令和元年度

新語対応による機械翻訳精度向上に関する調査事業

調査報告書

２０２０年3月

内容

１．本報告書の概要

特許庁、工業所有権情報・研修館は 2019 年 5 月、特許情報プラットフォーム（J-Plat Pat）の刷新に伴い、機械翻訳プラットフォーム（MTP）による日本特許庁の審査書類や日本公報等の日英翻訳の提供を開始した。MTP においては、最新の AI 技術を活用したニューラル機械翻訳による翻訳エンジンが用いられており、従来の機械翻訳と比較して格段に翻訳品質が向上している。

しかしながら、技術革新等により新たに生み出される技術用語（新語）など、MTP の学習データに含まれていない単語（未知語）を含む文章については、いかに翻訳精度の高いエンジンであったとしても、正確な機械翻訳を行うことは困難である。なおかつ、最新の審査書類や公報ほど外部ユーザーからのアクセス性向上が強く要望されるところ、日々生み出され増え続ける新語の辞書登録作業が行われなければ、より多くの新語を含むであろう最新の審査書類や公報ほど、その翻訳品質が低下することになりかねない。

そこで、本事業では、機械翻訳プラットフォーム（MTP）の翻訳品質を維持向上させることを目的として、新語に代表される未知語の英訳辞書を作成するとともに、未知語の辞書作成の効率化を図ることを目的として、未知語の出現傾向や辞書登録すべき未知語を効率的に選定する方法について調査を行った。

本報告書は、（１）翻訳対象語の特許公報中での出現状況や、翻訳対象語と特許庁から貸与された対訳辞書データとの関係を調査し、翻訳対象語を分類した結果（翻訳対象語の用語自体の調査）、（２）翻訳対象語が出現した文献の書誌情報を元に翻訳対象語の出現傾向を調査した結果（翻訳対象語が発見された公報の調査）、（３）これら調査結果を踏まえ、今後の効率的な新語対応のあり方を検討した結果（今後の効率的な新語対応のあり方についての検討）を報告するものである。

２．新語対訳辞書の作成

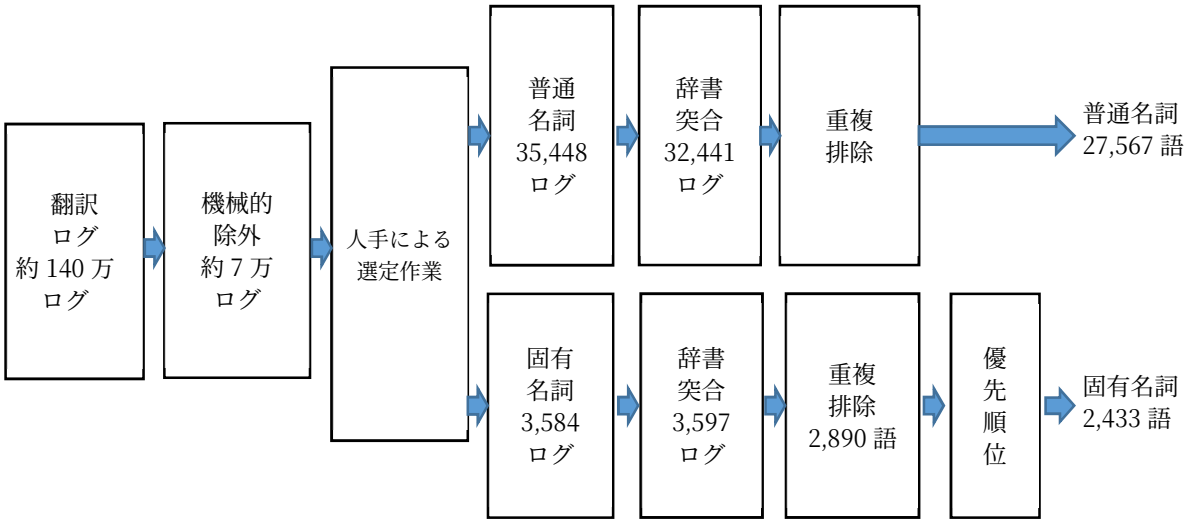
本事業における新語対訳辞書は、特許庁から貸与された翻訳ログから 30,000 語の翻訳対象語を選定し、これを日英翻訳することにより作成した。翻訳ログとは、MTP に対して特許公報等の一部の文章（「翻訳対象テキスト」）の翻訳要求がなされた際、その文章中に翻訳できない単語（「未知語」）が検出された場合に、関連する情報（翻訳対象テキストや未知語、翻訳対象テキストが記載されている公報の識別番号等）が対応付けられて作成されるログのことである（図表１参照）。以下では、翻訳ログから新語対訳辞書を作成する具体的な手順について説明する。

識別番号	翻訳対象テキスト	未知語
JPA 419528755	単離した新生児ラットの心筋細胞を、矩形に設計したＰＩＰＡＡｍグラフティングポリスチレン細胞培養皿上で培養し、そして温度を２０℃まで低下させることによって矩形の細胞シートとして剥離させた。	グラフティングポリスチレン
JPA 417220231	アリーロキシ基としては、フェノキシなどが挙げられ、スルホン酸含有基（-ＳO3Ra）としては、メタンスルホナト、ｐ-トルエンスルホナト、トリフルオロメタンスルホナト、ｐ-クロルベンゼンスルホナトなどが挙げられる。	クロルベンゼンスルホナト
JPA 426073107	これにより、VEGF 含浸ゼラチンパターンニング基材から放出された VEGF により、HUVEC が培養部位（マトリゲル層）中をＶＥＧＦの方向（下部）へ移動した。	ゼラチンパターンニング

図表１ 翻訳ログ（イメージ）

２．１　翻訳対象語の選定

翻訳ログにおける未知語には、様々な要因により、辞書登録すべき単語とはいえないものが多数含まれてしまっている。そこで、本事業では、まず、機械的に又は人手でこのような未知語に対応する翻訳ログを除外したり、単語の区切りが不適切と思われる未知語を人手で修補したりすることで、辞書登録を行う単語として適切なものを選定する作業を行った。図表２はこの作業の概略を示すものである。以下、それぞれの具体的な手順について詳述する。



図表 2 翻訳対象語選定手順

２．１．１　翻訳ログの機械的除外

以下の条件に該当する翻訳ログは、その未知語が辞書登録すべき単語とはいえない翻訳ログであると判断し、これを機械的に除外することとした。

（１）電子出願開始以前（平成４年まで）の公報に係るもの

電子出願開始以前の公報のテキストデータは、そのほとんどが自動文字認識によって作成されたものであり、自動文字認識の誤りに起因して未知語として検出されたものである可能性が高いため、このような翻訳ログを除外対象とした。図表 3 はこのような翻訳ログの例である。

翻訳対象テキスト	未知語
更にスケリソルブＢ中の７５%酢酸エチル５ｔと１００%酢酸エチル６ｔで相つぃで溶離をつづけ、対応する溶離液を２００ＴＬｌづつのフラノン	ぃで

図表 3 電子出願開始以前の公報に係る翻訳ログ例

（２）翻訳対象テキストを「みんなの自動翻訳@Textra®[1]」の翻訳エンジンを用いて機械翻訳を行った場合に未知語が検出されないもの

特許庁から貸与された翻訳ログにおいては、適切な形態素解析が行われなかったにより、誤った区切られ方をした単語が未知語として検出されている場合があると

[1] https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

ころ、形態素解析の精度がより高いことが確認されている「みんなの自動翻訳@Textra®」のエンジンにより未知語が検出されなかったものは、こうした形態素解析の誤りに起因して検出された未知語であるといえるため、このような翻訳ログを除外対象とした。図表 4 における未知語「クロルベンゼンスルホナト」は、機械翻訳結果内に未知語が検出されたため、翻訳ログから除外とならない。一方、未知語「ディープラーニング」は、機械翻訳結果内に未知語が検出されないため、翻訳ログから除外となる。

翻訳対象テキスト	機械翻訳結果	未知語
アリーロキシ基としては、フェノキシなどが挙げられ、スルホン酸含有基（-ＳＯ3Ra）としては、メタンスルホナト、ｐ-トルエンスルホナト、トリフルオロメタンスルホナト、ｐ-クロルベンゼンスルホナトなどが挙げられる。	Examples of the aryloxy group include phenoxy and the like, and examples of the sulfonic acid-containing group (-SO3Ra) include methane sulfonato, p-toluene sulfonato, trifluoromethane sulfonato, and p- クロルベンゼンスルホナト.	クロルベンゼンスルホナト
また、記録部 16 に記録される学習データ 17 は、学習済みのニューラルネットワークやディープラーニングのフレームワークと数値を含む。	The learning data 17 recorded in the recording unit 16 includes a learned neural network and a framework of a deep learning and numerical values.	ディープラーニング

図表 4 機械翻訳を行った場合に未知語が検出されない例

（３）以下のいずれかに該当するもの

① 文字数が１文字以下の未知語

② ひらがな、記号及び英数字のみからなる未知語

③ 特許庁から貸与された、作成済み対訳辞書データと一致するもの

①、②については、意味のある単語とはいえないため、③については、既に辞書登録済みの単語を再び登録する必要がないため、このような翻訳ログを除外対象とした。図表 5 はこのような翻訳ログの例である。

翻訳対象テキスト	未知語	備考
×:破断の発生若しくは成形体に大きなカールが発生したもの。	く	文字数が１文字以下の未知語
１０－”%～１．２２ｘ　１　（I’Ｔｏｒ　（出典はＳｉＯに同じ）でＳｉＯが格段に有利であることが分かる。	%～"	らがな、記号及び英数字のみからなる未知語
ケーブル送り込み装置及び剥線装置	剥線	作成済み対訳辞書データと一致

図表５ 翻訳ログの機械的除外例

２．１．２　人手による翻訳対象語選定作業

２．１．２．１　翻訳ログの人手による除外

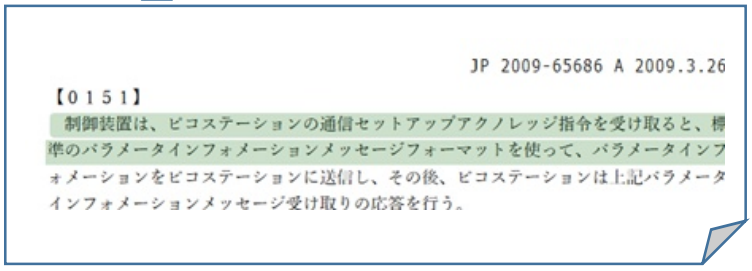
翻訳ログを人手で確認して以下の条件に該当すると判断した翻訳ログは、その未知語が辞書登録すべき単語とはいえない翻訳ログであると判断し、これを除外することとした。

（１）単語の途中で改行が挿入されることで本来１語として検出されるべき単語の一部分が未知語として切り出されたもの

公報のテキストデータ原文においては、文の途中で改行が挿入されていることがあり、これに起因して本来一語として扱われるべき単語の一部が切り出されて未知語として検出されることがある（図表６参照）。そこで、翻訳対象テキストの文頭又は文末で未知語が検出されているものについては公報原文を確認し、単語の途中で改行が挿入されているものと認められた翻訳ログは除外対象とした。図表７はこのような翻訳ログの例である。

翻訳対象テキスト	未知語
制御装置は、ピコステーションの通信セットアップアクノレッジ指令を受け取ると、標準のパラメータインフォメーションメッセージフォーマットを使って、パラメータインフ	パラメータインフ

公報原文を確認すると、本来１語となるべき「パラメータインフォメーション」の途中に改行が挿入されているものと認められるため、この翻訳ログを除外する。



図表６ 文頭・文末の未知語の人手確認例

翻訳対象テキスト	未知語
ードウェア又はソフトウェアモジュールを含む専用又は汎用コンピュータの使用を含み得る。	ードウェア
また、アルコキシシランで変性された重合体H、Ｉを用いた比較例３、４、及び７、８のゴム組成物は、カーボンブラッ	カーボンブラッ

図表７ 文頭又は文末で検出された未知語例

（２）以下のいずれかに該当するもの

① 名詞以外の単語

② 明らかな誤記

①については、新語対応においては名詞を優先的に辞書登録すべきと想定しているため、②については、いずれも辞書登録を行う意義のある単語とはいえないため、このような翻訳ログを除外対象とした。図表８はこのような翻訳ログの例である。

翻訳対象テキスト	未知語	備考
また、表示した候補からの選択指定も手入力による指定もなかった場合は、観測ユーザのみ（つまり仮想オブドェクトを表示する当人だけ）が共有ユーザに指定される。	オブドェクト	「オブジェクト」の誤記
また、夜のステージ画面を表示している状態から、表示画面を右にスクロールして、墓地が拡大されてキャラクタの執事が「オリャー」という叫び声と「ティロロローン」という効果音と共に、堀り起こされた画面中央の墓穴から主人公のキャラクタを登場させることにより、遊技者にとって激アツなプレミア演出としてもよい（図５２（Ｃ）参照)。	オリャー	名詞以外

図表８ 人手確認による翻訳ログ除外例

２．１．２．２　未知語の修補

項番２．１．１（２）でも言及したように、未知語は形態素解析の誤りを原因として検出されることがあるため、このような未知語については、翻訳対象テキストにおける未知語の前後の箇所を確認し、本来１語とされるべき単語に修補を行った。図表 9 は未知語の修補結果の例である。

翻訳対象テキスト	未知語	未知語の修補後	備考
「ＲＴ０」から「ＲＴ１」には、ＲＴ１図柄（いわゆるＡＴ役コボシ）を停止表示させた時に移行する。	コボシ	ＡＴ役コボシ	未知語「コボシ」は、それ自体では意味をなさない単語であるため、「ＡＴ役コボシ」と修補した。

図表 9 未知語の修補例

２．１．２．３　普通名詞・固有名詞の分類

新語対応においては、人名、法人名、製品名等の固有名詞でない普通名詞を優先的に辞書登録すべきと想定しており、後述の項番２．１．５における 30,000 語の選定の際に普通名詞を優先的に選定するため、修補後の各単語について普通名詞、固有名詞の分類分けを行った。

２．１．３　修補作後の未知語と作成済み対訳辞書データの突合

修補後の未知語について、作成済み対訳辞書データと一致するものは、項番２．１．１（３）と同様の理由により辞書登録は不要であるから、改めてこれに対応する翻訳ログを除外した。図表 10 はこのような翻訳ログの例である。

翻訳対象テキスト	未知語	備考
可視光の表示色としては、一般的には、発光表示装置における加色混法であればＲＧＢなどの原色系の３波長帯又は電子ペーパなどの反射型表示装置における減色混法であればシアン、マゼンダ、イエローなどの補色系の３波長帯などから選択することができる。	混法	修補後の用語「加色混法」が作成済み対訳辞書データと一致

図表 10 修補後の未知語と作成済対訳辞書データとの突合例

２．１．４　未知語の重複排除

項番２．１．３までの作業で抽出した翻訳ログの中には、同じ単語に関するものが含まれているため、重複排除を行った。図表 11 はこのような翻訳ログの例である。

翻訳対象テキスト	未知語
また、上述の生理用ナプキン１の実施形態において、防漏溝７は、リング状の全周防漏溝を形成しているが、長手方向の両側部に沿って延びる一対の溝であれば良い。	防漏溝
これにより、防漏溝の深さが浅くなり、該防漏溝を形成する圧着部における剥離が生じやすくなる場合がある。	防漏溝

図表 11 同じ単語に関する翻訳ログの例

２．１．５　翻訳対象語 30,000 語の選定

項番２．１．４で選定した単語の中から 30,000 語の翻訳対象語を決定した。本事業では、普通名詞 27,567 語をすべて翻訳対象語とするとともに、翻訳ログの件数が多い順に固有名詞 2,433 語を翻訳対象語とした（図表 12 参照)。翻訳対象語の内訳は図表 13 のとおりである。

翻訳ログ件数	固有名詞
9	ラピゾール
9	医薬審発
8	オプトロニックラボラトリーズ社

図表 12 選定した固有名詞例

語数：30,000 語(重複なし)、 35,567 語(重複あり)
語数（普通名詞）：27,567 語、32,441 語(重複あり)
語数（固有名詞）： 2,433 語、3,126 語(重複あり)
翻訳対象を構成する文字種
　カタカナ用語（1 文字以上のカタカナ文字を含む用語）　　21,189 語
　上記以外の用語　 8,811 語

図表 13 翻訳対象語の件数

２．２　翻訳対象語の翻訳

30,000 語の翻訳対象語の日英翻訳を行い、新語対訳辞書を作成した。なお、１つの翻訳対象語に対して複数の翻訳対象テキストが存在し、複数の訳語が得られるときは、より多くの翻訳対象テキストの文脈において適切な訳語になるよう翻訳を行った。図表 14 は翻訳結果の例である。

翻訳対象語	翻訳結果	備考
（フルオロ）アルキルホスナート	(fluoro)alkylphosphonate	普通名詞
ＬＥＤ熱伝導座	LED heat conductive seat	普通名詞
近畿圏	Kinki Area	固有名詞
グラスロンミルドファイバー	Glasslon milled fiber	固有名詞（「グラスロンミルドファイバー」は商品名）

図表 14 翻訳対象語の翻訳例

３．翻訳対象語の調査分析

３．１　翻訳対象語の用語自体の調査

本事業で選定した翻訳対象語にはどのような傾向があるのかを調査した。本事業では、翻訳対象語の特徴として想定される以下の４つの分類について調査を行った。なお、以下では、固有名詞は調査対象とはせず、普通名詞のみを調査対象としている。

・ 辞書未登録の新しい技術用語であるため未知語として検出されたもの（新語）
・ 辞書登録済みの単語に別の単語が組み合わさることで未知語として検出されたもの（複合語）
・ 辞書登録済みの単語に対して大文字・小文字等の表記ゆれがあるため未知語として検出されたもの（表記ゆれ）
・ １語に対して複数の英訳が考えられるもの（複数訳語）

３．１．１　新語

３．１．１．１　定義

本調査では、下記の①、②の両方を満たす翻訳対象語を「新語」と定義した。

① （2006 年～2019 年の翻訳対象語記載公報数）／（1993 年～2005 年の翻訳対象語記載公報数）が 42 以上、又は、1993 年～2005 年の翻訳対象語記載公報数が０であること。
② 1993 年～2019 年の翻訳対象語記載公報数が 50 以上であること。

※ 「翻訳対象語記載公報数」とは、その翻訳対象語が１箇所以上記載された特許・実用新案公報の数をいう。
※ 条件①ついて、特許庁から貸与された翻訳ログは 1993 年～2019 年発行の公報に関するものであることから、中間地点の 2006 年を基準とし、前半に対して後半の出現頻度が著しく増加している、もしくは後半のみに出現している翻訳対象語を近年になって使用されるようになった新しい技術用語であると判断することとした。
※ 条件②について、出現頻度がそもそも少なすぎる翻訳対象語については、一般的に用いられる用語とはいえず、調査を行う意義が乏しいと考えられるため、一定の使用頻度以上の翻訳対象語についてのみ調査を行うこととした。

３．１．１．２　翻訳対象語記載公報数の集計

（１）集計対象とした公報の種類及び公報発行日の範囲について

翻訳対象語記載公報数の集計は、好ましくは、ある翻訳対象語が初めて公報に

記載された回数を集計すべきであるから、原則として、1993 年 1 月 1 日から 2019 年 12 月 31 日に発行された特許・実用新案公報であって、その翻訳対象語が記載された最先のものの数を集計することとしている。しかしながら、使用データベース（Japio 世界特許情報全文検索サービス[2]（Japio-GPG/FX））の制約上、以下の場合には、最先公知文献による集計が行われない点に留意されたい。

- 1992 年までの集計期間外に公開公報が発行され、その後 1993 年 1 月 1 日から 1996 年 3 月 29 日の集計期間内に公告公報が発行された出願については、後に発行された公告公報が集計対象となる。
- PCT 出願がなされたものについては、先に発行された国際公開公報ではなく、後に発行された公表公報・再公表公報が集計対象となる。

（2）公報内の検索対象箇所について

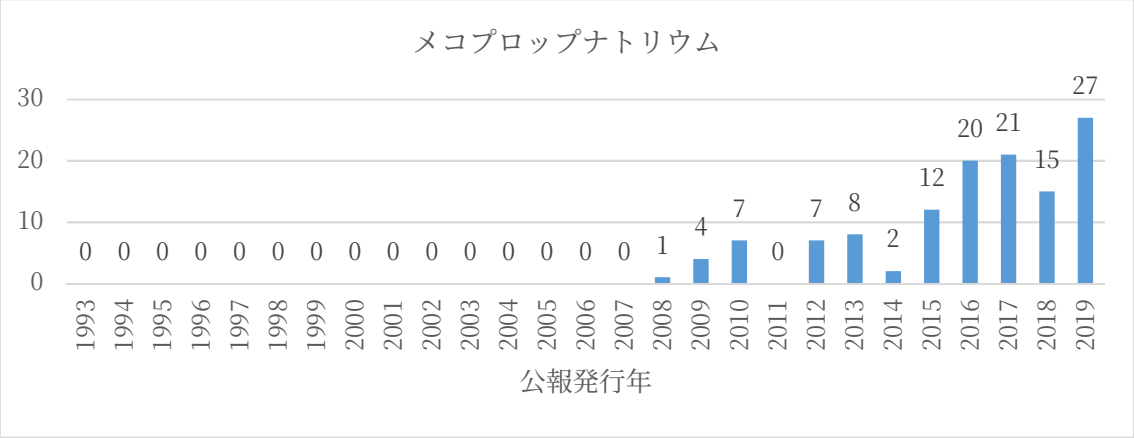
検索対象は書誌事項部分を含めず、発明の名称、要約、請求項、詳細な説明を検索対象とした。

（3）検索方式について

検索方式は、異表記展開を行わず、翻訳対象語そのものを含む文献数を集計した。

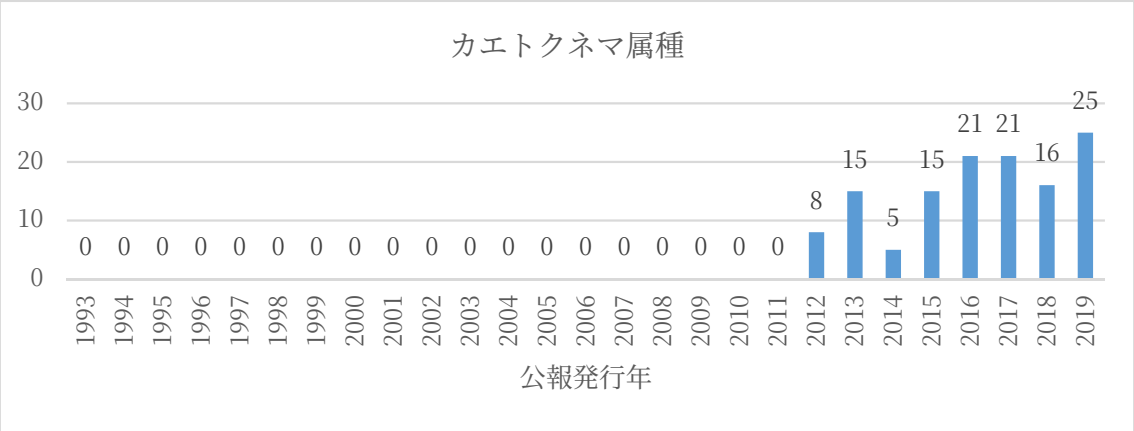
３．１．１．３　「新語」の具体例

図表 15 に「新語」に分類された翻訳対象語の翻訳対象語記載公報数の遷移を示す。また、「添付資料 新語例」にも「新語」に分類された翻訳対象語の翻訳対象語記載公報数の遷移を示す。



図表 15-1 新語の例

[2] https://gpgfx.japio.or.jp/



図表 15-2 新語の例

図表 15-3 新語の例

３．１．２　複合語

３．１．２．１　定義

本調査では、その単語の一部分が作成済み対訳辞書データの日本語の単語と一致する翻訳対象語を「複合語」と定義した。

３．１．２．２　「複合語」の具体例

図表 16 に「複合語」に分類された翻訳対象語の例を示す。

翻訳対象語 (下線は作成済み対訳辞書データとの一致部分)	備考
（メタ）アクリルオキシアルキルプロペナール	化合物名の一部が一致
アリルジグリシジルイソシアヌレート	化合物名の一部が一致

エピデルモフィトン・フロッコーサム	生物名「エピデルモフィトン」が一致

図表 16 複合語の例

３．１．３　表記ゆれ

３．１．３．１　定義

本調査では、各翻訳対象語と作成済み対訳辞書データの日本語の単語との双方に対して下記の表記統一を行った場合に、その単語の全部又は一部分が作成済み対訳辞書データの日本語の単語と一致する翻訳対象語を「表記ゆれ」と定義した。

① アルファベット、数字、記号を全て全角文字に統一

② 「 - 」(短いダッシュ)、「—」(ダッシュ)、「-」(ハイフン) を「ー」(マイナス) に統一

③ 拗音・促音部分の小さい文字を大きい文字に統一

３．１．３．２　「表記ゆれ」の具体例

図表 17 に「複合語」に分類された翻訳対象語の例を示す。

翻訳対象語	作成済み対訳辞書データ
エコフユーザー機能	「フューザ」→「フユーザ」と部分的に一致

図表 17 部分的な表記ゆれ例

３．１．４　複数訳語

本調査では、１つの翻訳対象語に対して複数の翻訳対象テキストが存在し、複数の訳語が得られるものを複数訳語とした（項番**２．２**参照)。複数訳語として翻訳対象語「お宅」の１語のみが検出された。「お宅」は図表 18 における翻訳対象テキスト１、２では「家」を意味する訳語 house が、翻訳対象テキスト３では相手を意味する you が適当な訳となる。そしてより多くの文で適訳となる house を最終的な訳語とした。

翻訳対象テキスト	お宅/ house
翻訳対象テキスト１	よく知られているように人の声紋は個人ごとに異なっており、マイク４，９から入力された特定の音声区間、例えば同一人を呼ぶ「誰々さん」「ごめんください」「誰々さんのお宅でしょうか（疑問)」といった音声信号をデジタル信号に変換し、これのスペクトル等の解析などから各個人の特徴パラメータを抽出し、予め格納している取得済み音声データの特徴パラメータを比較し、両者の特徴空間上での距離を測っ

	て、所定の範囲内であれば同一人であると推定する。
翻訳対象テキスト２	この結果、従来の安全性の低いクラス４の作業は、全て発振器製造工場で完結することができ、今までクラス４の作業であったレーザー発振器７１の交換作業、レーザーアライメント調整作業等は、全てクラス１の安全性に変わり、ユーザーのお宅で作業することもできる。
翻訳対象テキスト３	なお、怒り発話に偏って出現した単語としては、一人称の「私」や「わたしら」、二人称の「お宅」等がある。

図表 18 複数訳語の例

３．１．５　分類結果

翻訳対象語を分類した結果を図表 19 に示す。なお、「表記ゆれ」については、表記統一の結果、各翻訳対象語と作成済み対訳辞書データの日本語の単語とが完全一致するものは存在せず、全て部分一致するものであった。このような翻訳対象語は、表記が統一されていたとすれば、本来「複合語」として扱われるべき翻訳対象語といえるため、項番３．２以降の調査ではこれを「複合語」として扱うこととする。

分類	翻訳対象語数
新語	602
複合語	2,437
表記ゆれ	41
複数訳語	1

図表 19 翻訳対象語の分類と語数

３．２　翻訳対象語が発見された公報の調査

翻訳対象語の出現傾向を把握するため、翻訳対象語の検出元となった公報から、書誌情報を取得した[3]。そして取得した書誌情報に基づき、（１）技術分野ごとの翻訳対象語数、（２）出願年（優先年）／公報発行年ごとの翻訳対象語数、（３）出願人ごとの翻訳対象語数を調査した。

３．２．１　技術分野ごとの翻訳対象語数

技術分野別の調査では、第一分類のＩＰＣセクション単位で翻訳対象語数を集計したところ、セクションＣ，Ａの公報から検出された用語が多数を占める結果となった（図表 20 参照）。

ＩＰＣセクション	新語	複合語	翻訳対象語全体
A	415	344	8,101
B	41	301	3,602
C	376	1,605	9,731
D	7	55	710
E	2	7	726
F	3	13	1,094
G	101	265	4,491
H	120	354	3,983
分類不能	0	0	3

図表 20 技術分野別翻訳対象語

[3] 書誌情報は、基本的に DOCDB（欧州特許庁が提供する約 80 の国／機関で発行される特許文献の書誌情報等を含むデータベース）の情報を用いたが、本調査は日本の特許公報を対象とするため、出願人については、公報データ（SGML または XML 形式）の書誌情報の日本語表記の出願人を使用した。

３．２．２　出願年（優先年）ごとの翻訳対象語数

出願年別の調査では、翻訳対象語の件数は、全体としては 2008 年から、新語は 2014 年から、複合語は 2009 年から、その他は 2009 年から大きく増加する傾向が見られた（図表 21 参照）。

出願年[4]	新語	複合語	翻訳対象語全体	出願年	新語	複合語	翻訳対象語全体
1978	0	0	2	2000	0	41	360
1979	0	0	0	2001	0	56	504
1980	0	0	2	2002	2	46	533
1981	0	0	0	2003	3	59	491
1982	0	0	2	2004	2	99	586
1983	0	0	32	2005	10	30	458
1984	0	7	49	2006	3	40	417
1985	0	1	33	2007	16	83	950
1986	0	2	95	2008	57	183	2,205
1987	0	0	47	2009	62	392	2,933
1988	0	1	32	2010	43	240	2,553
1989	0	0	25	2011	68	224	2,367
1990	0	1	35	2012	91	209	2,776
1991	0	7	122	2013	99	258	3,397
1992	0	16	186	2014	201	301	3,625
1993	0	12	130	2015	180	185	2,807
1994	0	11	239	2016	27	69	704
1995	0	25	259	2017	94	110	1,192
1996	0	25	217	2018	91	140	1,092
1997	1	18	248	2019	15	14	154
1998	0	20	295	合計	1,065	2,944	32,441
1999	0	19	287				

図表 21 出願年別翻訳対象語

[4] 出願年は、優先権主張を伴う出願である場合、優先権主張の基礎出願の出願年で集計した。

３．２．３　公報発行年ごとの翻訳対象語数

公報発行年別の調査では、翻訳対象語の件数は、全体では 2009 年から、新語は 2013 年から、複合語は 2009 年から、その他は 2009 年から大きく増加する傾向が見られた（図表 22 参照)。

公報発行年	新語	複合語	翻訳対象語全体	公報発行年	新語	複合語	翻訳対象語全体
1993	0	13	328	2007	3	18	274
1994	0	16	179	2008	6	59	525
1995	0	16	163	2009	21	93	1,133
1996	0	14	194	2010	39	179	2,123
1997	1	25	250	2011	50	390	2,772
1998	0	12	183	2012	61	340	3,349
1999	0	11	192	2013	105	275	3,373
2000	0	11	192	2014	83	275	3,521
2001	0	20	234	2015	118	250	3,210
2002	0	26	266	2016	206	277	3,301
2003	0	52	373	2017	152	193	2,437
2004	0	26	301	2018	28	54	767
2005	0	39	322	2019	190	233	2,149
2006	2	27	330	合計	1,065	2,944	32,441

図表 22 公報発行年別翻訳対象語

なお、出願の 2016 年、公開の 2018 年の件数が、その前後の年と比べ各項目の件数が顕著に少ないが、これは、本事業で使用した翻訳ログにおいて、もともと公報発行年が 2018 年のデータ[5]の件数が極端に少なかったことに起因する。2018 年に公開される公報は、18 ヵ月前に出願されたものであるため出願年としては主に 2016 年に該当する。

[5] 翻訳ログ全体の発行年調査は、翻訳ログには公報発行日が含まれていないため、便宜的に翻訳ログの識別番号の年部にて集計を行った(2017 年 83,106 件、2018 年 23,367 件、2019 年 47,898 件)。

３．２．４　出願人ごとの翻訳対象語数

出願人別の調査では、翻訳対象語全体は、C08(有機高分子化合物；その製造または化学的加工；それに基づく組成物) (10 社中 6 社)、新語は A63(スポーツ；ゲーム；娯楽)(10 社中 4 社)、複合語は C08 (10 社中 6 社)の出願が多い企業であった（図表 23 参照)。

#	翻訳対象語全体			新語			複合語		
	出願人	件数	FI[6]	出願人	件数	FI	出願人	件数	FI
1	住友ベークライト株式会社	415	C08	住友化学株式会社	67	C08	住友ベークライト株式会社	180	C08
2	一丸ファルコス株式会社	324	A61	バイエル・クロップサイエンス・アクチェンゲゼルシャフト	66	C07	東洋インキＳＣホールディングス株式会社	71	C09
3	住友化学株式会社	306	C08	日本曹達株式会社	39	A01	三菱化学株式会社	71	C08
4	三菱化学株式会社	265	C08	株式会社三共	36	A63	ＪＳＲ株式会社	63	C08
5	花王株式会社	244	A61	東洋インキＳＣホールディングス株式会社	29	C09	株式会社ブリヂストン	59	B60
6	クゥアルコム・インコーポレイテッド	243	H04	バイエル・クロップサイエンス・アーゲー	22	G03	独立行政法人産業技術総合研究所	59	H01
7	ビーエーエスエフソシエタス・ヨーロピア	237	C08	京楽産業．株式会社	22	A63	住友化学株式会社	51	C08
8	キヤノン株式会社	222	H04	コニカミノルタ株式会社	21	A63	花王株式会社	43	A61
9	ＪＳＲ株式会社	219	C08	サミー株式会社	20	G02	株式会社日本触媒	41	C08
10	東レ株式会社	215	C08	富士フイルム株式会社	17	A63	信越化学工業株式会社	37	C08

図表 23 出願人頻出技術分野

[6] その企業を出願人とする 2015 年 1 月 1 日～2019 年 12 月 31 日発行の特許・実用新案公報において、最も多く付与された FI のサブセクションを意味する。

３．３　今後の効率的な新語対応のあり方についての検討

項番３．２に示した調査結果を踏まえて、次年度以降の MTP における効率的な新語対応のあり方について、（１）翻訳対象語の発見されやすい公報の特徴の分析、（２）翻訳対象語選定作業の効率化の検討を行った。

３．３．１　翻訳対象語の発見されやすい公報の特徴の分析

翻訳対象語の発見されやすい公報の特徴の分析では、（１）新語の傾向、（２）技術分野の傾向、（３）複数訳語の傾向、（４）頻度が多いが「新語」に該当しなかった翻訳対象語の考察、（５）古い年代の公報から作成された翻訳対象語の傾向について分析した。

３．３．１．１　新語の傾向

３．３．１．１．１　公報発行年別件数の傾向

2009 年以降の翻訳対象語、及びその中で新語に相当する語を含む公報件数の推移を図表 24 に示す。

公報発行年	翻訳ログ[7]		翻訳対象語（新語を含む）		新語	
	件数	割合	翻訳対象語数	割合	新語数	割合
2009	42,498	3.70%	1,133	3.49%	21	1.97%
2010	86,808	7.57%	2,123	6.54%	39	3.66%
2011	97,394	8.49%	2,772	8.54%	50	4.69%
2012	109,678	9.56%	3,349	10.32%	61	5.73%
2013	118,743	10.35%	3,373	10.40%	105	9.86%
2014	162,465	14.16%	3,521	10.85%	83	7.79%
2015	247,095	21.54%	3,210	9.89%	118	11.08%
2016	128,035	11.16%	3,301	10.18%	206	19.34%
2017	83,125	7.25%	2,437	7.51%	152	14.27%
2018	23,379	2.04%	767	2.36%	28	2.63%
2019	47,898	4.18%	2,149	6.62%	190	17.84%

図表 24　公報発行別翻訳対象語数(翻訳対象語全体/新語)[8]

翻訳対象語を含む公報数のピークは 2014 年（3,521 件）である。この年はもともと翻訳ログ数が対象期間中で最も多く、これに比例した妥当な結果といえる。翻訳対象語に関しては、2014 年以外も、おおむね翻訳ログ件数と同期した推移を示している。

これに対し、新語を含む公報数のピークは 2016 年（206 件）と 2 年のずれが生じている

[7] 翻訳ログ全体の発行年調査は、翻訳ログには公報発行日が含まれていないため、便宜的に翻訳ログの識別番号の年部にて集計を行った。

[8] 本表はすべて翻訳ログ単位に集計しており、表中の「翻訳対象語数」及び「新語数」は、同じ語であっても複数の翻訳ログに出現した場合はその都度カウントしている。

が、これは、そもそも本事業の新語の定義（項番３．１．１．１参照）自体が、いわば「2006年以降に急激に普及したか、新たに使われ始めた語」のみを対象としており、翻訳対象語全体に比べて期間の後半に件数が偏るのは当然といえる。

また、表中の「割合」欄（集計期間内の総件数に対する各年の件数の比率）を見ると、2009 年から 2014 年までは翻訳対象語数の比率のほうが新語数の比率よりも高いが、2015年以降はこれが逆転して新語数の比率のほうが高くなっており、年を追うごとに比率の差は広がっている。これも「新語は古い公報よりも新しい公報に多く含まれる」という当然の結果ではあるが、「新語を採取するには、できるだけ新しい公報を対象にするのが効率的である」ことが数値的にも確認できた。

なお、上記「割合」で考えると、今回の調査では、直近 5 年（2015～2019）で全体の約65%にあたる新語がカバーされている。2018 年の翻訳ログが少量であることによる特異値であることも考え合わせると、この 5 年間の本来の新語カバー率はさらに高いものと思われる。

もちろん、本集計では同じ語であっても公報が異なれば重複してカウントしており、かつ新語自体が「出現後しばらくは年を経るごとに該当公報数が増加していく」性質であることを考えると、単純にこれを「直近 5 年のみを対象にしても新語の 65%以上がカバーできる」と解釈することはできない。だが、重要性の高い語であれば、たとえ直近 5 年以前に頻度のピークが来ており現時点ではすでに「新語」とは言えなくなっているとしても、大抵の場合、それ以降のすべての年代の公報にも一定程度出現するものと考えられる。したがって、こうした語をカバーするためにあまりに過去の公報まで採取範囲を拡げていく必要はないと考える。

これに対し、最新の公報に含まれる「翻訳対象語」には、その時点では件数が足りず本調査で定義した「新語」に該当しなくとも、将来的に普及して「新語」となる可能性を有している。もちろん古い公報の翻訳対象語が将来「新語」となる可能性も皆無ではないが、その確率は相対的に低い。この観点からも、仮に同等規模の公報を対象（新語の採取源）とするのであれば、各年代を均等に対象とするよりも、極力最新のものに集中させるほうが、より効率的に重要性の高い語の収集を行えるであろう。

３．３．１．１．２　出願人別の傾向

1993 年以降に公開された公報から検出された翻訳対象語及び新語の数（ヒット数[9]及び語数）を、検出元の公報の出願人別に調査した（図表 25 参照）。翻訳対象語数上位 10 社は、項番３．２に前掲したとおりだが、新語数の上位 10 社は、1 社（住友化学株式会社）を除き、すべて異なる出願人となった。

#	翻訳対象語全体（新語を含む）			新語		
	出願人	ヒット数	語数	出願人	ヒット数	語数
1	住友ベークライト株式会社	415	368	住友化学株式会社	67	38
2	一丸ファルコス株式会社	324	174	バイエル・クロップサイエンス・アクチェンゲゼルシャフト	42	18
3	住友化学株式会社	306	200	日本曹達株式会社	39	17
4	三菱化学株式会社	265	238	株式会社三共	36	11
5	花王株式会社	244	188	東洋インキＳＣホールディングス株式会社	29	23
6	クゥアルコム・インコーポレイテッド	243	214	バイエル・クロップサイエンス・アーゲー	24	19
7	ビーエーエスエフソシエタス・ヨーロピア	237	223	コニカミノルタ株式会社	22	6
8	キヤノン株式会社	222	194	京楽産業．株式会社	22	14
9	ＪＳＲ株式会社	219	165	サミー株式会社	21	4
10	東レ株式会社	215	198	富士フイルム株式会社	20	18

図表 25 出願人別翻訳対象語数(翻訳対象語全体/新語)

[9] 「ヒット数」は、翻訳対象語又は新語の検出元公報単位に集計した数値である。これに対し「語数」は出願人単位で用語（翻訳対象語又は新語）の重複排除を実施した後の数値である。

新語の上位 10 社には、遊具の取り扱いを事業内容とする会社（株式会社三共、京楽産業.サミー株式会社）が 3 社ランクされた。翻訳対象語（すなわち未知語）の上位 10 社には遊具関連の出願人は見当たらず、新語特有の傾向といえる。

これら 3 者の出願から新語として採取された語は、それぞれ 11 語、6 語、4 語だが、このうち他の出願人にも使用されていた語は「激熱（3 社とも）」、「報知演出（京楽）」及び「3 択役（サミー）」の 3 語のみであり、それ以外は例えば「高確情報」、「電断復帰」及び「加減算玉数カウンタ」（三共）、「低確時短遊技」（京楽）、「ベルコボシ」（サミー）など、他の出願人が使用する可能性がきわめて低そうな、“造語”と見なせる語が大勢を占めた。これらの語は、当該一社のみによる、特定の時期における多数の出願で使用された結果、本事業の新語の条件に該当したものといえる。

もちろん、このように”特定の出願人のみが特定の時期に用いる”タイプの用語であっても、実際に未知語となっている件数自体は多く、辞書登録による改善効果は十分に期待できる。ただし、「新語」という言葉から本来イメージされる、”新たにスタンダードとして普及・定着してきた言葉”という概念とはやや異なる類のものであることも事実であり、かつ、こうした遊具関連技術はあまり諸外国では普及していない日本特有の技術といえ、このため翻訳もたとえば「操作順序不正解：incorrect answer to the operation order」（サミー）など、非常に説明的なものにならざるを得ない。

こうした状況を考慮すると、場合によっては、こうしたタイプの語の優先度を落とすため、本事業の新語の定義に「一定数以上の出願人に使用されているもの」という条件を加えることにも一考の余地がある。

なお、これら 3 社を除けば、新語数上位にランクされた各社は、傾向としてはおおむね翻訳対象語数と同様であった。具体的には「化学分野に偏っている」ことと、「外国籍の出願人が含まれる」ことである。

前者については、「（２）翻訳対象語が発見されやすい技術分野」で後述するため、ここでは後者について考察する。まず、外国籍出願人が上位にランクされた要因としては、「（日本の出願人が用いないような）特殊な日本語表現や、外国語をそのままカタカナ表記した表現」を多用している可能性が考えられる。

そこで、2 位にランクされたバイエル・クロップサイエンス・アクチェンゲゼルシャフト

と 6 位にランクされたバイエル・クロップサイエンス・アーゲー[10]それぞれの出願から検出された新語の内容を見てみると、前者は「ウロシスチス・オクルタ」「グイグナルジア・ビドウェリ」（ともに植物の一種）といったラテン語学名をカタカナ表記したものが大半（18 語中 12 語）、後者も「アトランス・テニュイッシマ」（植物の病気）、「エルシノエ・ファウセッチイ」（病原菌）など同傾向（19 語中 15 語）であった。

これらは、「外国語をそのままカタカナ表記した表現」ではあるものの、当初想定していたもの、つまり、日本語として表現できるものを外国語カナ表記で書いているものとは性質が異なる。少なくともこの 2 社において、当初想定していた「日本語に訳せるのに外国語カナ表記を用いている」ケースは、翻訳対象語全体に範囲を拡げても一切見られなかった。

今回多数検出された「ラテン語学名のカタカナ表記」については、そもそも和訳語が存在しないことも多く、その場合、多くの日本の出願人はそのままアルファベットで表記しているものと考えられる。外国籍出願人の場合、そのような慣習を知らずにカタカナ表記を使用することも多いようである。

こうした学名のカナ表記も、現状で未知語になっていることは事実であり、辞書登録に一定の意義はある。とはいえ、こちらも本来の意味での”新語”とはやや異質のものであり、今後、本来の新語と同等に扱うべきかについては、検討する必要があるだろう。学名カナ表記の大部分は、（当然ながら）カタカナのみで書かれるか、もしくは末尾のみ「属」「目」などの漢字が付随するかのいずれかであるが、前者については他種の語（たとえば化学物質名）にも多数当てはまるため特定条件としては不十分である。上記各例のように、中点（・）を含むものが多いこと、中点で分割した単位でも未知語となる確率が高いことなどである程度の絞り込みは可能であるが、機械的に完璧な取捨選択は難しいであろう。むしろ翻訳作業時に目視で判定し対象外とするほうが容易かつ効率的であるかもしれない。

なお、同 2 社による、カタカナを含まない新語は皆無であり、翻訳対象語全体に範囲を拡げても、121 語中「避陰」、「葉枯性」及び「弁鰓綱」の 3 語のみであった。これら 3 語とも、Google 検索では「避陰」が 15,300 件（主に「避陰反応」として）、「葉枯性」が 2,380 件、「弁鰓綱」も 169 件とそれぞれ一定の使用例があり、正式な技術用語の範疇であると考えられる。したがって、これらのいずれも当初懸念していた「（日本の出願人が用いない）特殊な日本語表現」には該当せず、少なくとも今回の調査からは、「正しくない日本語」を除外する目的で外国出願人の案件を対象外とすることは効果が薄いと結論できる。

---

[10] 両者はおそらく同一出願人と思われるが、本調査では出願人の異表記の統一は実施しておらず、個別に集計されている。

３．３．１．２　翻訳対象語が発見されやすい技術分野

項番３．２の調査では、翻訳対象語全体、そして新語に限定しても、いずれも第一分類の IPC セクションが C 又は A である公報から検出された翻訳対象語が多数を占めるという結果となった。ここでは、この上位２セクションについて、サブクラス単位に集計し、頻出するサブクラスの翻訳対象語を分析した。

＜C セクションの公報から作成された翻訳対象語の特徴＞

Ｃセクションの公報から検出された翻訳対象語を、検出元の公報のサブクラス別に集計すると、C08L(高分子化合物の組成物)、C07D(複素環式化合物)、C08F(炭素－炭素不飽和結合のみが関与する反応によってえられる高分子化合物)、C12N(微生物または酵素；その組成物)から検出された翻訳対象語が多い結果となった（図表 26 参照)。

#	技術分野	翻訳対象語数	割合 (C セクションの公報から検出された全翻訳対象語に対する割合)
1	C08L	1,178	12.11%
2	C07D	1,172	12.04%
3	C08F	822	8.45%
4	C12N	792	8.14%
5	C08G	762	7.83%
6	C07C	582	5.98%
7	C09D	462	4.75%
8	C09J	349	3.59%
9	C08J	314	3.23%
10	C07F	296	3.04%

図表 26 技術分野別翻訳対象語数(C セクション)

C セクションの公報から検出された翻訳対象語は、その大多数がカタカナで構成される化学物質名であった。また、用語の文字数に着目すると、例えば C08L に属する公報から採取した翻訳対象語は、平均文字数が 17.2 文字と非常に長大で、30 文字以上のものも存在する。本事業で検出した翻訳対象語全件の平均文字数は 12.8 文字であり、これ自体も通常の技術用語として想定される文字数より顕著に長いが、それだけこの「長大な化学物質名」の占める割合が大きいということを示している。図表 27 に C08L の公報から作成された翻訳対象語の例を、図表 28 に長い翻訳対象語の例を示す。

（メタ）アクリロキシメチルジメトキシメチルシラン
５,５-ビテトラゾールナトリウム塩
アクリロニトリロブタジエンゴム
アルキルアリールエーテルホスフェート

図表 27 IPC C08L の公報から作成された翻訳対象語の例

ビス（ジプロピルオキシモノエチルシリルペンチル）ジスルフィド
ビス（ジプロピルオキシモノブチルシリルプロピル）ジスルフィド
ビス（ジプロピルオキシモノブチルシリルペンチル）ポリスルフィド

図表 28 長い翻訳対象語の例

これらの化学物質名の大半は、完全に新規な語ではなく、「既存の化学物質名の新たな連結結果」と捉えるべきものである。たとえば図表 28 の 1 つめの実例「ビス（ジプロピルオキシモノエチルシリルペンチル）ジスルフィド」も、英訳結果 bis(dipropyloxy monoethyl silyl pentyl)disulfide を見れば明らかなように、「ビス」、「ジプロピルオキシ」、「モノエチル」、「シリル」、「ペンチル」、「ジスルフィド」の６語が連結されている。そして、連結された化学物質のそれぞれは、単独であれば Google 翻訳でも問題なく翻訳できる程度の一般的な化学物質名に過ぎない。

このような化学物質名が未知語扱いされる原因は、主として「現状の機械翻訳システムが長大なカタカナの羅列を適切な単位で単語（上例であれば「ビス」、「ジプルピルオキシ」、…）に分割して個別に翻訳することができない」ためであると考えられる。適切な分割ができていないため、化学物質名全体がひとまとめで未知語扱いされてしまっている。

もちろん、現状の機械翻訳システムではこうした長大な化学物質名を適切な単位に区切ることが難しい以上、これらを逐一辞書登録していくことは、ユーザー側で現状行える最も有効な対処策であるといえる。その意味では、こうした語を多数含む C セクションを優先的に対象とすることは十分な妥当性をもつ。

しかしその反面、こうした語を構成している個々の要素自体は一般的な化学物質名であり、システム側で適切な区切り方ができるようになれば、現在未知語となっている語もその大半は問題なく訳出されるようになるはずのものであるとも言える。したがって、システム開発側にこの課題を提示し、根本的な解決を図るべき、という考え方もできそうである。そして、この課題の解決の目算が立つのであれば、C セクションから採取される未知語の多くは「いずれ自動的に対処される可能性が高いもの」となるため、むしろ優先度は他の新語よりも低くするべき、という考え方も成立する。

＜Aセクションの公報から作成された翻訳対象語の特徴＞

A セクションに属する公報から検出された翻訳対象語を、検出元の公報のサブクラス別に集計した結果は図表 29 のとおりである。

#	技術分野	翻訳対象語数	割合 (A セクションの公報から検出された全翻訳対象語に対する割合))
1	A61K	3,310	40.86%
2	A01N	946	11.68%
3	A63F	696	8.59%
4	A61B	488	6.02%
5	A23L	394	4.86%
6	A61F	226	2.79%
7	A61L	184	2.27%
8	A61M	158	1.95%
9	A01G	145	1.79%
10	A01K	134	1.65%

図表 29 技術分野別翻訳対象語数(A セクション)

続いて、図表 30 に上位 3 つのサブクラスにおいて「新語」として採取された用語を頻度順に示す。

サブクラス	出現頻度	新語
A61K	4	エカムスル
	3	オキシポリエントキシデカン
	3	ガニツマブ
	3	クィーンスシード
	3	テシットデシチン
	3	ブレンツキシマブ
	3	メリンジョ
A01N	3	ガノデルマ・ボニネンセ
	4	トリクロピリカルブ
	4	フェンピラザミン
	3	プッシニア・レコンジテ
	4	フフェノジド
	4	プロパモカルブホセチレート
	3	リギドポルス・リグノスス

A63F	4	３択役
	3	押下位置不正解
	3	加減算玉数カウンタ
	3	甘デジタイプ
	3	高確情報
	3	終了報知演出
	3	上軸支金具
	3	低確時短遊技
	3	当選報知演出
	3	非当落乱数更新処理
	3	不正解役
	4	副確変フラグ
	3	報知演出
	4	鉤金具部

図表 30 A セクションの公報から作成された新語の例

これらの結果を見ると、全体的な傾向として、A61K(医薬用，歯科用又は化粧用製剤)は化学物質名が多い。リストアップした上記各語は頻度上位のものにつき長大な連結型のものは見当たらないが、大まかな傾向としてはＣセクションに準ずると考えられる。

続く A01N(人間または動物または植物の本体，またはそれらの一部の保存)では、項番３．３．１．１．２の出願人別の傾向で考察した「ラテン語学名のカタカナ表記」が多く見られた。一方、A63F(カードゲーム，盤上ゲーム，ルーレットゲーム；小遊技動体を用いる室内用ゲーム；ビデオゲーム；他に分類されないゲーム)では、同じく項番３．３．１．１．２で考察した「遊具業界の出願人」の公報由来の新語と同傾向の語が並んだ。

この結果から、これらの「未知語になりやすい特徴を有する語」（すなわち「(長大な)化学物質名」、「ラテン語学名」、「特定の出願人特有の用語」）はかなり件数が多く、かつ比較的特定の技術分野に偏りがちであることがわかる。したがって、技術分野を意識せずに全分野を対象に新語の採取を行うと、こうした特殊な性質の用語が大量に含まれることとなる。仮にこうした語を採取対象から除外する、優先度を下げる等の措置が必要となる場合、IPC セクション等の技術分野情報に基づく対象範囲の限定や優先順位付けも、作業量との精度とのバランスに鑑み実用的であろう。

３．３．１．３　複数訳語の傾向

項番３．１の調査結果では、本調査で複数訳語が存在する用語は、１語（「お宅」）のみで

あった。そのため、この一語から複数訳語の全体の傾向を分析することは難しい。ここでは、この 1 語のみの限定な調査になるが、この翻訳対象語が、特許の中でどの程度使用されているかを調査した。

「お宅」を J-PlatPat でテキスト検索したところ、全ての期間（2019 以前）に発行され公報は 178 件と、少ない特許での使用となっているため、本事業の新語対応への影響は限定的と考えられる。

３．３．１．４　頻度が多いが「新語」に該当しなかった翻訳対象語の考察

項番３．１．１．１に示したとおり、本調査では「新語」の定義を「①（2006 年～2019 年の翻訳対象語記載公報数）／（1993 年～2005 年の翻訳対象語記載公報数）が 42 以上、又は、1993 年～2005 年の翻訳対象語記載公報数が０であること」と、「②1993 年～2019 年の翻訳対象語記載公報数が 50 以上であること」の双方を満たす語とした。

このうち②を満たさない（すなわち出現頻度が 50 件未満である）語は「そもそも滅多に使用されない未知語」であるため対処の優先度は低いが、②を満たす語、つまり出現頻度が 50 件以上である語は、たとえ①に抵触し本調査では「新語」に該当しなくとも、機械翻訳全体の精度を考えると、（少なくとも①を満たすが②を満たさない語よりも）優先的に辞書登録すべきものであると考える。

この観点から、本調査で採取した翻訳対象語のうち「②を満たし、かつ①を満たさない語」の実例を確認した。図表 31 にカウント期間（1993～2019 年）中の出現頻度がちょうど 50 語であったものからサンプルを 10 語（カタカナ語、漢字語各 5 語）示す。

		翻訳対象語	出現時期	2006 年以降の頻度
カタカナ語	#1	アセトナフテニル基	1993-2017	42
	#2	アルカンスルホネートアニオン	2001-2019	40
	#3	イソブテニルフェノール	1993-2017	27
	#4	エアサスペンシヨン	1993-2016	1
	#5	オーナードライバー	1993-2018	3
漢字語	#6	掛け流し方式	1995-2019	29
	#7	舷上	1993-2019	29
	#8	縦断裁	1993-2019	30
	#9	奪水剤	1995-2019	24
	#10	脱刷	1993-2019	39

図表 31 出現頻度 50 件で「新語」に該当しなかった未知語の例

上表に例示した各語を見ると、どれも特に「新語」であるという印象は受けないものの、大半の語（※カタカナ語の#4～#5 を除く）は 1993～5 年から 2016～9 年にかけて安定的に使用されており、今後の特許文献に使用される可能性も高いと思われる。したがって、これらは各語とも辞書登録の優先度は比較的高い。

本調査の定義では、これらの語は「新語」とは見なされなかった。だが、項番３．３．１．１で提案した「新語の採取の際には新しい公報を優先すべき」という方針を採用したとしても、これらの「新語でない頻出語」が対象外となる懸念は小さい。むしろ、こうした語の大半は「最新の公報を優先」する方針であってもカバーされる可能性が高いといえる。

本調査では、各翻訳対象語について、実際の特許公報での使用頻度を個別にカウントし、各語が新語に該当するか否かを判定した。しかしながらこの作業はあくまで新語の存在状況を把握する調査目的であり、今後、新語の採取作業を行う際、採取された翻訳対象語に対して逐一このような広範囲の公報件数カウントを行い、新語の条件に合致するか否かで取捨選択を行うことは現実的ではない。実作業時は、採取源となる公報の範囲は本調査で得た知見等に基づき絞り込むものの、その結果採取された未知語は、原則「新語である蓋然性が高い」ものとして一括で採用することになると思われる。したがって、たとえ本調査の定義では新語に該当しなかったとしても、こうした語の多くは「最新の公報」でも使用されている可能性が高く、そうである限り他の未知語と同様に採取の対象となる。上に例示した各語や、これと同等レベルの語であれば、最新の公報においても少なからず使用されることはほぼ確実であろう。

つまり、上掲の各語のような、過去から現在に至るまで平均的に使用されてきており、その結果「新語の条件②を満たすが①に抵触する」未知語をカバーするために、あえて採取源の年範囲を過去に広げる必要はない、という結論となる。

なお、カタカナ語の#4「エアサスペンシヨン」、#5 の「オーナードライバー」は、1993 年以降の使用頻度は他と同様に 50 回であるが、2006 年以降の頻度はそれぞれ 1 回、3 回とわずかであり、「最新の公報」を対象とする方針では、採取される確率は低くなる。

だが、これら 2 語は、少なくとも使用頻度的には「新語」とは対極の「死語」というべき語といえる。前者「エアサスペンシヨン」はごく一般的な語であるが、拗音促音の表記が「ション」ではなく「シヨン」となっているため未知語となったものである。このような旧い表記法は過去には頻繁に見られたが（項番３．３．１．５にて後述）、最近はほぼ使われなくなった印象で、頻度情報もこれを裏付けている。一方、#5「オーナードライバー」は表

記自体に古さはないが、たしかに最近はあまり耳にせず、実際の使用頻度を見ても、特許業界ではほぼ死語になっているようである。

これら２語のような「死語」は、「最新の公報」のみを対象とする方針では新語候補として採取されない可能性が高いが、もともと「新語」を優先的に採取するという方針は、裏を返せば「死語」の優先度は下げるということであり、意図どおりの適切な選別が実現することとなる。

３．３．１．５　古い年代の公報から作成された翻訳対象語の傾向

項番３．１．１．１では、翻訳対象語のうち、一定以上の使用実績があって、新しい年代に発行された公報から急に出現するようになった翻訳対象語を、新語と定義した。ここでは、こうした「新語」とは対照的な、古い年代のみに出現する翻訳対象語について、その傾向を分析した。

1993 年～2005 年の古い年代に発行された公報に 10 件以上出現する翻訳対象語を調査したところ、拗音・促音が小さい文字でなく、通常の文字で表現された翻訳対象語が 52 語存在した（図表 32 に例を示す。）。このことから、古い年代の公報に出現する翻訳対象語は、拗音・促音に特徴があると考えられる。なお、これらの用語の拗音・促音部分を小さい文字に変更し、みんなの自動翻訳@Textra®で翻訳を行ったところ、適切な訳語が得られることが確認できた（図表 33 参照）。

ホームポジシヨンセンサ
アウトプツトシヤフト
ストツプバルブ
バブルジエツト
コイルユニツト
コレツトチヤツク
パイロツトポート

図表 32　古い年代の公報に出現する翻訳対象語の例

図表 33　拗音・促音の機械翻訳（みんなの自動翻訳@Textra®）例

３．３．２　翻訳対象語選定作業の効率化の検討

３．３．２．１　単語の途中で改行が挿入されることにより検出された未知語の扱い

項番２．１．２．１（１）で述べたとおり、本調査では、翻訳対象テキストの文頭又は文末で検出された未知語について、公報原文を参照し、単語の途中で改行の挿入による未知語と判定された場合に、その未知語に対応する翻訳ログを除外する作業を行った（図表 34 参照）。

本事業において文頭、文末から検出された未知語は、およそ 14,000 件存在した。しかし、これらから選定された翻訳対象候補は、244 語(1.7%)にすぎず、選定作業のために人手による大規模な公報原文確認作業を行ったにも関わらず、翻訳対象候補が少数しか得られない結果となった。

今後、同様の選定を行う際には、該当する翻訳ログの未知語１件毎に、未知語抽出元の公報の中から、該当箇所を探すという公報原文確認作業を行うのは手間がかかるため、この作業の効率化が有効と考えられる。例えば、確認作業者が作業する前に、あらかじめ、翻訳ログにおける翻訳対象テキストの前後の文を自動的に取得し、その内容を容易に確認できる仕組みを用意することなどが考えられる。

また、文末に未知語を含む翻訳対象テキストに続く文は、文頭に未知語を含む文となる可能性が高いため、同一文献に発生する文頭、文末の未知語の確認を一括で行う仕組みも有効と考える。

未知語	パラメータインフ
翻訳対象テキスト	制御装置は、ピコステーションの通信セットアップアクノレッジ指令を受け取ると、標準のパラメータインフォメーションメッセージフォーマットを使って、パラメータインフ
未知語	ォメーション
翻訳対象テキスト	ォメーションをピコステーションに送信し、その後、ピコステーションは上記パラメータインフォメーションメッセージ受け取りの応答を行う。

図表 34　文頭・文末の未知語例

３．３．２．２　漢字から成る未知語の扱い

翻訳対象語は、翻訳ログの未知語から採取／作成したものである。翻訳ログの未知語には、図表 35 に示すように、未知語全体がカタカナで構成されているものが大勢を占める一方で、図表 36 に示すように、漢字 2 文字から成る未知語も 9,694 件と多数存在した。ただし、これら漢字２文字から成る未知語は、そのまま翻訳対象語として採用できず、項番２．１．２．２の未知語の修補の対象となるケースが多かった。たとえば図表 36 の一例目では、翻訳ログ上の未知語は「鈑又」という漢字二文字であったが、この語は翻訳対象語として適切な単位ではなく、修補作業により「カラー鋼鈑」と修正されている。このような状況を踏

まえ、漢字 2 文字からなる未知語について調査を行った。

翻訳対象語	未知語	翻訳対象テキスト
チタンアミノエチルアミノエタノレート	チタンアミノエチルアミノエタノレート	上記の好ましい金属種を代表表記すれば、例えば、チタンラクテート、チタントリエタノールアミネート、チタンラクテートアンモニウム塩、チタンジエタノールアミネート、チタンアミノエチルアミノエタノレート、塩化ジルコニル化合物、ジルコニウムラクテートアンモニウム塩等があげられる。

図表 35 未知語（カタカナ）例

翻訳対象語	未知語	翻訳対象テキスト（下線は未知語を表す）
カラー鋼鈑	鈑又	～～～カラー鋼鈑又は～～～
Ｈｅｂｂ則理論	則理	～～～たとえば、Ｈｅｂｂ則理論による～～～
ＨＦＣ冷媒	媒時	～～～ＨＦＣ冷媒時より配管径を大きく～～～
アルミニウム箔	箔独	～～～アルミニウム箔独特のグレー色、～～～

図表 36 未知語（漢字 2 文字）例

これら漢字 2 文字から成る未知語の傾向を調査したところ、先頭の文字が「謂」(168 件)、「該」（137 件）である未知語が多数存在した。そこで「謂」、「該」で始まる未知語を含む翻訳対象テキストを確認したところ、図表 37 に示すように、「謂」は本来「所謂」という単位で取り扱われるべきところ、機械翻訳システムがこれを正しく認識できず、「所」と「謂」とに区切って扱った結果、「謂」が単独では成立しない語であるために強制的に直後の文字と連結され、「謂暗」や「謂掛」といった意味のない用語として扱われ、未知語が発生したと考えられる。「該」に関しては、図表 37 に示すように、本来は単独で成立する語であるが機械翻訳システム上ではそうなっておらず、このため「謂」と同様の結果が生じている。これらの未知語への対応としては、本事業で行ったように未知語を修補し翻訳対象語を作成する方法の他に、機械翻訳システムにて「所謂」や「該」を適切に分割できるようにするという根本的な解決策も検討すべきであろうと考えられる。

未知語	翻訳対象テキスト
謂暗	～～～圧調整器が作動すると所謂暗電流が大きくなり～～～
謂掛	この実施例では所謂掛け流し方式としている。
該咥	～～～咥胴５で咥えられた折丁は一回転する前に該咥胴から離れ～～～

該擂	該凹状溝を該擂鉢状凹部の周縁部より～～～
該収	実施時、該収納体４０は、～～～

図表 37 「謂」、「該」で始まる漢字２文字の未知語例

なお、漢字 2 文字から成る未知語以外にも、図表 38 に示すように、3 文字以上の漢字から成る未知語も存在したが、各未知語を含む公報数を J-PlatPat で確認したところ、いずれの未知語を含む公報も 100 件に満たなかった。MTP で用いているニューラル機械翻訳は、機械翻訳が取り扱うことのできる用語数に上限があるため、頻度の高い用語を優先して翻訳のための機械学習を行う仕組みになっている。そのため、このような低頻度の用語は機械学習対象とされず、未知語になったと推測される。また用語自体は一般語であるため、一般の文書を学習した、みんなの自動翻訳@Textra®の汎用 NMT で翻訳したところ、以下のような訳語が得られた。

得られた訳語を確認すると、「祭囃子」の訳語は「Matsuribayashi」とローマ字表記に訳されたが、英語ネイティブが内容を理解するためには、本事業の翻訳作業で作成した「japanese festival music」の活用が有効と考える。

未知語	未知語を含む文献数	機械翻訳結果（みんなの自動翻訳@Textra®の汎用 NM）
貴婦人	30	Lady
原生林	74	Primeval forest
最優秀賞	12	Top Prize
祭囃子	8	Matsuribayashi
参院選	14	House of Councillors
終戦記念日	8	Anniversary of the end of the war

図表 38 漢字からなる未知語（２文字以外）例

３．３．２．３　文字数が長い翻訳対象語の扱い

翻訳対象語の中には、図表 39 に示すような、比較的長い文字数の翻訳対象語が存在した。

このような長い文字数の翻訳対象語の多数は、複合語であると考えられるため、適切な単位の単語に分割することで、適切に翻訳できる可能性がある。そこで、翻訳対象語の英語訳を参考に、翻訳対象語を分割して再度翻訳したところ、正しい訳語が得られた。

このように、翻訳対象語の訳語を参考に、未知語を分割して翻訳することで未知語が改善する可能性があるため、翻訳対象語の英語訳を利用し、原文となる日本語文の分割を行い、機械翻訳することが効果的と考える。日本語の文章を分割する方式は、英語訳のような、日本語以外の情報を参考にして行う方式以外に、日本語のデータのみから分割を行う方式も

存在する。

この方式の一つに、Google の翻訳で採用されている SentencePiece[11]がある。SentencePiece は対象とする文書を構成する部分文字列を機械学習し、学習結果を使い、文書の分割を行う方式である。

翻訳対象語	シクロドデシリデンビスフェノールビスクロルホルメートポリカーボネートオリゴマー
翻訳対象語の英語訳	cyclododecylidene bisphenol bischloroformate polycarbonate oligomer
英語訳を参考に/部分で分割	シクロドデシリデン / ビスフェノール / ビスクロルホルメート / ポリカーボネート / オリゴマー
分割した単位の機械翻訳(みんなの自動翻訳@Textra®の特許 NMT)	CYCLODODECIDENE BISPHENOL BISCHLOROFORMATE POLYCARBONATE OLIGOMER

図表 39 翻訳対象語の分割例

[11] https://github.com/google/sentencepiece

添付資料 新語例

翻訳対象語	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	合計
ブタテッショウウイルス	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	8	3	9	16	24	44	109
トリメチロールプロパンフォルマル（メタ）アクリレート	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	9	8	8	11	15	16	18	12	21	122
イムガツズマブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	13	24	32	51	123
メコプロップナトリウム	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	7	0	7	8	2	12	20	21	15	27	124
カエトクネマ属種	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	15	5	15	21	21	16	25	126
ティガツズマブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	10	7	10	23	37	33	127
シアントラニルプロール	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	1	10	6	17	32	19	45	133
トリメチルスクシノニトリル	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	10	15	14	12	17	20	18	23	134
ファーレツズマブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	4	22	31	28	39	134
キノフメリン	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	10	44	89	146
イッテンオオメイガ	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	1	3	6	8	3	7	13	7	17	23	20	34	147
テッショウウイルス	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	4	6	9	7	15	27	30	49	149
メフェントリフルコナゾール	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	11	44	105	163
パクリチニブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	11	20	33	47	60	174
イキサゾミブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	36	46	80	178
アドバンテーム	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	6	16	20	31	57	49	185
フォスタマチニブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	11	11	17	42	47	52	188
エスタフェナトクス	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	4	16	10	41	26	42	51	193
アマツキシマブ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	9	37	36	46	70	201