対訳辞書データの語数を調査した。調査結果を表 4.2-4 に示す。

表 4.2-4 中国語／日本語見出し語重複結果

新語データ上位 100 語の中国語と重複する平成 24 年度辞書データの語数	38 語
新語データ上位 100 語の日本語と重複する平成 24 年度辞書データの語数	30 語

調査結果から、新語上位 100 語の内およそ 4 割にあたる 38 語が平成 24 年度中日対訳辞書データと中国語または日本語見出しの重複が存在した。

中国語と重複する用語例として、新語“显示单元”（中国語)、“表示部”（日本語）の中国語と重複する平成 24 年度中日対訳辞書データは“ディスプレイユニット”、“ ディスプレイ部”、“ 表示セル”、“ 表示ユニット”、“ 表示手段”、“ 表示装置”が存在した。

また日本語と重複する用語例として、新語“支持部”（日本語)、“支撑部”（中国語）の日本語と重複する、平成 24 年度中日対訳辞書データは“支承部”、“支持部分”、 “支撑单元”、“支撑构件”、 “支撑部分”が存在した。

（４）初出新語データ件数の推移

新語データ全件 100 万語を対象に、2010 年、2011 年、2012 年各年で初めて出現した語数の集計を行った。2010 年の初出語数については平成 24 年度事業の成果物である対訳辞書候補データ[16]の有無により集計を行った。集計結果を表 4.2-5 に示す。

16 平成 24 年度事業の対訳辞書公報データは 2005 年～2009 年に公開された中国公開公報を使用しているため本調査では、このデータと重複する用語は 2009 年以前に初出のデータであるとした。

表 4.2-5　初出新語データ件数及び中国公開公報件数

公開年	初出新語	初出新語のうち当該年のみに出現した語数	中国公開公報件数
2010 年	379,134 語	151,672 語	315,836
2011 年	250,387 語	170,566 語	368,434
2012 年	202,372 語	202,372 語	543,296

表 4.2-5 の初出年 2010 年の初出新語は新語データのうち平成 24 年度調査で作成した「対訳辞書候補データ」に存在せず、2010 年に出現した語数を集計した。2011 年の初出新語は、新語データのうち 2010 年には出現せず、2011 年に出現した語数を集計した。2012 年の初出新語は、2010 年、2011 には出現せず 2012 年に出現した語数を集計した。

集計結果は、2010 年の初出新語がおよそ 38 万件で最も多く、2011 年の初出新語が 25 万件、2012 年の初出新語は 20 万語となった。

また、各年の初出新語の中にはその年だけに出現した用語が含まれていた。そのような語を初出新語のうち当該年にのみ出現した語数として集計した。例えば 2011 年の初出新語 250,387 語のうち、68%にあたる 170,566 語は 2011 年のみに存在し、2010 年と 2012 年には出現しない用語であった。

次に初出新語数と各年における中国公開特許公報発行件数との関係を調査した。2010 年から 2012 年の中国公開特許公報[17]の件数は 31 万 5 千件、36 万 8 千件、54 万 3 千件となり、年々公開公報件数が増加する傾向が見られた。初出新語のうち当該年のみに出現した語数と中国公開特許公報発行件数との関係は、2010～2012 年にかけて増加傾向が見られ、公報発行件数との間に関連性があると考えられる。

（５）平成２４年度辞書データとの中国語／日本語見出し重複の割合

本調査の新語について、中国語のユニークな語数と日本語のユニークな語数を集計した。次に別訳語を持つ中国語の語数と別訳語を持つ日本語の語数を集計した。また、平成 24 年度中日対訳辞書データについても同様の調査を行った。さらに、新語データと平成 24 年度中日対訳辞書データを合わせたデータについても同様に調査した。調査結果を表 4.2-6 に示す。

17 ここでの公報発行件数は、本調査で使用した中国公開公報の XML 形式データのファイル数を公報発行件数とした。

表 4.2-6 中国語／日本語見出し重複の割合

	中国語の別訳語(中:日=1:n)			日本語の別訳語(日:中=1:n)		
	(A)中国語見出し語数	(B)別語をもつ語数	(C)全体に占める割合(B/A)	(D)日本語見出し語数	(E)別語をもつ語数	(F)全体に占める割合(E/D)
新語データ	867,852	86,691	9.9%	884,077	99,106	11.2%
平成 24 年度中日対訳辞書データ	829,063	108,356	13.1%	791,713	148,982	18.8%
新語と平成 24 年度中日対訳辞書データ	1,572,305	232,663	14.7%	1,530,581	300,144	19.6%

新語データのユニークな中国語の内、別訳語を持つのは 86,691 語(9.9%)で、ユニークな日本語のうち、別訳語をもつのは 99,106 語(11.2%)であった。同様の調査を平成 24 年度中日対訳辞書データに対して行った。さらに、新語データと平成 24 年度中日対訳辞書データを合わせたものについて同様の調査を実施した結果、別訳語を持つユニークな中国語の語数の割合が 14.7%と別誤訳を持つユニークな日本語の語数の割合が 19.6%となった。

これらの結果から、新語データと平成 24 年度中日対訳辞書データの間で共通する中国語見出し語が存在することが分かった。言い換えると本調査で作成した新語のうちいくつかは、新しい用語というよりも既存の見出し語に対する訳のバリエーションの追加が含まれていると考えられる。以下、表 4.2-7、表 4.2-8 に別訳語の例を示す。これらの例から、別訳語の中には訳語のゆれに近い表現が多く含まれていると考えられる。

表 4.2-7 中国語の別訳語例

中国語見出し語	別訳語（日本語）
连接杆	インタフェースレバー ||| コネクティングロッド ||| タイバー ||| リンクバー ||| リンクレバー ||| リンクロッド ||| レバーリンク ||| 取り付け棒 ||| 引張棒||| 接続バー ||| 接続リンク ||| 接続レバー ||| 接続ロッド ||| 接続棒 ||| 連係ロッド ||| 連接棒 ||| 連結バー ||| 連結リンク ||| 連結レバー ||| 連結ロッド ||| 連結桿 ||| 連結棒 ||| コネクションバー ||| コンロッド ||| リンク杆 ||| 接続支柱 ||| 接続杆 ||| 結合ロッド ||| 締結ロッド ||| 連接ロッド ||| 連結杆 ||| 連絡バー

表 4.2-8 中国語別訳語例

日本語見出し語	別訳語（中国語）
シャッタ部材	卷帘部件 ||| 开闭器部件 ||| 开闭部件 ||| 快门部件 ||| 挡板构件 ||| 遮板件 ||| 遮档板构件 ||| 闸片部件 ||| 闸门构件 ||| 闸门部件 ||| 挡板部件 ||| 滑门构件 ||| 遮挡构件 ||| 遮蔽构件 ||| 闸板部件

（６）平成２４年度辞書データの、中日対訳コーパスにおける出現頻度カウント

平成 24 年度に作成した中日対訳辞書データ（100 万語）について、本調査で作成した中日対訳コーパス（絞り込み後）[18]における中日出現頻度の集計を行った。その結果、出現頻度１以上の用語が約 54 万語存在した。これらの用語について平成 24 年度の出現頻度と今回カウントした出現頻度の比較を行った。比較は本調査での出現頻度値が、平成 24 年度の出現頻度より 5 倍以上増加したもの、逆に 1/5 以下に減少したもの、その他のものに分け用語数を集計した。これは出現頻度カウントに用いた中日対訳コーパスの文数が、平成 24 年調査では約 1,500 万文、本調査で作成した中日対訳コーパス（絞り込み後）はおよそ 2.5 倍の 3,800 万件の文であり、単純に出現頻度値の比較は行えないため、2.5 倍のさらに倍以上の差があれば出現頻度が急に増加もしくは減少したと考えこのような集計を行った。集計結果を表 4.2-9 に示す。

[18] 項番 3.3 抽出結果-(2)中日対訳コーパスの解析（対訳辞書候補データの作成）-(a)中日対訳コーパス(絞り込み後)を参照のこと。

表 4.2-9 本調査で作成した中日対訳コーパスにおける出現頻度と
平成 24 年度中日対訳辞書データの出現頻度比較

項目	語数	構成比
出現頻度 5 倍以上増加	79,516	8.0%
出現頻度が 1/5 から 5 倍未満	416,609	41.7%
出現頻度が 1/5 以下に減少	46,155	4.6%
出現頻度ゼロ	457,720	45.8%
合計	1,000,000	100%

集計結果から、出現頻度ゼロの用語が全体の 45.8%であることから、平成 24 年度辞書データのうちおよそ半数以上は 2010 年～2012 年までの中日対訳コーパスにおいても出現する用語であることが分かった。出現頻度が 5 倍以上出現した用語が 8%弱存在し、逆に出現頻度が 5 分の 1 以下に減少した用語が 4.6%弱存在した。これらの用語例を表 4.2-10、表 4.2-11 に示す。

表 4.2-10 出現頻度が 5 倍以増加した用語例

日本語	中国語	頻度 H24 H24 年度中日対訳辞書データ内の出現頻度	頻度 H25 2010～2012 年中日対訳コーパスの出現頻度	頻度 H25 ÷ 頻度 H24
基准帧	参照フレーム	3	11,151	3,717.0
接收品质信息	受信品質情報	5	7,052	1,410.4
替换记录	交替記録	2	2,748	1,374.0
管理信息列表	管理情報リスト	3	3,916	1,305.3

表 4.2-11 出現頻度が 5 分の 1 以下に減少した用語例

日本語	中国語	頻度 H24 H24 年度中日対訳辞書データ内の出現頻度	頻度 H25 2010～2012 年中日対訳コーパスの出現頻度	頻度 H25 ÷ 頻度 H24
折射率各向异性材料	屈折率異方性材料	348	1	0.003
演算控制装置	演算制御装置	345	1	0.003
射出侧偏振片	射出側偏光板	341	1	0.003
金属支持基板	金属支持基板	312	1	0.003

平成 24 年度中日対訳辞書データの 2005～2009 年の中日対訳コーパスにおける出現頻度と 2010～2012 年の対訳コーパスにおける出現頻度を比べて、1/5 以下に減少または出現頻度ゼロになってしまう用語が全体の 50.4%あるが、これらの用語は用語を採取した年範囲より新しい 3 年間の中日対訳コーパスにおいてほとんど使われなかった用語である。前述表 4.2-11 の出現頻度が 1/5 以下に減少した語を見ると、これらの語の H24 の出現頻度は 300 回程度とそれなりに高く、用語も特異な用語とは思えない。このような年により出現頻度が著しく増減する語であっても出現頻度が多い年代の文献の翻訳に必要な語であり、辞書化する必要がある。そして、このようなある年代に高頻度で現れる用語は今後も発生し得るため、、新しい中日対訳コーパスがある程度蓄積される度に、定期的に新語の抽出を行うことが望ましい。

５．中日対訳辞書データの作成

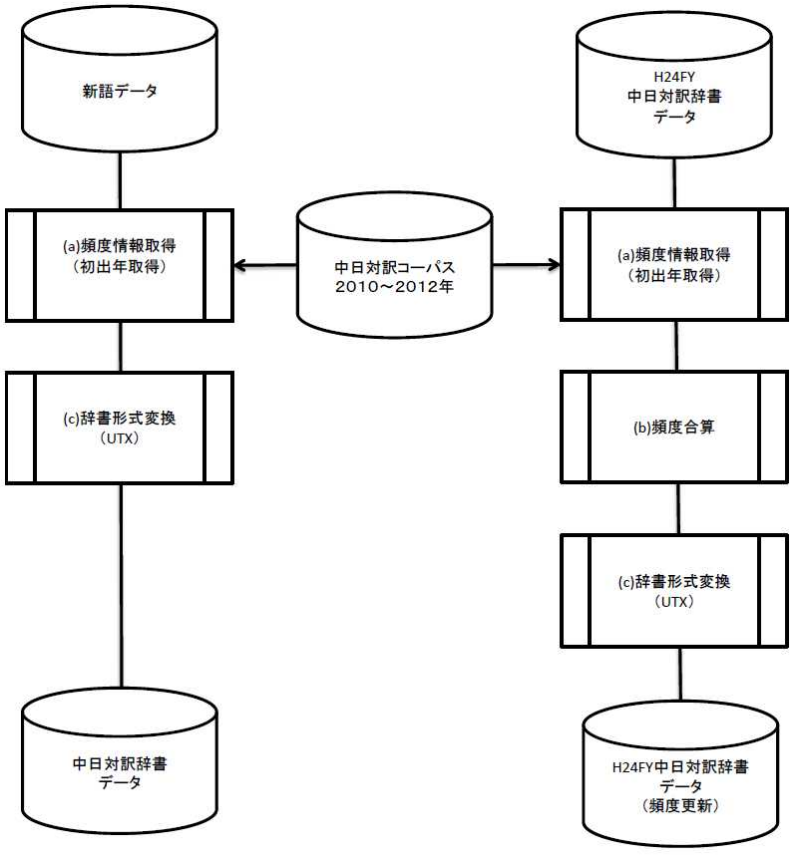
５．１　作成手法

前述項番３.新語の抽出で作成した 100 万語の新語データについて、「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））、出現頻度情報から中日対訳辞書データ[19]を作成した。

出現頻度情報は、化学分野、電気分野、機械分野、物理分野および全分野に属する中日対訳コーパスにおいて、中国語単独で出現する頻度、日本語単独で出現する頻度、中国語と日本語に同時に出現する頻度を作成した。また、中日対訳辞書データは UTX 形式(UTX1.11)で作成した。（詳細な中日対訳辞書データの形式は「添付資料５．１中日対訳辞書データレイアウト説明」を参照のこと。）

さらに、平成 24 年度に作成した中日対訳辞書データについて、前述と同様の出現頻度情報を取得し、平成 24 年度の中日対訳辞書データに付されている出現頻度情報に加算したデータを作成した。

図 5.1 中日対訳辞書データ概要図



[19] 詳細は、添付資料 5.1 中日対訳辞書データレイアウト説明を参照のこと。

(a) 出現頻度情報取得

出現頻度情報取得は、前述項番３．新語の抽出で使用した 2010 年～2012 年の中日対訳コーパスにおける技術分野別出現頻度を取得した。技術分野は、中日対訳コーパスの国際特許分類（IPC）を基に、化学分野、電気分野、機械分野、物理分野の４分野[20]に分類し使用した。

(b) 出現頻度合算

平成 24 年度の中日対訳辞書データには 2005 年～2009 年までの中日対訳コーパスにおける出現頻度情報が付されている。出現頻度合算では、この出現頻度情報と上述(a)で取得した 2010 年～2012 年の出現頻度情報を合算し 2005 年～2012 年までの出現頻度情報を作成した。

(c)辞書形式変換(UTX)

辞書形式変換では、「添付資料５．１中日対訳辞書データレイアウト説明」で示す形式にデータレイアウトの変換を行った。

[20] 詳細は、添付資料 5.2 ４分野と IPC の関係を参照のこと。

６．中日対訳コーパスの高精度化の検証

前述項番 ３．新語の抽出 において中日自動文アライメントツールを用いて中日対訳コーパスを作成した。この自動文アライメントツールは中国語と日本語の対訳辞書に基づいて対応する日本語と中国語の対訳文を推定する。この自動文アラインメントツールが読み込む（学習する）中国語と日本語の対訳辞書を学習用辞書という。

ここで自動文アライメントによって作成される中日対訳コーパスの品質、言い換えると作成された中日対訳コーパスが正しい対訳文対をどれくらい含むかは、

・自動文アライメントの分析対象となった「対訳関係にある２つの文書の文章」の内容
・自動文アライメントツールのアルゴリズム
・中日自動文アライメントツールが学習する対訳辞書データの品質と語数

に依存する。

本調査では、中日自動文アライメントツールの学習用辞書を語数により３種類用意してそれぞれの学習用辞書で作成した中日対訳コーパスの対訳文としての正解率（精度）がどのように変化するか検証した。

検証の結果、中日自動文アライメントツールに中日対訳辞書データを追加することで中日対訳コーパスが高精度化することが確認できた。一方、辞書の語数が最も多いときの精度が二番目の結果となり、辞書の語数の増加による精度向上が頭打ちになる傾向が見られた。

６．１　検証の環境

（１）学習用辞書

中日自動文アライメントツールが利用する学習用辞書は中国語を見出し語とし、日本語を訳語としたデータである。中日自動文アライメントツールはこの学習辞書を使い、入力の中国語文を辞書引きして日本語の単語列に変換し、対訳候補の日本語と突き合わせて用語の一致率を見て、対訳文とするか否かを判定する。学習用辞書例を下記図 6.1-1 に示す。

図 6.1-1　学習用辞書例

模板 ID 信息 テンプレート ID 情報
墨馈出通路 インク 送出 路
苯乙烯类单体 スチレン 系 モノマー
HSV 疾病 HSV 疾病
位置信息请求消息 位置 情報 要請 メッセージ

学習用辞書の項目は左から中国語見出し と 複数の日本語用語で構成される。日本語用語は単語に分かち書きしたデータとなる。例えば、上記学習辞書例の中国語 ” 模板 ID 信息” に対応する日本語は”テンプレート”、”ＩＤ”、”情報”の３単語に分かち書きしたデ

ータとなる。

検証データ１、２、３を作成するため、表 6.1-1 に示す３種類の学習用辞書データを作成した。

表 6.1-1 検証データと検証データ作成に使用した学習用辞書との関係

検証データ	学習用辞書
検証データ１	中日自動文アライメントツール付属の対訳辞書（語数 156,845 語）
検証データ２	検証データ１の対訳辞書に平成 24 年度調査で作成した中日対訳辞書データを追加したもの（合計語数：973,865 語）
検証データ３	検証データ２の対訳辞書に、本調査で作成した中日対訳辞書データ[21]を追加したもの（合計語数： 1,704,744 語）

＊学習用辞書の語数は中国語のユニークな語数である。

（２）検証データ

前述（１）で作成した３種類の学習用辞書を中日自動文アライメントツールに学習し、各検証用データを作成した。検証データは特許公報データのタイトル、要約、請求項、明細書に分けて作成した。作成箇所別の検証データの件数を表 6.1-2 に示す。

表 6.1-2 検証データ件数

検証データ	タイトル	要約	請求項	明細書	合計
検証データ１	347,960	936,112	3,430,641	61,088,869	65,803,582
検証データ２	347,960	936,824	3,435,270	61,119,760	65,839,814
検証データ３	347,960	937,103	3,435,790	61,141,517	65,862,370

６．２　検証の手法

検証は、上述検証の環境で説明した各検証データに対して、下記(1)から(3)の検証をした。

（１）文対応スコア分布の詳細比較

[21] 検証データ３で使用した中日対訳辞書データは、調査工程の都合から最終的な人手確認前の辞書データを使用した。

（2）人手によるサンプルチェック
（3）高精度な対訳文数の推定

（1）文対応スコア分布の詳細比較

各検証データについて、文の対応度合いを示す文対応スコア毎に対訳文の件数を集計し、比較を行った。また、文対応スコア毎の文数を累積し、検証データに占める割合を集計し、その結果の比較を行った。

（2）人手によるサンプルチェック

人手によるサンプルチェックは、検証データを文対応スコアの降順（最大スコアが先頭）にソートし、所定順位に該当する 20 件を人手によるサンプルチェックし、その結果から各中日対訳コーパス中の正確とみなせる対訳文数を推定・比較した。所定順位は前述(1)文対応スコア分布の調査結果を参考に、文対応スコアが 0.20，0.18，0.16，0.14，0.12，0.10，0.09，0.08，0.07，0.06，0.05，0.04，0.03 の 13 スコアについて人手によるチェックを行った。

（サンプルチェックの詳細は、添付資料６．２人手によるサンプルチェック方法を参照のこと。）

（3）高精度な対訳文数の推定

高精度な対訳文数を推定は、前述(2)人手によるサンプルチェック結果を基に各検証データに正しい対訳文対がどの程度の分量含まれるか解析した。具体的には検証データ毎の文対応スコアと、サンプルチェックから得たその文対応スコアの正解率から近似曲線[22]を示す式を算出した。そして、この算出した近似曲線の数式に前述(1)で得た所定の文対応スコアの文数を代入することで、所定の文対応スコアの内の正しい文対応の分量を推定した。

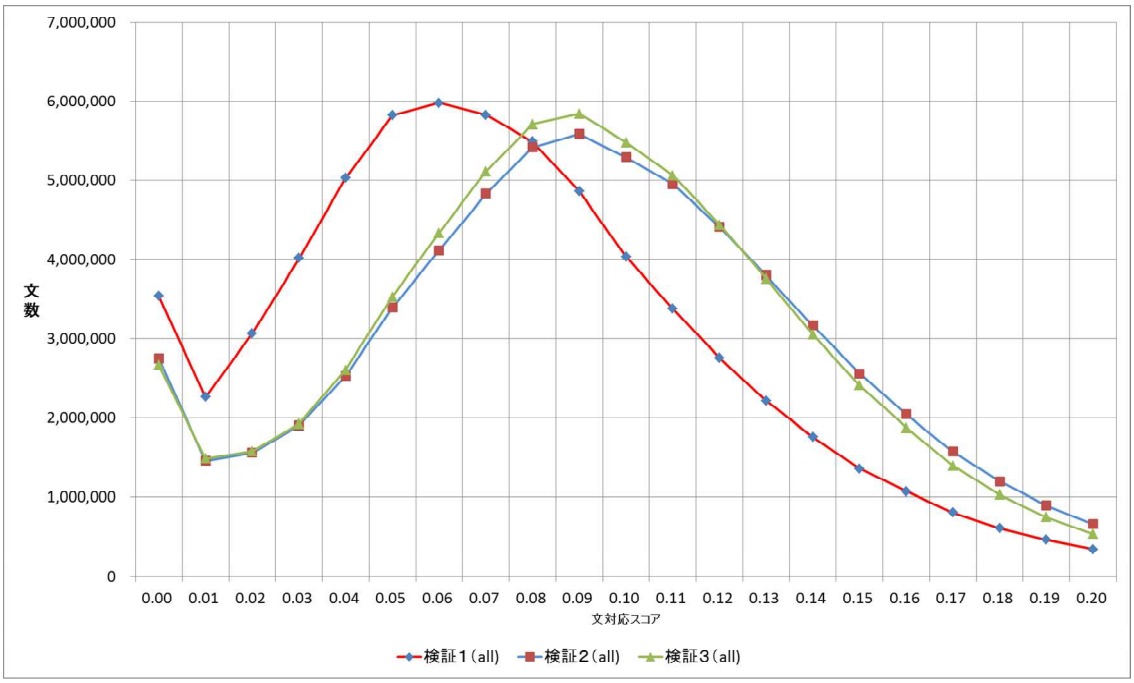
[22] 近似曲線はサンプルチェック結果の傾向を示す曲線となる。この極性は特定の式で表現することが可能であり、この式を用いてデータの予測の問題を分析することが可能となる。（回帰分析ともいう）

６．３　検証結果

（１）文対応スコア分布の詳細比較

各検証データの文対応スコア毎の文数の分布を調査した。調査結果を図 6.3-1 に示す。

図 6.3-1 文対応スコア分布



調査結果から、検証データ１と比べ、検証データ２、検証データ３は文対応スコアの値の大きい文数が多くなる傾向が見られた。この傾向は、対訳辞書増加により日本語文と中国語文の用語の対応が多くなり、その結果文対応スコアが高くなったと考えられる。検証データ２と検証データ３との比較では、わずかに検証データ２の方がスコア値の大きい文数が多くなった。

文対応スコア分布の差の要因を確認するため、同一の文の検証データ１と、検証データ２との違いを調査した。調査結果を表 6.3-1，表 6.3-2 に示す。

表 6.3-1 確認例１（検証データ１の文対応スコア ＜ 検証データ２の文対応スコア）

日本語文	相互パイロット結合推定を判断することは、測定受信電力とアクセス端末ロケーション情報とを使用することをさらに含む。
中国語文	确定导频交叉耦合估计还包括使用测得的收到功率和接入终端定位信息。
文対応スコア（検証データ1）	0.03
文対応スコア（検証データ2）	0.14
備考	検証データ2で使用した学習用辞書により、日本語文に該当する辞書引きが行われたことにより、検証データ2の文対応スコア値が 0.11 上昇した。 導频 → パイロット 接入终端 → アクセス端末 定位 → ロケーション

確認例１は、検証データ１から文対応スコアの値が上昇した例である。備考欄にあるように学習用辞書の用語の増加により、辞書引き結果と日本語文の対応数が多くなり、文対応スコア値が大きくなることが確認できた。

表 6.3-2 確認例２（検証データ１の文対応スコア ＞ 検証データ２の文対応スコア）

日本語文	よって、同じ位置で停止することにより生じる永久ひずみを防止することができる。
中国語文	因此，能防止因停止在相同位置而引起的永久变形。
文対応スコア（検証データ1）	0.09
文対応スコア（検証データ2）	0.05
備考	検証データ１で使用した学習用辞書による辞書引きでは、 相同 → 同じ 位置 → 位置 と二語の辞書引きが行われた。 一方、検証データ2で使用した学習用辞書による辞書引きでは、 相同位置 → 同一 場所 の一語の辞書引きとなり、“同一 場所”が日本語文に存在しなかったため、検証データ2の文対応スコア値が小さくなった。 つまり、辞書を追加することにより辞書引きされる単位が長くなった結果の副作用と考えられる。

確認例 2 は、検証データ２の文対応スコア値が検証データ１より小さくなった例である。備考欄にあるように、中国語の用語“相同位置”が登録されたことにより日本語文中にある“同じ”や、“位置”の辞書引きがされず、その結果文対応スコアが低くなった。但し、確認例２で示すような文対応スコアが低くなるものは、下記図 6.3-2 文対応スコア分布から少数であると考えられる。

次に、この文対応スコア毎の文数を累積し、所定の文対応スコアまでの文数を調査した。調査結果を図 6.3-2 および表 6.3-3 に示す。

図 6.3-2 文対応スコア分布（累積）

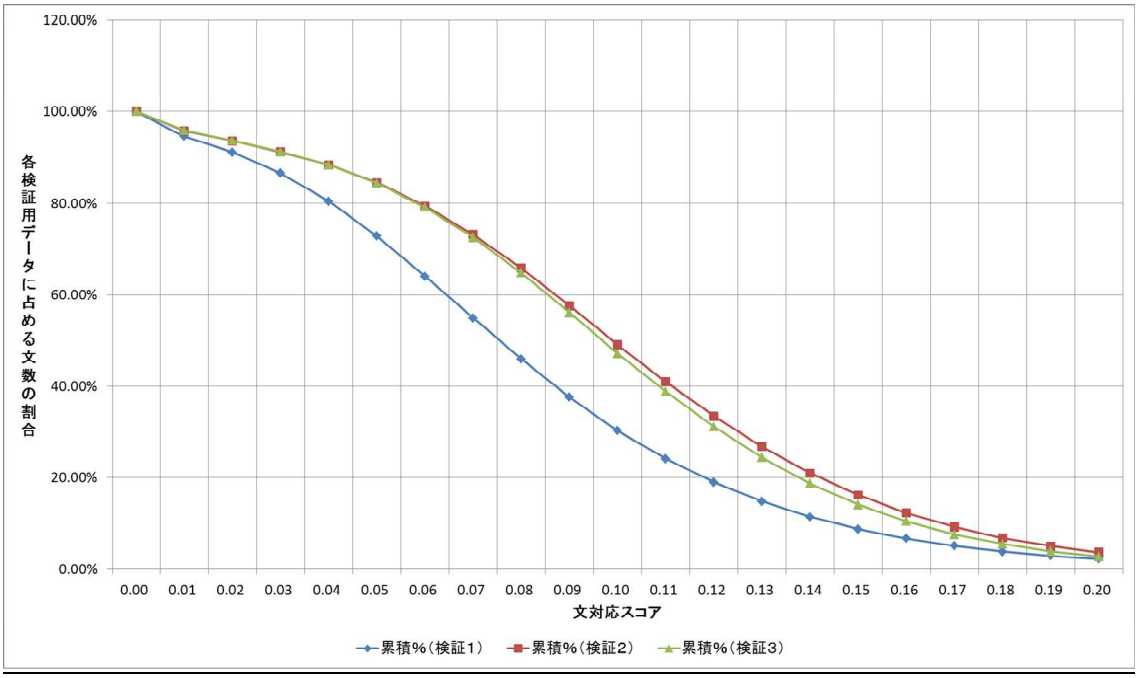


表 6.3-3 所定の文対応スコアまでの文数

文対応スコア	文 数（構成比）		
	検証データ１	検証データ２	検証データ３
1～0.20	1,468,757（ 2.23%）	2,405,000（ 3.65%）	1,824,303（ 2.77%）
1～0.18	2,532,138（ 3.85%）	4,490,327（ 6.82%）	3,596,989（ 5.46%）
1～0.16	4,405,160（ 6.69%）	8,106,639（12.31%）	6,863,890（10.42%）
1～0.14	7,522,158（11.43%）	13,834,333（21.01%）	12,323,862（18.71%）
1～0.12	12,495,496（18.99%）	22,047,365（33.49%）	20,524,003（31.16%）
1～0.11	15,875,308（24.13%）	26,997,617（41.01%）	25,584,543（38.85%）
1～0.10	19,908,406（18.99%）	32,290,871（49.04%）	31,069,039（47.17%）
1～0.09	24,772,149（24.13%）	37,877,254（57.53%）	36,906,865（56.04%）
1～0.08	30,267,007（30.25%）	43,297,755（65.76%）	42,612,467（64.70%）
1～0.07	36,090,184（37.65%）	48,125,615（73.10%）	47,727,646（72.47%）
1～0.06	42,068,342（46.00%）	52,241,949（79.35%）	52,063,810（79.05%）
1～0.05	47,888,596（54.85%）	55,635,358（84.50%）	55,595,263（84.41%）
1～0.04	52,918,999（63.93%）	58,166,000（88.34%）	58,202,769（88.37%）
1～0.03	56,935,678（72.78%）	60,068,946（91.23%）	60,129,936（91.30%）
1～0.00	65,803,582（ 100% ）	65,839,814（ 100% ）	65,862,370（ 100% ）

集計結果から、検証データ１の集計値と検証データ２の集計値を比べると、文対応スコア 1～0.20 では 6.6 倍、文対応スコア 1～0.16 では 1.8 倍の差が見られた。検証データ２と検証データ３との比較では文対応スコア 1～0.20 では 0.75 倍、1～0.16 では 0.84 倍検証データ３の件数が減少する傾向が見られた。

（2）人手によるサンプルチェック

文対応スコアと対訳文の正誤の関係（以降、正解率という）を見るため、検証データ１、２、３から文対応スコア毎に 30 文ずつ抽出して中国文と日本語文の対応の正誤を人手によりチェックをした。チェック結果を表 6.3-4 に示す。

表 6.3-4 人手によるサンプルチェック結果（正解率）

文対応スコア	検証データ１	検証データ２	検証データ３
0.03	50.0%	36.7%	36.7%
0.04	53.3%	56.7%	43.3%
0.05	86.7%	66.7%	56.7%
0.06	83.3%	73.3%	60.0%
0.07	83.3%	73.3%	70.0%
0.08	70.0%	93.3%	90.0%
0.09	80.0%	96.7%	100.0%
0.10	90.0%	86.7%	90.0%
0.12	96.7%	93.3%	96.7%
0.14	86.7%	96.7%	96.7%
0.16	100.0%	100.0%	90.0%
0.18	93.3%	96.7%	96.7%
0.20	86.7%	93.3%	96.7%

表 6.3-4 人手によるサンプルチェック結果（正解率）、文対応スコア毎に 30 文のサンプルデータの正解率を集計した。例えば、検証データ１の文対応スコア 0.03 の対訳文をチェックした結果、30 文中 15 文が正しい文対応データであったので、正解率は 50.0%とした。

サンプルチェックの結果から、検証データ１，２，３とも文対応スコアの値が大きくなるに従って、正解率が高くなる傾向が見られた。一方、検証データ１では文対応スコア 0.10 から正解率が 90%を超えるものの、文対応スコア 0.20 で 86.7%となり正解率が下がる場合があった。検証データ２と３は共に、文対応スコア 0.08 以上で正解率 90%を超える傾向が見られた。

文対応スコアの間の正解率の上昇具合を見ると検証データ１では、文対応スコア 0.04 と 0.05 の間で正解率が大きく上昇する傾向が見られた。検証データ２と３は共に、文対応スコア 0.07 と 0.08 の間で正解率が大きく上昇する傾向が見られた。

（３）高精度な対訳文数の推定

ここまで調査した結果を用いて、

(a) 文対応スコアと対訳コーパスの正解率の関係 と

(b) 対訳コーパスの正解率（精度）と対訳コーパスの文数の関係

を算出した。

(a) 文対応スコアと対訳コーパスの正解率の関係

前述表 6.3-4 人手によるサンプルチェック結果（正解率）から文対応スコアと対訳文の精度、そして対訳文の精度から得られる対訳文数を下記の手順で推定した。

なお、p(x) は文対応スコア x の時の正解率とし、近似曲線は Excel の機能で算出した。

① y＝1.0 - p(x) のグラフを作成して人手によるサンプルチェック結果をプロット

② Excel で①のプロットの近似曲線を算出（式 6.3-1 近似曲線式）

③ ②で算出した近似式を y=p(x)の形に変換してグラフ化（図 6.3-3 近似曲線）

③が求める文対応スコアと対訳コーパスの正解率の関係である。順に説明する。

①y＝1.0 - p(x) のグラフ作成では、横軸を文対応スコア、縦軸は人手によるサンプルチェック結果を y=1.0-p(x)としてプロットした。

②Excel で①のプロットに対して指数関数と累乗関数による近似曲線を試した結果、よりよく相関する累乗関数による近似曲線を採用し､式 6.3-1 に示す近似曲線の式が得られた。

式 6.3-1　近似曲線式

検証データ１　$p(x) = 1.0 - 0.00634^{x-1.22764}$
検証データ２　$p(x) = 1.0 - 0.00125^{x-1.79565}$
検証データ３　$p(x) = 1.0 - 0.00131^{-x-1.80684}$
x は文対応スコアを表す。
p(x)は文対応スコア x の時の正解率の近似を表す。

③式 6.3-1　近似曲線式の文対応スコア x を 0.03～0.20 まで変化させて y=p(x)の値を算出して図 6.3-3 近似曲線と②の人手によるサンプルチェックの値と共にグラフ化した。

図 6.3-3 近似曲線（文対応スコアと対訳コーパスの正解率の関係）

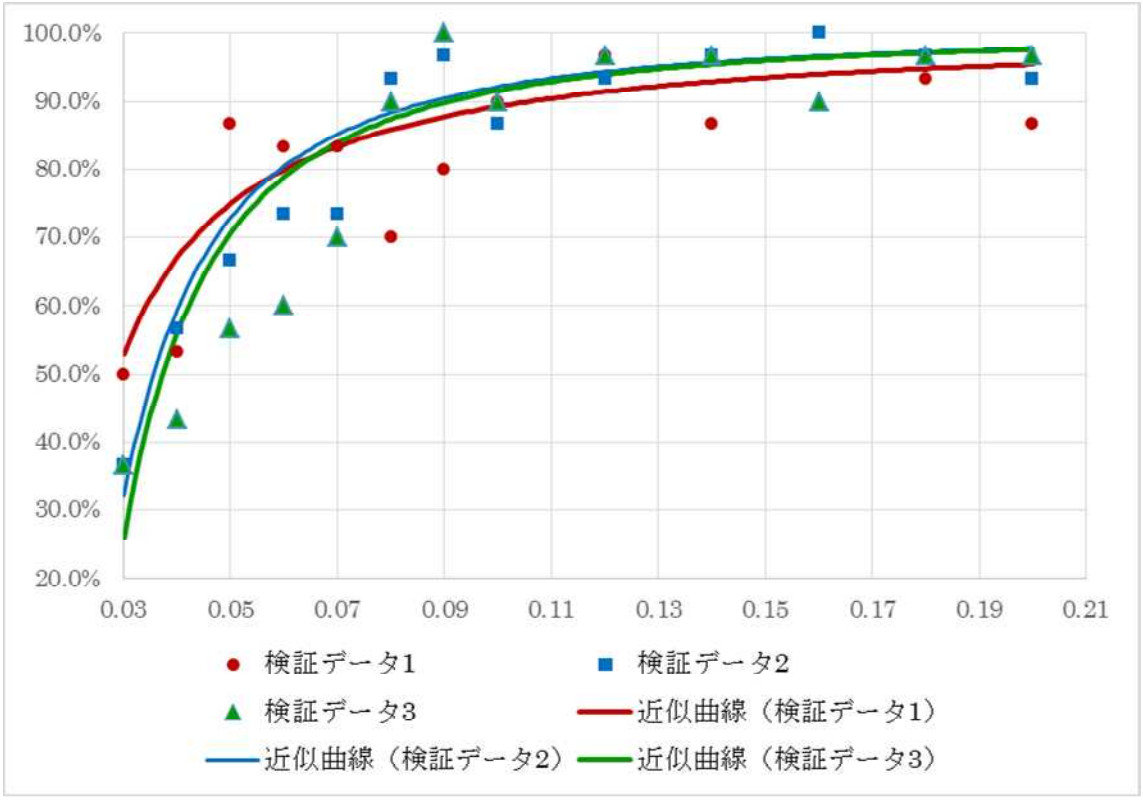


図 6.3-3 の検証データ１，２，３の近似曲線を比較すると、検証データ１の曲線は検証データ２、３の曲線と異なる傾向が見られた。検証データ１の曲線の形状は検証データ２，３の曲線と比べ、縦軸方向の変化の幅が狭く、曲線のカーブが緩やかな曲線形状となった。一方、検証データ２と３の近似曲線は文対応スコアの大小に従って、急激に近似曲線の正解率が変化する傾向が確認できた。つまり、検証データ２、３の対訳コーパスのほうが、文対応のスコア値で中日対訳コーパスの正解、不正解を判断する性能が良いことが分かった。

(b) 対訳コーパスの正解率（精度）と対訳コーパスの文数の関係

対訳コーパスはその利用用途により必要とされる対訳コーパスの正解率（以降、精度という）が異なる。例えば、対訳コーパスを統計用翻訳で使用する場合、高精度の対訳コーパスが求められる。一方、対訳コーパスを人手翻訳の参考情報として参照するなどの用途での精度は、統計用翻訳の用途ほど精度が高くなくとも対訳文数が多い方が有益である。そこで、本調査では対訳文の精度を１つに定めず、中日対訳コーパスの正解率を 95%以上、90%以上、85%以上、80%以上、75%以上、70%以上としたときに必要な文対応のスコア値を

推定して、そのスコア値をもとにして利用可能な中日対訳コーパスの文数を推定した。具体的には下記の様に推定した。

① 式 6.3-1 近似曲線式のスコア値 x を 0.0300～0.2000 まで 0.0001 刻みで計算してコーパスの正解率 y とスコア値 x の表を作成
② ①の表から y が所定の正解率 95%～70%に最も近くなるスコア値を読み取り
③ ②スコア値以上の各中日対訳コーパス数をカウント

こうして表 6.3-5 精度毎の推定文数を作成した。

表 6.3-5 精度毎の推定文数

対訳コーパスの精度	推定文数（文対応スコア）		
	検証データ１	検証データ２	検証データ３
精度 95%以上	222 万文 (0.1860)	1,869 万文 (0.1282)	1,510 万文 (0.1332)
精度 90%以上	1,760 万文 (0.1057)	3,943 万文 (0.0871)	3,645 万文 (0.0900)
精度 85%以上	3,260 万文 (0.076)	48,324 万文 (0.0695)	46,432 万文 (0.0725)
精度 80%以上	4,200 万文対 (0.0601)	5,250 万文 (0.0592)	5,126 万文 (0.0619)
精度 75%以上	4,782 万文 (0.0501)	5,485 万文 (0.0523)	5,395 万文 (0.0547)
精度 70%以上	5,131 万文 (0.0432)	5,634 万文 (0.0473)	5,575 万文 (0.0494)
0%以上（全件）	6,580 万文	6,584 万文	6,586 万文

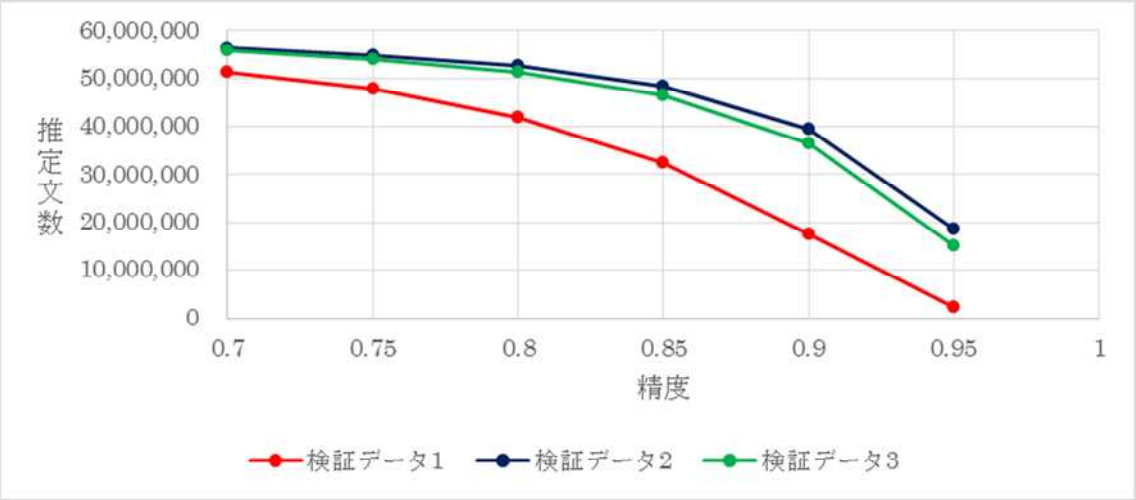
表 6.3-5 精度毎の文数（推定）は、精度 95%以上の検証データ１の推定文数の場合、近似曲線の式より 95%以上の精度を得るには文対応スコア 0.1860 以上であることが分かる。次に前述の表 6.3-3 所定の文対応スコアまでの文数と同様にカウントして文対応スコア 1～0.1860 までの検証データ１の文数が 222 万文であると推定した。

推定文数結果から、検証データ１と検証データ２とを比較すると精度 95%の文数が検証データ１ではおよそ 222 万文、検証データ２では 1,869 万文となり検証データ２は検証データ１に比べ 8 倍以上の文数が得られる推定結果となった。 90%, 85%も同様に検証２データの方が多くの文数が得られる結果となった。

検証データ２と検証データ３との推定文数の結果を比較すると精度 95%の文数が検証データ２ではおよそ 1,869 万文、検証データ３ではおよそ 1,510 万文となり、検証データ３がおよそ 360 万文少なくなる推定結果となった。

また、表 6.3-5 の結果を基に推定文数と精度の関係をグラフ化した。(図 6.3-4)このグラフを参照することで、必要とする対訳コーパスの件数から、利用可能な対訳コーパスの精度を推定することが可能となる。

図 6.3-4 推定文数と精度



以上の結果から、所定精度の文数は検証データ２が最も多く得られる結果となった。但し、前述の図 6.3-3 近似曲線を見て分かるように、検証データ２と検証データ３の近似曲線は非常に近いことが分かった。

今回利用したツールと同様に対訳辞書を用いた自動文アライメントツールに Champollion というツールがある。このツールに関する論文では中国語と英語の国連文書 1,461 文の対応付けにおいて、対訳辞書を 4,000 語から 58,000 語に増やしても精度(Precision/Recall)が 0.96～0.97 で頭打ちになることが報告されている[23]。この頭打ちとなる精度は今回検証した精度(Precision)とほぼ同じ値である。そのため、特許文献においては検証データ２の 97 万語程度で精度が頭打ちになっている可能性が考えられる。

より正確な推定結果を得るにはサンプル調査の文数を増やし、同様の調査を行うことが考えられるが、図 6.3-3 近似曲線のグラフにおいて文対応スコアが高い方は検証データ２と検証データ３の近似曲線がほぼ重なっており、正確な推定による効果はスコアが小さい場

23 https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2006-champollion-sentence-aligner.pdf

合の文アラインメントの正誤の判別精度向上の確認にとどまる可能性がある。なお、辞書の用語数が増加したことにより精度が落ちる例は表 6.3-2 を参照のこと。

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　対訳辞書候補データレイアウト説明
添付資料５．１　中日対訳辞書データレイアウト説明
添付資料５．２　４分野と IPC の関係
添付資料６．１　高精度化した中日対訳コーパスレイアウト説明
添付資料６．２　人手によるサンプルチェック方法
添付資料６．３　人手によるサンプルチェックデータ（抜粋）

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

family_numlist.tab.gz

（補足）ファイルは gnuzip で圧縮されている。

２．ファイル形式

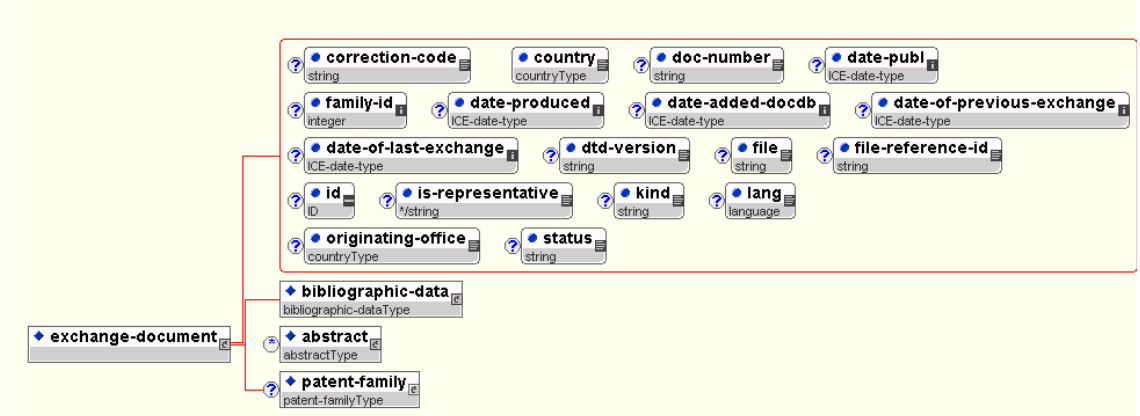
対応中国・日本公開特許公報番号リストは、Unix(Linux)で作成したテキストデータである。各行の項目はタブで区切られている。

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

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

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 IPC１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.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	CR+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	中国語	
2	日本語	
3	出現頻度	中日対訳コーパスの中国語、日本語の共起出現頻度が設定される。

４．サンプル

此	この	1724638
本发明	発明	1507145
我	本発明	1473880
本发明	本発明	1444398
通过	扱	1401614
通过	範	1401614
可以	よ	1347096
装	装置	1312248

添付資料５．１　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-h25fy.utx

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	頻度（全分野）中国語	C'freq-all-src' 固定
6	頻度（全分野）日本語	C'freq-all-tgt' 固定
7	頻度（全分野）中国語・日本語	C'freq-all-src-tgt' 固定
8	頻度（化学分野）中国語	C'freq-c-src' 固定
9	頻度（化学分野）日本語	C'freq-c-tgt' 固定
10	頻度（化学分野）中国語・日本語	C'freq-c-src-tgt' 固定
11	頻度（電気分野）中国語	C'freq-e-src' 固定
12	頻度（電気分野）日本語	C'freq-e-tgt' 固定
13	頻度（電気分野）中国語・日本語	C'freq-e-src-tgt' 固定
14	頻度（機械分野）中国語	C'freq-m-src' 固定
15	頻度（機械分野）日本語	C'freq-m-tgt' 固定
16	頻度（機械分野）中国語・日本語	C'freq-m-src-tgt' 固定
17	頻度（物理分野）中国語	C'freq-p-src' 固定
18	頻度（物理分野）日本語	C'freq-p-tgt' 固定
19	頻度（物理分野）中国語・日本語	C'freq-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	頻度（全分野）中国語	
6	頻度（全分野）日本語	
7	頻度（全分野）中国語・日本語	
8	頻度（化学分野）中国語	
9	頻度（化学分野）日本語	
10	頻度（化学分野）中国語・日本語	
11	頻度（電気分野）中国語	
12	頻度（電気分野）日本語	
13	頻度（電気分野）中国語・日本語	
14	頻度（機械分野）中国語	
15	頻度（機械分野）日本語	
16	頻度（機械分野）中国語・日本語	
17	頻度（物理分野）中国語	
18	頻度（物理分野）日本語	
19	頻度（物理分野）中国語・日本語	

５．サンプル[24]

#UTX 1.11; zh/ja; 2014-02-01; JPO										
#src	tgt	src:pos	tgt:pos	freq-all-src		freq-all-tgt		freq-all-src-tgt		
	freq-c-src	freq-c-tgt	freq-c-src-tgt		freq-e-src	freq-e-tgt	freq-e-src-tgt		freq-m-src	
	freq-m-tgt		freq-m-src-tgt		freq-p-src	freq-p-tgt	freq-p-src-tgt			
组合物	組成物	noun	noun	176846	180073	170343	147482	148526	142337	9210
	10114	8716	7295	7798	6845	12859	13635	12445		
端部	端部	noun	noun	157989	190932	144120	38678	53250	34139	41815
	46058	38201	51570	61617	47738	25926	30007	24042		
控制部	制御部	noun	noun	114633	133468	110810	15896	17003	15438	40023
	48781	38919	18672	20262	17887	40042	47422	38566		
反应混合物		反応混合物		noun	noun	111709	108845	105387	109382	106635
	103290	536	504	482	499	487	476	1292	1219	1139

[24] サンプルは紙面の制約から折り返し表示となっている。

添付資料５．２　４分野と IPC の関係

分野	IPC
C:化学	A01～24 C12、13 C12N A43～47 B68 A45D、A47J、L、 B26B、D05B、D06F A61～62 B01～09 C02 B21、22 B23K、B30 C01、03～06、30 C07 A01N、P、A61K、Q C40 C08～11、14 B22F C21～25 A41～42 D
M:機械	A63 B23～26,81 (B23K,B26B→C:化学) B82 B27～44 (B30→C:化学) B60～64 (B60L→E:電気、B60W→P:物理) B65B、C、D、 B67 B65F、G、H、B66 E01～03、21 E04～06 F01～04 F15～17 F22、23、28～42 F24～27
P:物理	G01、04、12、21 G02、03、09、10 G05～08 B60W G06F、G06Q G11
E:電気	F21、H01B、H、J、K、M、R、T H05 H01C、F、G、L、S、 H05K H01P、Q、H03、04 B60L、H02

添付資料６．１　高精度化した中日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

h25fy_jpo_cj_corpus_tit.txt.gz

h25fy_jpo_cj_corpus_abs.txt.gz

h25fy_jpo_cj_corpus_clm.txt.gz

h25fy_jpo_cj_corpus_des_xx.gz (xx は'01'から'07')

である。

（補足）ファイルは gnuzip 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=0.966666666666667
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.

４．レコードサンプル

0.134713835333333 ||| CNA101336677_JPA22009011234_des.txt ||| 1 ||| 0.350255965583756 ||| 58 ||| 60 ||| 0.966666666666667 ||| 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６．２　人手によるサンプルチェック方法

人手によるサンプルチェックは各検証データから文対応スコア毎に 30 文のサンプルデータを抽出し、チェックを行った。サンプルデータの構成はタイトル、要約、請求項、明細書別のデータ件数を踏まえ､明細書から作成した文を 26 文､請求項から作成した文を 2 文、要約、タイトルから作成した文を 1 文とした。

また、サンプルデータに使用するデータの文の長さは作成箇所毎最も頻出する文の長さ付近のデータを使用した。検証データの文字数を集計した結果を下記表に示す。

表　検証データの文字数集計

作成箇所	言語	検証データ１	検証データ２	検証データ３
		平均	平均	平均
		頻出	頻出	頻出
タイトル	日本語	20.1	20.1	20.1
		15.0	15.0	15.0
	中国語	14.9	14.9	14.9
		13.0	13.0	13.0
要約	日本語	113.3	113.2	113.2
		50.0	50.0	50.0
	中国語	83.1	83.0	83.0
		30.0	30.0	30.0
請求項	日本語	149.8	149.5	149.5
		58.0	58.0	58.0
	中国語	118.7	118.6	118.5
		48.0	48.0	48.0
明細書	日本語	82.3	82.2	82.2
		48.0	49.0	49.0
	中国語	66.0	66.0	66.0
		33.0	33.0	33.0

チェックは、サンプルデータの中国語文と日本語文が翻訳関係であるものを正解とした。翻訳関係は中国語文と日本語文は用語単位で対応することを条件とした。但し、以下に示す中国語文と日本語文の表記の差は不一致の対象から除外することとした。

例外1. 中国語”一种”の扱い

中国特許公報上には” 一种～”といった表現が頻出する。日本語での意味は”１つの”となるが、日本の公報では表現されない場合がほとんどである。

よって、今回のサンプルチェックでは” 一种”に対応する日本語の有無は不問とした。

【例】
ナフタレン系減水剤の製造方法
一种萘系减水剂的制备方法

例外2. 図番号の表記相違の扱い

公報には構成要素に番号などを付けて表現することがある。中国の公報と日本の公報で同じ番号でも片方が丸括弧付きで表現し、他方が括弧なしで表現される場合がある。

今回のサンプルチェックでは、構成要素に付されている番号の記号の表記の違いは不問とした。

【例】
低圧蒸気タービン１０のロータ軸２６で、外部車室１２とグランド部３２との間をシールするシール機構４０Ａを設ける。
低压汽轮机（10）的转子轴（26）上，设有对外部机室（12）和密封压盖部（32）之间进行密封的密封机构（40A）。

例外3. 請求項番号の扱い

日本語の公報は XML の構造上、請求項の番号はタグ内の属性で表現されている。一方中国語公報の XML では、タグの外にテキストデータの一部として請求項番号が表現されている。このため、請求項の対訳コーパスでは中国語文の先頭にある請求項番号の有無は不問とした。

【例】
測定される負荷電ポリマーが、ＨＰＳ－Ｉ、ＡＥＣ、及びＡＰＥＳからなる群から選択される、請求項１２記載の薄膜センサー。
20.权利要求 12 的膜传感器，其中所述带负电荷的待测聚合物选自 HPS-I、AEC 和 APES。

例外4. 文区切り記号の扱い

中日自動文アライメントツールは分単位の比較を行うため、入力データ文の単位に分割する前処理を行う。そして中日自動文アライメントツールは、この前処理の後に文対応を行う。日本語文と中国語文のデータでは前処理の文分割処理が異なるため、下記例のように日本語１文に対して中国語が２文で判断されることがある。

このような分対応のデータのとき、中日自動文アライメントツールは文と文の間に記号"

/// "を挿入する仕様となっている。今回の調査では、この文区切り記号を比較対象とはせずに評価を行った。

【例】
［工程（ｉｉ）：乾燥工程］ 工程（ｉｉ）は、上記工程（ｉ）で得られた原料スラリーを噴霧乾燥して、乾燥粉体を得る工程である。
[工序(ii)：干燥工序] /// 工序(ii)是喷雾干燥上述工序(i)中获得的原料浆液来获得干燥粉体的工序。

例外5. 不自然な文区切りの扱い

下記例のように、閉じ括弧からはじまるような、文の途中から抽出された文対応であっても、その内容が対応している場合は正しいものと判断した。

【例】
) 上記の表から理解される如く、ドライ対物レンズを使用した場合でも、発光粒子濃度に対応して発光粒子の光信号の数が増大することが確かめられた。
) /// 表 1 /// 如从上表可以理解到的，已经确认的是，即使是在使用干式物镜时，发光粒子的光信号的数量也与发光粒子浓度一起增加。

添付資料６．３　人手サンプルチェックデータ（抜粋）

人手サンプルチェックデータは、作成元、文対応スコア、IPC、チェック結果、日本語文、中国語文から構成される。

項番	項目名	内容
1	作成元	“tit” タイトル部分から作成したデータ “abs” 要約部分から作成したデータ “clm” 請求項部分から作成したデータ “des” 明細書部分から作成したデータ
2	文対応スコア	文アライメントツールが算出した文対応度合いを示す値
3	IPC	作成元の文献に付与されている IPC
4	チェック結果	“1” 中国語文と日本語文が対応している。 “0” 上記以外
5	日本語文	サンプルチェック対象の日本語文
6	中国語文	サンプルチェック対象の中国語文

（空白）

作成元	文対応スコア	IPC	チェック結果	日本語文	中国語文
tit	0.033058	G06F17/30	0	適応検索のための方法および技法	自适应搜索结果用户界面
abs	0.030205	G06T13/00	1	ある実施形態では、シーングラフ・パラメータを制御するためのユーザー・インターフェースが生成される。	在一个实施例中，产生用于控制场景图参数的用户界面。
clm	0.030087	C09J5/00	1	前記接着剤（3）によって接続される前記2つの物体（1，2）の接触面が、拡散防止されている、請求項1記載の結合手段。	2.根据权利要求 1 的紧固件，其特征在于，物体(1、2)与粘合剂(3)连接的界面是防扩散的。
clm	0.03006	G05B23/02	1	前記第2通信リンクは無線リンクであり、前記第2プロセス制御プロトコルは無線通信をサポートする、請求項42に記載の方法。	44.根据权利要求 42 所述的方法，其特征在于，所述第二通信链路是无线链路，并且其中所述第二过程控制协议支持无线通信。
des	0.030016	G09G3/34	1	これらのピークは、以下で説明するように図8B－8Dにおいて明らかである。	这些峰值在如下文所论述的图 8B-8D 中显而易见。
des	0.030009	G04B19/253	0	図5は、実施形態の電子時計のCPUによる日付表示処理の制御手順を示すフローチャートである。	所述日期显示装置中，使用所述旋转板和所述固定板来同时显示表示所述连续的预定数个日期的预定数个数字，
des	0.030012	G06F1/16	1	例えば、剛性でありかつ撓みに耐える形状を有する埋め込みバッテリは、耐荷重性であると考えられる。	例如，具有刚硬且抗挠曲的形状的嵌入式电池可被视为承载构件。
des	0.030008	H01J37/244	0	本発明によって検出され修正することができるその他のアーチファクトは、ドリフト（drift）である。	可以通过本发明检测及校正的另一种伪像是漂移。

des	0.030013	A61B19/00	0	この第１の連結アセンブリは、上記連結ベースによって旋回可能に支持され、そして第１の外側ハウジンクを有する。	第一联动装置被联动装置底座转动支撑并且具有第一外壳。
des	0.03001	G02B5/30	1	より詳しくは、光学色素は、配列装置との直接的および／または間接的な相互作用によって配列されることができる。	更具体地，该光学染料可通过与该配向机构直接和/或间接相互作用来配向。
des	0.030001	H04M1/00	0	色が違っても良い。 /// 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。	在一些实施例中，设备 110 的总体形状可以与所示出的不同。
des	0.030004	H04W28/12	1	相互パイロット結合推定を判断することは、測定受信電力とアクセス端末ロケーション情報とを使用することをさらに含む。	确定导频交叉耦合估计还包括使用测得的收到功率和接入终端定位信息。
des	0.030012	G01J1/02	1	検知対象の発熱が検知され、そのレベルが比較的深刻でない場合には、電子機器3の駆動を省電力モードとする制御が行われる。	当检测到检测对象的发热并且其水平不是非常严重时，电子设备 3 被控制为进入节能模式。
des	0.030008	A61K8/44	0	1H—NMR(CD3OD): δ 4．02（2H、m）、4．74（1H、m）、7．81（2H、d）、8．07（2H、d）.	1H-NMR (CD3OD): δ 4.02 (2H, m), 4.74 (1H, m), 7.81 (2H, d), 8.07 (2H, d)。 ///
des	0.030015	H04N13/04	0	このデータアイランド区間では、補助データのうち、制御に関係しないデータである、例えば、音声データのパケット等が伝送される。	在数据岛时间段中，例如发送与控制无关的辅助数据（如，音频数据分组）。

des	0.030015	G02F1/1335	0	本発明では、2つのロールの材質は金属であることが好ましく、より好ましくはステンレスであり、表面をメッキ処理されたロールも好ましい。	另外，除了将所供给的热塑性树脂组合物的熔融物在 2 个辊的表面上连续地夹压以成形为薄膜状的现有方法以外，还优选对辊间施加 5～500MPa 的压力。 /// 更优选的压力为 20～300MPa，进一步优选为 25～200MPa，特别优选为 30～150MPa。 /// 本发明中，2 个辊的材质优选为金属，更优选为不锈钢，还优选表面经过镀覆处理的辊。
des	0.030013	H04N13/00	0	上記オーバーラップしている視野は、上記第1の画像の第1の端部と、上記第2の画像の上記第1の端部とは反対側の第2の端部に位置してもよい。	重叠的区域可以分别在第一和第二图像的对端(opposite end)处。
des	0.03	G06F9/44	1	アクセス層１００２は1つまたは複数のコンポーネントを備え、それらのコンポーネントを介してディレクトリ・スタックを公開することができる。	访问层 1002 包括一个或多个组件，可通过这些组件展示目录栈。
des	0.030006	G03G9/09	0	＜3．粉砕＞ クラッシャーで0．1～1．0mmに粉砕した後に、水を加え湿式ボールミルで0．1～0．5μmに微粉砕し、フェライトスラリーを得た。	<3.粉碎> /// 将所述铁氧体用破碎机粉碎成各自具有粒径 0.1 至 1.0mm 的颗粒。 /// 其后，将水加入颗粒中，并将得到的颗粒用湿球磨机细碎成各自具有粒径为 0.1 至 0.5μm 的颗粒，从而得到铁氧体浆料。 /// <4.造粒>
des	0.030009	G01N21/78	1	好ましくは、第1および第2の測定チャネルの少なくとも一方は、アナライトと相互作用するための少なくとも1つの中間試薬を更に含むように構成されている。	优选地，第一和第二测量通道中的至少一个被配置为进一步至少包括用于与被分析物发生反应的中间反应物。

des	0.03001	H04B1/7103	0	このように、所定のスロットの間、複数のタグがアクセスポイントに対して送信をおこなう確率に基づく送信スロットの所定数を待つように、タグを構成することができる。	因此，可以将标签配置为基于在给定的时隙期间多个标签会向所述接入点发送的可能性，等待预先定义数目的传输时隙。
des	0.030001	A61K31/713	0	オリゴヌクレオチドは、非経口投与のために、リポソームもしくはポリエチレングリコールで修飾されたリポソーム中へ組み込んでも、またはカチオン性脂質と混合してもよい。	可将寡核苷酸整合入脂质体或用聚乙二醇修饰的脂质体中或与阳离子脂质混合用于肠胃外施用。
des	0.030016	G11C15/04	0	そこで、全て一致と等価な参照セル、すなわち、磁化状態が平行の時の抵抗を有するトランジスタのみで構成されたセルを設けておき、これらの電流を比較することで当たり判定ができる。	因此，命中率(hitting)可基于这些电流的比较来确定，所述比较比如通过提供等效于全部匹配的参考单元。 /// 这样的参考单元可由具有平行磁化电阻的晶体管形成。
des	0.030011	C08J5/18	1	対照的に、重合体はその懸垂基の全てが連鎖の同一側に配置されている時には「アイソタクチック」でありそしてその懸垂基が連鎖の反対側に交互にある時には「シンジオタクチック」である。	与此不同，当聚合物的所有侧基都排列在链的同一侧时该聚合物是“等规的”，而当聚合物的侧基交替排列在链的相对侧时该聚合物是“间规的”。
des	0.030008	H04N7/30	0	さらに詳細に説明するように、拡張レイヤの係数ベクトルのそれぞれの非ゼロ係数は任意の後続の係数、すなわち現在復号されている非ゼロ係数に続く任意の係数の知識なしに符号化され得る。	如将进一步详细描述，可在不知晓任何后续系数(即，当前正被译码的非零系数之后的任何系数)的情况下编码加强层的系数向量的每一非零系数。

des	0.030016	B23K9/095	1	前記溶接システム１４は、溶接電圧及び電流を生成するための装置、前記溶接電圧及び電流を制御するための溶接制御装置及び前記溶接電圧及び電流をモニターするためのモニターリングシステムを含む。	所述焊接系统14包括用于产生焊接电流和焊接电压的焊接设备、用于控制所述焊接电流和所述焊接电压的焊接控制系统，以及用于监控所述焊接电流和所述焊接电压的监控系统。
des	0.030007	H04W72/04	1	既存の無線通信システムにおける基地局は、一つの周波数チャネルを支援する一つのPHYエンティティ２２０を有し、一つのPHYエンティティ２２０を制御する一つのMAC制御器２１０を備えることができる。	在传统的无线通信系统中使用的基站(BS)包括支持单个频率信道的单个 PHY 实体 220，并且可以为基站(BS)提供用于控制该单个 PHY 实体 220 的单个 MAC 控制器。
des	0.030015	G06F3/048	1	一実施形態では、一部のウィジェットは、サーバなどのリモート情報ソースと対話して、情報を提供することができ、例えば、気象モジュールは、リモート・サーバからライブの気象データを取得することができる。	在一个实施方式中，某些控件可以与诸如服务器这样的用于提供信息的远程信息源交互；例如，天气模块可以从远程服务器取回直播天气数据。
des	0.030016	H04W72/04	1	メモリユニットは、プロセッサ内に又はプロセッサの外部に実装することができ、プロセッサの外部に実装する場合は、当業において知られる様々な方法で通信可能な形でプロセッサに結合させることが可能である。	存储器单元可实施于处理器内部或处理器外部，在此情况下，其可经由此项技术中已知的多种方法以通信方式耦合到处理器。

des	0.030007	G03G9/087	0	－第一の結着樹脂（A）－ 本発明のトナーに含まれる前記第一の結着樹脂（A）は、有機溶媒中で活性水素基含有化合物（A1）と、該化合物（A1）と反応可能な部位を有する樹脂（A2）とを反応させてなる。	调色剂 /// 本发明的调色剂含有粘合剂树脂、着色剂和脱模剂，所述粘合剂树脂含有第一粘合剂树脂 A 和第二粘合剂树脂 B。 /// 所述第一粘合剂树脂 A 是通过使具有活性氢基团的化合物 A1 与具有与所述化合物 A1 反应的部分的树脂 A2 在有机溶剂中反应制造的。 /// 所述树脂 A2 是通过使在其主链中具有多羟基羧酸骨架的非结晶聚酯树脂“a”与具有与具有活性氢基团的化合物 A1 反应的部分的化合物反应形成的。