令和３年度

機械翻訳プラットフォームの中日翻訳における

未知語選定等に関する調査研究事業

調査報告書

令和４年　３月２４日

東芝デジタルソリューションズ株式会社

目　次

1. 本報告書の概要

特許庁、（独）工業所有権情報・研修館は、特許情報プラットフォーム（J-PlatPat）を通じて中国の公開特許公報の日本語機械翻訳文を提供している。そして、当該機械翻訳文は、令和 2 年 5 月から機械翻訳プラットフォーム（MTP）によって作成しており、その翻訳品質は、ニューラル機械翻訳等の各種翻訳エンジンの使い分け等によって、従来の機械翻訳と比較して格段に高いものとなっている。

しかしながら、技術革新等により登場した新技術用語等の、MTP 搭載の翻訳エンジンの学習データに含まれていない用語（未知語）は、MTP によって正確に翻訳することは困難である。したがって、新技術用語等の未知語について、MTP 搭載の翻訳エンジンに辞書登録又は学習させ、機械翻訳の品質低下を防止する取組みが必要である。

本事業では、今後、継続的かつ効率的に MTP の翻訳品質向上を行うため、新技術用語に代表される未知語の中日対訳辞書／対訳コーパスを一定数作成することにより、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分析を行った。これを通じて、MTP の誤訳を防止するために取り組むべき課題について考察した。

本報告書は、（1）未知語を調査し、翻訳対象語として選定された用語の分類、（2）翻訳対象語が出現した公報の書誌情報を基に翻訳対象語の出現傾向の調査、（3）これら分類・調査結果を踏まえ、今後、継続的かつ効率的に MTP の翻訳品質を向上させるための課題について検討結果を報告するものである。2 章では、対訳辞書及び対訳コーパスの作成について述べる。3 章では、翻訳対象語の出現傾向について述べる。4 章では、対訳辞書又は対訳コーパスを作成すべき翻訳対象語を効率的に選定する方法について述べる。5 章では、翻訳校閲者による効果的又は効率的な校閲方法について述べる。6 章では、本事業で作成した辞書を MTP に適用した結果について述べる。7 章では、MTP の翻訳品質を向上させるための課題についてまとめる。

表 1-1　用語の定義

用語	定義
MTP	機械翻訳プラットフォーム
未知語	機械翻訳システムが翻訳できなかった語句のこと。技術革新等により登場した新技術用語のほか、形態素解析での誤解析等により翻訳できない文字列も含む。
翻訳対象候補	中日の対訳辞書に採用する候補とする用語（原語）の集合。翻訳ログから機械的に不要なログを除外したもの。

用語	定義
翻訳対象語	中日の対訳辞書に採用する用語として、翻訳対象候補から優先順位に従って選定された、中日 6,000 語（原語）なお、本来の単語として適切ではない単位で出力された未知語は辞書登録単語として適切な単位（単語）に修補した。
翻訳対象テキスト	未知語が検出された中国語原文のこと。
翻訳ログ	MTP への翻訳要求のうち、未知語情報のみを抽出したログ。 未知語、翻訳対象テキスト、未知語が検出された公報の識別番号等が対応づけられたログ。 本事業開始時に中日翻訳の翻訳ログを受領した。（約 30 万ログ）
形態素解析	自然言語で書かれた文を単語に分割し、品詞情報を判別すること。MTP のニューラル機械翻訳エンジンでは、入力原文の単語分割のみを行っており、品詞情報は判別しない。

2. 対訳辞書／対訳コーパスの作成

本章では、対訳辞書及び対訳コーパスの作成について述べる。特許庁から貸与された翻訳ログ（約 30 万件）から、6,000 語の翻訳対象語を選定し、これを中日翻訳することにより対訳辞書を作成し、さらに翻訳対象語を含む 6,000 文を中日翻訳することにより対訳コーパスを作成した。翻訳ログとは、MTP に対して中国特許公報の一部の文章（「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（表 2-1 参照）。

表 2-1　翻訳ログ（イメージ）

識別番号	翻訳対象テキスト	未知語
CN1020210005378 971631316359938 0000748125	可治疗肝风夹痰，惊痫抽搐，小儿急惊风，破伤风，中风口㖞，风热头痛，目赤咽痛，风疹瘙痒，发颐疟腮等。	㖞
CN1020210005988 851631131516312 0000515901	c）选择“实时模式”，温度传感器将环境温度信号发送至 PLC，PLC 计算系统无因次炯利润率，将计算结果发送至触控显示屏并进行判定，若无因次炯利润率大于 0，系统开关闭合；否则系统开关断开；每隔 n 小时再次进行判定，直至更改模式。	炯
CN2020200033056 681631068912126 0001769831	其中，冷却箱 1 的侧壁开设有出渣口 10，出渣口 10 与出渣斗 11 对应，具体的，出渣口 10 的底沿与拨料盘 82 的底边和出渣斗 11 内壁料道的顶沿平齐，即可出渣口 10 将料渣导入出渣斗 11 内。	底沿

対訳辞書及び対訳コーパス作成の概略を図 2-1 に示す。以下では、翻訳ログから対訳辞書及び対訳コーパスを作成する具体的な手順について説明する。

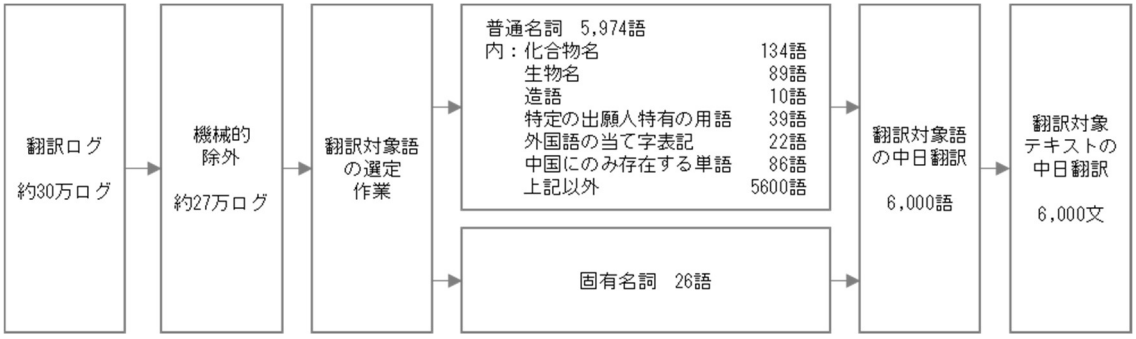

図 2-1　対訳辞書及び対訳コーパス作成の概略

2.1. 翻訳対象候補の決定

2.1.1. 翻訳ログの機械的除外

(1) 重複行の除外

翻訳ログは翻訳要求単位で記録されており、同一公報・同一翻訳対象テキストが複数行含まれている。そのため、重複する翻訳対象テキストを機械的に除外した。表 2-2 はこのような翻訳ログの例である。

表 2-2 重複行の例

識別番号	翻訳対象テキスト	未知語	翻訳要求日時
CN1120190000890371631089305468 0001819280	（254）5-｛9-氟-5-［（1S）-1-（哌啶-4-基）乙基］-2,3,4,5-四氢-1,5-苯并氧氮杂䓬-7-基｝-1,3,4-噁二唑-2（3H）-酮。	䓬	2021/09/08 17:30:07
CN1120190000890371631089305468 0001819280	（254）5-｛9-氟-5-［（1S）-1-（哌啶-4-基）乙基］-2,3,4,5-四氢-1,5-苯并氧氮杂䓬-7-基｝-1,3,4-噁二唑-2（3H）-酮。	䓬	2021/09/08 18:11:21

(2) 漢字以外を含む未知語の除外

未知語には、中日の対訳辞書の見出し語としてふさわしくない、記号、英数字又はこれらの組合せからなるものが存在するため、これらを機械的に除外した。表 2-3 は、このような未知語の例である。

表 2-3 漢字以外を含む未知語の例

翻訳対象テキスト	未知語
表示运算符(例如，‘&’、‘+’)的 IR 的每个计算操作被映射到实施该运算符的一个或多个 LUT 上。	’&’、’+’)
利用公式 XHu＝(Gu×c1+Tu×c2)/(c1+c2)计算得到变频设备的损耗值 XHu，式中 c1 和 c2 均为比例系数固定数值，且 c1 和 c2 的取值均大于零；	Gu×c

2.1.2. 優先順位づけ

「2.1.1. 翻訳ログの機械的除外」までの作業で抽出した翻訳ログについて、「高い頻度で参照されやすい用語」、「文献を理解する上で重要な用語」及び「最新の未知語」を選定するため、以下の優先順位で翻訳対象語の選定作業の優先度をつけた。

- 第一優先項目：複数文献における出願回数が多いもの
- 第二優先項目：同一文献における出現回数が多いもの
- 第三優先項目：未知語が検出された公報の出願日が新しいもの

2.2. 翻訳対象語の選定

2.2.1. 未知語の修補

MTP で採用しているニューラル機械翻訳エンジンでは、入力された原文が複数の単語に分割（以下、「形態素解析」という。）されてそれ以降の翻訳処理が行われる。この単語分割は、エンジンがあらかじめ持っている単語集合（辞書）によって行われるため、辞書に含まれない未知語については、本来の単語として適切ではない単位で分割され、未知語として検出されることがある。このため、検出された未知語について、翻訳対象テキストの文脈を目視で確認し、辞書登録単語として適切な単位（単語）に修補した。表 2-4 は、このような未知語の例である。

表 2-4　未知語の修補の例

翻訳対象テキスト	未知語	未知語の修補	参考：和訳
3. 根据权利要求 1 所述的评估方法，其中，在所述步骤 2 中，采用**德尔菲法**确定分区的权重。	尔菲	德尔菲法	デルファイ法
或/与, 在形成所述第二效应氧化层的步骤中, 所述第二效应氧化层具体为**栅氧化层**, 以热氧化或热氧化加上淀积方式形成所述**栅氧化层**于所述沟槽剩余空间的内壁与所述处理表面上;	栅	栅氧化层	ゲート酸化層

2.2.2. 不要語の除外

本事業の対訳辞書は MTP の新技術用語などへの対応を目的としており、人名、法人名、製品名等の固有名詞以外の普通名詞を優先的に辞書登録すべきと考えた。また、ニューラル機械翻訳の辞書登録において、名詞以外の単語や単語単位が不適切な語を登録した場合、適切に訳文に反映されない、又は翻訳品質が悪化する可能性があることが知られている。

上記を踏まえて、以下の条件に合致する未知語を翻訳対象語候補から除外した。

- 普通名詞以外の単語
- 人名、法人名、製品名等の固有名詞[1]
- 原文の不適切な改行に起因して発生したもの
- それ自体で意味をなさない、又はそれの意味が理解できないもの
- 既存辞書との重複

表 2-5 は固有名詞の例と名詞以外の単語の例である。

[1] ただし、中日特許翻訳者が重要と判断した 26 語については、除外せず翻訳対象用語として選定した。

表 2-5　固有名詞の例と名詞以外の単語の例

翻訳対象テキスト	未知語	除外理由
值得注意的是，本实施例中所公开的控制开关组 3 控制电机 82 工作采用现有技术中常用的方法，电机 82 可选用东莞市威邦机电有限公司型号为 51K120RGN-CF 型电机。	威邦	法人名（东莞市威邦机电有限公司）のため除外
半靴体的内部具有弹性垫层，不会硌伤患者的脚部，提高使用的舒适性。	硌伤	動詞（傷つける）のため除外

2.2.3. 未知語の重複除外

「2.2.2. 不要語の除外」までの作業で抽出した翻訳ログについて、未知語の重複除外を行った。

2.2.4. 翻訳対象語 6,000 語の選定

「2.2.3. 未知語の重複除外」までの作業で抽出した翻訳ログから翻訳対象語 6,000 語を選定した。選定した翻訳対象語の語数は表 2-6 のとおりである。

表 2-6　翻訳対象語種別ごとの語数

普通名詞／固有名詞	語数
普通名詞	5,974
固有名詞	26
合計	6,000

2.3. 翻訳対象語の翻訳

「2.2. 翻訳対象語の選定」までで選定した 6,000 語の翻訳対象語の中日翻訳を行い、対訳辞書を作成した。翻訳結果の例を表 2-7 に示す。

表 2-7　翻訳対象語の翻訳結果の例

翻訳対象テキスト	翻訳対象語	翻訳結果
5.根据权利要求 3 所述的一种用于水下潜航器的灯光指示系统，其特征在于，还包括：	水下潜航器	水中潜水艦
(3)预先将 PCR 仪降至 16℃，将上述反应管转移至 PCR 仪上，按以下程序进行反转录：16℃，30min；42℃，30min；85℃，5min；4℃forever。	PCR 仪	PCR 機器
安卓框架中包括访问控制、权限管理以及 UID 和权限对应表。	安卓框架	Ａｎｄｒｏｉｄフレームワーク

2.4. 翻訳対象テキストの翻訳

「2.2. 翻訳対象語の選定」までで選定した 6,000 語の翻訳対象語を含む翻訳対象テキストの人手による中日翻訳を行い、対訳コーパスを作成した。翻訳対象テキストに誤記が存在する場合は、誤記を修正して翻訳を行った。翻訳結果の例を表 2-8 に示す。

表 2-8　翻訳結果の例

翻訳対象テキスト	翻訳対象テキストの修補	翻訳結果
进一步地，所述弹性囊为橡胶材质，所述弹性囊靠近空心薄板一侧呈半球型，所述弹性囊远离空心薄板一侧呈向内凹陷型，多个所述散热条材质为环氧树脂**氧树脂**，所述弹性囊靠近转子线圈一侧的凹面呈凸梯型，内凹陷圆型排除与线槽内壁之间的空气相互吸合，增加防甩包装置对转子线圈的稳定性。	进一步地，所述弹性囊为橡胶材质，所述弹性囊靠近空心薄板一侧呈半球型，所述弹性囊远离空心薄板一侧呈向内凹陷型，多个所述散热条材质为环氧树脂，所述弹性囊靠近转子线圈一侧的凹面呈凸梯型，内凹陷圆型排除与线槽内壁之间的空气相互吸合，增加防甩包装置对转子线圈的稳定性。	さらに、前記弾性バッグはゴム製であり、前記弾性バッグの中空シートに近い側は半球形であり、前記弾性バッグの中空シートから離れた側は内側に凹んでおり、複数の前記放熱ストリップはエポキシ製であり、前記弾性バッグの回転子コイルに近い側の凹面は、凸台形であり、内側凹状円形は配線溝の内壁との間の空気が排除されて相互に吸着し、それによって振り防止装置による回転子コイルに対する安定性を高める。
面膜基布层的表面还设置有储囊，储囊设置在鼻部开口的外缘和眼部开口的内**预**缘之间，储囊内填充有天然矿物粘土泥浆，储囊靠近天然矿物粘土层的一侧设置有易开口；	面膜基布层的表面还设置有储囊，储囊设置在鼻部开口的外缘和眼部开口的内缘之间，储囊内填充有天然矿物粘土泥浆，储囊靠近天然矿物粘土层的一侧设置有易开口；	マスク基布層の表面に収納バッグがさらに設けられ、収納バッグが鼻部開口の外縁と眼部開口の内縁との間に設けられ、収納バッグ内に天然鉱物粘土モルタルが充填されており、収納バッグの天然鉱物粘土層に近い側に開けやすい開口が設けられる。
路面体按结构层次自上而下可分为面层、基层、垫层或联结层等，桥隧工程：**桥隧工程：**是高等级公路中的重要组成部分。	路面体按结构层次自上而下可分为面层、基层、垫层或联结层等，桥隧工程：是高等级公路中的重要组成部分。	路面体は、構造階層に基づいて上から下まで表面層、基層、クッション層又は連結層等に分けられることができ、橋とトンネル工事はハイグレード道路では重要な構成要素である。

3. 翻訳対象語の出現傾向

本章では、2 章で選定した翻訳対象語の出現傾向について述べる。翻訳対象語が示す対象物の種類及び翻訳対象語の成り立ち（語源）などを特徴とした種別に分類した上で、技術分野や出願人の観点で出願傾向を調査・分析した。

3.1. 翻訳対象語の種別の調査

本事業で選定した翻訳対象語を表 3-1 に示す種別に分類した。表 3-1 に翻訳対象語種別ごとの語数、表 3-2 に翻訳対象語種別ごとの翻訳対象語の例を示す。なお、「その他」は、どの種別にも当てはまらない単語である。

表 3-1 に示すとおり、「その他」の出現頻度が最も高く、次に、「化合物名」、「生物名」、「中国語にのみ存在する単語」が高い傾向である。

表 3-1　翻訳対象語種別ごとの語数

翻訳対象語種別	語数
化合物名	134
生物名	89
造語	10
特定の出願人特有の用語	39
外国語の当て字表記	22
中国にのみ存在する単語	86
その他	5,594
固有名詞	26
合計	6,000

表 3-2　翻訳対象語種別の例

翻訳対象語の種別	翻訳対象語	翻訳結果	備考
化合物名	聚芳醚砜	ポリアリールエーテルスルホン	－
生物名	篱边粘褶菌	キカイガラタケ	－
造語	银企	銀行と企業	銀行を表す「银行」と企業を表す「企业」を合わせた造語。
特定の出願人特有の用語	磷浮选给矿	リン浮選供給鉱	中南大学の出願に見られる用語。
外国語の当て字表記	默克尔树	メルケルツリー	「メルケル」を当て字表記とした用語。

翻訳対象語の種別	翻訳対象語	翻訳結果	備考
中国にのみ存在する単語	白虎历节	関節リウマチ	中国医学で使われる用語。類義語である「関節リウマチ」を訳語とした。
その他	午饭 货仓 烘干仓	昼食 倉庫 乾燥室	種別に分類できない一般的な単語。学習データの不足や翻訳エンジンの問題により、形態素解析の誤解析や文章の依存関係を適切に捉えられなかったことで出力されたと推測される。
固有名詞	景岳全书	景岳全書	－

3.2. 技術分野ごとの翻訳対象語数

次に、技術分野（IPC 分類）における翻訳対象語の出現傾向を調査した。IPC 分類（A～H セクション）における各種別の翻訳対象語の出現頻度を表 3-3 に示す。B セクション（処理操作;運輸）が 1,426 件と最も多く、次いで A セクション（生活必需品）が 1,270 件、G セクション（物理学）が 1,099 件となり、上位 3 セクションの合計で全体の 63%を占めている。

表 3-3　技術分野ごとの翻訳対象語数

セクション	翻訳対象語の種別の件数								合計
	化合物名	生物名	造語	特定の出願人特有の用語	外国語の当て字表記	中国にのみ存在する単語	その他	固有名詞	
A	28 (21%)	52 (58%)	0 (0%)	15 (38%)	7 (32%)	55 (64%)	1,108 (20%)	5 (19%)	1,270 (21%)
B	13 (10%)	5 (6%)	3 (30%)	7 (18%)	1 (5%)	9 (10%)	1,388 (25%)	0 (0%)	1,426 (24%)
C	62 (46%)	28 (31%)	0 (0%)	3 (8%)	5 (23%)	2 (2%)	500 (9%)	6 (23%)	606 (10%)
D	5 (4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (5%)	108 (2%)	0 (0%)	117 (2%)
E	0 (0%)	0 (0%)	1 (10%)	2 (5%)	0 (0%)	3 (3%)	621 (11%)	2 (8%)	629 (10%)
F	1 (1%)	1 (1%)	0 (0%)	1 (3%)	0 (0%)	0 (0%)	363 (6%)	0 (0%)	366 (6%)
G	10 (7%)	2 (2%)	6 (60%)	9 (23%)	5 (23%)	13 (15%)	1,043 (19%)	11 (42%)	1,099 (18%)
H	15 (11%)	1 (1%)	0 (0%)	2 (5%)	4 (18%)	0 (0%)	463 (8%)	2 (8%)	487 (8%)
合計	134 (100%)	89 (100%)	10 (100%)	39 (100%)	22 (100%)	86 (100%)	5,594 (100%)	26 (100%)	6,000 (100%)

(1) 化合物名

化合物名の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 28 クラスで、出現件数上位 10 クラスを表 3-4 に示す。C セクション（化学；冶金）以外にも、A61（医学または獣医学；衛生学）が 19 件、H01（基本的電気素子）が 14 件であり、上位 3 位で 37%を占めた。

表 3-4 化合物名の出現傾向（上位 10 クラス）

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	A61	医学または獣医学；衛生学	19	14%	度伐鲁单抗 怀俄昔 哈巴俄昔	デュルバルマブ ワイオシン ハルパゴシド
2	C07	有機化学	17	13%	二甲基砜 奥泰朱单抗 阿特利朱单抗	ジメチルスルホン オテリキシズマブ アテゾリズマブ
3	H01	基本的電気素子	14	10%	碲镉汞 氢氧化钛 氢氧化镍	テルル化カドミウム水銀 水酸化チタン 水酸化ニッケル
4	C09	染料；ペイント；つや出し剤；天然樹脂；接着剤；他に分類されない組成物；他に分類されない材料の応用	13	10%	苯四酸二酐 碳化硼 油胺	ベンゼンテトラカルボン酸二無水物 炭化ホウ素 オレイルアミン
5	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	7%	卡卢酸钾 聚芳醚砜 多胺	ペルオキシ一硫酸カリウム ポリアリールエーテルスルホン ポリアミン
6	B01	物理的または化学的方法または装置一般	9	7%	格蓬酯 四氢苯酐 顺酐	アリル・アミル・グリコール酸塩 テトラヒドロフタル酸無水物 無水マレイン酸
7	G01	測定；試験	8	6%	氟砜灵 邻苯二酚 乙醛	フルエンスルホン ピロカテコール アセトアルデヒド
8	C12	生化学；ビール；酒精；ぶどう酒；酢；微生物学；酵素学；突然変異または遺伝子工学	7	5%	噬铁素 木糖苷 D-甘油醛	フェロトーシス グルコシド Ｄ－グリセルアルデヒド

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
9	A01	農業；林業；畜産；狩猟；捕獲；漁業	6	4%	吡蚜酮 亚磷酸钾 硫酸锰	ピメトロジン 亜リン酸カリウム 硫酸マンガン
10	C01	無機化学	4	3%	二硒化钒 铯铅卤素 钛酸锂镧	バナジウムジセレニド セシウム鉛ハロゲン チタン酸リチウムランタン

(2) 生物名

生物名の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 12 クラスで、出現状況を表 3-5 に示す。A01（農業；林業；畜産；狩猟；捕獲；漁業）が全体の 42%を占めた。

表 3-5 生物名の出現傾向（クラス別）

No.	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	A01	農業；林業；畜産；狩猟；捕獲；漁業	37	42%	大叶黄杨叶 准噶尔 灰葡萄孢	マサキの葉 ジュンガル ボトリチスシネレア
2	C12	生化学；ビール；酒精；ぶどう酒；酢；微生物学；酵素学；突然変異または遺伝子工学	17	19%	食源性致病寄生虫 罗尔伏革菌 卢克诺文思金孢子菌	食物媒介病原性寄生虫 白絹病菌 クリソスポリウムラクノウェンス
3	C07	有機化学	10	11%	似食酪螨 嗜粘阿克曼菌 单孢锈菌属	ホウレンソウケナガコナダニ アッカーマンシア・ムシニフィラ さび病属
4	A61	医学または獣医学；衛生学	9	10%	冀地鳖 脱氮嗜脂环物菌 奴卡菌	ステレオファーガブランシー アリシクリフィラス・デニトリフィカン ノカルジア
5	A23	食品または食料品；他のクラスに包含されないそれらの処理	6	7%	赤苍藤 野艾蒿 蝉花	エリスロパラムスカンデンス ヒメヨモギ セミタケ
6	B01	物理的または化学的方法または装置一般	2	2%	复端孢菌 香樟	セファロテシウム クスノキ

No.	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
7	B65	運搬；包装；貯蔵；薄板状または線条材料の取扱い	2	2%	菱瓜 黄褐毛忍冬	ウシノシイ 黄褐毛ニンドウ
8	G06	計算または計数	2	2%	米氏凯伦藻 灰飞虱	カレニアミキモトイ ヒメトビウンカ
9	B02	破砕，または粉砕；製粉のための穀粒の前処理	1	1%	蕈菌	キノコ
10	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	1	1%	岩藻聚糖	フコイダン

(3) 造語

造語の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 5 クラスで、出現状況を表 3-6 に示す。G セクション（物理学）が全体の 60%を占めた。

表 3-6 造語の出現傾向（クラス別）

No.	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	G06	計算または計数	5	50%	银企 侦控打 侦控打评	銀行と企業 偵察制御打撃能力 偵察制御打撃評価
2	B65	運搬；包装；貯蔵；薄板状または線条材料の取扱い	2	20%	兰熏 淘宝机	蘭薫 淘宝業務
3	B08	清掃	1	10%	侦控	偵察制御
4	E04	建築物	1	10%	梯跑	ラダーランニング
5	G02	光学	1	10%	纬排线	緯フラットケーブル

(4) 特定の出願人特有の用語

特定の出願人特有の用語の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 39 クラスで、出現件数上位 10 クラスを表 3-7 に示す。全体的に分野ごとの傾向はみられない。

表 3-7　特定の出願人特有の用語の出現傾向（上位 10 クラス）

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	A61	医学または獣医学；衛生学	5	13%	前小叶 药坨 匙羹	前尖 薬物の塊 スプーン
2	A63	スポーツ；ゲーム；娯楽	4	10%	火攻策略 普攻策略 选择和棋按键	火による攻撃戦略 一般的な攻撃戦略 持碁選択ボタン
2	G06	計算または計数	4	10%	有序告警对 硬围攻策略 虾沟	順序付けられた警報対 強行包囲攻撃戦略 エビ溝
4	A47	家具；家庭用品または家庭用設備；コーヒーひき；香辛料ひき；真空掃除機一般	3	8%	粪尿分离式旱厕 五子棋按键 军棋按键	糞尿分離型乾式便所 五目並べボタン 軍棋ボタン
4	G01	測定；試験	3	8%	个人辐射沾污监测模式 组织提篮拉杆 先韧后脆特征	個人用放射線汚染監視モード 組織手かごプルロッド 最初に強靭で次に脆いという特性
6	B03	液体による，または，風力テーブルまたはジグによる固体物質の分離；固体物質または流体から固体物質の磁気または静電気による分離，高圧電界による分離	2	5%	磷浮选给矿 磷浮选尾矿	リン浮選供給鉱 リン浮選尾鉱
6	B23	工作機械；他に分類されない金属加工	2	5%	托爪让位缺口 珠盖	支持ジョー逃がし切り欠き ボールカバー
6	C04	セメント；コンクリート；人造石；セラミックス；耐火物	2	5%	铆结 铆结力	リベット留め リベット留め力
6	G05	制御；調整	2	5%	相对姿态律 参数更新律	相対姿勢法則 パラメータ更新法則
6	H01	基本的電気素子	2	5%	暖宝 电池暖宝	ウォーマー バッテリー式ウォーマー

(5) 外国語の当て字表記

外国語の当て字表記の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 10 クラスで、出現状況を表 3-8 に示す。A61（医学または獣医学；衛生学）における出現頻度が高い傾向が見られた。

表 3-8 外国語の当て字表記の出現傾向（クラス別）

No.	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	A61	医学または獣医学；衛生学	7	32%	非勃发朋 他拉唑帕尼 培西达替尼	フィブロネクチン タラゾパニブ ペシダチニブ
2	C07	有機化学	3	14%	奎扎替尼 悦色盘长孢 替沃扎尼	ケザチニブ 炭疽病 チボザニ
3	G06	計算または計数	3	14%	安卓运行环境 梅克尔树 默克尔树根	Android 動作環境 メルケルツリー メルケルツリールート
4	H04	電気通信技術	3	14%	逆威沙特 默克尔树 梅克尔树根	逆ウィシャート メルケルツリー メルケルツリールート
5	B60	車両一般	1	5%	珮珀尔幻象	ペッパーズゴースト
6	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	1	5%	寇夫醛	コアボーンアルデヒド
7	C12	生化学；ビール；酒精；ぶどう酒；酢；微生物学；酵素学；突然変異または遺伝子工学	1	5%	萨纳瑞西斯假丝酵母	カンジダソニアヤエンシス
8	G08	信号	1	5%	哈达玛积	アダマール積
9	G16	特定の用途分野に特に適合した情報通信技術［ＩＣＴ］	1	5%	德布莱英图	DeBrying ダイアグラム
10	H03	基本電子回路	1	5%	博德-范诺法	ボードーファノ法

(6) 中国にのみ存在する単語

中国にのみ存在する単語の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 21 クラスで、出現件数上位 9 クラスを表 3-9 に示す。A セクション（生活必需品）が大部分を占め、A61 クラス（医学または獣医学；衛生学）が 36 件と全体の 42%を占めている。

表 3-9 中国のみに存在する単語の出現傾向（上位 9 クラス）

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	A61	医学または獣医学；衛生学	36	42%	症候评分 白虎历节 寒饮喘咳	症候群採点 関節リウマチ 寒飲によるあえぎと咳
2	G06	計算または計数	9	10%	中院 国庆 抗疫保电	中級人民法院 国慶節 防疫電力保持
3	A23	食品または食料品；他のクラスに包含されないそれらの処理	8	9%	奔豚 寒泻 烫炙	賁豚 冷え性下痢 温灸
4	A01	農業；林業；畜産；狩猟；捕獲；漁業	5	6%	宋代 西藏四宝 构菌	宋王朝 チベットの四宝 構茸
4	B65	運搬；包装；貯蔵；薄板状または線条材料の取扱い	5	6%	巷子 米筛花 麻糍泻	路地 米篩花 麻糍瀉
6	A47	家具；家庭用品または家庭用設備；コーヒーひき；香辛料ひき；真空掃除機一般	3	3%	祭祖仪式 公筷公勺 花饽饽	祖先崇拝儀式 取り箸とサービススプーン 蒸し菓子
6	D02	糸；糸またはロープの機械的な仕上げ；整経またはビーム巻き取り	3	3%	湘绣 粤绣 蜀绣	湘繍 広東刺繍 四川刺繍
6	E04	建築物	3	3%	祭祖区 祭祖亭 祭祖联	祖先崇拝エリア 祖先崇拝あずまや 祖先崇拝対聯
9	G01	測定；試験	2	2%	胸痞胀满 驿道	胸痞膨満 駅道

(7) その他

その他の単語の出現傾向を IPC 分類のクラスごとに調査した。出現したクラスは全 116 クラスで、出現件数上位 10 クラスを表 3-10 に示す。A セクション（生活必需品）、B セクション（処理操作；運輸）、E セクション（固定構造物）、G セクション（物理学）で 74%を占めている。クラスについては G06（計算または係数）、A01（農業；林業；畜産；狩猟；捕獲；漁業）、A61（医学または獣医学；衛生学）など、上記(1)~(6)の観点でも 1 位に該当するクラスが上位に来ているものの、G08 の 8%が最大となっており、技術分野に大きな偏りはみられない。

表 3-10　その他の単語の出現傾向（上位 10 クラス）

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
1	G06	計算または計数	434	8%	丧偶 样品 全仓	配偶者死亡 サンプル 全室
2	A01	農業；林業；畜産；狩猟；捕獲；漁業	418	7%	燃烧仓 井窖 疏伐前	燃焼室 井戸 間伐前
3	G01	測定；試験	382	7%	样本 缝洞 称重仓	試料 スリット 計量ビン
4	A61	医学または獣医学；衛生学	350	6%	饭后 储液仓 血氧仪	食後 液体貯蔵ビン オキシメータ
5	B65	運搬；包装；貯蔵；薄板状または線条材料の取扱い	210	4%	智能仓 粮仓 放置仓	スマート倉庫 穀倉 収容室
6	H04	電気通信技術	166	3%	诊断仪 夜视仪 公厕	診断機器 暗視装置 公衆トイレ
7	H01	基本的電気素子	161	3%	侦查仪 电势 弹力囊	検出器 電位 弾性カプセル
8	E02	水工；基礎；土砂の移送	160	3%	矩形井 液压坝 混凝土坝	長方形井戸 油圧ダム コンクリートダム

No	クラス	クラスの説明	件数	割合	翻訳対象語の例	翻訳対象語の和訳
9	B01	物理的または化学的方法または装置一般	151	3%	存储仓 密封仓 泥污	貯蔵室 密封室 泥汚れ
10	B23	工作機械；他に分類されない金属加工	140	3%	出料仓 巷道 刻度尺	排出ビン 坑道 スケール

3.3. 出願人ごとの翻訳対象語数

次に出願人における翻訳対象語の出現傾向を調査した。出願人の上位 10 者を表 3-11 に示す。

表 3-11　出願人別の出現傾向（上位 10 者）

No.	出願人名	個人／法人	翻訳対象語数 (6000 語中の割合)
1	中南大学	法人	20(0.33%)
2	武汉大学	法人	14(0.23%)
3	江苏富路建设有限公司	法人	13(0.22%)
4	阿里巴巴集团控股有限公司	法人	11(0.18%)
5	浙江大学	法人	11(0.18%)
6	天津大学	法人	10(0.17%)
7	广东工业大学	法人	10(0.17%)
8	腾讯科技（深圳）有限公司	法人	10(0.17%)
9	安徽农业大学	法人	9(0.15%)
10	中国农业大学	法人	9(0.15%)

3.3.1. 個人又は法人の別

出願人を個人と法人で分類し、翻訳対象語 6000 語と翻訳ログについての出現傾向を調査した。個人又は法人別の翻訳対象語数を表 3-12 に示す。

表 3-12　個人／法人別の出現傾向

個人／法人	翻訳対象語数 （割合）	翻訳ログの件数[2] （割合）
法人	5,389(90%)	281,524(91%)
個人	611(10%)	26,211(9%)

[2] DOCDB から出願人情報が取得できなかった 3,202 件は除く。

表に示すとおり、翻訳ログ中の個人又は法人の別の割合と、翻訳対象語 6000 語における個人又は法人の別の割合の傾向は同じとなっていた。

次に出願人の個人又は法人の別と IPC セクション関連を図 3-1 に示す。

個人の出願人の場合、A セクション（生活必需品）、B セクション（処理操作；運輸）、E セクション（固定構造物）、G セクション（物理学）が上位 4 セクションであり、84%を占めている。また、法人の出願人の場合、B セクション、A セクション、G セクション、E セクションの上位 4 セクションで 73%を占めている。

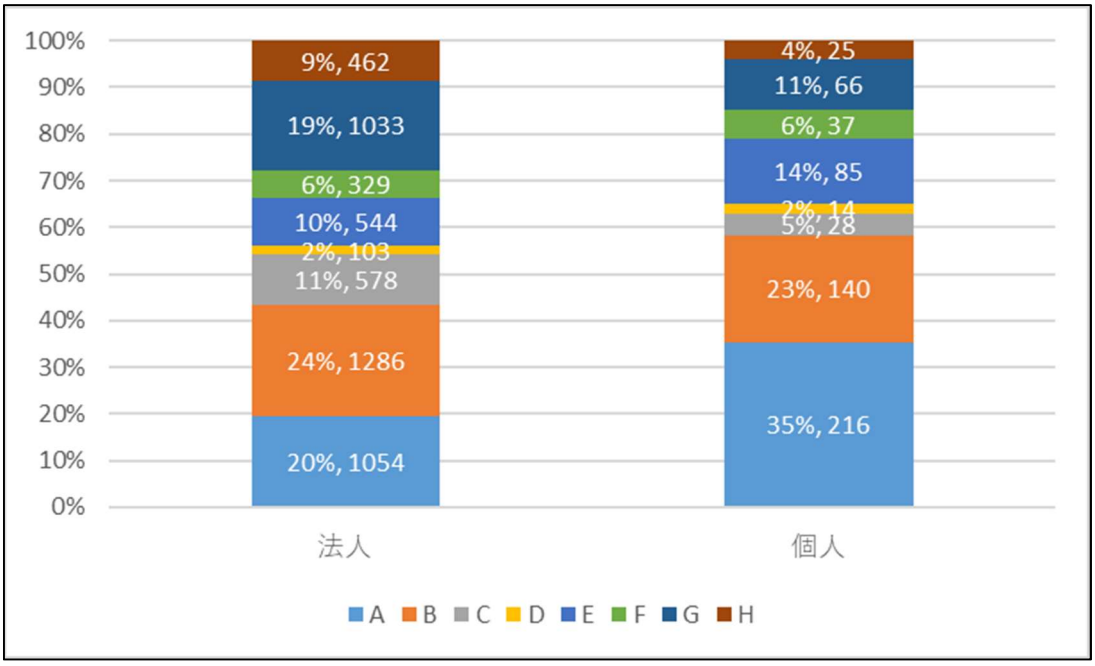

図 3-1　個人／法人の別と IPC セクションの関連

3.3.2. 優先権主張の国又は地域別

優先権主張の国又は地域[3]別の出願傾向を調査した。優先権主張の国又は地域別の翻訳対象語数を表 3-13 に示す。中国が 5,756 語と最も多く、次いでアメリカ合衆国が 121 語、欧州特許庁が 36 語、日本が 35 語となり、上位 4 つの国又は地域で 99%を占める。

[3] DOCDB の<exch:priority-claims>タグ内の<country>タグから抽出したデータを優先権主張の国又は地域として分析した。そのため、この報告書では中国のみに出願されたものを、優先権主張の国又は地域が中国であると表記する。

表 3-13　優先権主張の国又は地域別の出現傾向

No.	優先権主張の国又は地域	件数 (6000 語中の割合)	件数 （法人）	件数 （個人）
1	中国（CN）	5756(95.93%)	5153	603
2	アメリカ合衆国（US）	121(2.02%)	117	4
3	欧州特許庁（EP）	36(0.60%)	36	0
4	日本（JP）	35(0.58%)	35	0
5	連合王国（イギリス）（GB）	11(0.18%)	11	0
6	大韓民国（KR）	9(0.15%)	8	1
7	ドイツ（DE）	6(0.10%)	6	0
8	フランス（FR）	5(0.08%)	5	0
9	台湾、中華民国（TW）	5(0.08%)	4	1
10	オーストラリア（AU）	4(0.07%)	4	0

3.3.3. 出願人の属性別

語数の多い出願人の上位 10 者の属性別の出願傾向を分析した。大学（副部級大学や国家重点大学、公立大学)、企業（大企業、中小企業）に属す法人が上位を占め、頻出セクションは法人ごとに特性がみられた。しかし、出願人別に翻訳対象語に占める割合をみると最も多い中南大学で 0.33%であり、出願人による偏りはみられない。表 3-14 に上位 10 者の属性、翻訳対象語の語数、出願人別の頻出セクションの割合及び翻訳ログの件数を示す。また、表 3-15 に出願人の属性の説明を示す。

表 3-14　出願人上位 10 者の企業規模と件数

No.	出願人名	属性 （資本金）	翻訳対象語数 （割合）	頻出セクション （割合）	翻訳ログの件数（割合）
1	中南大学	副部級大学	20(0.33%)	B：35%	364(1.09%)
2	武汉大学	副部級大学	14(0.23%)	G：64%	478(1.43%)
3	江苏富路建设有限公司	中小企業 (1,000 万元)	13(0.22%)	E：100%	294(0.88%)
4	阿里巴巴集团控股有限公司	大企業 (15,298 万ドル)	11(0.18%)	G：64%	1,024(3.06%)
5	浙江大学	副部級大学	11(0.18%)	G：64%	413(1.24%)
6	天津大学	副部級大学	10(0.17%)	G：90%	474(1.42%)
7	广东工业大学	公立大学	10(0.17%)	A,C：30%	546(1.63%)
8	腾讯科技（深圳）有限公司	中小企業 (200 万ドル)	10(0.17%)	G：60%	797(2.38%)

No.	出願人名	属性 （資本金）	翻訳対象語数 （割合）	頻出セクション （割合）	翻訳ログの件数（割合）
9	安徽农业大学	国家重点大学	9(0.15%)	B：78%	130(0.39%)
10	中国农业大学	副部級大学	9(0.15%)	G：56%	143(0.43%)

※　頻出セクション：当該出願人の公報から選定した翻訳対象語のうち、頻出セクションとそのセクションの占める割合(%)

表 3-15　出願人の属性の説明

副部級大学	国家重点大学のうち、中国共産党中央により直接支配される大学のこと。
国家重点大学	権威ある大学であると中国政府が認定し、予算の優先配分などの支援を行うものとして、設置者の別を問わず選定された大学のこと。
公立大学	中国各州に設置された大学のこと。
大企業	資本金 3 億円超※
中小企業	資本金 3 億円以下※

※　中小企業庁の中小企業定義
（https://www.chusho.meti.go.jp/faq/faq/faq01_teigi.htm#q1）

4. 辞書又は対訳コーパスを作成すべき翻訳対象語を効率的に選定する方法の検討

本章では、翻訳対象語を効率的に選定する方法について述べる。3 章で調査・分析した翻訳対象語が発見されやすい技術分野や出願人を検討し、2 章で述べた翻訳対象語の選定作業を更に効率的に行うための方法を検討した。

4.1. 翻訳対象語が発見されやすい公報の分析

4.1.1. 翻訳対象語が発見されやすい技術分野の傾向

本事業で選定した翻訳対象語 6000 語を IPC セクションごとに調査・分析した結果を表 4-1 に示す。B セクション（処理操作；運輸）が 1,426 語と最も多く、次いで、A セクション（生活必需品）が 1,270 語、G セクション（物理学）が 1,099 語で、全体の 63%を占めている。翻訳ログに占める IPC セクションの割合と選定した翻訳対象語 6000 語の IPC セクションに大きなずれはなく、翻訳対象語の選定については提供された翻訳ログに含まれる IPC 情報の分布を反映した選定となっていた。

表 4-1　IPC セクションごとの翻訳対象語数

セクション	翻訳対象語数（割合）	翻訳ログの件数（割合）
A	1,270(21%)	69,354(22%)
B	1,426(24%)	59,533(19%)
C	606(10%)	38,699(12%)
D	117(2%)	7,205(2%)
E	629(10%)	25,160(8%)
F	366(6%)	17,176(6%)
G	1,099(18%)	66,558(21%)
H	487(8%)	27,252(9%)
合計	6,000(100%)	310,937(100%)

翻訳対象語を選定する際の各セクションの優先順位については、翻訳品質の改善方針に合わせて検討する必要がある。

方針１．すべてのセクションの翻訳品質の底上げ

翻訳ログにおけるセクションの比率と同じ割合で翻訳対象語を選定して辞書や対訳コーパスを整備することがすべてのセクションの翻訳品質を底上げする上で有効であ

る。これは、全セクションを同じ件数とした場合、翻訳ログに占める割合の少ないセクションについては品質向上の度合が相対的に高くなる可能性があるためである。

方針２．翻訳要求の絶対量が多いセクションの翻訳品質の向上

翻訳要求の絶対量（翻訳ログの件数）が多いセクションを優先して翻訳対象語を選定し、辞書やコーパスの整備をすることが当該セクションの翻訳品質を重点的に向上させる上で有効である。

翻訳対象種別ごとにみると、「3.1. 翻訳対象語の種別の調査」で示すとおり、「その他」の翻訳対象語数が 5,594 語と圧倒的に多かった。また、「その他」に分類された単語は、一般的に使われる単語であった。これは、学習データの不足により、形態素解析を行う際に参照される単語集合（辞書）に一般的に使われる単語が登録されていないことや、翻訳エンジンの問題により、文章の依存関係を適切に捉えられていないことが原因と考えられる。

「その他」を除く翻訳対象語数が多い翻訳対象種別は、化合物が 134 件と最も多く、次いで、生物名が 89 件、中国にのみ存在する単語が 86 件で、翻訳対象種別の大半を占めている。以降、それぞれの翻訳対象種別ごとの詳細について述べる。

(1) 化合物

化合物として出現した翻訳対象語をセクション別にみると表 3-3 に示すとおり、C セクション（化学；冶金）が 46%を占めるが、クラス別にみると、表 3-4 のとおり、A61（医学または獣医学；衛生学）が 19 件、C07（有機化学）が 17 件、H01（基本的電気素子）が 14 件であった。このため、化合物を含む文の翻訳品質を効率的に向上させるためには、C セクション、A61、H01 を優先することで効率的な辞書、コーパス整備を行うことができる。

(2) 生物名

生物名として出現した翻訳対象語をセクション別にみると表 3-3 に示すとおり、A セクション（生活必需品）が 58%を占め、かつ、クラス別にみても、表 3-5 のとおり、A01（農業；林業；畜産；狩猟；捕獲；漁業）が 41%を占めた。このため、生物名を含む文の翻訳品質を向上させるためには、A セクション、特に A01 を優先することで効率的な辞書、コーパス整備を行うことができる。

(3) 造語

造語として出現した翻訳対象語をセクション別にみると表 3-3 に示すとおり、G セクション（物理学）が 60%を占め、かつ、クラス別にみても表 3-6 のとおり、G06（計算または計数）が 50%を占めた。このため、造語を含む文の翻訳品質を向上させるためには、G セクション、特に G06 を優先することで効率的な辞書、コーパス整備を行うことができる。

(4) 特定の出願人特有の用語

特定の出願人特有の用語として出現した翻訳対象語をセクション別にみると表 3-3 に示すとおり、A セクション（生活必需品）が 38%、次いで G セクション（物理学）が 23%、B セクション（処理操作；運輸）が 18%と上位 3 セクションで 79%を占めている。

クラス別でみても、表 3-7 のとおり、本調査範囲においては大きな差はない。このため、特定の出願人特有の用語を含む文の翻訳品質を向上させるためには、A セクション、G セクション、B セクションを優先することで効率的な辞書、コーパス整備を行うことができる。

(5) 外国語の当て字表記

外国語の当て字表記の用語として出現した翻訳対象語をセクション別にみると表 3-3 に示すとおり、A セクション（生活必需品）が 32%、C セクション（化学；冶金）が 23%、G セクション（物理学）が 23%と上位 3 セクションで 78%を占めている。

クラス別でみても表 3-8 のとおり、A61（医学または獣医学；衛生学）が 32%、C07（有機化学）が 14%、G06（計算または計数）が 14%と A セクション、C セクション、G セクションのクラスが上位となっている。

このため、外国語の当て字表記を含む文の翻訳品質を向上させるためには、A セクション、C セクション、G セクションを優先することで効率的な辞書、コーパス整備を行うことができる。

(6) 中国のみに存在する単語

中国のみに存在する単語として出現した技術分野をセクション別にみると表 3-3 に示すとおり、A セクション（生活必需品）が 64%を占め、クラス別にみても、表 3-9 のとおり、A61（医学または獣医学；衛生学）が 42%を占めた。中国のみに存在する単語は表 4-2 に示すとおり、中国医学に関するものが多く、上記クラスにおける出現傾向が高い傾向にあると考えられる。このため、中国のみに存在する単語を含む文の翻訳品質を向上させるためには、A61 を優先することで効率的な辞書、コーパス整備を行うことができる。

表 4-2 中国のみに存在する単語の例

翻訳対象語	和訳
寒疝	寒疝（かんせん）
白虎历节	関節リウマチ
盐炙	塩灸（しおやいと）

4.1.2. 翻訳対象語が発見されやすい技術分野と個人又は法人の別の傾向

表 3-12 に示したとおり、翻訳対象語 6000 語の個人又は法人の別の割合と翻訳ログの個人又は法人の別の割合は個人が約 10%、法人が約 90%であり、翻訳対象語も翻訳ログも同様の傾向であった。

また、図 3-1 に示したとおり、個人の出願人の上位 4 セクションと法人の出願人の上位 4 セクションは同じ A セクション（生活必需品）、B セクション（処理操作；運輸）、E セクション（固定構造物）、G セクション（物理学）となっている。

このため、翻訳対象語の効率的な選定においては「4.1.1.翻訳対象語が発見されやすい技術分野の傾向」の方針にのっとることで出願人の個人又は法人の別は特段の考慮を要しないと考える。

4.1.3. 翻訳対象語が発見されやすい技術分野と国又は地域の傾向

(1) 技術分野（IPC セクション）と国・地域の関連

本事業で選定した翻訳対象語を発見した技術分野（IPC セクション）別の優先権主張の国・地域を図 4-1 に示す。

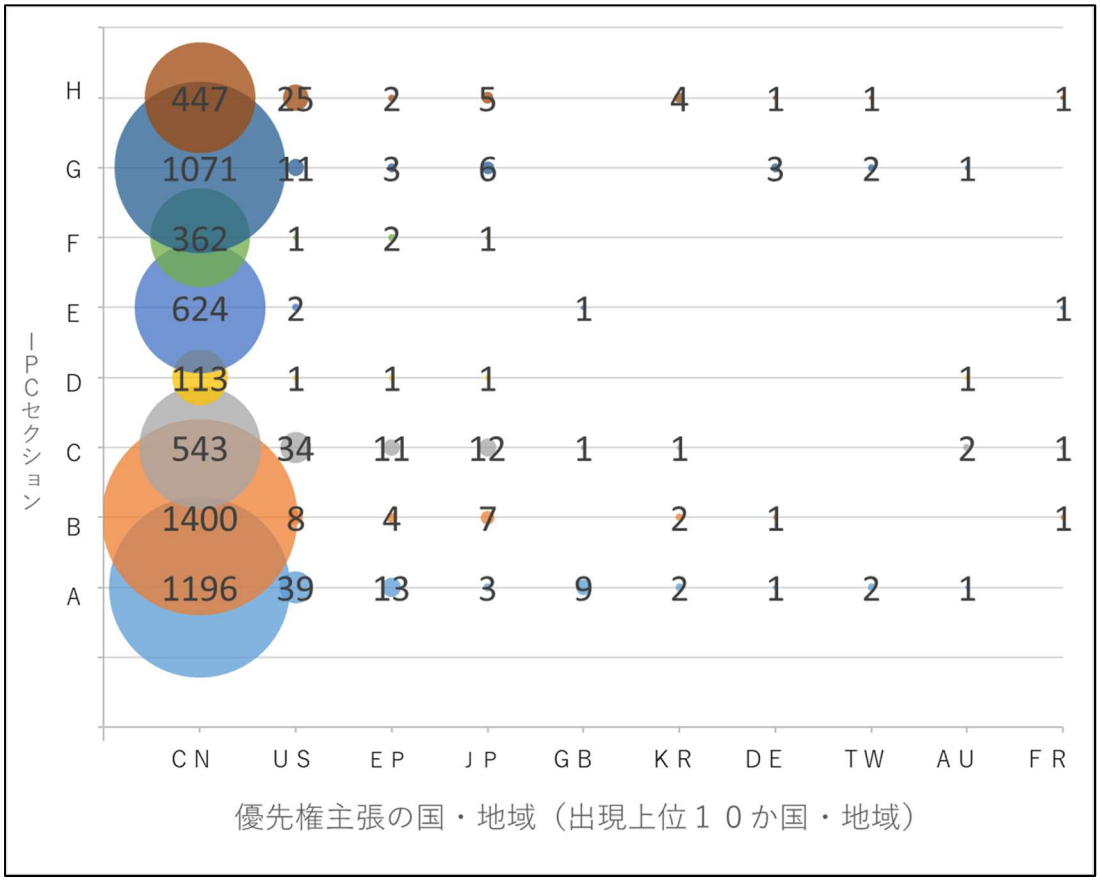

図 4-1 IPC セクションと優先権主張の国・地域の件数

優先権主張の国・地域が中国（CN）である公報が圧倒的に多く、中国（CN）内では A セクション（生活必需品）、B セクション（処理操作；運輸）、G セクション（物理学）で 63%を占めている。その他、上位 4 つの国・地域については A セクション、B セクション、C セクション（化学；冶金）、G セクション、H セクション（電気）が上位となっている。これ

らは翻訳ログにおける件数上位のセクションと一致していることから、国・地域を考慮したIPC セクションの優先順位についても「4.1.1. 翻訳対象語が発見されやすい技術分野の傾向」と同様に翻訳品質の改善方針に合わせて検討する必要がある。

方針１．すべてのセクションの翻訳品質の底上げ

翻訳ログにおけるセクションの比率と同じ割合で翻訳対象語を選定して辞書や対訳コーパスを整備することがすべてのセクションの翻訳品質を底上げする上で有効である。これは、全セクションを同じ件数とした場合、翻訳ログに占める割合の少ないセクションについては品質向上の度合が相対的に高くなる可能性があるためである。

方針２．翻訳要求の絶対量が多いセクションの翻訳品質向上

翻訳要求の絶対量（翻訳ログの件数）が多い、A セクション、G セクション、B セクション、C セクションを優先して翻訳対象語を選定し、辞書やコーパスの整備をすることが当該セクションの翻訳品質を重点的に向上させる上で有効である。この際、国・地域に合わせて対象とするセクションを切り替えることも効率向上につながる可能性がある。

(2) 翻訳対象種別と国・地域の関連

翻訳対象語を発見した翻訳対象種別ごとの優先権主張の国・地域を図 4-2 に示す。なお、「その他」は翻訳対象語数が多いため、バブルの縮尺を変えて図 4-3 に示す。

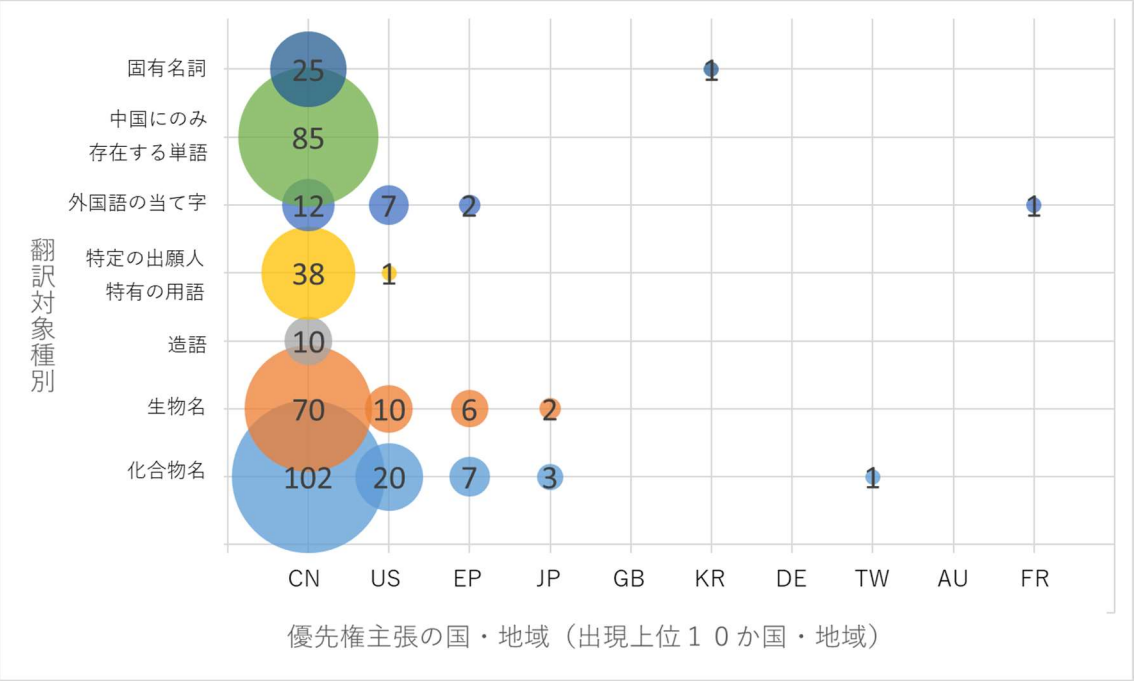

図 4-2 翻訳対象種別ごとの優先権主張の国・地域の件数

図 4-3　翻訳対象種別（その他）の優先権主張の国・地域の件数

優先権主張の国・地域が中国（CN）である公報が圧倒的に多いが、化合物名、生物名、外国語の当て字表記は、アメリカ合衆国（US)、欧州特許庁（EP）でも割合が高い。これは化合物名、生物名、外国語の当て字表記は外来語由来である場合が多いことが要因と考えられる。このため、化合物名、生物名、外国語の当て字表記などの外来語由来の単語を含む中国特許請求項の文の翻訳品質を向上するには優先権主張の国・地域が中国（CN）である公報のほか、アメリカ合衆国（US)、欧州特許庁（EP）に優先権主張した公報から辞書候補やコーパス候補を抽出し、整備することが有効である。

また、中国にのみ存在する単語は、優先権主張の国・地域が中国（CN）である公報でのみ抽出されている。このため、中国にのみ存在する単語を含む文の翻訳品質を向上させるには、優先権主張の国・地域が中国（CN）である公報から辞書候補やコーパス候補を抽出し、整備することが有効である。

4.1.4.　翻訳対象語が発見されやすい技術分野と出願人属性の傾向

表 3-14 に示したとおり、出願人上位 10 者の頻出セクションは出願人ごとに 30%～100%を占め、一定の傾向がみられる。しかし、出願人別の翻訳対象語は、最も多い中南大学で 0.33%であり、出願人の偏りはみられない。

このため、翻訳対象語の効率的な選定においては「4.1.1. 翻訳対象語が発見されやすい技術分野の傾向」の方針にのっとることで出願人の属性は特段の考慮を要しないと考える。

4.2. 翻訳対象語選定作業の効率化の検討

4.2.1. 翻訳対象候補の優先付けの効果

本事業では、人手による翻訳対象語の選定作業を効率化するため、翻訳対象語として選定すべきと考える「高い頻度で参照されやすい用語」、「文献を理解する上で重要な用語」、「最新の未知語」の観点で優先順位をつけた上で翻訳対象語の選定作業を行った（優先順位の条件は、「2.1.2. 優先順位づけ」参照のこと。）。

選定した翻訳対象語の IPC セクションの割合を、翻訳ログの IPC セクションの割合と比較した結果を図 4-4 に示す。

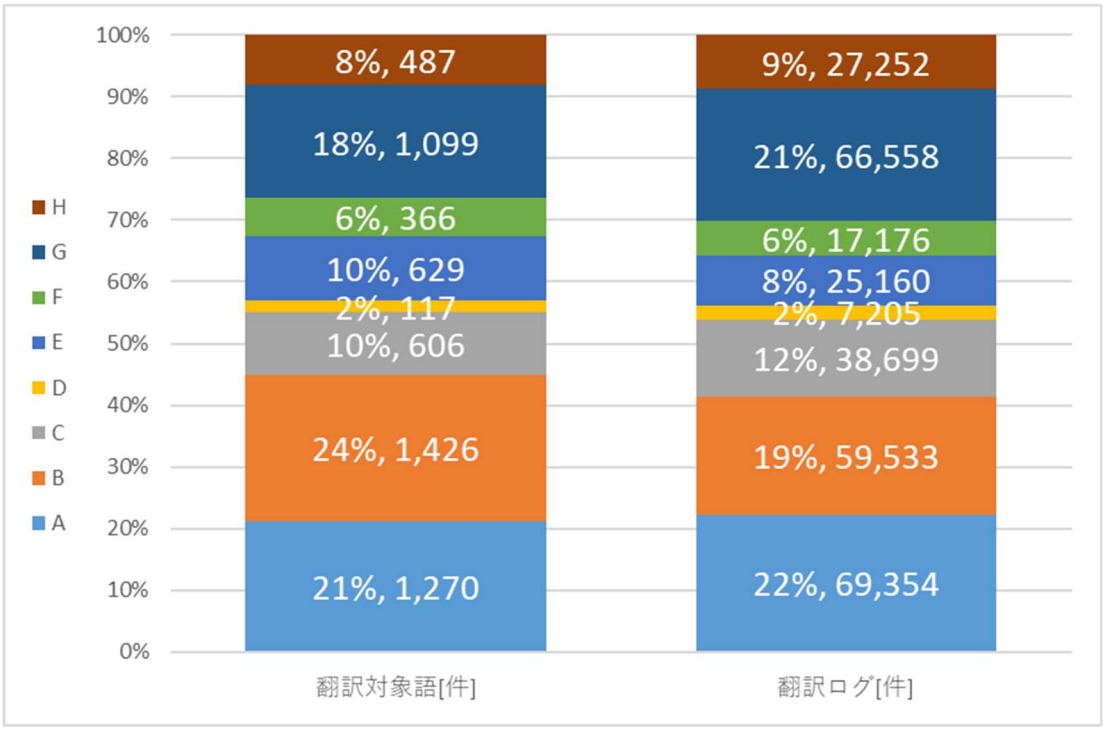

図 4-4　翻訳対象語と翻訳ログの IPC セクションの比較

図 4-4 のとおり、全翻訳ログに占める IPC セクションの割合と選定した翻訳対象語 6000 語の IPC セクションに大きなずれはなく、翻訳対象語の選定については提供された翻訳ログに含まれる IPC 情報の分布を反映した選定となっていた。

翻訳対象語選定作業の効率化においては、翻訳品質の改善方針に合わせて各セクションから翻訳対象語を選定する際の優先順位について検討する必要がある。

<u>方針１．すべてのセクションの翻訳品質の底上げ</u>

翻訳ログにおけるセクションの比率と同じ割合で翻訳対象語を選定して辞書や対訳コーパスを整備することがすべてのセクションの翻訳品質を底上げする上で有効である。これは、全セクションを同じ件数とした場合、翻訳ログに占める割合の少ないセク

ションについては品質向上の度合が相対的に高くなる可能性があるためである。

方針２．翻訳要求の絶対量が多いセクションの翻訳品質向上

翻訳要求の絶対量（翻訳ログの件数）が多いセクションを優先して翻訳対象語を選定し、辞書やコーパスの整備をすることが当該セクションの翻訳品質を重点的に向上させる上で有効である。

4.2.2. 翻訳対象語選定作業の課題

翻訳対象語選定作業の課題としては、未知語の修補を行った翻訳対象語が多いことが挙げられる。「2.2.1. 未知語の修補」で述べたとおり、本事業では形態素解析での誤解析等により、辞書登録単語として不適切な単位で検出されることがあるため、辞書登録単語として適切な単位（単語）に修補した。表 4-3 に示すとおり、翻訳対象語の選定に際し、全体の90%の未知語を修補しており、また、文字数が少ない未知語ほど修補率が高かった。

表 4-3　未知語修補の件数と修補率

未知語の文字数	未知語の件数	修補した件数	修補率 (修補数/未知語数)
1 文字	41	41	100%
2 文字	5,920	5,350	90%
3 文字	34	14	41%
4 文字	5	1	20%
合計	6,000	5,406	90%

中国語は、漢字 1 文字で成立する単語が数多く存在する。このため、形態素解析において、本来 1 つの単語として捉えるべき文字列を分割して誤解析してしまう。表 4-4 に誤解析した未知語の例を示す。

このような問題の解決策としては、機械翻訳システムで未知語を出力する際、前後の単語も併せて出力することが挙げられる。前後の単語も併せて未知語とすることで、修補を行うべき件数を減らすことが可能と考える。

表 4-4　形態素解析で誤解析した未知語の例

未知語	未知語の修補	翻訳対象テキスト	参考：和訳
尔菲	德尔菲法	3.根据权利要求 1 所述的评估方法，其中，在所述步骤 2 中，采用德**尔菲**法确定分区的权重。	デルファイ法
芦单	曲罗芦单抗	在一些实施方案中，IL-13 抑制剂是曲罗**芦单**抗(CAT-354)或其变体(Brightling 等,Lancet 3(9):692-701,2015；	トラロキヌマブ

5. 翻訳校閲者による効果的又は効率的な校閲方法の検討

本章では、翻訳校閲者による校閲方法を効果的又は効率的に行う方法について述べる。翻訳校閲者は、翻訳者が作成した翻訳対象語の和訳語と翻訳対象テキストの和訳を校閲し、誤りがある場合は修正する役割である。本事業では、翻訳時の誤り混入を最小限にすることで、翻訳校閲者による修正を減らし、校閲を効果的又は効率的に行うための方法を検討した。

5.1. 校閲結果のフィードバック

翻訳者による翻訳対象語及び翻訳対象テキストの和訳作業において、是正を要するような誤訳や作業漏れはなかった。しかし、以下のような軽微な誤記（誤字、誤訳、作業漏れ）などの誤りが混入したため、翻訳者へより注意して作業するように指導を行った。表 5-1 に誤り「誤字」が混入した例を示す。本事業では発生していないが、翻訳校閲者が翻訳対象語の和訳語や翻訳対象テキストの和訳訳文を修正した場合、他の作業者にも展開すべき誤りがある場合は、修正理由とともに、翻訳者にフィードバックすることで、その後の翻訳作業での誤り混入を最小限にし、効果的な校閲が可能となる。

表 5-1　翻訳時の誤り混入の例（誤字）

翻訳対象テキスト	优选的，所述研磨球的内侧环形间隙设置有研磨锥，所述研磨锥远离研磨球内侧侧壁一端固定安装有冲击板，所述冲击板靠近研磨球内侧侧壁一端固定安装有辅助弹簧，辅助弹簧远离冲击板一端固定安装在研磨球的对应位置壁面上，且研磨球侧壁对应研磨锥位置开设有与研磨锥相适配槽孔。
翻訳者による和訳（誤字の混入）	好ましくは、前記研磨ボールの内側の環状隙間に研磨コーンが設けられ、研磨ボールの内側側壁から離れた前記研磨コーンの端部には、衝撃板が固定して取り付けられ、研磨ボールの内側側壁に近い前記衝撃板の端部には、補助スプリングが固定して取り付けられ、衝撃板から離れた補助スプリングの端部は、研磨ボールの対応する位置の壁に固定して取り付けられ、**、あた、**研磨コーンに対応する研磨ボールの側壁には、研磨コーンに適合する溝が開けられる。
翻訳校閲者による修正	好ましくは、前記研磨ボールの内側の環状隙間に研磨コーンが設けられ、研磨ボールの内側側壁から離れた前記研磨コーンの端部には、衝撃板が固定して取り付けられ、研磨ボールの内側側壁に近い前記衝撃板の端部には、補助スプリングが固定して取り付けられ、衝撃板から離れた補助スプリングの端部は、研磨ボールの対応する位置の壁に固定して取り付けられ、**また、**研磨コーンに対応する研磨ボールの側壁には、研磨コーンに適合する溝が開けられる。

6. 辞書適用の効果

本事業で作成した辞書の効果を図るため、2022 年 3 月時点の MTP システムにおいて辞書の適用前後で未知語検出文を翻訳し、未知語の出力状況を確認した。辞書の適用前後での未知語の数を表 6-1 に示す。

なお、本事業で貸与された翻訳ログは 2021 年 9 月に翻訳要求されたものであり、9 月時点では未知語出力状況を確認した全ての文で未知語が検出されている。辞書適用前の未知語数が 5,881 語に減っているのは、2021 年 9 月から 2022 年 3 月までに追加された辞書によるものと考えられる。

表 6-1　辞書適用前後での未知語の数

未知語の数	
辞書適用前	辞書適用後
5,881	493

上記のとおり、辞書の適用前後で未知語の数が減少しており、未知語を減らす効果が確認できた。

一方で、辞書適用後に発生した未知語 493 語の内訳は以下のとおりであった。

辞書適用前から未知語であった語の数　　：489 語

辞書適用後に新たに未知語となった語の数：　4 語

辞書適用により未知語にならなくなった例を表 6-2 に、辞書適用前後どちらでも未知語となっている語の例を表 6-3 に、辞書適用後に新たに未知語となった例を表 6-4 に示す。

表 6-2　辞書適用により未知語とならなくなった語の例

辞書登録の中国語	辞書登録の訳語
俄罗斯语	ロシア語
三角架式	三脚型
仓振动摇筛	ダブルトレイ振動揺動篩

表 6-3　辞書適用前後どちらでも未知語となっている語の例

未知語	和訳	補足
中午饭	昼食後	形態素解析の誤解析（中午饭后）
传统舞	舞台	形態素解析の誤解析（舞台）
加快轴	軸受	形態素解析の誤解析（轴承）

表 6-4　辞書適用後に新たに未知語となった語の例

未知語	和訳	補足
介尺	メソスケールリン石膏	形態素解析の誤解析（介尺度磷石膏）
婀娜	艶やか	辞書登録により単語分割の区切り方が変わったことによるもの
疆北	ジュンガリア	形態素解析の誤解析（新疆北部）

辞書の適用により、5,392 語の未知語が減少したが、新たな未知語が 4 語発生している。これは入力された原文の単語分割が辞書の適用前後で変化したことにより、分割箇所の変更で未知語となってしまった語が発生したものと考えられる。

また、表 6-5 に辞書適用後の機械翻訳文と訳質の判定結果を示す。No.1 は、マーカーで示す箇所に辞書登録の悪影響による訳出箇所の誤りが生じている。また No.2、No.3 は、辞書に登録した語は正しく訳出できているが、そのほかの部分の誤訳は改善していない。

表 6-5　辞書適用後の機械翻訳文と訳質の判定

No.	翻訳対象テキスト	機械翻訳文 （辞書登録後）	人手和訳	辞書登録語 （訳語）	判定
1	优选地，所述第二驱动机构包括固定连接在所述箱体上端的第二动力电机，所述第二动力电机的输出轴末端固定连接有**往复丝杠**，所述**往复丝杠**远离所述第二动力电机的一端贯穿移动板，且转动连接在所述箱体的内底部，所述移动板与所述**往复丝杠**相配合。	好ましくは,前記の第 2 駆動機構は前記ボックス本体の上端に固定接続される第 2 動力モータを含み,前記の第 2 動力モータの出力軸末端に**レシプロスクリュー**が固定接続され,前記**レシプロスクリュー**は前記から第 2 動力モータの一端まで離れて移動板を貫通し,且つ前記の内底部に回転接続され,移動板はに係合する。ボックス本体前記前記**レシプロスクリュー**。	好ましくは、前記第２の駆動機構は、前記ボックス本体の上端に固定的に接続された第２の動力モーターを含み、前記第２の動力モーターの出力シャフトの端部に前記レシプロスクリューが固定的に接続され、前記レシプロスクリューの前記第２の動力モーターから離れた一端は、前記レシプロスクリューに係合する移動板を貫通して前記ボックス本体の内底部に回転可能に接続される。	往复丝杠(レシプロスクリュー)	×

No.	翻訳対象テキスト	機械翻訳文 （辞書登録後）	人手和訳	辞書登録語 （訳語）	判定
2	优选的，所述**出料仓**的下方设置有手旋螺栓，且所述手旋螺栓贯穿**出料仓**的内壁且与**出料仓**螺纹连接，所述手旋螺栓的顶端固定安装有密封板。	好ましくは,前記**排出ビン**の下に手回転ボルトが設置され,且つ前記手回転ボルトが**排出ビン**の内側壁を貫通し且つ**排出ビン**ねじ山に接続され,前記ハンドが回転する先端に密封板が固定して取り付けられる。ボルト。	好ましくは、前記排出ビンの下方には手締めボルトが設けられ、前記手締めボルトは、排出ビンの内壁を貫通し、排出ビンに螺着され、前記手締めボルトの上端には封止板が固定的に取り付けられる。	出料仓 (排出ビン)	×
3	**冬枣**还含有钾、钠、铁、铜等多种微量元素以及抗癌物质环磷酸腺苷、环磷酸鸟苷等，现有的**冬枣**采摘方法大多数为耗时、耗力的人工采摘，不仅费时费力而且效率极低，在进行采摘后通常需要再按**冬枣**的大小进分选，大大降低**冬枣**的加工效率。	**冬ナツメ**はさらにカリウム、ナトリウム、鉄、銅等の複数種の微量元素及び抗癌材料の環状リン酸アデノシン、環状グアノシン一リン酸等を含み,従来の**冬ナツメ**摘み取り法は多くの場合時間がかかり、力を消費する人工摘み取りであり,時間がかかり且つ効率が非常に低く,摘み取りを行った後に一般的に**冬ナツメ**の大きさに応じて選別する必要があり,**冬ナツメ**の加工効率を大幅に低下させる。	冬ナツメはカリウム、ナトリウム、鉄、銅などの多種の微量元素及び抗がん物質である環状アデノシンリン酸、環状グアノシンリン酸などをさらに含み、従来の冬ナツメの摘み取り方法の多くは時間と労力を要する手動による摘み取りであり、時間と労力を費やすだけでなく、効率が非常に低く、摘み取った後に一般的には冬ナツメの大きさに応じて選別する必要があり、冬ナツメの加工効率を大幅に下げる。	冬枣 (冬ナツメ)	×
4	在磁力振荡器上振荡使存活细胞与 MTT 反应产物**甲臜**充分溶解，放入酶标仪中测定结果。	磁気発振器で発振して生存細胞と MTT 反応生成物**ホルマザン**を十分に溶解させ,マイクロプレートリーダーに入れて測定結果を測定する。	磁気発振器で振動させて生存細胞とＭＴＴ反応生成物のホルマザンを完全に溶解し、マイクロプレートリーダーに入れて結果を測定する。	甲臜 (ホルマザン)	○
5	图 4 示出了根据本公开一示例性实施例的**解耦双同步坐标系**的电流检测原理图；	図 4 は、本発明の一実施形態に係る**デカップリングデュアル同期座標系**の電流検出原理を説明するための図である。	図 4 は、本開示の例示的な実施例に係るデカップリングデュアル同期座標系の電流検出原理図を示す。	解耦双同步坐标系 (デカップリングデュアル同期座標系)	○

このため、辞書の適用に際しては辞書適用により減少する未知語と新たに発生する未知語の関係や辞書登録による悪影響に注意する必要がある。

7. まとめ

本章では、3 章で述べた翻訳対象語の出現傾向と 4 章で述べた辞書又は対訳コーパスを作成すべき翻訳対象を効率的に選定する方法、6 章で述べた本事業で作成した辞書を MTP に適用した結果から、MTP における中日翻訳の翻訳品質を向上させるための課題について報告する。

本事業では、技術革新等により登場した新語が MTP の学習データに含まれないことから誤訳につながっているという仮説により、MTP の未知語から新語を含む対訳辞書及び対訳コーパスを作成した。しかし、「4.1.1. 翻訳対象語が発見されやすい技術分野の傾向」のとおり、本事業で翻訳対象語とした単語の大部分は一般的な用語であり、新語が未知語につながっているケースはわずかである。

また、6 章で述べたとおり、対訳辞書を適用しても、辞書登録により別の単語が未知語となったり、辞書登録語以外の箇所の誤訳が残ったりすることから、辞書の充実化による翻訳品質の向上は限定的である。

MTP の中日翻訳で検出された未知語は、「4.2.2. 翻訳対象語選定作業の課題」で述べたとおり、単語として成立しない不適切な単位で検出されたものが大半を占めた。この原因は複数考えられ、漢字 1 文字で成立する単語が数多く存在するという中国語の言語特性により、形態素解析において本来 1 つの単語として捉えるべき文字列を分割してしまうこと、学習データの不足により一般的な用語がニューラル翻訳エンジンの辞書に登録されておらず形態素解析で単語として捉えられないこと、MTP のニューラル翻訳エンジンで採用する形態素解析エンジンにより、低頻度語が未知語となってしまうこと等が挙げられる。

中国語の言語特性と学習データの不足による形態素解析の誤解析に関しては、ニューラル翻訳エンジンの追加学習により、一般的に使われる高頻度語がニューラル翻訳エンジンの辞書に登録されることで改善すると考えられ、それにより、未知語の減少及び翻訳品質の改善が期待できる。しかし、ニューラル翻訳エンジンの語彙の上限に達している場合、他の高頻度語が辞書から押し出されてしまうため、根本的な解決とならない。このような課題を解決する手法としてサブワードによる分割方式がある。サブワード方式を用いることで、単語を更に小さい単位で分割することで語彙の量を圧縮し、扱える語彙サイズを増やすことができる。サブワードの手法として、SentencePiece[4]のように単言語コーパスから分割単位を機械的に学習することができるものがある。例えば、未知語の検出が顕著であった技術分野の特許公報を用いて分割単位を再学習することにより、当該技術分野に応じた適切な単位で単語分割ができ、翻訳品質向上につながる可能性がある。

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

対訳辞書の作成に当たっては、6 章で述べたとおり、辞書適用による悪影響に注意する必要がある。作成した対訳辞書を登録した機械翻訳システムで、未知語の前後の文字列を含めた翻訳対象テキストを再翻訳し、文意に誤りが発生していないかを確認し、辞書登録語以外の箇所への悪影響が発生していないことを確認することで、悪影響が発生するおそれのある単語を除外することができると考えられる。

以上