

審査書類・審決の機械翻訳に関する
解析及びデータ作成事業

審査書類・審決の機械翻訳精度の調査・分析

最終報告書

令和4年3月22日

一般財団法人 日本特許情報機構

内容

報告書概略	1
I. 事業実施報告	39
1. 本事業の概要	41
1.1. 本事業の目的	41
1.2. 本事業の内容	42
1.2.1. 審査書類対訳コーパスの作成	42
1.2.2. 審決対訳コーパスの作成	42
1.2.3. 機械翻訳の調査・分析	43
1.3. 本事業の成果物	43
1.3.1. 審査書類対訳コーパス	43
1.3.2. 審査書類対訳辞書	43
1.3.3. 審決英訳文データ	43
1.3.4. 審決対訳コーパス	43
1.3.5. 審決対訳辞書	44
1.3.6. 審決汎用文例リスト	44
1.3.7. 英訳書誌記入済審決リスト	44
1.3.8. 審決逆翻訳コーパス	44
1.3.9. 審決文の要点の英訳	45
II. 調査報告	47
2. 調査の概要	49
2.1. 調査の目的と内容	49
2.2. 調査対象の機械翻訳エンジン	49
2.2.1. 「みんなの自動翻訳」のバージョンアップについて	49
2.2.2. 「みんなの自動翻訳」の大規模アップデート	50
2.3. 機械翻訳システムの学習データと学習のタイミング	51
2.4. 機械翻訳精度の評価	53
2.5. 評価用コーパス	53
2.5.1. 文の種類の偏り回避	54
2.5.2. 技術分野の偏り回避	55
2.5.3. 極端な長文と短文の排除	55
2.5.4. 審決の評価用コーパスについて	55
3. 審査書類①システムの評価結果とその分析	57
3.1. 審査書類①システムの評価の概要	57
3.2. 審査書類①システムの自動評価結果	58

3.2.1. BLEU と RIBES	58
3.2.2. 審査書類①全体の自動評価結果	59
3.2.3. 審査書類ごとの自動評価結果	62
3.2.4. 技術分野間の比較	66
3.2.5. 文の種類間の比較	68
3.2.6. 文単位のスコア向上／低下カウント結果.....	71
3.2.7. 実例を用いた機械翻訳文の定点観測	75
3.2.7.1. 実例 3-1：「拒絶査定」由来、文番号 A02-#1047 の分析.....	75
3.2.7.2. 実例 3-2：「前置報告書」由来、文番号 A913-#314 の分析	87
3.2.7.3. 実例 3-3：「補正却下の決定」由来、文番号 A191- # 38 の分析	99
3.2.7.4. 実例 3-4：「拒絶理由通知書」由来、文番号 A131-#68 の分析	105
3.2.7.5. 実例を用いた定点観測のまとめ.....	112
3.3. 審査書類①システムの人手評価結果.....	114
3.3.1. 人手評価の概要	114
3.3.2. 人手評価観点	114
3.3.3. 人手評価者	114
3.3.4. 人手評価の対象文	115
3.3.5. 各評価項目の集計結果.....	115
3.3.5.1. 内容伝達レベルの評価	115
3.3.5.2. 学習前後の機械翻訳文の相対評価.....	120
3.3.5.3. エラーのカテゴリ別カウント	124
3.4. 審査書類①システムに関するまとめ.....	159
3.4.1. 審査書類対訳コーパスの学習による翻訳品質の改善	159
3.4.2. 特許・審査書類関連用語の早期からの改善と安定	160
3.4.3. 特許・審査書類関連用語以外の翻訳品質について	161
3.4.4. 技術／一般用語及び文構造のランダム的な変化.....	162
3.4.5. 各種審査書類由来の学習データ間の高い互換性.....	165
3.4.6. 文種間の翻訳難易度の違いについて	166
3.4.7. 審査書類①に残存する課題.....	168
3.4.7.1. カギカッコで括られた原文において翻訳文が途絶するエラー	168
3.4.7.2. カンマで区切られた複数の数字が合計 3 つ以上のときに連結されるエラー	169
4. 審査書類②システムの評価結果とその分析	171
4.1. 審査書類②システムの評価の概要	171
4.2. 審査書類②システムの自動評価結果.....	172
4.2.1. 審査書類②全体の自動評価結果	172
4.2.2. 技術分野間の差異	176

4.2.3. 文の種類による差異	177
4.2.4. 文単位のスコア向上／低下カウント結果.....	178
4.2.5. 実例を用いた機械翻訳結果の定点観測	180
4.2.5.1. 実例 4-1：「意見書」由来、文番号 A53-#94 の分析.....	180
4.2.5.2. 実例 4-2：「意見書」由来、文番号 A53- # 55 の分析.....	187
4.2.5.3. 無学習時の機械翻訳文で「未知語」が検出された実例の分析	192
4.3. 審査書類②システムの人手評価結果.....	195
4.3.1. 人手評価の概要.....	195
4.3.2. 各評価項目の集計結果.....	195
4.3.2.1. 「内容伝達レベル」の評価	195
4.3.2.2. 学習前後の翻訳文の相対評価	198
4.3.2.3. エラーのカテゴリ別カウント	199
4.4. 審査書類②システムに関するまとめ.....	203
4.4.1. 意見書で多用される丁寧表現の影響.....	203
4.4.2. 学習データ量が比較的少量であることの翻訳品質への影響	203
4.4.3. 審査書類②の機械翻訳品質.....	204
5. 審決③／審決④システムの評価結果とその分析.....	207
5.1. 審決③／④システムの評価の概要	207
5.1.1. 審決逆翻訳コーパス	208
5.2. 審決③／④システムの自動評価結果.....	209
5.2.1. 審決③／④の自動評価結果.....	209
5.2.1.1. 各回学習後のスコア変動.....	210
5.2.1.2. 審決③と審決④のスコア比較	211
5.2.1.3. 審査書類①／②とのスコア差	212
5.2.2. 技術分野間の差異	214
5.2.3. 文の種類による差異	217
5.2.4. 文単位のスコア向上／低下カウント結果.....	218
5.2.5. 実例を用いた機械翻訳結果の定点観測	220
5.2.5.1. 実例 5-1：特許審決由来、文番号#1736 の分析	220
5.2.5.2. 実例 5-2：特許審決由来、文番号 # 1277 の分析.....	224
5.2.5.3. 実例 5-3：商標審決由来、文番号 # 122 の分析.....	229
5.2.5.4. 実例 5-4：商標審決由来、文番号 # 159 の分析.....	234
5.2.5.5. 実例の定点観測のまとめ.....	237
5.3. 審決③／④システムの人手評価結果.....	239
5.3.1. 人手評価の概要.....	239
5.3.2. 人手評価観点	239

5.3.3. 人手評価の対象文	239
5.3.4. 各評価項目の集計結果.....	240
5.3.4.1. 「内容伝達レベル」の評価	240
5.3.4.2. 学習前後の翻訳文の相対評価	243
5.3.4.3. エラーのカテゴリ別カウント	245
5.3.4.4. 審決③と審決④の相対評価.....	259
5.4. 審決③／④システムに関するまとめ.....	261
5.4.1. 審決③／④の機械翻訳品質.....	261
5.4.2. 審決逆翻訳コーパスについて	261
5.4.3. 審決の翻訳難易度について.....	263
5.4.4. 固有名詞の翻訳品質について	265
5.4.5. 同一案件内の訳語ゆれの発生について	267
5.4.6. 最終学習後に検出された主なエラー	268
5.4.6.1. 同一文でランダム的に発生する途絶エラー	268
5.4.6.2. 和暦→西暦の誤変換.....	269
6. 各システムの統合の可能性について	271
7. 本調査のまとめ	277
7.1. 各システムで達成された翻訳水準について.....	277
7.2. 各種対訳コーパスの潜在的な翻訳改善効果について	278
7.3. 各システムで未解決の課題について.....	279
7.3.1. 最終学習後も残存した特徴的なエラー	279
7.3.1.1. カギカッコで括られた原文において翻訳文が途絶するエラー	280
7.3.1.2. 同一文でランダム的に発生する途絶エラー	281
7.3.1.3. カンマで区切られた複数の数字が合計3つ以上のときに連結されるエラー	283
7.3.1.4. 和暦→西暦の誤変換.....	283
7.3.2. 同一案件内の訳語ゆれの発生	285
7.3.3. 固有名詞の翻訳難易度の高さ	285

報告書概略

本事業の目的

近年、我が国企業の研究開発活動や企業活動のグローバル化が大きく進展し、国内出願の件数は比較的落ち着きを見せる中、日本から海外への出願割合は増加傾向にあり、海外での早期かつ安定的な権利化を可能とする環境整備が求められている。我が国企業の海外での円滑な権利取得を可能とするためには、日本の高品質な審査・審判結果を外国特許庁に対し正確に伝達すべく、審査・審判書類の高精度な英語翻訳文を提供する必要がある。

現在、特許情報プラットフォーム（J-PlatPat）やワン・ポータル・ドシエ（OPD）の審査書類情報照会システムを通じて、機械翻訳により英語に翻訳された日本の審査・審判書類が外国特許庁で照会可能となっている。これらの機械翻訳文は AI を活用したニューラル機械翻訳等の新技術を用いて提供されているため、従来よりも翻訳品質は大幅に改善されているものの、機械翻訳の精度は今後も改善していく必要がある。

ニューラル機械翻訳により高品質な機械翻訳を提供するためには、翻訳対象とする書類、すなわち日本の審査・審判書類から採取した十分な分量の原文と、その翻訳文とからなる対訳文データ（対訳コーパス。本報告書では単に「コーパス」ともいう。）を翻訳エンジンに学習させる必要がある。そして、翻訳文の精度がコーパスの学習を通じた機械翻訳品質に直結することから、これらの大量のデータを人手翻訳により作成する必要がある。

そこで、本事業では、日本の審査書類・審決について、日英対訳コーパスの作成や、機械翻訳における課題分析を実施するなど、日本の審査書類・審決の高品質な機械翻訳文提供に向けた取組を行う。これにより、日本の高品質な審査・審判情報の対外発信力を強化し、我が国企業の海外での円滑な権利取得を可能とする環境整備を目指す。

事業内容

本事業では、大別して「審査書類対訳コーパスの作成」、「審決対訳コーパスの作成」、「機械翻訳の調査・分析」の三種の作業を実施した。各作業の概要を以下に記す。

〔審査書類対訳コーパスの作成作業〕

日本の特許審査書類のうち「拒絶査定」「拒絶理由通知書」「補正却下の決定」「前置報告書」「意見書」のそれぞれから日本語原文を合計 2,000,515 文選定し、これを人手で英訳して、日英対訳文形式の「審査書類対訳コーパス」を作成した。

各書類から選定する文の比率は拒絶理由通知書から 7、それ以外の 4 種の書類からそれぞれ 1 とした（7：1：1：1：1）。コーパス作成対象とする文は各書類とも以下の条件に基づき選定した。

- ・書類の本文部分から抽出する
- ・1文あたりの平均文字数を100文字／文程度とする
- ・極端な短文（20文字未満）や長文（250文字以上）は除外する
- ・同一書類から既に選定している文パターン¹は除外する（重複排除）
- ・過去事業で作成された「作成済み審決コーパス（特許庁貸与物）」に含まれる文パターンは除外する

上記条件により、各書類の審査書類対訳コーパスとも、文字数が21文字～249文字の範囲の非重複文で構成した。書類種別ごとの審査書類対訳コーパスの作成件数を下表に示す。

書類種別ごとの「審査書類対訳コーパス」作成件数

書類種別		コーパス作成件数
(1)	拒絶査定	181,845 文対
(2)	拒絶理由通知書	1,273,011 文対
(3)	補正却下の決定	181,896 文対
(4)	前置報告書	181,933 文対
(5)	意見書	181,830 文対
合 計		2,000,515 文対

[審決対訳コーパスの作成作業]

特許庁の選定した特許、実用新案、意匠、商標の審決全300案件（各年度100案件）の全文を人手翻訳して「審決英訳文データ」を作成し、当該データからさらに日英対訳文形式の「審決対訳コーパス」全103,564文対を作成した。以下、詳細を記す。

「審決対訳コーパス」内訳

四法種別		案件数	コーパス作成件数
(1)	特許	253 案件	94,884 文対
(2)	実用新案	0 案件	0 文対
(3)	意匠	17 案件	4,625 文対
(4)	商標	30 案件	4,055 文対
合 計		300 案件	103,564 文対

¹ 完全一致文に加え、変数（数値部分）のみが異なる同一文パターンについても除外した。次行記載の「作成済み審決コーパス」との重複排除も同様。

前述の審査書類対訳コーパスが文長 21 文字～249 文字の範囲の非重複文のみで構成されるのに対し、審決対訳コーパスは審決全文を対訳コーパス化しているため、文字数は制限されず、また重複文も含んでいる。

本作業ではさらに、審決対訳コーパスから審決文に特有、かつ、特許庁より貸与された既存の「審決対訳辞書」中に存在しない語句 3,060 語を抽出して作成した「審決対訳辞書」、同コーパスから汎用的な文例 364 件を新たに追加した「審決汎用文例リスト²」、翻訳対象とした全審決を掲載した「英訳書誌記入済審決リスト」、そして海外審決から抽出した英文と、これを英日機械翻訳して取得した日本語文とを対訳文形式データに編集した「審決逆翻訳コーパス」約 250 万件を作成した。

〔機械翻訳の調査・分析作業〕

本事業で作成した審査書類対訳コーパス及び審決対訳コーパスによる機械翻訳の品質改善効果を検証するため、各コーパスを機械翻訳エンジンサービスサイト「みんなの自動翻訳@TexTra[®]」が提供するニューラル機械翻訳エンジン「特許 NT」に段階的に学習させ、各学習段階における翻訳品質改善の度合いや機械翻訳文の変化の傾向、実用にあたっての課題等について調査分析を行った。

本調査分析では、審査書類対訳コーパスのうち拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書に由来する文対を一括で学習させた審査書類①システム、意見書に由来する文対のみを学習させた審査書類②システム、審決対訳コーパスを学習させた審決③システム、そして審決対訳コーパスと審決逆翻訳コーパスの双方を学習させた審決④システムと、それぞれ異なるコーパスを学習させた 4 種のシステムを構築した。各システムについて、各学習段階の機械翻訳品質を「自動評価」と「人手評価」により評価するとともに、学習データの増量に伴う機械翻訳文の変化の状況を観測して、コーパスの学習による品質改善の状況やその具体的な傾向、学習後も残存する課題について分析した。

以下、各システムの品質評価結果及びそれに基づく調査分析内容をまとめる。

² 貸与された既存の「審決汎用文例リスト」に対し、本事業で新たに 364 件を追加した。追加後の収録文例数は 2,164 件となった。

審査書類①システムの機械翻訳品質の調査分析結果

審査書類①システムは、「みんなの自動翻訳」の特許 NT エンジンに拒絶査定、拒絶理由通知書、補正却下の決定及び前置報告書の 4 書類に由来する審査書類対訳コーパスを一括で学習させたものである。

学習は、上記コーパスを計 10 万文対作成する都度段階的に実施し、計 18 回、1,810,685 文対³を学習させた。各回の学習データにおける 4 種の審査書類の構成比率は常時 1 : 7 : 1 : 1 であり、拒絶理由通知書由来のデータが全体の 70%、残る 3 種がそれぞれ 10% ずつを占める。

評価には、自動評価指標 BLEU 及び RIBES による自動評価と、評価者が目視で行う人手評価とを併用した。自動評価は学習の都度、人手評価は学習データを 50 万文対適用する都度（つまり学習 5 回ごとに）実施した。

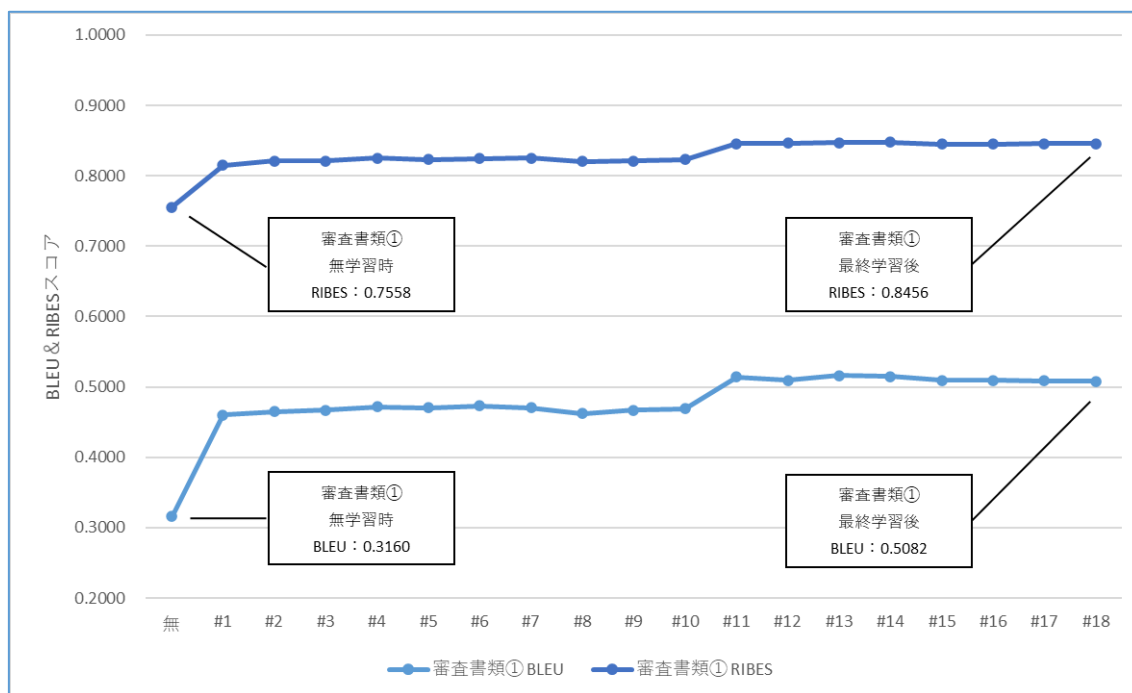
評価にあたっては、初回作成分の審査書類対訳コーパス 10 万文対から各書類 2,000 文対ずつ、計 8,000 文対を「評価用コーパス」に選定した。これを各学習段階の審査書類①システムで機械翻訳し、出力された翻訳文を評価した。ただし、人手評価は各書類とも評価用コーパスから 100 文対、計 400 文対のみを選定して評価対象とした。

[自動評価結果]

審査書類①の各学習段階における自動評価結果を示す。自動評価は無学習時と全 18 回の学習の都度、計 19 回実施した。評価指標には BLEU と RIBES を併用し、評価用コーパス各文の評価スコアを文単位で取得した。以下、審査書類①の各学習段階における評価用コーパス 8,000 文対の各学習回の全件平均スコアの推移を示す。

³ 初回学習のみ、作成した 10 万文対から評価用コーパス 8,000 文対（2,000 文対×4 書類）を選定したため、これらを除く 92,000 文対を学習させた。また、最終学習時は 118,685 文対を学習させた。

審査書類① 評価用コーパス全件の BLEU、RIBES 平均スコア推移



上記グラフ中、主な学習回の BLEU スコア（「B」と表記する。）と RIBES スコア（「R」と表記する。）を下表に示す。また、審査書類①を構成する 4 種の審査書類ごとの平均スコアも併せて掲載する。

審査書類① BLEU、RIBES 全件平均スコア一覧（主要学習回を抜粋）

学習回		無学習	第 1 回	第 5 回	第 10 回	第 11 回	第 15 回	第 18 回
審査書類①	B	0.3160	0.4602	0.4704	0.4692	0.5142	0.5096	0.5082
	R	0.7558	0.8150	0.8235	0.8232	0.8458	0.8448	0.8456
拒絶査定	B	0.3387	0.4736	0.4861	0.4821	0.5344	0.5280	0.5259
	R	0.7669	0.8156	0.8261	0.8246	0.8479	0.8451	0.8449
拒絶理由 通知書	B	0.3298	0.4550	0.4559	0.4600	0.4936	0.4934	0.4938
	R	0.7668	0.8168	0.8218	0.8218	0.8416	0.8430	0.8427
補正却下 の決定	B	0.2658	0.4061	0.4117	0.4150	0.4511	0.4527	0.4506
	R	0.7254	0.7950	0.8041	0.7993	0.8239	0.8237	0.8268
前置 報告書	B	0.3296	0.5059	0.5240	0.5197	0.5779	0.5644	0.5625
	R	0.7641	0.8329	0.8470	0.8471	0.8697	0.8674	0.8679

審査書類対訳コーパス全件学習後の審査書類①の全体平均スコアは、BLEU が 0.5082（無学習時から **+0.1922**）、RIBES が 0.8456（同 **+0.0898**）と、両指標とも顕著に向上した。グ

ラフに示したとおり、特に第 1 回学習後に大幅なスコア向上があり、以後の各回はおおむね同水準内で微増と微減を繰り返す形となったが、第 11 回学習後のみは両指標とも明瞭なスコア向上（BLEU が前回比+0.0450、RIBES が+0.0226）が見られた。これは当時実施された「みんなの自動翻訳」の大規模アップデートの効果と考えられる。それ以降は第 11 回時に上昇したスコア水準が最終学習後まで維持された。

4 種の書類別の平均スコアは、各書類で評価対象文（評価用コーパス）が異なることもあり、無学習の時点から書類間でバラつきが見られた。ただし、全期間での各書類のスコアの向上度は表の掲載順に BLEU が+0.1872、+0.1640、+0.1848、+0.2329、RIBES が+0.0780、+0.0759、+0.1014、+0.1038 といずれも大きく向上しており、各書類とも十分な品質改善効果が得られたことがわかる。

【人手評価結果】

人手評価では、審査書類①に属する 4 書類の評価用コーパスから各 100 文、全 400 文を評価対象とした。評価は無学習時の 1 回と学習データを約 50 万文対適用する都度の 3 回（第 5 回、第 10 回、第 15 回学習後）、そして最終学習回である第 18 回学習後の 1 回の計 5 回実施した。以下、主な評価観点の評価結果を示す。

【内容伝達レベルの評価】

機械翻訳文から原文の内容が正確に把握できるかを以下の 5 段階で評価し、各段階の数値を評価スコア値として集計した。

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

下表に審査書類①全体（400 文）と、書類別（各 100 文）の各回における「内容伝達レベル」評価スコアの平均値を示す。

審査書類①の内容伝達レベル集計結果

システム	無学習	第 5 回	第 10 回	第 15 回	第 18 回	向上度
審査書類①	3.14	4.09	4.20	4.67	4.69	+1.55
拒絶査定	3.08	4.13	4.16	4.61	4.69	+1.61
拒絶理由通知書	3.44	4.28	4.45	4.85	4.86	+1.42
補正却下の決定	3.20	4.09	4.18	4.63	4.62	+1.42
前置報告書	2.84	3.85	4.01	4.58	4.59	+1.75

審査書類①の人手評価対象文 400 文全件の内容伝達レベルの平均値は無学習時の 3.14 から第 18 回学習後には+1.55 となる 4.69 へと大きく向上した。学習初期（第 5 回時）に最も大幅な改善を見せ（+0.95）、「みんなの自動翻訳」の大規模アップデートを経た第 15 回学習後にもそれに次ぐ改善があり（+0.47）、以降もその水準が維持された。こうした各学習段階における推移は自動評価スコアのものと一致している。なお、書類ごとの集計結果も各書類でおおむね全体平均と同程度の向上を果たし、同程度の品質水準に達している。

【エラーのカテゴリ別カウント】

機械翻訳結果に頻出する典型的なエラーを 10 種類にカテゴリライズし、該当するエラーが発生しているか否かを文単位でカウントした。以下、審査書類①全 400 文におけるカテゴリ別のエラー発生文数の集計結果を示す。

審査書類①におけるエラーのカテゴリ別カウント結果

カテゴリ	無学習	第 5 回	第 10 回	第 15 回	第 18 回	エラー増減
特許誤訳	287	54	34	16	11	-276
技術／一般誤訳	108	55	56	20	25	-83
未知語	4	3	0	0	0	-4
係り受けエラー	84	70	73	42	39	-45
英語表現エラー	110	96	88	95	91	-19
湧き出しエラー	75	79	94	1	1	-74
欠落エラー	186	72	93	16	23	-163
数値エラー	32	33	23	5	5	-27
記号エラー	25	16	15	12	11	-14
その他エラー	13	9	16	1	1	-12

上表のとおり、各カテゴリともコーパスの学習の結果、無学習時に比べてエラーの発生文数が減少した。なかでも特許・審査書類用語が誤訳された文をカウントした「特許誤訳」は、

無学習の時点では評価対象 400 文のうち 287 文（約 72%）と大半の文において発生していたが、最終学習後には 11 文（5.5%）のみの発生と著しく減少した。

【評価結果の分析及び考察】

上掲した自動／人手評価結果および実際の機械翻訳文の観測により、審査書類①システムの機械翻訳品質および現状の課題に関して以下の各点が把握された。

【全体的な品質改善効果について】

審査書類①システムの機械翻訳品質は、審査書類対訳コーパスの学習により顕著に改善した。

評価用コーパス全件（2,000 文×4 書類）を対象とした自動評価では、二種の自動評価指標 BLEU、RIBES の双方で全件平均スコアが大きく向上した。評価用コーパスの一部（全 400 文）を対象に実施した人手による「内容伝達レベル」の評価結果も自動評価と完全に整合し、その妥当性を裏付けた。

さらに、内容伝達レベルの評価結果から、審査書類対訳コーパス学習前後の「みんなの自動翻訳」の具体的な翻訳品質も把握することができた。内容伝達レベルのレベル定義に照らすと、コーパス学習前（無学習時）の「みんなの自動翻訳」で審査書類を翻訳して得られる平均的な品質はレベル 3.15 相当、すなわち「半分以上の重要情報は正確に伝達されている（50%～）」程度にとどまっていた。これに対し、コーパス全件学習後の翻訳品質はレベル 4.67、すなわちレベル 5「すべての重要情報が正確に伝達されている（100%）」とレベル 4「ほとんどの重要情報は正確に伝達されている（80%～）」の中間のレベルまで向上しており、機械翻訳として極めて高い精度に到達したことが示された。

本事業は、日本の審査書類の機械翻訳品質が現状では不十分であること、そしてニューラル機械翻訳に審査書類由来の対訳コーパスを大量に学習させることで高品質な機械翻訳を得ることを目的として実施した。本調査分析の結果から、本事業で作成した審査書類対訳コーパスが企図したとおりの品質改善効果を奏することが実証された。

【特許・審査書類用語の訳質改善について】

審査書類対訳コーパスを作成することの主目的は、学習データとして機械翻訳システムを審査書類特有の用語や表現に対応させることにある。この目的がどの程度達成されたかは、人手評価における「エラーのカテゴリ別カウント」の「特許誤訳」の改善度によって具体的に把握できる。

「特許誤訳」に該当するエラーとしては、特許・審査書類用語及び表現が誤訳されている文をカウントした。無学習の時点の審査書類①では、評価対象全 400 文中の約 72%に相当する 287 文において何らかの「特許誤訳」が発生していた。これに対し、コーパス全件を学習させた第 18 回学習後に「特許誤訳」が検出された文は 11 文（5.5%）のみと大幅に減少した。

各学習回の機械翻訳結果の定点観測でも、ほぼ全ての特許・審査書類用語は学習初期から審査書類対訳コーパスで使用されている適訳語に改善され、以降の各学習回でも常時その訳語が採用され続けた。こうした特許・審査書類用語の適訳語への改善が、前項に示した審査書類①全体の顕著な品質改善の最大の理由である。各文中の特許・審査書類用語がコーパス学習初期に一斉に適訳語に改善し、その後も安定的に採用された結果、自動評価、人手評価とも学習初期にスコアが大幅に向上し、以降も常時その水準が維持された。

【特許・審査書類関連用語以外の翻訳品質について】

エラーのカテゴリ別カウントにおいて、「特許誤訳」以外の各カテゴリのエラーは無学習時から比較的少量であり、かつ、コーパス学習初期にもさほどの改善が見られないものが多かった。これらのエラーが審査書類用語・表現に直接関連しておらず、エンジン自体のロジックや基礎学習内容に起因するものが大半であったためと考えられる。

ただし、コーパスの学習では改善困難であったこれらのエラーも、その多くは第 11 回学習前に実施された大規模アップデートを機に顕著に改善された。その結果、学習初期以降現状維持状態であった機械翻訳品質はさらに向上し、翻訳品質の絶対評価である内容伝達レベルの評価では最終学習後のレベルが 5 段階評価で 4.69 と、機械翻訳として極めて高い水準に達した。この結果は、審査書類①システムが特許・審査書類用語だけでなく技術／一般用語や文法も正確に翻訳できており、かつ、「欠落」や「湧き出し」といった突発的なエラーも少数に抑えていることを示すものであり、現「みんなの自動翻訳」⁴が技術／一般用語や文法に関しても優れた翻訳精度を有している証左といえる。

⁴ 大規模アップデート以降の、バージョン 3.9 の「みんなの自動翻訳」を指す。以下同じ。

【技術／一般用語及び文構造の不安定な訳語変化について】

機械翻訳文の実例定点観測の結果、審査書類①システムでは、特許・審査書類用語の訳語は各文とも早期から適訳語で安定する一方、技術／一般用語の訳語や翻訳文の文構造⁵は各回の学習前後で不安定な変化を繰り返していることが判明した。こうした訳語・文構造の変化の多くは同義の範囲内に収まっていたが、学習前期においては突発的に不適訳語に変化して実質的な翻訳品質を低下させているケースも散見された。

「みんなの自動翻訳」において技術／一般用語の不安定な訳語変化が生じる理由には、(基礎学習データを含めた)学習データ中に複数の訳語候補を持つことと、ニューラル翻訳方式における訳語選定のメカニズムとが挙げられる。一般的にニューラル翻訳方式では、学習に際し、ある用語について学習データ中に複数の訳語が存在する場合、それぞれの訳語に対し、その訳語の各文例での周辺情報(周辺に存在する語句等の情報)に基づいた多次元ベクトルを付与する。そして、翻訳時には、入力文における用語の周辺情報と、学習データ中の各訳語候補のベクトルとを照合して、最もフィットすると判断された訳語が採用される。このため、複数の訳語候補を持つ技術／一般用語では、同じ用語であっても入力文の内容によって異なる訳語が採用される可能性がある。つまり同一用語の訳ゆれは、「みんなの自動翻訳」に限らず現在のニューラル機械翻訳の共通の課題といえる。

なお、用語に付与されるベクトルは、学習データ全体の全ての用語の相互の関連に基づくため、学習データが追加される都度、全ての用語について再計算される。大規模アップデート以前の旧「みんなの自動翻訳」では、このベクトルが再計算の都度大きく変化する傾向にあったと見られる。このため入力文が同一である(つまり入力文中の各語の周辺情報は変化していない)にもかかわらず、多数の技術／一般用語の訳語が学習の前後で不安定に変化したものと推察される。特許・審査書類用語の訳語はそれでも常時安定していたが、これは学習データ中で訳語が統一されていて他の訳語候補が存在しないため、ベクトルが変動しても採否に影響しなかったのであろう。

こうした技術／一般用語や文構造の不安定性は、第11回学習後に実施された「みんなの自動翻訳」の大規模アップデート以降、かなり抑制された。具体的には、各実例において学

⁵ 同じ訳語を用いて同じ内容を表した文であっても、各訳語の配置は一意には定まらない。例えば「請求項6には発明の詳細な説明に記載も示唆もない事項が記載されている」という文は、The matters that are not described or suggested in the detailed description of the invention are described in Claim 6.と訳しても、Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention.と訳しても、伝わる内容は同じであり、かつ、使われている訳語も同じである。本報告書では、翻訳文におけるこうした各語の配置を「文構造」と称し、上記のように訳語が変化せず各訳語の配置が同義の範囲内で変化している状況を「文構造の変化」と称した。

習データ中で最有力と見られる適訳語が採用される確率が高まり、変化の頻度や変化する語数が明らかに減少するとともに、不適訳語の突発的な採用が顕著に減少した。これは、大規模アップデートの際に「みんなの自動翻訳」のベクトル付与／再計算のロジックが改良されたためと推察される。特許・審査書類用語が早期から適訳語で安定していたにもかかわらず大規模アップデートを機に自動評価スコアや内容伝達レベルがさらに一段階向上したのも、ロジックの改良によって技術／一般用語に不適訳語が採用されたり、不適切な文構造が採用されたりするケースが減少したことが主な理由と見られる。

【各種審査書類由来の学習データの互換性について】

審査書類①の学習データでは、4種の審査書類（拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書）それぞれに由来するコーパスの構成比が1：7：1：1と、拒絶理由通知書の学習データ量を他の3書類に比して多く選定した。

しかしながら、書類別の自動評価スコアや内容伝達レベルの集計結果からは、学習データ量の違いにもかかわらず、拒絶理由通知書の評価スコアの推移や最終的に到達した品質水準に、他の3書類との顕著な差異は見られなかった。内容伝達レベルは拒絶理由通知書が4書類中で最上位ではあるが、これは無学習時の順位がそのまま維持された結果であり、学習データの多寡が反映されるべき学習前後のスコアの向上度は4書類中で最も低かった。

一方、拒絶理由通知書以外の3書類については、全学習データ中、自身に由来するデータはそれぞれ全体の10%に過ぎない。つまり学習データの90%が他種の書類に由来するデータとなるが、このことによる翻訳品質への悪影響は見られず、各書類とも拒絶理由通知書と同等以上の品質改善が果たされている。

こうした結果から、審査書類①に属する4種の審査書類由来の学習データ間には高い互換性があると結論される。4種の審査書類とも同じ制度に基づく審査手続の過程で特許庁審査官の手により作成される書類であり、使われる審査書類用語や表現もおおむね共通していることを考えれば、妥当な結果といえる。

審査書類②システムの機械翻訳品質の調査分析結果

審査書類②システムは、審査書類①と同じく「みんなの自動翻訳」の特許 NT エンジンベースとするが、学習データには審査書類対訳コーパスのうち「意見書」由来の文対のみを用いている。

審査書類②の学習は、上記コーパスを 1 万文対作成する都度段階的に実施し、計 18 回、179,830 文対⁶を学習させた。審査書類①と比べると約 10 分の 1 の学習データ量となる。

評価は、審査書類①と同様、BLEU 及び RIBES の自動評価と、内容伝達レベル等の人手評価を併用した。各評価の回数及びタイミングも同じである。評価用コーパスには意見書由来の審査書類対訳コーパスから 2,000 文対を選定し、自動評価ではその全件、人手評価ではここから 100 文対を選定して評価対象としている。

審査書類①に属する 4 種の書類がいずれも特許庁審査官が作成する書類であったのに対し、審査書類②システムが翻訳対象とする意見書は出願人が作成する書類である。このため他の審査書類に比べて丁寧表現が多用される傾向がある。加えて学習データ量も審査書類①に比べて大幅に少なく、これらの要因により機械翻訳文の品質や傾向が審査書類①と異なっている可能性がある。

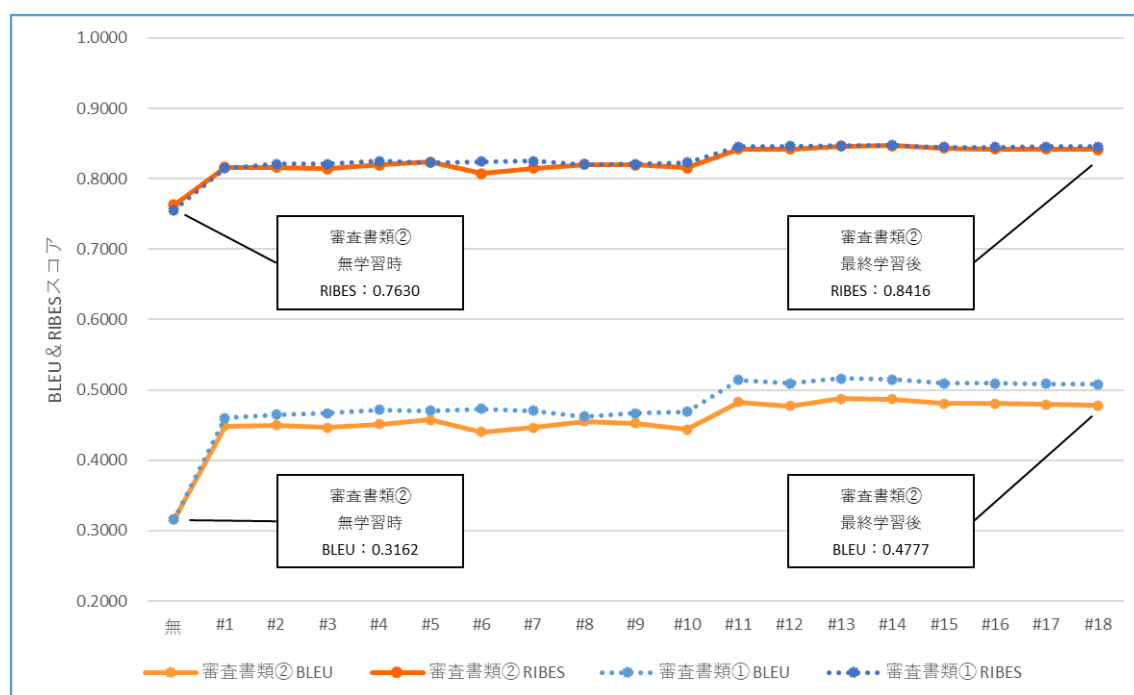
この観点のもと、本章では審査書類②の品質評価結果について、前章に示した審査書類①システムの評価結果との比較を交えて分析・考察する。

⁶ 初回学習のみ、作成した 1 万文対から評価用コーパス 2,000 文対を選定したため、これらを除く 8,000 文対を学習させた。また最終学習時は 11,830 文対を学習させた。

[自動評価結果]

審査書類②の各学習段階における自動評価結果を示す。自動評価は無学習時と全 18 回の学習の都度、計 19 回実施した。学習・評価のタイミングは各回とも審査書類①と同時である。以下、審査書類②の評価用コーパス 2,000 文対の、BLEU 及び RIBES の各回の全件平均スコアの推移を示す。

審査書類② 評価用コーパス全件の BLEU、RIBES 平均スコア推移



上記グラフ中、主な学習回の BLEU、RIBES スコアを下表に示す。参考として審査書類①の各回のスコアも併記した。

審査書類② BLEU、RIBES 全件平均スコア一覧（主要学習回を抜粋）

学習回		無学習	第 1 回	第 5 回	第 10 回	第 11 回	第 15 回	第 18 回
審査書類②	B	0.3162	0.4483	0.4577	0.4434	0.4825	0.4808	0.4777
	R	0.7630	0.8163	0.8238	0.8155	0.8426	0.8438	0.8416
[参考] 審査書類①	B	0.3160	0.4602	0.4704	0.4692	0.5142	0.5096	0.5082
	R	0.7558	0.8150	0.8235	0.8232	0.8458	0.8448	0.8456

コーパス全件学習後の審査書類②の全体平均スコアは、BLEU が 0.4777（無学習時から +0.1615）、RIBES が 0.8416（同 +0.0786）といずれも大きく向上した。審査書類①と同様、初回学習時に最も顕著にスコアが向上し、その後は現状維持状態となったものの、学習前に

「みんなの自動翻訳」の大規模アップデートが実施された第 11 回学習後に再び大きな向上があり、その水準が最終学習後まで維持されている。

各回のスコアを審査書類①と比較すると、BLEU スコアはコーパス初回学習以降常時やや劣っており、最終学習後も **-0.0305** と有意に低いスコアとなった。一方、RIBES では、最終学習後の審査書類①とのスコア差は **-0.0040** と誤差レベルの小差にとどまっており、双方の自動評価結果の示す状況にやや矛盾が生じている。

〔人手評価結果〕

審査書類②の人手評価結果として、審査書類①と同じく「内容伝達レベル」の評価とエラーのカテゴリ別カウント結果を示す。

【内容伝達レベルの評価】

審査書類②の評価対象文 100 文の各回における「内容伝達レベル」評価スコアの平均値を下表に示す。なお、参考として審査書類①全体の平均スコアを併記している。

審査書類②の内容伝達レベル集計結果

システム	無学習	第 5 回	第 10 回	第 15 回	第 18 回	向上度
審査書類②	2.99	3.89	4.10	4.68	4.70	+1.71
参考：審査書類①	3.14	4.09	4.20	4.67	4.69	+1.55

審査書類②の内容伝達レベルの平均値は無学習時の 2.99 から第 18 回学習後には 4.68 (**+1.71**) へと向上した。学習初期に最も大きく改善し (**+0.90**)、大規模アップデートを経た第 15 回時にもそれに準ずる向上を見せた (**+0.58**) という推移を含め、審査書類①とほぼ同様の推移を見せており、最終学習後に到達したレベルも審査書類①と同水準であった。

【エラーのカテゴリ別カウント】

審査書類②全 100 文におけるカテゴリ別のエラー発生文数のカウント結果を下表に示す。

審査書類②におけるエラーのカテゴリ別カウント結果

カテゴリ	無学習	第 5 回	第 10 回	第 15 回	第 18 回	エラー増減
特許誤訳	61	13	8	4	5	-56
技術／一般誤訳	31	20	25	5	7	-24
未知語	4	0	0	0	0	-4
係り受けエラー	9	5	9	10	8	-1
英語表現エラー	42	39	33	24	23	-19
湧き出しエラー	18	26	22	2	2	-16
欠落エラー	46	25	30	6	4	-42
数値エラー	7	6	4	1	1	-6
記号エラー	7	5	5	3	3	-4
その他エラー	0	2	0	0	0	±0

審査書類②のカウント結果でも、コーパスの学習による「特許誤訳」の顕著な改善や、技術／一般誤訳、湧き出し、欠落といったエンジンのロジックに起因する各種エラーの大規模アップデート後（第 15 回学習後）の急激な減少など、審査書類①と同傾向の結果となった。

【評価結果の分析及び考察】

上掲した各種品質評価結果及び実際の機械翻訳文の観測により、審査書類②システムの機械翻訳品質及び現状の課題に関して以下の各点が把握された。

【審査書類②の機械翻訳品質について】

審査書類②は、対象とする書類が意見書であること、そして学習データ量が大幅に少ないことから、審査書類①に比べて機械翻訳品質が劣る懸念があった。しかしながら、上掲の各評価結果から、審査書類②の機械翻訳品質は審査書類①と同水準に達していると結論される。

審査書類②の評価結果は、自動評価の BLEU スコアのみ審査書類①に比べて有意に低かったが、それ以外の各評価結果では遜色なく、最終学習後に到達した内容伝達レベル 4.70 も審査書類①の 4.69 と同等であった。このことから、審査書類②の機械翻訳品質も、審査書類①と同じく、機械翻訳として極めて高水準に達していると判断される。

なお、審査書類①において解決すべき課題として指摘した「同一案件における技術用語の訳ゆれ」、「本文全体がカギカッコで括られた原文において翻訳文が途絶するエラー」及び「カンマで区切られた 3 桁以上の数値が連結されるエラー」はいずれも「みんなの自動翻訳」のエンジン自体に起因するものであると考えられるため、審査書類②でも同様に発生している。

【意見書で多用される丁寧表現の影響について】

審査書類②で翻訳対象とした意見書は、出願人が審査官に宛てて作成する書類であり、丁寧表現が多用される。こうした丁寧表現への対応が必要である点において、意見書は他の審査書類よりも機械翻訳難易度が高くなる可能性があり得た。

事実、実例を調査したところ、無学習時の「みんなの自動翻訳」では丁寧表現に起因する未知語がいくつか発生しており、また内容伝達レベルも審査書類①よりもやや低い水準となっていた。これらから、デフォルトの「みんなの自動翻訳」は丁寧表現への対応（より厳密には、丁寧語と審査書類用語とが同じ文で使われている状況への対応）が不十分であり、審査書類②の機械翻訳品質を低下させる一因となっていることがわかった。

ただし、意見書由来の審査書類対訳コーパスを学習させた結果、当初未知語扱いされていた丁寧表現の大多数は適切に翻訳されるようになり、これに伴い審査書類②の内容伝達レベル、各種エラーの発生件数なども、学習後期には審査書類①と同等となった。

このことから、意見書で多用される丁寧表現は、機械翻訳の難易度を高める一因ではあるが、意見書由来の学習データを機械翻訳エンジンに学習させることで十分に対応可能であることが検証された。ごく一部、最終学習後も未知語扱いされる丁寧表現も存在するものの、全体的には、丁寧表現が機械翻訳品質に及ぼす悪影響は、意見書由来のコーパスを十分な量学習させることで、ほぼ完全に解消できると結論される。

【学習データ量が比較的少量であることの翻訳品質への影響】

審査書類②に学習させた意見書由来のコーパスの総量は約 18 万件であり、審査書類①の学習データ（約 181 万件）に比べて約 10 分の 1 と少量である。この学習データ量の多寡により、審査書類②の翻訳品質が十分に改善されない懸念があった。

審査書類対訳コーパスの学習量に不足があったならば、これはコーパス学習の主目的である特許・審査書類用語の翻訳品質、すなわち「特許誤訳」の発生件数に顕著に表れるはずである。また、審査書類②の場合、丁寧表現への対応の不十分さとしても検出される可能性もある。そして、こうした不備が発生すれば、それは内容伝達レベルにも反映される。

しかしながら、ここまで述べてきたとおり、審査書類②の内容伝達レベルや各種エラーの発生件数は学習後期には審査書類①と同等となっており、特許誤訳の発生件数にも優劣は生じていない。丁寧表現への対応の不足を示す事例も学習後期にはほぼ見られなくなり、審査書類②の学習量の顕著な不足を示す事象は検出されなかった。

学習データの効果は、その書類において頻出の用語であるほど早期から得られる。そして、頻出の用語が一斉に改善されるため、学習初期ほど全体的な品質改善の度合いも大きくなる。これに対し、学習が進み、主要な用語が網羅的にカバーされるにつれ、学習データを増量させても、改善される語は出現頻度の少ないマイナーなものに限られるようになる。このため、全体に占める改善の度合いは極めて小さくなり、巨視的には現状維持と映る状態に移行していくことは避けられない。

審査書類②も、大規模アップデートが実施された第 11 回以降は学習データを増量しても自動評価スコア、内容伝達レベルとも現状維持状態となっており、審査書類①と同様、意見書で用いられる主要な特許・審査書類用語はおおむねカバーされた状態に達していることが見て取れる。

ただし、一般に学習データの総量が多くなるほど、そこに含まれる用語や表現のバリエーションは増加する。その意味では、第 11 回以降の学習データにも、それ以前の学習データではカバーされていなかったマイナーな用語や表現が一定量含まれていると考えられる。こうしたマイナーな用語・表現は無数に存在し、評価用コーパス中に含まれるのはそのごく一部にすぎないため、学習した内容が必ずしも本調査分析の評価結果に直結するとは限らないが、学習データを増量するほどマイナーな用語・表現に幅広く対応できるようになることが期待される。

審決③／④システムの機械翻訳品質の調査分析結果

審決③／④システムは、審査書類①／②と同じく「みんなの自動翻訳」特許 NT エンジン

をベースとするが、学習データには「審決」由来の各種対訳コーパス、具体的には本事業で作成した「審決対訳コーパス」及び過去事業で作成された既存の「作成済み審決コーパス」（特許庁貸与物）を用いて審決の翻訳に特化させている。また、審決④には更に約 250 万文対の「審決逆翻訳コーパス⁷」も追加学習させた。

審決③／④システムの学習は審決対訳コーパスを審決 100 案件分作成すると、全 3 回実施した。各回の学習データは下表のとおりである。

審決③／④システムの各回の学習データ量

システム	学習	審決コーパス※	作成済コーパス	逆翻訳コーパス	総計
審決③	第 1 回学習	32,601 文対	56,964 文対	—	86,565 文対
	第 2 回学習	+35,089 文対			121,654 文対
	第 3 回学習	+31,498 文対			156,152 文対
審決④	第 1 回学習	32,601 文対	56,964 文対	2,521,792 文対	2,611,357 文対
	第 2 回学習	+35,089 文対		2,521,790 文対	2,646,444 文対
	第 3 回学習	+31,498 文対		2,521,826 文対	2,677,978 文対

※各回の新規データ数のみを示した。各回の学習データには過去回の分も含まれる。

評価には自動評価と人手評価を併用した。評価用コーパスには審決対訳コーパスの初回作成分から 2,024 文を案件単位で選定し、自動評価ではその全件、人手評価では 100 文のみを評価対象とした。なお、審決③／④は学習回数が 3 回と少ないため、人手評価も自動評価と同様に全 3 回の学習後と無学習時の計 4 回実施した。

審査書類対訳コーパスと異なり、審決対訳コーパスでは実用新案、意匠、商標の案件由来の文もコーパス作成対象となる⁸。このことが審決③／④の機械翻訳品質に影響を及ぼしている可能性がある。また、審決③と審決④の品質を比較することで、審決逆翻訳コーパスの品質改善効果の程度が測れる。これらが調査分析の主な焦点となる。

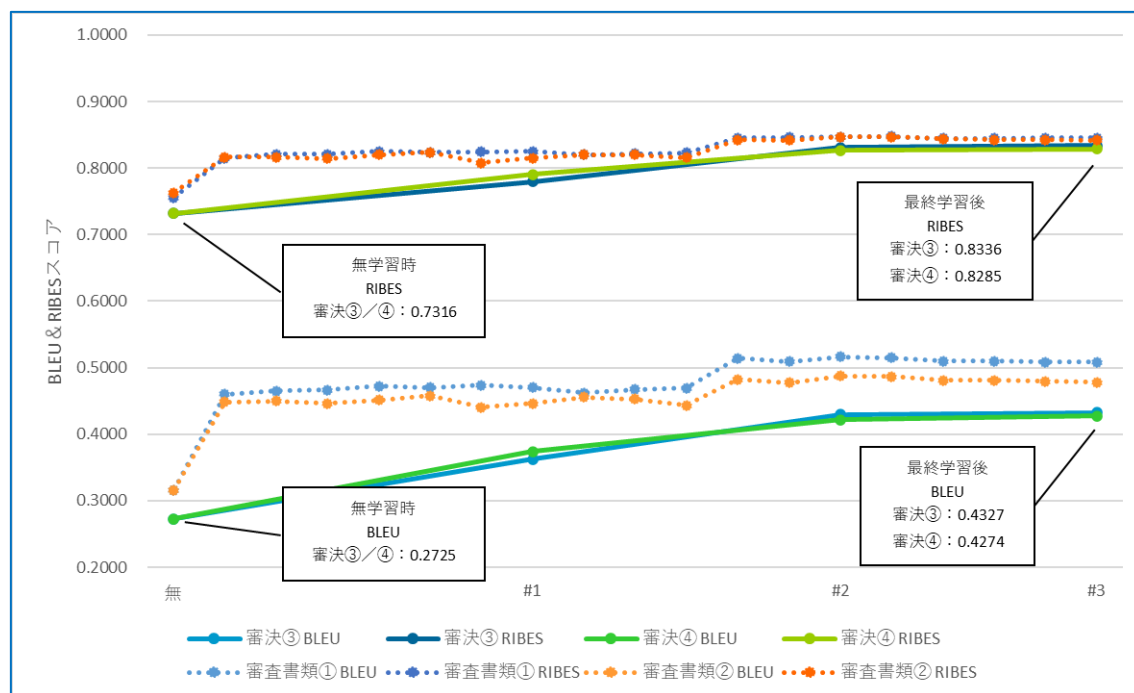
⁷ 海外審決由来の約 250 万件の英語審決文について、各学習回の時点の審決③システムを逆方向（英→日）で使用して日本語文を取得し、対訳コーパス化したものである。そのため各英文に対応する日本語翻訳文は各回同一とは限らない。また、各回とも英日機械翻訳処理時に若干数の翻訳不能文が発生したため、各回の文数にも僅かな相違がある。

⁸ ただし、本事業で作成対象として選定された 300 案件に実用新案は含まれなかった。

[自動評価結果]

審決③／④の各学習段階における自動評価結果を示す。学習は全 3 回実施し、それぞれ審査書類①／②の第 7 回、第 13 回、第 18 回と同時期にあたる。以下、審決③／④の評価用コーパス 2,000 文の、各回学習後の BLEU 及び RIBES の全件平均スコアの推移を示す。

審決③／④ 評価用コーパス全件の BLEU、RIBES 平均スコア推移



審決③／④ BLEU、RIBES 全件平均スコア一覧

学習回		無学習 ⁹	第 1 回	第 2 回	第 3 回
審決③	B	0.2725	0.3627	0.4297	0.4327
	R	0.7316	0.7793	0.8310	0.8336
審決④	B	0.2725	0.3740	0.4217	0.4234
	R	0.7316	0.7904	0.8269	0.8285

審決③／④の最終学習後（第 3 回）の全体平均スコアは、審決③は BLEU が 0.4327（無学習時から +0.1602）、RIBES が 0.8336（同 +0.1020）、審決④は BLEU が 0.4234（同 +0.1509）、RIBES が 0.8285（同 +0.0969）となり、両システム、両指標とも無学習時から大きく向上した。審決対訳コーパスの初回学習後（第 1 回）に最も顕著にスコアが向上し、

⁹ 審決③／④は同じ評価用コーパスを使用しており、このため無学習の時点の機械翻訳文は同一となる。このため評価スコアも同値である。

「みんなの自動翻訳」の大規模アップデート直後となる第 2 回学習後もそれに準ずる向上があり、その後は最終学習後（第 3 回）までその水準が維持された、という推移は、審査書類①／②の品質改善状況と同様であった。

審決③／④の自動評価スコアは審査書類①／②よりも有意に低く、例えば審決③の最終学習後のスコアを審査書類①／②と比較すると、両システムに対して BLEU では-0.0755／-0.0450、RIBES では-0.0120／-0.0080 といずれも劣る。ただし、このスコア差は無学習時から存在しており、当時の差は BLEU では-0.0435／-0.0437、RIBES では-0.0242／-0.0314 と、審査書類①の BLEU スコアを除けば最終学習後と同程度か、より大きかった。つまりこのスコア差は、審決③／④の評価用コーパス自体が、審査書類①／②のものよりも翻訳難易度が有意に高かったためといえる。

なお、審決③と審決④の比較では、第 1 回学習後は審決逆翻訳コーパスを学習させた審決④が審決③を上回ったものの、第 2 回学習後以降は審決逆翻訳コーパスを学習させていない審決③のほうが若干高いスコアを得ている。

【人手評価結果】

審決③／④の人手評価結果として、審査書類①／②と同じく「内容伝達レベル」の評価とエラーのカテゴリ別カウント結果を示す。

【内容伝達レベルの評価】

審決③／④の評価対象文 100 文の各回における「内容伝達レベル」評価スコアの平均値を下表に示す。

審決③／④の内容伝達レベル集計結果

システム	無学習	第 1 回	第 2 回	第 3 回	向上度
審決③	2.52	3.31	4.37	4.42	+1.90
審決④		3.32	4.24	4.31	+1.79

審決③／④の内容伝達レベルの平均値は、無学習時の 2.52 から最終学習となる第 3 回学習後にはそれぞれ 4.42 (+1.90)、4.31 (+1.79) へと向上した。全期間を通じての向上度、無学習時から最終学習後に至るまで審査書類①／②よりも評価値が有意に低い点、そして学習が進むにつれ審決③が審決④よりも僅かに高評価となる点とも、自動評価結果と一致した。

【エラーのカテゴリ別カウント】

審決③／④の人手評価対象文全 100 文におけるカテゴリ別のエラー発生文数のカウント結果を下表に示す。

審決③／④におけるエラーのカテゴリ別カウント結果

カテゴリ		無学習	第 1 回	第 2 回	第 3 回	エラー増減
特許誤訳	審決③	61	17	5	4	-56
	審決④		12	6	5	-55
技術／一般誤訳	審決③	25	24	8	10	-15
	審決④		18	10	9	-16
未知語	審決③	2	0	0	0	-2
	審決④		0	0	0	-2
係り受けエラー	審決③	34	30	6	8	-26
	審決④		37	9	9	-25
英語表現エラー	審決③	39	43	37	39	±0
	審決④		39	41	35	-4
湧き出しエラー	審決③	41	34	6	2	-39
	審決④		35	6	2	-39
欠落エラー	審決③	57	35	20	13	-44
	審決④		41	16	11	-46
数値エラー	審決③	18	20	6	6	-12
	審決④		20	9	7	-11
記号エラー	審決③	1	1	1	1	±0
	審決④		1	1	0	-1
その他エラー	審決③	0	0	0	0	±0
	審決④		1	0	0	±0

審決③／④とも、コーパスの学習による「特許誤訳」の顕著な改善や、「技術／一般誤訳」、「湧き出し」、「欠落」といったエンジンのロジックに起因する各種エラーの大規模アップデート直後（第 2 回学習後）の大幅な減少など、審査書類①／②とほぼ同じ傾向を示した。各カテゴリのエラーの発生文数も同水準であった。

また、審決③と審決④の間にも、エラーの発生件数や各学習段階での改善状況に特段の差異は見られなかった。

〔評価結果の分析及び考察〕

上掲した各種品質評価結果及び実際の機械翻訳文の観測により、審決③／④システムの機械翻訳品質及び現状の課題に関して以下の各点が把握された。

【審決③／④の機械翻訳品質について】

審決③／④の最終学習後の内容伝達レベルの平均値は、審決③が4.42、審決④が4.31であった。これは審査書類①／②の最終学習後の内容伝達レベル（4.69、4.70）に比べるとやや低い。審決③／④は自動評価スコアも審査書類①／②より有意に低く、全体的な品質は若干劣ると結論される。

ただし、審決③／④がコーパス学習後に到達した品質水準は内容伝達レベル 4（「ほとんどの重要情報は正確に伝達されている（80%～）」）を上回っており、機械翻訳としては十分に高い品質水準といえる。

審決③／④の内容伝達レベルや自動評価スコアは無学習の時点から低く、したがって審査書類①／②との品質差は、対象とする書類自体の難易度の差と見られる。審決③／④では商標審決由来の文も翻訳対象¹⁰としており、そこで多用される固有名詞の存在などにより、評価用コーパス全体の翻訳難易度が高まったものと推察される。

固有名詞は無限に存在し、かつ、案件ごとに全く異なる語が使用されるため、学習データで十分に網羅することは事実上不可能である。このため固有名詞の訳質改善は機械翻訳全般の課題となっているが、現時点では画期的な解決策は存在しない。

審決③／④の自動評価スコア及び内容伝達レベルは、審査書類①／②と同じく、最終学習後の時点では既に現状維持状態に移行していた。この状況は、審決③／④においても、審決対訳コーパスでカバーされるべき主要な審査書類用語・表現はおおむねカバーされていることを示している。したがって、審査書類①／②との品質差は学習データ量の顕著な不足によるものとは考えにくい。審決③／④の翻訳品質は、審決由来の対訳コーパスを学習させることで到達可能な上限にほぼ達しており、審査書類①／②と同様、今後の改善の余地はほぼマイナーな用語・表現に限られつつあると結論される。ただし、審決③／④では、固有名詞という典型的なマイナー用語が商標審決で多用される傾向があるなど、マイナーな用語・表現をより多く含む分、審査書類①／②よりも品質水準がやや劣ったと考えられる。

¹⁰ 審査書類①／②は原出願が特許の出願書類のみを翻訳対象とした。

【審決逆翻訳コーパスの要否について】

本調査分析では、審決③と審決④の品質評価結果は学習後期にはほぼ同等となり、審決逆翻訳コーパスの追加学習による特段の品質改善効果は見られなかった。したがって、審決対訳コーパス全件を学習データに使用する前提であれば、逆翻訳コーパスの追加学習は不要と判断される。

各システムの統合の可能性について

本調査分析では、審査書類対訳コーパス及び審決対訳コーパスを使い分け、審査書類①、審査書類②、そして審決③／④と、それぞれ特定の書類を対象としたシステムを個別に構築した。だが、書類ごとに個々のシステムを構築するよりも、各システムの学習データを統合し、単独のシステムで全ての書類に対応できるならば、より効率的である。

その可否を調査すべく、本調査分析では、審査書類①、審査書類②、そして審決③の全学習データを一括で学習させた全書類統合システムを構築し、この統合システムを用いて個々のシステムの評価用コーパスを改めて機械翻訳した。そして、得られた翻訳文の自動評価スコアを個々のシステムの最終学習後の自動評価スコアと比較することで、全書類統合システムで得られる機械翻訳文の翻訳品質が、個別のシステムで得られる品質と比較してどの程度であるかを把握した。比較結果を下表に示す。

全書類統合システムによる各評価用コーパスの機械翻訳結果の自動評価スコア一覧

評価用コーパス	指標	全書類統合	個別（最終学習後）	スコア差
審査書類①	BLEU	0.5059	0.5082	-0.0023
	RIBES	0.8447	0.8456	-0.0009
審査書類②	BLEU	0.4770	0.4777	-0.0007
	RIBES	0.8423	0.8416	+0.0007
審決③	BLEU	0.4207	0.4327	-0.0120
	RIBES	0.8280	0.8336	-0.0056

上表に示したとおり、審査書類①、審査書類②に関しては、全書類統合システムによる機械翻訳結果に対する自動評価スコアは個々のシステムの最終学習後のスコアと同水準であり、統合システムでも同等の翻訳品質が得られることが示された。一方、審決③では、統合システムによる機械翻訳文の自動評価スコアは個別システムに比べて BLEU では-0.0120、RIBES でも-0.0056と明らかに低下しており、品質がやや劣っていることが示された。

この結果を受け、審決③の評価用コーパス 2,024 文において特にスコアが下落していた数文について、最終学習後の機械翻訳文と全書類統合システムの機械翻訳文とを比較し、スコア下落の要因を調査したが、各例ともスコア低下の主な原因は正解訳文と語順が異なるという表層的なものであり、採用されている訳語は審決③システムの最終学習後の翻訳文とほぼ同一であった。学習データを統合したことによる実質的な品質低下要因は検出されず、これら数例を見る限り、全書類統合システムを用いたことによるスコア低下は、実質を伴わない表層的なものである可能性がある。

とはいえ、一定規模の文集合の平均値による自動評価スコアの相対比較結果に高い信頼性があることも事実である。このため、特筆すべき品質低下要因は存在しないにせよ、品質低下自体はスコアが示すとおり実際に生じている可能性はある。

全書類統合システムによる審決③の評価用コーパスの自動評価スコアは BLEU が 0.4207、RIBES が 0.8280 であった。このスコアは、審決④システムの第 2 回学習後のスコアに最も近く、その差異は BLEU が **-0.0010**、RIBES が **+0.0011** と誤差レベルである。そして、審決④の第 2 回学習後の内容伝達レベルが 4.24 であることから、全書類統合システムの実質的な翻訳品質もこれと同程度、すなわち 4.24 相当と推定される。個別システム（審決③）での最終学習後の内容伝達レベルは 4.42 であり、これとの比較で **-0.18** 程度の品質低下が生じていることになる。

以上の調査結果から、本事業で作成した全ての対訳コーパスを一括で学習させた全書類統合システムでは、審査書類①及び審査書類②に属する各書類に関しては個別システムと同水準の翻訳品質が得られるが、審決のみは若干の品質低下が生じる可能性があるとは結論される。具体的には、個別システム（審決③）との比較で内容伝達レベル **-0.18** 相当の品質低下が生じる懸念がある。

ただし、全書類統合システムによる審決の翻訳品質は内容伝達レベルで 4.24 相当と推定され、依然としてレベル 4 を大きく上回っており、機械翻訳としては十分に高品質といえる。また、いくつかの事例を見る限り、スコア低下は表層的な要因によるものであり、実質的な品質低下が実際に生じているケースは確認できていない。

全書類統合システムの可否は、こうした審決における若干の品質低下の可能性と、複数のシステムを統合することによるコスト削減とのトレードオフとして判断すべきである。

本事業のまとめ

〔各システムで達成された翻訳水準について〕

本事業は、日本の審査書類及び審決の機械翻訳品質にまだ改善の余地があること、そしてニューラル機械翻訳に審査書類／審決由来の対訳コーパスを大量に学習させることで高品質な機械翻訳が得られることを前提に実施した。本報告書に記した各システムの調査分析結果により、これら本事業の前提が的確であり、かつ、本事業で作成した審査書類対訳コーパス及び審決対訳コーパスが企図したとおりの品質改善効果を奏することが実証された。

本調査分析では、前掲したとおり、調査対象とした4種のシステム（審査書類①／②、審決③／④）の全てが、本事業で作成した対訳コーパスの学習により、内容伝達レベル4、すなわち「重要情報の80%以上が正確に伝達される」レベルを大きく上回る機械翻訳品質を達成した。機械翻訳としては極めて高品質、かつ、各システムとも無学習時から大幅なレベル向上を果たしており、本事業で作成した各種コーパスが高い翻訳品質改善効果を奏し、本事業の目指す日本の審査書類及び審決の高品質な機械翻訳文の提供を実現するものであることを示す結果といえる。

また、本調査分析で構築した4種のシステムとも学習後期には品質評価結果が現状維持状態に移行しており、品質向上の最大の要素であった特許・審査書類用語への対応がほぼ網羅的に完了していることがうかがえた。このことから、各システムの翻訳品質は審査書類／審決対訳コーパスの学習によって達成可能な上限に達しつつあり、後はマイナーな用語・表現に改善の余地を残すのみの状態であると結論される。

〔各種対訳コーパスの潜在的な翻訳改善効果について〕

本事業で構築し、段階的な学習を施した4種のシステムは、いずれも学習初期に大幅にスコアが向上したが、それ以降は、大規模アップデート直後を除いてほぼ現状維持状態で推移した。各システムは学習データの量が大きく異なるが、各システムが現状維持状態に移行したタイミングはほぼ同時期であった。

この結果から、各システムとも学習初期の時点で、それぞれが対象とする書類に頻出の特許・審査書類用語や表現の大部分がカバーされたものと見なせる。具体的には、人手翻訳評価で内容伝達レベルが4を超えた時点、すなわち審査書類①／②であれば第5回学習後、審決③／④であれば第2回学習後の学習データ量であれば、各種の審査書類／審決をレベル4程度の高品質で機械翻訳するために必要な学習データ量を確実に満たしているといえる。具体的には、審査書類①が492,000件、審査書類②が48,000件、審査書類③／④が124,654件（審決逆翻訳コーパスは除く。）となる。

ただし、このことは、上記データ量に達した後に作成された各種対訳コーパスの意義が減じたことを意味するものではない。機械翻訳エンジンの学習において、学習初期にまず頻出の用語が一斉に改善されて全体的な翻訳精度が大きく向上し、その後は、そこから漏れたマイナーな用語の改善に移行していくことは必然である。このため、同量の学習データを追加しても、全体品質に表れる改善効果が急激に小さくなることも、本調査分析に限らず必至である。しかし、学習早期に達成される頻出の用語・表現のカバーは、機械翻訳の利用性においては最低限の要件にすぎず、さらなる改善のためには、こうしたマイナーな用語も地道に改善していかなければならない。

本調査分析において、こうしたマイナーな用語の改善が全体品質、すなわち評価スコアに反映されるためには、まずは限られた文数の評価用コーパス中で使用されているマイナーな用語が、学習データ中にも含まれ、両者が合致する必要がある。そして、仮に合致して翻訳品質が改善したとしても、マイナーな用語であり、かつ、評価用コーパス自体が非重複文で構成されているため、改善は局所的であり、全体品質に与える改善効果は非常に見えにくくなる。

本事業において、審査書類対訳コーパスを非重複文のみとしたのは、限られた文数で、可能な限り広範囲の用語・表現を効率的にカバーするためである。このため、これを評価用コーパスと学習データとに分割して使用した本調査分析では、両者の合致率が低くなることは避けられない。実用時においては、審査書類対訳コーパスに含まれる文は全て非重複文であることから、基本的には全ての文が、それぞれ何らかのマイナーな用語・表現の訳質改善に寄与する可能性を等しく有している。そして、こうした文は、可能な限り広範に作成しておくほど、それを学習させた機械翻訳エンジンの潜在的な翻訳性能が向上することは確実である。

その意味において、本事業は、非重複文のみで対訳コーパスを大量に作成するという方針により、限られた文数で最も広範囲の潜在的な翻訳改善効果を効率的に取得することができたと結論される。

【各システムで未解決の課題について】

本調査分析で構築した4種のシステムは、審査書類／審決対訳コーパスの学習によって到達可能なほぼ上限の品質水準に到達したと見られる。ただし、各システムの調査分析では、対訳コーパスの学習では改善困難であったいくつかの課題が検出された。以下、これらの課題についてまとめる。

【最終学習後も残存した特徴的なエラー】

各システムの最終学習後の機械翻訳文で検出されたエラーの中には、複数の文で発生しており、発生する原文の条件や、発生したエラーの状況から特定のパターンに括られるものが見られた。以下に列挙する。

① カギカッコで括られた原文において翻訳文が途絶するエラー

原文全体がカギカッコで括られている場合に、機械翻訳文が文頭から254～255文字に達した時点で途絶してしまうエラーである。本項ではこれ以降、「カギカッコ途絶エラー」と称する。実例を以下に示す。

A913-#1083

原文	・「炭酸を含有する電子材料用洗浄水の余剰分を戻すポイントは、脱炭酸装置の上流であれば特に制限はないが、特に前処理工程と一次純水製造部との中間に位置する水槽が適当である。」(段落【0007】)
機械翻訳文 (最終学習後)	"The point for returning the surplus of the washing water for electronic materials containing carbonic acid is not particularly limited as long as it is upstream of the decarboxylation device, but in particular, a water tank located between the pretreatm

上例では、原文の「前処理工程」に相当する pretreatment が pretreatm と単語の途中で途絶し、それ以降の内容が失われている。出力された文字数は254文字である。

こうした「カギカッコ途絶エラー」は、審査書類①／②では第3回学習後以降、審決③／④では第2回学習後以降に発生するようになり、以降、最終学習後まで各システムで複数件検出されている。学習データの内容や量が大きく異なる全システムで共通して発生していることから、「みんなの自動翻訳」自体のロジックに起因するものと考えられる。

本エラーは、原文が以下のいずれかの条件に該当し、かつ、機械翻訳文が254文字以上となる場合に発生している。

- ・タイプ A：冒頭に中点（・）が打たれ、直後の本文全体がカギカッコで括られ、その後にカッコ補記が存在する

（※上例が該当）

- ・タイプ B：上記タイプ A と同様だが、カッコ補記を伴わない

「すなわち、本願新請求項 1 の発明は、走行体にカーブを走行させることなく、…給電部を配置することが技術的な特徴です。」

- ・タイプ C：本文全体がカギカッコで括られ、その後にカッコ補記が存在する

「交換機または回線に呼が集中すると過負荷となり、通信網が異常となって、疎通が不可能な輻輳状態になる。…このような通信網の品質低下を抑えるために輻輳時の対策が必要である。」（段落 0 0 0 2）

- ・タイプ D：上記タイプ C と同様だが、カッコ補記を伴わない

「よって、本願発明では、モータの電氣的時定数を転流間隔時間の $1/6$ にするための…請求項 1 の従属項として定義したものである。」

「みんなの自動翻訳」で見られた通常の「欠落エラー」には、学習データの不足に起因するものと、それまで問題なく訳されていた語が突発的に欠落するものとが混在していた。本調査分析では、後者をエンジンのロジックに起因する「欠落エラー」として扱った。ただし、これらのエラーにおいても、上例のように単語が途中で途絶したり、特定の文字数以降の翻訳文が全て消失したりするような現象は見られない。したがって、「カギカッコ途絶エラー」は、カテゴリとしては「エンジンのロジックに起因する欠落エラー」ではあるが、他の「欠落エラー」とは根本的に性質が異なるものといえる。

審査書類には出願書類や先行技術文献などからの引用文も多く、引用であることを示すために上記の各タイプの様式を用いることも多い。このため、比較的影響範囲の広いエラーである。

② 同一文でランダム的に発生する途絶エラー

上記「カギカッコ途絶エラー」と同様、翻訳文が 254 文字に達した時点で途絶してしまうエラーである。以下、一例を示す。

【審決-#910】

原文	引用発明における「前記マルチフィラメントと前記被覆層との間に配置されたホットメルト接着剤 A-1：三井化学（株）製の「アドマー QE-060」で形成された接着層」は、本願補正発明における「前記補強コードと前記被覆用樹脂層との間に配置されて前記補強コードと前記被覆用樹脂層とを接合する接合用樹脂層」に相当する。
正解訳文	"An adhesive layer made from a hot melt adhesive A-1 of 'ADMER QE-060' manufactured by Mitsui Chemicals, Inc. disposed between said multifilament and said covering layer" of the Cited Invention corresponds to "a resin layer for adhesion disposed between said reinforcing cord and said resin layer for covering to adhere said reinforcing cord and said resin layer for covering" of the Amended Invention.
審決③ 第3回学習後 欠落：有	"A hot melt adhesive a- 1 arranged between the multifilament and the covering layer : An adhesive layer formed of" ADMER QE 060 "manufactured by Mitsui Chemicals, Inc." in the Cited Invention corresponds to "a bonding resin layer arranged between the rei
審決④ 第3回学習後 欠落：無	The "hot melt adhesive a- 1 disposed between the multifilament and the coating layer : An adhesive layer formed of' ADMER QE - 060' manufactured by Mitsui Chemicals, Inc." in the Cited Invention corresponds to the "bonding resin layer disposed between the reinforcing cord and the coating resin layer to bond the reinforcing cord and the coating resin layer" in the Amended Invention.

本例も、審決③の最終学習後の機械翻訳文において、翻訳文が 254 文字で途絶し、原文の「補強コード」に相当する reinforcing cord が rei と、単語の途中までしか出力されていない。なお、審決④では、同じタイミングで同じ原文を機械翻訳しているにもかかわらず、文末まで翻訳文が出力されている。

上例をはじめ、本エラーが発生する各文ともカギカッコは使用されているが、前項に示した「カギカッコ途絶エラー」の発生条件（タイプ A～D）には該当しておらず、発生条件が特定できない。また、「カギカッコ途絶エラー」が条件に合致する文で毎回確実に発生して

いたのに対し、本エラーは、上例のとおり、同一の入力文でもシステムによってエラーが発生する場合としない場合があり、かつ、同じシステムでも学習前後でエラーが発生したり解消したりすることもあるなど、明らかにランダム性が強い。実際、上例も、前回（第2回）学習後は審決③では翻訳文は途絶せず、審決④で途絶するという、第3回学習後とは正反對の結果となっている。

こうした状況から、本エラーは、翻訳文が254文で途絶するという状況自体は「カギカッコ途絶エラー」と同じだが、本質的には別種のエラーであると考えられる。翻訳文が254文字で途絶する現象は、単に「みんなの自動翻訳」が、機械翻訳処理が何らかの理由により完了できなかった場合に、容量254文字相当の作業用バッファの内容を出力する構造であるためと推察される。

つまり、本エラーは、必ずしも発生条件が一意に定まるものではなく、ニューラル機械翻訳のランダム性に起因する種々の要因で機械翻訳処理が完了できなかった場合に発生するものであって、その際の出力文字数が共通しているにすぎない可能性が高い。

③ カンマで区切られた複数の数字が合計3つ以上のときに連結されるエラー

原文中で複数の数値がカンマで区切られて並び、数字が合計で3つ以上である場合に、3桁以上の1つの数値として誤訳されてしまうエラーである。例えば「請求項14, 16」であれば、Claims 14, 16ではなく、カンマで区切られた4つの数字1、4、1、6が連結されて Claim 1416 と1つの数値として連結される。以下、実例を示す。

A131-#377

原文	したがって、請求項1～7, 10～12, 18に係る発明以外は、発明の単一性の要件以外の要件についての審査対象としない。
機械翻訳結果 【第18回学習後】 内容伝達レベル：4 数値エラー：有	Therefore, the inventions other than the inventions according to Claims 1 to 710 and 1218 are not subject to examination for requirements other than the requirement of unity of invention. (BLEU : 0.3874, RIBES : 0.8160)

上例では、原文「請求項1～7, 10～12, 18」の「7, 10」と「12, 18」が、カンマで区切られた複数の数字が合計3つ以上という発生条件に合致しており、それぞれ710、1218と連結されてしまっている。

本エラーは各システムで大規模アップデート直後から突如発生しており、条件に合致する文では確実に発生していることから、「みんなの自動翻訳」のエンジン自体のロジックに

起因するエラーと見られる。

④ 和暦→西暦の誤変換

「みんなの自動翻訳」では原文中の和暦を西暦に変換して翻訳文に反映しているが、まれに誤った西暦年が採用されることが判明した。これらの和暦→西暦の誤変換は、同一文中において数値（すなわち西暦年）の配置の入れ替わりと、日本語文で多用される「同年」という表現の翻訳難易度に起因するものと考えられる。

文中の数値が入れ替わるエラー自体は、ニューラル機械翻訳では比較的良好に見られる。文中に年が複数含まれている場合、それらが入れ替わって結果的にそれぞれが誤った西暦となることがある。一方、「同年」という表現は、英文では具体的な西暦年で表すのが一般的であるが、その際に誤った西暦年が採用される場合がある。和暦→西暦の誤変換は、これら二つの要因でそれぞれ発生するため、発生数が多くなっているものと考えられる。以下、実例を示す。

【文番号：#626】

原文	その後、平成 27 年 5 月 11 日付けで審理事項通知がされ、平成 27 年 5 月 24 日付け（同年 5 月 25 日差出し）で被請求人より上申書が提出され、平成 27 年 6 月 15 日付けで被請求人より口頭審理陳述要領書が提出され、平成 27 年 6 月 22 日差出しで請求人より口頭審理陳述要領書が提出された。
正解訳文	Notification of matters to be examined were notified as of May 11, 2015. A written statement was submitted by the demandee as of May 24, 2015 (sent on May 25, 2015). An oral proceedings statement brief was submitted by the demandee as of June 15, 2015. An oral proceeding statement brief was sent and submitted by the demandant on June 22, 2015.
審決③ 第 3 回学習後 数値エラー：有	Thereafter, a notice of matters to be examined was issued on May 11, 2015, a written statement was submitted by the demandee on May 24, 2015 (submitted on May 25, 2015), an oral proceedings statement brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2019.
審決④ 第 3 回学習後 数値エラー：有	After that, a notification of matters to be examined was issued on May 11, 2015, a written statement was submitted by the demandee on May 24, 2015 (delivered on May 25, 2016), an oral proceedings statement

	brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2015.
--	--

上例では、原文に「(同年 5 月 25 日差出し)」という文言が存在する。ここでの「同年」は、直前の「平成 27 年」を指し、翻訳文では西暦 2015 年に変換される。審決③はこの箇所を(submitted on May 25, 2015)と正しく訳しているが、審決④では、(delivered on May 25, 2016)と、西暦年が誤変換されている。

原文には「平成 27 年」が 4 回、「同年」が 1 回出現するが、全て西暦 2015 年に相当し、審決④で採用された 2016 年にあたる要素は原文中に存在しない。このことから、この誤変換は、「同年」を「平成 27 年」を指すものと解釈できず、何らかの理由で 2016 年と解釈した結果であると考えられる。

「同年」という語は、英語に直訳すると of the same year 等となるが、この訳語を上例の「同年 5 月 25 日」のような日本語文での省略用法に適用すると、May 25 of the same year となり、具体的な西暦年で訳すよりも大幅に冗長なものになってしまう。このため、英文では、既出の年を再度述べる文脈でも具体的な西暦年を用いるのが一般的である。よって、こうしたデータを学習した機械翻訳エンジンも、「同年」を具体的な西暦で翻訳する傾向が強くなる。

しかしながら、この場合、機械翻訳エンジンは、「同年」が具体的に指す西暦年を、文ごとに正確に選定しなければならない。現状においては、この選定を誤る場合があり、このため、西暦の誤変換が散見されるものと推察される。

なお、上例では、審決③にも和暦→西暦の誤変換が見られる。具体的には、文末の「平成 27 年 6 月 22 日差出し」が June 22, 2019 と誤訳されている。これは一見すると「平成 27 年」が西暦 2019 年と誤訳されたように見える。だが、これは実際には、審決④と同じく「同年」が誤解釈されて 2019 となり、かつ、ニューラル機械翻訳特有の数値の入れ替わりによって、これと文末の「平成 27 年」に対応する 2015 とが入れ替わって出力されたものと考えられる。

文中における数値の入れ替わり、そして「同年」に対応する西暦年の正確な判定とも、学習データでの対応は困難である。ニューラル機械翻訳技術のさらなる発展が待たれる。

【同一案件内の訳語ゆれの発生】

ニューラル機械翻訳では、各用語の訳語は、大量の学習データ中で、その用語が入力文と最も近い使われ方をしている文例の対訳文で採用されている訳語が採用される傾向が強い。これは、同じ用語でも文脈に応じて最適の訳語に訳し分けることができるという、ニューラル機械翻訳の利点に通じる。ただし、その反面、同じ案件に繰り返し出現し、同じ訳語を採用すべき語が、文によって異なる訳語で訳されるという状況、すなわち用語の訳ゆれが発生する可能性もはらんでいる。

本調査分析でも、審決③／④の最終学習後の機械翻訳文を確認した結果、審判番号 2017-800128 号(評価用コーパス文番号#402～#617)の審決における「突出部」(56 文中、projecting portion が 27 文、protrusion portion が 17 文、protrusion が 11 文、protrusion part が 1 文)や 2017-006367 号(同#746～#938)における「タイヤ骨格体」(tire frame が 14 文、tire skeleton が 4 文)、2018-700001 号(同#1230～#1313)における「被覆層」(covering layer が 15 文、coating layer が 13 文)、2017-012323 号(同#1563～#1738)における「泡立ち」(foaming が 20 文、bubbles が 5 文、bubbling が 4 文)など、同一学習回・同一案件内での技術用語の訳ゆれが複数件検出された。

上記 4 案件で採用された各訳語を見ると、おおむね同義の範囲内かつ頻用の訳語ではある。それでも、特に技術用語は、同じ用語が案件内で複数の訳語に訳ゆれしていると、読み手を無用に混乱させる場合が多く、好ましくはない。

こうした訳語変化は、冒頭で述べたニューラル機械翻訳エンジンの訳語選定メカニズムに起因するものであり、本調査分析で多数見られた学習前後の技術／一般用語の不安定な訳語変化と同根である。このため、上述した審決③／④システムに限らず、全システムで発生する可能性が高い。

【固有名詞の翻訳難易度の高さ】

審決③／④の評価用コーパスには商標審決由来の文が含まれており、これらの文の翻訳精度が、特許審決由来の文に比して有意に低かった。その一因として、商標審決における固有名詞の多用が挙げられる。

固有名詞は無限に存在し、学習データでその全てを網羅することは事実上不可能である。このため、いかなる機械翻訳方式であっても、固有名詞の機械翻訳品質がその他の語句に比べて低くなることは、少なくとも現状においては、不可避といえる。

そして、商標審決は、その性質上、固有名詞が多用される傾向が他の書類よりも強い。商

標は、それ自体が固有名詞であることが多く、つまり審理の主題となることが多いため、審決内で繰り返し言及される傾向にある。その分、商標審決は、特許を対象とした審決や審査書類に比べて、固有名詞の翻訳難易度の高さの影響を受ける度合いが強くなる。商標審決をはじめとする商標関連の書類に機械翻訳を導入する際には、この点に留意が必要である。

以 上

I. 事業実施報告

1. 本事業の概要

本報告書は、2019年6月から2022年3月にかけて実施された「審査書類・審決の機械翻訳に関する解析及びデータ作成事業（以下、「本事業」という。）」の事業報告書である。本章において、本事業の概要をまとめる。

1.1. 本事業の目的

近年、我が国企業の研究開発活動や企業活動のグローバル化が大きく進展し、国内出願の件数は比較的落ち着きを見せる中、日本から海外への出願割合は増加傾向にあり、海外での早期かつ安定的な権利化を可能とする環境整備が求められている。我が国企業の海外での円滑な権利取得を可能とするためには、日本の高品質な審査・審判結果を外国特許庁に対し正確に伝達すべく、審査・審判書類の高精度な英語翻訳文を提供する必要がある。

現在、特許情報プラットフォーム（J-PlatPat）やワン・ポータル・ドシエ（OPD）の審査書類情報照会システムを通じて、機械翻訳により英語に翻訳された日本の審査・審判書類が外国特許庁で照会可能となっている。これらの機械翻訳文はAIを活用したニューラル機械翻訳等の新技術を用いて提供されているため、従来よりも翻訳品質は大幅に改善されているものの、機械翻訳の精度は今後も改善していく必要がある。

ニューラル機械翻訳により高品質な機械翻訳を提供するためには、翻訳対象とする書類、すなわち日本の審査・審判書類から採取した十分な分量の原文と、その翻訳文とからなる対訳文データ（対訳コーパス。本報告書では単に「コーパス」ともいう。）を翻訳エンジンに学習させる必要がある。そして、翻訳文の精度がコーパスの学習を通じた機械翻訳品質に直結することから、これらの大量のデータを人手翻訳により作成する必要がある。

そこで、本事業では、日本の審査書類・審決について、日英対訳コーパスの作成や、機械翻訳における課題分析を実施するなど、日本の審査書類・審決の高品質な機械翻訳文提供に向けた取組を行う。これにより、日本の高品質な審査・審判情報の対外発信力を強化し、我が国企業の海外での円滑な権利取得を可能とする環境整備を目指す。

1.2. 本事業の内容

本事業では、大別して「審査書類対訳コーパスの作成」、「審決対訳コーパスの作成」、「機械翻訳の調査・分析」の三種の作業を実施した。本項で各作業の詳細を記す。

1.2.1. 審査書類対訳コーパスの作成

審査書類対訳コーパスの作成作業では、日本の特許審査書類のうち「拒絶査定(A02)」「拒絶理由通知書(A131)」「補正却下の決定(A191)」「前置報告書(A913)」「意見書(A53)」のそれぞれから日本語原文を合計 2,000,515 文選定し、これを人手で英訳して、日英対訳文形式の「審査書類対訳コーパス」を作成した。

各書類から選定する文の比率は拒絶理由通知書から 7、それ以外の 4 書類からそれぞれ 1 とした (7 : 1 : 1 : 1 : 1)。対訳コーパスの作成対象とする文は以下の条件に基づき選定した。

- ・書類の本文部分から抽出する
- ・原文 1 文あたりの平均文字数を 100 文字／文程度とする
- ・極端な短文 (20 文字未満) や長文 (250 文字以上) は除外する
- ・各書類とも既にコーパス作成対象とした文パターン¹¹は除外する (重複排除)
- ・過去事業で作成された「作成済み審決コーパス (特許庁貸与物)」に含まれる文パターンは除外する

上記条件により、各書類の審査書類対訳コーパスとも、文字数が 21 文字～249 文字の範囲の非重複文で構成した。

1.2.2. 審決対訳コーパスの作成

審決対訳コーパスの作成作業では、特許庁の選定した特許、実用新案、意匠、商標の審決全 300 案件 (各年度 100 案件) の全文を人手翻訳して「審決英訳文データ」を作成し、当該データから日英対訳文形式の「審決対訳コーパス」を作成した。なお、対訳コーパスの作成対象とする文は以下の条件に基づき選定した。

- ・図表訳は除外する
- ・審決文の末尾に記載された、審判長及び特許庁審判官の人名、「行政事件訴訟法第 46 条に基づく教示」の文、及び「〔審決分類〕」、「〔決定分類〕」の分類記号は除外する
- ・日本語訳文が「当審訳」として併記された外国文献からの引用英文箇所は除外する

¹¹ 完全一致文に加え、変数部分 (数値部分) のみが異なる同一文パターンについても除外した。

1.2.3. 機械翻訳の調査・分析

本事業で作成した審査書類対訳コーパス及び審決対訳コーパスをニューラル方式の機械翻訳エンジン「みんなの自動翻訳@TexTra®」に段階的に学習させ、学習の進行に伴う翻訳品質の改善の度合いや変化の傾向、実用にあたっての課題等について調査分析を行った。調査分析結果については、第Ⅱ部（p.47～）で報告する。

1.3. 本事業の成果物

本事業で作成した成果物は以下のとおりである。

1.3.1. 審査書類対訳コーパス

日本の審査書類 5 種から合計 2,000,515 文を選定し、日本語原文と人手翻訳した英文とによる対訳文形式の「審査書類対訳コーパス」を作成した。審査書類別の作成件数は以下のとおりである。

表 1.3.1. 書類種別ごとの「審査書類対訳コーパス」作成件数

書類種別		コーパス作成件数
(1)	拒絶査定(A02)	181,845 文対
(2)	拒絶理由通知書 (A131)	1,273,011 文対
(3)	補正却下の決定 (A191)	181,896 文対
(4)	前置報告書 (A913)	181,933 文対
(5)	意見書 (A53)	181,830 文対
合 計		2,000,515 文対
(原文の平均文字数)		(100.10 文字)

1.3.2. 審査書類対訳辞書

審査書類対訳コーパスから名詞に相当する語句を抽出して「審査書類対訳辞書」を作成した。言語処理技術を用いて名詞に相当する対訳語の候補を自動抽出し、出現頻度の多い順に目視による採否判断を行って、採用された上位 30,000 語を審査書類対訳辞書とした。

1.3.3. 審決英訳文データ

特許庁が選定した特許、実用新案、意匠、商標の審決を対象に全文の人手翻訳を行い、各年度 100 案件、全 300 案件の「審決英訳文データ」を作成した。

1.3.4. 審決対訳コーパス

前項で作成した審決英訳文データ全 300 案件について、文単位で日本語原文との日英対訳文形式に加工し、「審決対訳コーパス」として全 103,564 文対を作成した。以下、詳細を

記す。

表 1.3.2.「審決対訳コーパス」内訳

四法種別		案件数	コーパス作成件数
(1)	特許	253 案件	94,884 文対
(2)	実用新案	0 案件	0 文対
(3)	意匠	17 案件	4,625 文対
(4)	商標	30 案件	4,055 文対
合 計		300 案件	103,564 文対
(原文の平均文字数)			(54.18 文字)

なお、前述の審査書類対訳コーパス（⇒1.2.1.）が 21 文字～249 文字の範囲の非重複文のみで構成されるのに対し、審決対訳コーパスは審決全文を対訳コーパス化しているため、文字数は制限されず、また重複文も含んでいる。

1.3.5. 審決対訳辞書

前項で作成した「審決対訳コーパス」をソースとして「審決対訳辞書」を作成した。具体的には、本事業で翻訳対象とした審決 300 案件の各案件から、審決文に特有、かつ、特許庁より貸与された既存の「審決対訳辞書」中に存在しない語句を優先して各案件から 10 語程度を対訳語として人手で抽出し、全 3,060 語の新規「審決対訳辞書」を作成した。

1.3.6. 審決汎用文例リスト

審決対訳辞書に記載の用語のうち、汎用性が高いと考えられる用語について、特許庁から貸与された「審決汎用文例リスト」への追加・更新を行った。「審決汎用文例リスト」は、最新のものを翻訳者及び校閲者間で随時共有し、訳語表現の統一に利用した。本事業でリストに新規登録／更新した文例は 364 件であり、これにより、「審決汎用文例リスト」の収録文例は全 2,164 件となった。

1.3.7. 英訳書誌記入済審決リスト

貸与された審決リストに、各案件の日本語原文の参考文字数、四法、審判種別等を追記して「英訳書誌記入済審決リスト」を作成した。

1.3.8. 審決逆翻訳コーパス

海外審決から抽出した英文 2,527,476 文と、これを英日機械翻訳して取得した日本語文とを対訳文形式データに編集し、「審決逆翻訳コーパス」を作成した。

審決逆翻訳コーパスの詳細及び本事業での用途は、第 5.1.1.項を参照されたい。

1.3.9. 審決文の要点の英訳

貸与された審決リストに記載の「要約（審決文の要点）」を人手翻訳した。その後、事前チェックを経て確定した訳文を同リストの所定欄に追記し、各年度 100 案件、全 300 案件の「審決文の要点の英訳」を作成した。

Ⅱ. 調 査 報 告

2. 調査の概要

本章では、本事業の一環として実施した機械翻訳の調査・分析作業の概要を記す。

2.1. 調査の目的と内容

本調査分析は、本事業で作成した「審査書類対訳コーパス」及び「審決対訳コーパス」を学習データに用いた際の機械翻訳の品質改善効果を検証するとともに、日本の審査書類、審決を機械翻訳する際の課題を把握し、その解決策を模索することを目的に実施された。

調査分析では、各コーパスを種別ごとにニューラル機械翻訳エンジンに一定量ずつ段階的に学習させ、学習前後の機械翻訳結果の品質を「自動評価」及び「人手評価」で評価して、学習データの増量に伴う翻訳品質の変化を把握した。

さらに、各学習段階における機械翻訳文の変化の状況を観測し、各文における品質改善の具体的な内容やその傾向、最終的に到達した翻訳品質や残存する課題について分析した。

2.2. 調査対象の機械翻訳エンジン

本調査分析に使用した機械翻訳エンジンは、「みんなの自動翻訳@TexTra®」¹²が提供するニューラル機械翻訳エンジン「特許 NT」である¹³。本エンジンのデフォルト（無学習）状態を起点として、各種コーパスを段階的に学習させていき、各回学習前後の機械翻訳文の内容や品質評価結果を比較した。

2.2.1. 「みんなの自動翻訳」のバージョンアップについて

「みんなの自動翻訳」は定期的にバージョンアップされる。本調査分析期間中も全6回のバージョンアップが実施された。バージョンアップの有無は、各回の機械翻訳処理時に得られるバージョン情報から把握できる。

バージョンアップに伴い旧バージョンは使用不可能となるため、本調査分析では常にその時点で最新のバージョンのエンジンを使用した。以下、調査期間中に実施されたバージョンアップの一覧を示す。

¹² <https://mt-auto-minhon-mlt.ucri.jgn-x.jp/>

¹³ これらは現行サービスの名称であり、本事業開始当初は別名称の旧エンジン（特許 NMT）を用いていた。本報告書では混乱を避けるため、全編にわたり特許 NT で統一した。

表 2.2.1. 「みんなの自動翻訳（特許 NT）」バージョンアップ一覧

バージョン情報	審査書類* 対象学習回	審決* 対象学習回
GPMT-3.8_190930_nmt	無学習～第 2 回	—
GPMT-3.8_200131_nmt	第 3 回～第 7 回	無学習時～第 1 回
GPMT-3.8_200531_nmt	第 8 回～第 10 回	—
GPMT-3.9_200930_nmt	第 11 回～第 12 回	—
GPMT-3.9_210131_nmt	第 13 回～第 14 回	第 2 回
GPMT-3.9_210531_nmt	第 15 回～第 16 回	—
GPMT-3.9_210930_nmt	第 17 回～第 18 回	第 3 回

* 「審査書類」と「審決」の学習頻度の違いについては、後述の第 2.3.項を参照されたい。

2.2.2. 「みんなの自動翻訳」の大規模アップデート

「みんなの自動翻訳」のバージョン情報は、例えば「GPMT-3.8_190531_nmt」のように 2 つの数値を含むが、定期的なバージョンアップでは後半部の 6 桁の日付部分のみが更新される。これに対し、第 11 回学習前に実施されたバージョンアップは、上表のとおり日付部分に加えて前半の数値も 3.8 から 3.9 に更新されており、通常より上位のバージョンアップであることがわかる。

通常の定期的なバージョンアップでは、その前後で機械翻訳の傾向に特段の違いは見られなかったが、この第 11 回学習前のバージョンアップのみは、その後の機械翻訳文の品質や訳語選定の傾向に明白な変化が生じた。本報告書でもその影響について各所で言及することとなるため、本報告書では第 11 回学習前のバージョンアップのみ、他の定期的なバージョンアップと区別して「『みんなの自動翻訳』の大規模アップデート」と称する。

2.3. 機械翻訳システムの学習データと学習のタイミング

「みんなの自動翻訳」の学習では、本事業で作成した審査書類対訳コーパス、審決対訳コーパス及び審決逆翻訳コーパス（⇒1.3.8.）を種別・内容に応じて使い分けた。デフォルト状態の特許 NT エンジンをベースに、学習させるコーパスごとに下表に示す 4 種の機械翻訳システムを構築した。

表 2.3.1. 機械翻訳システム名称と学習に使用した対訳コーパスの対応表

システム名	学習コーパス
審査書類①	審査書類コーパス A (拒絶査定+拒絶理由通知書+補正却下の決定+前置報告書)
審査書類②	審査書類コーパス B (意見書)
審 決③	審決コーパス +作成済み審決コーパス (特許庁貸与物)
審 決④	審決コーパス 審決逆翻訳コーパス +作成済み審決コーパス (特許庁貸与物)

審査書類①システムには審査書類対訳コーパスの対象である 5 種類の審査書類のうち、意見書を除く 4 書類（拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書）のコーパスを一括で学習させた（「審査書類コーパス A」）。一方、審査書類②システムには審査書類対訳コーパスのうち、意見書由来のコーパス（「審査書類コーパス B」）のみを学習させた。

審決③及び審決④システムには審査書類対訳コーパスは使用せず、本事業で作成した審決対訳コーパスと、特許庁貸与物である「作成済み審決コーパス」を学習させて、審決専用チューニングした。なお、審決④システムのみ、さらに「審決逆翻訳コーパス」も学習させた。

機械翻訳システムの学習は、それぞれ学習させるコーパスが所定の文対数作成される都度、段階的に実施した。具体的には、下表に示す文対数（審決は案件数）を基準とした。

表 2.3.2. 機械翻訳システムごとの学習のタイミング及び学習回数

システム名	学習タイミング	学習回数
審査書類①	審査書類コーパス A を 10 万文対作成ごと	18 回
	[内訳] ・拒絶査定……………1 万文対 ・拒絶理由通知書…7 万文対 ・補正却下の決定…1 万文対 ・前置報告書……………1 万文対	
審査書類②	審査書類コーパス B を 1 万文対作成ごと [内訳] ・意見書……………1 万文対	18 回
審 決③	審決コーパスを 100 案件分作成ごと	3 回
審 決④	審決コーパスを 100 案件分作成ごと	3 回

なお、「みんなの自動翻訳」では 1 回学習させたエンジンに対して再度追加の学習を行うことは不可能であるため、各システムとも、学習データを増量する都度、デフォルト状態のエンジンに対して過去分の学習データと一括で学習させ直す形とした。

2.4. 機械翻訳精度の評価

機械翻訳精度の評価は、自動評価指標 BLEU 及び RIBES を用いた「自動評価」と（BLEU 及び RIBES の詳細については 3.2.1. を参照。）、評価者が目視で行う「人手評価」の双方を実施した。各システムの自動評価及び人手評価の実施タイミングは下表のとおりである。なお、各システムともコーパス学習前のデフォルト状態（無学習時）の機械翻訳結果についても評価しており、下表の評価回数はその 1 回分を含んでいる。

表 2.4.1. 機械翻訳システムごとの評価のタイミング及び評価回数

システム名	評価タイミング		評価回数
審査書類①	自動	審査書類コーパス A を 10 万文対学習の都度	19 回
	人手	審査書類コーパス A を 50 万文対学習の都度	5 回 ¹⁴
審査書類②	自動	審査書類コーパス B を 1 万文対学習の都度	19 回
	人手	審査書類コーパス B を 5 万文対学習の都度	5 回
審 決③	自動	審決コーパスを 100 案件分学習の都度	4 回
	人手		
審 決④	自動	審決コーパスを 100 案件分学習の都度	4 回
	人手		

2.5. 評価用コーパス

本調査分析で評価する機械翻訳文には、書類種別ごとに選定した「評価用コーパス」を各学習段階のシステムで機械翻訳した結果を用いた。

評価用コーパスは、審査書類対訳コーパスに含まれる 5 種の審査書類と審決の計 6 種類を作成した。具体的には、各審査書類／審決コーパスの初回作成分からそれぞれ 2,000 文対程度を選定した。選定した評価用コーパスは、各システムの無学習時から最終学習後まで全回の評価に使用した。

なお、本調査分析では審査書類対訳コーパス及び審決対訳コーパスは原則として全件を学習データとして使用したが、評価用コーパスとして選定した文対は学習データから除外している。

¹⁴ 審査書類①及び②の人手評価は、無学習時の 1 回、所定件数ごとの 3 回に加え、全件完成時（審査書類コーパス A：約 181 万文対、同 B：約 18 万文対）に最終評価を実施した。このため全 5 回となる。

2.5.1. 文の種類の変り回避

評価用コーパスの要件として、文の種類（認定・対比・判断・その他）に偏りが無いよう選定することが求められる。

「偏りが無い」とは、必ずしも認定・対比・判断の比率が同等（1：1：1）であることを意味せず、実際の分布状況から著しく逸脱していないことを指す。また、各書類とも、これら3文種（認定・対比・判断）のいずれにも属さない文も多数含んでおり、これらについても本事業の評価対象として同等に取り扱う必要がある。

これに鑑み、評価用コーパスの選定においては、各書類から2,000文ずつを無作為抽出することを基本方針とした。これが、各書類における「文の種類（認定・対比・判断・その他）」に関して、コーパス全体での構成比率を反映させ、偏りなく選定する手法として最も蓋然性が高いと考えた。

ただし、上記手法を用いても、偶然の結果として、特定の文種が極端に多くなる又は少なくなるなど、偏った選定内容となる可能性も皆無ではない。そこで、この可能性を排除するために、無作為抽出した各評価用コーパス（2,000文）について、それぞれの40%（800文）程度を目視で判定し、認定／対比／判断／その他の比率を調査することにした¹⁵。以下、その結果を示す。

表 2.5.1. 文の種類の変り判定の結果

評価用コーパス名	認定	対比	判断	その他	合計
拒絶査定	225	70	438	67	800
拒絶理由通知書	251	69	446	34	800
補正却下の決定	224	80	429	67	800
前置報告書	211	61	469	59	800

サンプリングの結果、上表のとおり各書類とも「判断」が過半数を占め、「認定」が3割程度、「対比」と「その他」は1割弱という、きわめて相似した構成比率となった。文の種類ごとに顕著な多寡は見られるものの、各書類とも同様の構成比率であることから、それぞれの文種の性質に起因した、実際の構成比率を反映したものであると判断した。

したがって、各書類とも無作為抽出した約2,000文をそのまま評価用コーパスとして採

¹⁵ 「意見書」に関しては、その性質上「認定・対比・判断」という論理構成を採らないと考えられることから、サンプル判定は行わず、無作為抽出の結果をそのまま採用した。

用した。

2.5.2. 技術分野の偏り回避

評価用コーパスの無作為抽出は、IPC 情報に基づく技術分野ごとに実施し、コーパスが特定の技術分野に偏らないよう配慮した。

2.5.3. 極端な長文と短文の排除

本事業では、日英対訳コーパスの作成にあたり、あらかじめ極端な長文と短文を排除しており、具体的には原文文字数が 20 文字以上～249 文字以下の範囲に属する文のみを作成対象とした（⇒1.2.1.項）。これは、極端な長文は機械翻訳エンジンへのアダプテーションが物理的に不可能であり、また、極端な短文は見出しや文献番号、数値データなど、やはり学習データとしての利用性の低い文である確率が高いことが主たる理由である。

評価用コーパスの選定に際しては文長に関する特段の調整は行っていないものの、上記により、評価用コーパスの全ての文は原文文字数 20 文字以上～249 文字以下の範囲内のものとなる。

2.5.4. 審決の評価用コーパスについて

審決の品質評価は、既存及び新規の審決対訳コーパスのみを学習させた審決③システムと、これらに加えて更に審決逆翻訳コーパスも学習させた審決④システムの 2 種のシステムの機械翻訳文が評価の対象となる。

2 種のシステムを用いるのは、審決逆翻訳コーパスの使用による翻訳品質の違いを測るためである。このため、審決③と審決④には共通の評価用コーパスを用い、正確な相対評価が行えるようにした。

また、本事業で作成する審決対訳コーパスは、審査書類対訳コーパスと異なり、案件単位で作成される。したがって、評価用コーパス選定時に、案件中の一部の文のみを選定すると、同一案件由来の文が評価用コーパスと学習データとに併存するという、実用時には想定されない状況となり、正確な品質評価を妨げかねない。つまり同一案件由来の文は用語や表現が共通することが多く、翻訳精度面で不当に有利となる。このため、審決③／④の評価用コーパスは原則として案件単位で選定した。

以上、本事業で実施した調査分析の概要を記した。次章より、調査対象の機械翻訳システムごとに、第 3 章で審査書類①システム、第 4 章で審査書類②システム、そして第 5 章で審決③／④システムの評価・分析結果をまとめる。

3. 審査書類①システムの評価結果とその分析

本章で、審査書類①システムの機械翻訳品質の評価結果についてまとめる。

3.1. 審査書類①システムの評価の概要

審査書類①システムは、「みんなの自動翻訳」の特許 NT エンジン¹⁶をベースとし、これに「審査書類コーパス A」、すなわち本事業で作成した審査書類対訳コーパスのうち「拒絶査定(A02)」、「拒絶理由通知書(A131)」、「補正却下の決定(A191)」及び「前置報告書 (A913)」の 4 書類に由来するコーパスを一括で学習させたものである。

学習は、審査書類コーパス A を 10 万文対作成する都度¹⁶段階的に実施し、計 18 回、最終的には 1,810,685 文を学習させた。各回の学習データにおける 4 種の審査書類の構成比率は常時 1 : 7 : 1 : 1 であり、拒絶理由通知書由来のデータが全体の 70%、残る 3 種がそれぞれ 10%ずつを占める。

評価には、自動評価指標 BLEU 及び RIBES による自動評価と、評価者が目視で行う人手評価とを併用した。自動評価は学習の都度、人手評価は学習データを 50 万文対適用する都度（すなわち学習 5 回おき）に実施した。

評価にあたっては、審査書類①に属する 4 種の審査書類それぞれについて 2,000 文対、計 8,000 文対を「評価用コーパス」に選定し、これを各学習段階の審査書類①システムで機械翻訳した結果を評価対象とした。ただし、人手評価は各書類とも評価用コーパスのうち 100 文対、計 400 文対のみを対象とした。

¹⁶ 初回学習時は 10 万文対のうち 8,000 文対（2,000 文対×4 書類）を評価用コーパスとして選定しており、これらを除いた 92,000 文対を学習させた。また、最終学習時は 118,685 文対を学習させた。

3.2. 審査書類①システムの自動評価結果

本項では、審査書類①システムに対して実施した自動評価の結果を記す。

3.2.1. BLEU と RIBES

自動評価に使用した二種類の指標 BLEU¹⁷及び RIBES¹⁸は、いずれもあらかじめ与えられた「正解訳文」を評価の基準とし、評価対象の機械翻訳文がこれとどの程度一致しているかを、0.0000（最低スコア）～1.0000（最高スコア）の範囲内の評価スコアで示す仕組みである。

本調査分析に使用した評価用コーパスは審査書類対訳コーパスから抽出しており、日本語原文と、本事業で人手翻訳した英文との対訳文形式のデータである。このデータの日本語原文を機械翻訳の入力文とし、英文を自動評価の「正解訳文」として使用した。

図 3.2.1. BLEU、RIBES の自動評価スコア算出方法の概要

評価用コーパス (A02-#1)		機械翻訳結果 (特許 NT 第 18 回学習後)
原文	人手翻訳文	
また、同下記刊行物 4 の段落 0 0 0 5 等には、エアバッグセンサを支持する部材の剛性を強化することによって、エアバッグセンサの検出性能を向上させるという技術思想が開示されている。	Furthermore, paragraph 0005 etc. in Publication 4 discloses the technical idea that the detection performance of the air bag sensor is improved by strengthening the rigidity of the member supporting the air bag sensor.	In addition, paragraph 0005, etc. of the following Publication 4 disclose a technical idea of improving the detection performance of the airbag sensor by strengthening the rigidity of the member supporting the airbag sensor.

評価用コーパスの「人手翻訳文」と原文の「機械翻訳結果」とを比較し、その一致度に基づいてスコアを算出する。

BLEU:0.3978、RIBES : 0.8946

¹⁷ Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu : BLEU : a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

¹⁸ 平尾努、磯崎秀樹、須藤克仁、Duh Kevin、塚田元、永田昌明：語順の相関に基づく機械翻訳の自動評価法、自然言語処理、Vol.21, No.3, pp.421–444, 2014.

3.2.2. 審査書類①全体の自動評価結果

自動評価は、無学習時（すなわちデフォルト状態の「みんなの自動翻訳」特許 NT エンジン）の 1 回と、審査書類コーパス A を 10 万文対ずつ段階的に学習させる都度の 18 回、計 19 回実施した。

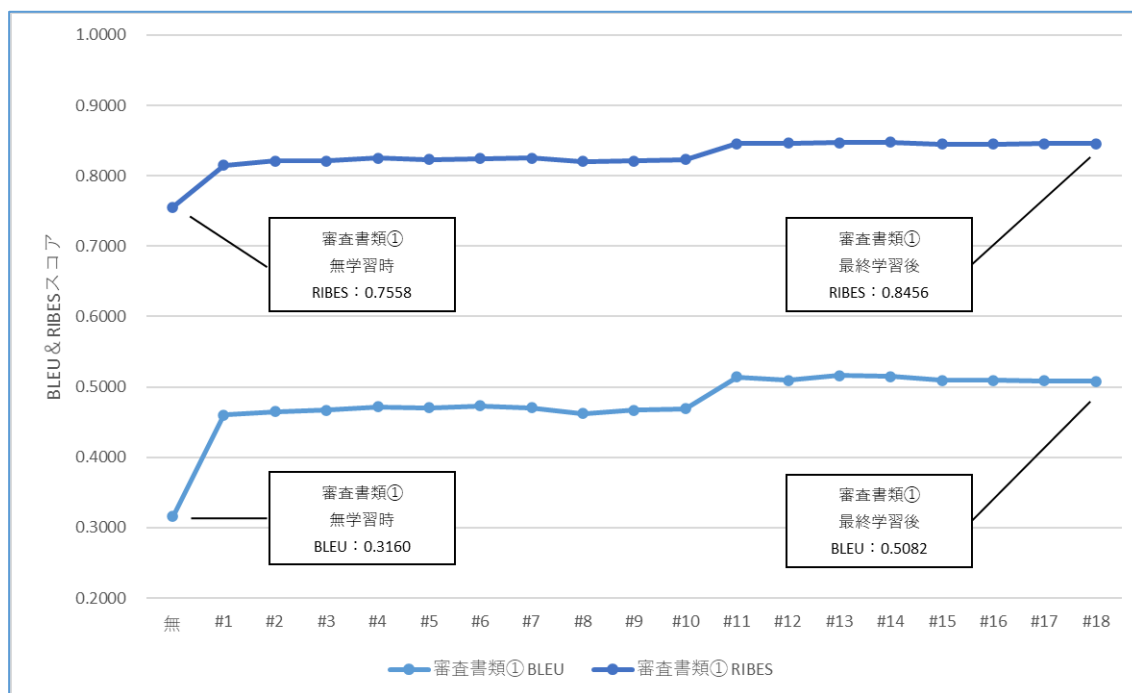
審査書類①システムには 4 種の審査書類（拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書）に由来する対訳コーパスを一括で学習させたが、評価対象となる評価用コーパスは書類別に 2,000 文対ずつ選定している。このため、スコアを書類ごとに集計することも可能であるが、本項ではまず審査書類①全体、すなわち 4 書類の評価用コーパス全 8,000 文の平均スコアを分析する。下二表に BLEU、RIBES それぞれの集計結果を示す。

表 3.2.1. 審査書類①（8,000 文）の BLEU/RIBES 全件平均スコア

学習回数	学習データ量	バージョン ¹⁹	BLEU		RIBES	
無学習時	0	3.8_190930	0.3160	前回比	0.7558	前回比
第 1 回学習後	92,000	3.8_190930	0.4602	+0.1442	0.8150	+0.0592
第 2 回学習後	192,000	3.8_190930	0.4648	+0.0046	0.8210	+0.0060
第 3 回学習後	292,000	3.8_200131	0.4668	+0.0020	0.8209	-0.0001
第 4 回学習後	392,000	3.8_200131	0.4721	+0.0053	0.8253	+0.0044
第 5 回学習後	492,000	3.8_200131	0.4704	-0.0017	0.8235	-0.0018
第 6 回学習後	592,000	3.8_200131	0.4734	+0.0030	0.8245	+0.0010
第 7 回学習後	692,000	3.8_200131	0.4702	-0.0032	0.8251	+0.0006
第 8 回学習後	792,000	3.8_200531	0.4620	-0.0082	0.8205	-0.0046
第 9 回学習後	892,000	3.8_200531	0.4671	+0.0051	0.8213	+0.0008
第 10 回学習後	992,000	3.8_200531	0.4692	+0.0021	0.8232	+0.0019
第 11 回学習後	1,092,000	3.9_200930	0.5142	+0.0450	0.8458	+0.0226
第 12 回学習後	1,192,000	3.9_200930	0.5093	-0.0049	0.8461	+0.0003
第 13 回学習後	1,292,000	3.9_210131	0.5163	+0.0070	0.8472	+0.0011
第 14 回学習後	1,392,000	3.9_210131	0.5151	-0.0012	0.8476	+0.0004
第 15 回学習後	1,492,000	3.9_210531	0.5096	-0.0055	0.8448	-0.0028
第 16 回学習後	1,592,000	3.9_210531	0.5096	±0	0.8450	+0.0002
第 17 回学習後	1,692,000	3.9_210930	0.5085	-0.0011	0.8458	+0.0008
第 18 回学習後	1,810,685	3.9_210930	0.5082	-0.0003	0.8456	-0.0002
スコア向上度（無学習時と第 18 回後の差）			+0.1922		+0.0898	

¹⁹ 「みんなの自動翻訳」のバージョン情報は正確には“GPMT-3.8_1905930_nmt”などのフォーマットであるが（⇒2.2.1.）、表中ではスペースの都合上、中盤の数値部分のみ表示している。

図 3.2.1. 審査書類①（8,000 文）の BLEU、RIBES 全件平均スコア推移



上掲のとおり、審査書類①システムの機械翻訳文に付された自動評価スコア（評価用コーパス 8,000 文の全件平均）は、無学習時と最終学習後の比較で BLEU が+0.1922、RIBES が+0.0898 といずれも顕著に向上した。

学習回ごとのスコアの変動を見ると、両評価指標とも初回学習後にスコアが最も大きく向上しており（BLEU が前回比+0.1442、RIBES が+0.0592）、それぞれ全期間を通じたスコア伸び幅の 75%、66%を占めた。

その後、第 2 回学習後以降は両指標とも誤差レベルの微増微減を繰り返す傾向に転じ、第 2 回学習後から第 10 回学習後までのスコアの合計伸び幅は BLEU が+0.0090、RIBES が+0.0082 と、初回学習後に比べるとごく小規模な向上にとどまった。

このように第 2 回以降長期にわたり微動傾向が続くなか、第 11 回学習後のみは BLEU が+0.0450、RIBES が+0.0226 と初回時に次ぐ顕著なスコア向上が見られた。これは第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）の影響が大きいと考えられる。自動評価スコアからは品質改善の具体的な内容までは読み取れないが、第 3.2.7.項で後述する実例の定点観測や第 3.3.項にまとめた人手評価結果では、この大規模アップデート以降、学習データの内容がよりの確に翻訳文に反映されるようになるとともに、

旧エンジンで発生していた学習データでは改善不可能なタイプのエラーが多数解消されたことが確認されている。

その後、第 12 回学習後以降の全体平均スコアは、第 11 回学習後に向上した水準を保ちつつ再び微動傾向となった。これは、大規模アップデート時に得られた品質改善効果が一時的なものではなく、最終学習後まで常時維持されたことを示している。

以上、各回学習後の BLEU 及び RIBES の自動評価スコアの集計値の推移を概観した。両指標とも審査書類対訳コーパスの学習によってスコアが顕著に向上しており、本事業で作成したコーパスが高い翻訳品質改善効果を有していることが示された。

3.2.3. 審査書類ごとの自動評価結果

本項では、審査書類①に属する 4 種の審査書類ごとに自動評価スコアを集計し、その平均値を比較した。書類ごとの評価用コーパス各 2,000 文の BLEU と RIBES の平均スコア、及び各学習回における 4 書類間の相対順位を下記三表に示す。

表 3.2.2. 審査書類ごと（4 書類、各 2,000 文）の BLEU 平均スコア

学習回数	拒絶査定(A02)		拒絶理由(A131)		補正却下(A191)		前置報告(A913)	
無学習時	0.3387	前回比	0.3298	前回比	0.2658	前回比	0.3296	前回比
第 1 回学習後	0.4736	+0.1349	0.4550	+0.1252	0.4061	+0.1403	0.5059	+0.1763
第 2 回学習後	0.4748	+0.0012	0.4586	+0.0036	0.4090	+0.0029	0.5167	+0.0108
第 3 回学習後	0.4771	+0.0023	0.4603	+0.0017	0.4153	+0.0063	0.5144	-0.0023
第 4 回学習後	0.4852	+0.0081	0.4660	+0.0057	0.4127	-0.0026	0.5245	+0.0101
第 5 回学習後	0.4861	+0.0009	0.4599	-0.0061	0.4117	-0.0010	0.5240	-0.0005
第 6 回学習後	0.4864	+0.0003	0.4611	+0.0012	0.4159	+0.0042	0.5301	+0.0061
第 7 回学習後	0.4802	-0.0062	0.4611	±0	0.4171	+0.0012	0.5222	-0.0079
第 8 回学習後	0.4729	-0.0073	0.4545	-0.0066	0.4085	-0.0086	0.5119	-0.0103
第 9 回学習後	0.4749	+0.0020	0.4584	+0.0039	0.4161	+0.0076	0.5190	+0.0071
第 10 回学習後	0.4821	+0.0072	0.4600	+0.0016	0.4150	-0.0011	0.5197	+0.0007
第 11 回学習後	0.5344	+0.0523	0.4936	+0.0336	0.4511	+0.0361	0.5779	+0.0582
第 12 回学習後	0.5313	-0.0031	0.4911	-0.0025	0.4482	-0.0029	0.5667	-0.0112
第 13 回学習後	0.5368	+0.0055	0.4998	+0.0087	0.4512	+0.0030	0.5775	+0.0108
第 14 回学習後	0.5353	-0.0015	0.5013	+0.0015	0.4533	+0.0021	0.5705	-0.0070
第 15 回学習後	0.5280	-0.0073	0.4934	-0.0079	0.4527	-0.0006	0.5644	-0.0061
第 16 回学習後	0.5282	+0.0002	0.4946	+0.0012	0.4546	+0.0019	0.5612	-0.0032
第 17 回学習後	0.5239	-0.0043	0.4948	+0.0002	0.4545	-0.0001	0.5607	-0.0005
第 18 回学習後	0.5259	+0.0020	0.4938	-0.0010	0.4506	-0.0039	0.5625	+0.0018
スコア向上度	+0.1872		+0.1640		+0.1848		+0.2329	

表 3.2.3. 審査書類ごと（4 書類、各 2,000 文）の RIBES 平均スコア

学習回数	拒絶査定(A02)		拒絶理由(A131)		補正却下(A191)		前置報告(A913)	
無学習時	0.7669	前回比	0.7668	前回比	0.7254	前回比	0.7641	前回比
第 1 回学習後	0.8156	+0.0487	0.8168	+0.0500	0.7950	+0.0696	0.8329	+0.0688
第 2 回学習後	0.8221	+0.0065	0.8198	+0.0030	0.7992	+0.0042	0.8427	+0.0098
第 3 回学習後	0.8224	+0.0003	0.8230	+0.0032	0.7965	-0.0027	0.8416	-0.0011
第 4 回学習後	0.8261	+0.0037	0.8266	+0.0036	0.8014	+0.0049	0.8470	+0.0054
第 5 回学習後	0.8261	±0	0.8218	-0.0048	0.8041	+0.0027	0.8421	-0.0049
第 6 回学習後	0.8285	+0.0024	0.8194	-0.0024	0.8021	-0.0020	0.8481	+0.0060
第 7 回学習後	0.8252	-0.0033	0.8258	+0.0064	0.8041	+0.0020	0.8451	-0.0030
第 8 回学習後	0.8220	-0.0032	0.8209	-0.0049	0.7964	-0.0077	0.8427	-0.0024
第 9 回学習後	0.8221	+0.0001	0.8223	+0.0014	0.7977	+0.0013	0.8432	+0.0005
第 10 回学習後	0.8246	+0.0025	0.8218	-0.0005	0.7993	+0.0016	0.8471	+0.0039
第 11 回学習後	0.8479	+0.0233	0.8416	+0.0198	0.8239	+0.0246	0.8697	+0.0226
第 12 回学習後	0.8489	+0.0010	0.8406	-0.0010	0.8267	+0.0028	0.8680	-0.0017
第 13 回学習後	0.8508	+0.0019	0.8422	+0.0016	0.8258	-0.0009	0.8701	+0.0021
第 14 回学習後	0.8503	-0.0005	0.8432	+0.0010	0.8283	+0.0025	0.8688	-0.0013
第 15 回学習後	0.8451	-0.0052	0.8430	-0.0002	0.8237	-0.0046	0.8674	-0.0014
第 16 回学習後	0.8453	+0.0002	0.8429	-0.0001	0.8249	+0.0012	0.8668	-0.0006
第 17 回学習後	0.8442	-0.0011	0.8443	+0.0014	0.8268	+0.0019	0.8679	+0.0011
第 18 回学習後	0.8449	+0.0007	0.8427	-0.0016	0.8268	±0	0.8679	±0
スコア向上	+0.0780		+0.0759		+0.1014		+0.1038	

表 3.2.4. 審査書類間の自動評価スコア相対順位（上段：BLEU、下段：RIBES）

	無	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
拒絶査定	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
(A02)	1	3	2	2	3	2	2	3	2	3	2	2	2	2	2	2	2	3	2
拒絶理由	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
(A131)	2	2	3	3	2	3	3	2	3	2	3	3	3	3	3	3	3	2	3
補正却下	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
(A191)	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
前置報告	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
(A913)	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

審査書類①の学習データ（審査書類コーパス A）は4種の審査書類由来の文で構成されるが、その構成比は均等ではなく、1：7：1：1と拒絶理由通知書のみ他の3書類の7倍のデータ量となっている²⁰。この学習データ量の多寡によって拒絶理由通知書の評価スコアが他と異なる傾向を見せるかが本項の焦点となる。

無学習時は各書類とも特許 NT エンジンでデフォルト状態で使用しており、審査書類コーパス A は未使用である。したがって、書類間のスコア相対順位は、それぞれの評価用コーパス 2,000 文の機械翻訳難易度の差と見なせる。無学習時における拒絶理由通知書の書類間のスコア順位は BLEU、RIBES とともに首位と僅差の第 2 位であった。

審査書類コーパス A の初回学習以降は、各回で拒絶理由通知書は 7 万文対、他の 3 書類は 1 万文対ずつ学習データが増量される²¹。このため、学習データの比率としては常に 7 倍、絶対量では回を重ねるごとに 6 万文対ずつ拒絶理由通知書が他を引き離すことになる。しかしながら、こうした状況にもかかわらず、拒絶理由通知書の相対順位はコーパス学習後にむしろ低下する傾向を見せた。BLEU では初回学習後に第 3 位に下降し、最終学習後までその順位にとどまった。RIBES では第 2 位を維持した回数も数回あったが、こちらも学習が進むにつれてほぼ第 3 位に定着した。

一方、他の 3 書類に関しては、それぞれの学習データ中での比率は常時同一であったが、各書類間の相対順位を見ると、無学習時に第 3 位であった前置報告書が初回学習後以降は第 1 位に定着し、無学習時に第 1 位であった拒絶査定はコーパス学習後に第 2 位へ後退するなど、こちらもコーパス学習前後で順位の変動が生じている（ただし、補正却下の決定は常時第 4 位であった。）。

こうした相対順位の推移を見る限り、書類ごとの学習データ中の比率の多寡と品質改善の度合いとの間に相関性は見いだせない。拒絶理由通知書に由来する学習データは他の書類の 7 倍を占めたが、評価スコアは全ての書類でおおむね同程度に向上しており、拒絶理由通知書のみが突出することはなかった。むしろ無学習時にスコアが低かった書類ほど全学習期間を通じてのスコアの伸びは大きくなっており、学習データの多寡と無関係に書類間の品質の均質化が進んでいる。

²⁰ この比率は本事業の調達仕様に定められたものである。

²¹ ただし、初回学習時のみは各書類とも評価用コーパスに選定した 2,000 文対が除外されるため、拒絶理由通知書が 6.8 万文対、他の 3 書類がそれぞれ 0.8 万文対となる。

また、拒絶理由通知書以外の 3 書類に関しては、学習データ中に占める比率はそれぞれ 10%であり、学習データの 90%が他種の書類由来の文であったことになるが、自動評価スコアを見る限り、このことが翻訳品質に悪影響を及ぼした形跡は見られない。

こうした状況から、審査書類①に属する各書類由来のコーパスは、学習データとして互換性が高いと考えられる。これは、4 種の審査書類が同じ特許制度に基づく審査手続の過程で特許庁審査官が作成する書類であり、記載内容は異なるにせよ互いに悪影響を及ぼすような性質の違いがないこと、かつ、使われる専門用語や表現もおおむね共通していることを考えれば、妥当な結果といえる。

なお、各書類のスコアを相対比較すると、補正却下の決定のみ無学習時から最終学習後まで全期間を通じてスコア水準がほかよりやや低い。これは、補正却下の決定の評価用コーパスが、他の 3 書類の評価用コーパスよりも機械翻訳難易度が高かったことを示している。各書類の評価用コーパスの平均文長や長文の含有率に差はなく、補正却下の決定という書類自体の記載内容が、概して他の 3 書類よりも機械翻訳難易度が高いという可能性がある。この点については、後述する人手評価結果も踏まえて結論する（⇒3.3.5.1.）。

3.2.4. 技術分野間の比較

本項では、審査書類①の自動評価スコアを技術分野ごとに集計した結果を分析する。4 書類の評価用コーパス全 8,000 文を原出願の筆頭 IPC に基づいて「化学」、「電気物理」、「機械」の三分野に分類し、それぞれを集計した。結果を下表に示す。

表 3.2.5.審査書類①技術分野ごとの BLEU、RIBES 平均スコア

学習回	全件		化学		電気物理		機械	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
無学習	0.3160	0.7558	0.3094	0.7534	0.3218	0.7589	0.3175	0.7551
第 1 回	0.4602	0.8150	0.4564	0.8134	0.4613	0.8185	0.4625	0.8133
第 2 回	0.4648	0.8210	0.4621	0.8209	0.4686	0.8227	0.4638	0.8192
第 3 回	0.4668	0.8209	0.4610	0.8206	0.4685	0.8207	0.4708	0.8210
第 4 回	0.4721	0.8253	0.4683	0.8235	0.4736	0.8260	0.4744	0.8264
第 5 回	0.4704	0.8235	0.4672	0.8248	0.4730	0.8239	0.4713	0.8217
第 6 回	0.4734	0.8245	0.4689	0.8237	0.4782	0.8257	0.4735	0.8241
第 7 回	0.4702	0.8251	0.4646	0.8252	0.4753	0.8269	0.4713	0.8234
第 8 回	0.4620	0.8205	0.4590	0.8223	0.4646	0.8216	0.4623	0.8177
第 9 回	0.4671	0.8213	0.4628	0.8203	0.4681	0.8207	0.4705	0.8228
第 10 回	0.4692	0.8232	0.4642	0.8228	0.4722	0.8254	0.4714	0.8216
第 11 回	0.5142	0.8458	0.5103	0.8459	0.5161	0.8462	0.5162	0.8450
第 12 回	0.5093	0.8461	0.5053	0.8463	0.5122	0.8477	0.5105	0.8441
第 13 回	0.5163	0.8472	0.5130	0.8473	0.5173	0.8480	0.5184	0.8462
第 14 回	0.5151	0.8476	0.5109	0.8461	0.5168	0.8494	0.5175	0.8475
第 15 回	0.5096	0.8448	0.5056	0.8433	0.5123	0.8473	0.5108	0.8442
第 16 回	0.5096	0.8450	0.5057	0.8445	0.5113	0.8487	0.5118	0.8423
第 17 回	0.5085	0.8458	0.5030	0.8453	0.5117	0.8490	0.5105	0.8433
第 18 回	0.5082	0.8456	0.5032	0.8448	0.5098	0.8490	0.5111	0.8435
向上	+0.1922	+0.0898	+0.1938	+0.0914	+0.1880	+0.0901	+0.1936	+0.0884

審査書類には技術用語・表現も多く用いられ、かつ、あらゆる技術分野が対象となり得る。このため技術用語全般への対応度も機械翻訳品質に大きく影響する。一方、審査書類対訳コーパスは、技術用語も多数含んではいるが、あらゆる技術用語を網羅的にカバーする性質のものではない。よって、技術用語の翻訳品質は、機械翻訳エンジンのデフォルト性能、つまり基礎学習データの充実度に委ねられる部分が大きくなる。本項に示した技術分野ごとの集計結果も、「みんなの自動翻訳」のデフォルトでの各技術分野の翻訳品質に偏りがいないか、

顕著に品質が劣る分野が存在しないかを確認することが主な目的となる。

各分野の特許・審査書類用語の翻訳難易度を同程度と仮定すれば、上表に示した各分野間のスコア差が、「みんなの自動翻訳」による技術用語の翻訳品質の差と見なせる。その観点で各分野のスコアを見ると、無学習時、最終学習後とも各技術分野のスコアはおおむね同水準の範囲内にあり、スコアが顕著に劣る分野は存在しない。最終学習後の相対順位を見ても、BLEU では「機械」→「電気物理」→「化学」、RIBES では「電気物理」→「化学」→「機械」と両者異なり、各分野の品質が僅差で拮抗していることがうかがえる。

分野ごとに異なる文を比較しているため一定の誤差が存在すると考えると、本項に示した集計結果からは、「みんなの自動翻訳」の基礎学習内容に技術分野間の顕著な偏りはなく、各技術分野がおおむね均等にカバーされていると結論できる。

ただし、「みんなの自動翻訳」における技術用語の翻訳品質が実用に足るレベルであるかは、自動評価スコアからは直接判断できない。また、自動評価は文全体に対して評価を行うため、仮にスコアが低調であっても、それが技術用語の問題なのか、特許・審査書類用語など他の理由によるものかは判別できない。この点に関しては、後述する人手評価結果などから判断する必要がある。

3.2.5. 文の種類間の比較

審査書類①の評価に用いた4種類の評価用コーパスでは、それぞれ2,000文中の800文について「認定」「対比」「判断」「その他²²」という文の種類の特定（ラベリング）を行っている。下二表に、これら各書類800文の自動評価スコアを文の種類ごとに集計した結果を示す。各文種とも1行目に無学習時、2行目に初回である第1回学習後、3行目に最終回である第18回学習後のスコアを示した。4行目には無学習時と第18回学習後のスコアの差分（向上度）、5行目には800文に占める各文種の文数を示している。スコアに添えられたカッコ数値は、各書類における4文種間の相対順位である。

表 3.2.6. 審査書類①文種ごとの自動評価結果の推移（BLEU）

書類		全件	認定	対比	判断	その他
拒絶査定	無	0.3387	0.3557(2)	0.4164(1)	0.3305(3)	0.2658(4)
	#01	0.4736	0.4706(2)	0.5079(1)	0.4687(3)	0.4448(4)
	#18	0.5259	0.5369(2)	0.5685(1)	0.5074(4)	0.5084(3)
	向上	+0.1872	+0.1812	+0.1521	+0.1769	+0.2426
	文数	2,000 文	225 文	70 文	438 文	67 文
拒絶理由 通知書	無	0.3298	0.3636(1)	0.3340(2)	0.3282(3)	0.2464(4)
	#01	0.4550	0.4614(1)	0.4558(2)	0.4518(3)	0.4017(4)
	#18	0.4938	0.5143(2)	0.5368(1)	0.4821(3)	0.4334(4)
	向上	+0.1640	+0.1507	+0.2028	+0.1539	+0.1870
	文数	2,000 文	251 文	69 文	446 文	34 文
補正却下 の決定	無	0.2658	0.2826(2)	0.3149(1)	0.2690(3)	0.2578(4)
	#01	0.4061	0.4060(3)	0.4482(1)	0.4128(2)	0.4025(4)
	#18	0.4506	0.4590(3)	0.4661(1)	0.4651(2)	0.4348(4)
	向上	+0.1848	+0.1764	+0.1512	+0.1961	+0.1770
	文数	2,000 文	224 文	80 文	446 文	34 文
前置 報告書	無	0.3296	0.3385(2)	0.4262(1)	0.3258(3)	0.2575(4)
	#01	0.5059	0.4680(3)	0.5722(1)	0.5142(2)	0.4652(4)
	#18	0.5625	0.5380(4)	0.6187(1)	0.5597(2)	0.5410(3)
	向上	+0.2329	+0.1995	+0.1925	+0.2339	+0.2835
	文数	2,000 文	211 文	61 文	469 文	59 文

²² 「その他」に該当する文は多岐にわたるが、一例としては、出願人への助言（「なお、審判請求される場合には下記の点についても留意されたい。」）や、事実の説明（「なお、請求項8－10については、平成23年11月14日付け手続補正によって新たに追加された請求項であるから、拒絶査定の対象とはなっていない。」）、原出願からの引用文、ただし書などが挙げられる。

表 3.2.7.審査書類①文種ごとの自動評価結果の推移 (RIBES)

書類		全件	認定	対比	判断	その他
拒絶査定	無	0.7669	0.7728(2)	0.7999(1)	0.7624(3)	0.7301(4)
	#01	0.8156	0.8303(2)	0.8346(1)	0.7999(3)	0.7929(4)
	#18	0.8449	0.8612(2)	0.8653(1)	0.8317(3)	0.8288(4)
	向上	+0.0780	+0.0884	+0.0654	+0.0693	+0.0987
	文数	2,000 文	225 文	70 文	438 文	67 文
拒絶理由 通知書	無	0.7668	0.8072(1)	0.7823(2)	0.7522(3)	0.7110(4)
	#01	0.8168	0.8494(1)	0.8345(2)	0.7937(4)	0.7954(3)
	#18	0.8427	0.8715(2)	0.8771(1)	0.8274(4)	0.8369(3)
	向上	+0.0759	+0.0643	+0.0948	+0.0752	+0.1259
	文数	2,000 文	251 文	69 文	446 文	34 文
補正却下 の決定	無	0.7254	0.7591(1)	0.7388(2)	0.7234(3)	0.7225(4)
	#01	0.7950	0.8152(2)	0.8126(3)	0.7875(4)	0.8175(1)
	#18	0.8268	0.8384(2)	0.8299(3)	0.8235(4)	0.8414(1)
	向上	+0.1014	+0.0793	+0.0911	+0.1001	+0.1189
	文数	2,000 文	224 文	80 文	446 文	34 文
前置 報告書	無	0.7641	0.7918(2)	0.8239(1)	0.7560(3)	0.7474(4)
	#01	0.8329	0.8295(3)	0.8812(1)	0.8258(4)	0.8504(1)
	#18	0.8679	0.8804(2)	0.8975(1)	0.8589(4)	0.8755(3)
	向上	+0.1038	+0.0886	+0.0736	+0.1029	+0.1281
	文数	2,000 文	211 文	61 文	469 文	59 文

上表のとおり、文種間の相対順位は書類ごとに異なり、また、学習前後でも順位が入れ替わるなど、文種間の翻訳品質の優劣に明瞭な傾向は見いだせない。ただし、全体的には、「認定」と「対比」が上位、「判断」と「その他」が下位となる傾向がうかがえる。

各書類の最下段（5 行目）に示したとおり、800 文における文種ごとの構成比率は各書類おおむね同様である。各書類とも「判断」が 450 文前後と過半数を占め、続いて「認定」が 200～250 文前後でこれに続き、「対比」と「その他」はそれぞれ 80～61 文、67～34 文と少数である。文種をラベリングした 800 文は無作為に選定しており、審査書類対訳コーパス全体においてもおおむね同様の構成比率である蓋然性が高い²³。

²³ ただし、審査書類対訳コーパスのこうした構成比は、対象文の選定時に文長制限や重複文の排除などの条件を設けた結果であり（⇒1.2.1.）、各審査書類そのものがこの構成比率であるとは限らない。

その場合、学習データ中でも「判断」文が過半数を占め、「対比」と「その他」はごく少数ということになる。しかしながら、実際の自動評価スコアの集計結果では「判断」は中位から下位、「対比」が最上位となるなど、学習データの構成比と評価スコアとの間に相関は見いだせない。むしろ各書類とも、無学習時における文種間の相対順位がコーパス学習後もそのまま維持され、かつ、無学習時の相対順位が低い文種ほど最終学習後までのスコアの伸び幅が大きく、文種間のスコア差が縮まる傾向が見て取れる。これは、第3.2.3.項で述べた審査書類別の集計結果の傾向と同様である。

以上、文種別の集計結果における傾向は、サンプル数が少量なこともありばらつきや例外も多いが、大枠としては、学習データ中の（推定）構成比率の違いにかかわらず、各文種ともおおむね同等にスコアが向上し、かつ、無学習時に比べて文種間の品質差は均質化されていると総括できる。そして、この状況から、審査書類対訳コーパス最終学習後の時点で品質が顕著に劣り、重点的な対処が必要な文種は存在しないと結論される。この結論に関しては、後述の人手評価結果を用いてさらに検証する。

3.2.6. 文単位のスコア向上／低下カウント結果

ここまで各項に示した自動評価スコアの動向は、評価用コーパス全体又は種別・分野ごとの全体平均スコアの動向であって、評価用コーパス中の個々の文が一律に平均スコアと同じ動向を示しているとは限らない。むしろ全体平均スコアが向上していても、個々の文レベルではスコアが低下したものや変化しなかったものも少なからず存在すると考えられる。この点に鑑み、本項では、審査書類別の評価用コーパス各 2,000 文について、各回の学習前後で BLEU 及び RIBES の評価スコアが向上した文、変化しなかった文、低下した文をそれぞれカウントした。以下、集計結果を示す。

表 3.2.8. 審査書類別 文単位 BLEU スコア向上／不変／低下カウント結果

	BLEU											
	拒絶査定			拒絶理由通知書			補正却下の決定			前置報告書		
	向上	不変	低下	向上	不変	低下	向上	不変	低下	向上	不変	低下
第 1 回	1,629	6	365	1,638	7	355	1,739	9	252	1,727	5	268
第 2 回	967	70	963	993	58	949	1,002	71	927	1,030	86	884
第 3 回	979	71	950	995	70	935	1,001	73	926	976	78	946
第 4 回	1,018	85	897	993	77	930	942	81	977	1,008	77	915
第 5 回	966	115	919	884	91	1,025	958	94	948	952	103	945
第 6 回	954	94	952	962	86	952	982	115	903	987	127	886
第 7 回	878	122	1,000	936	101	963	958	122	920	876	138	986
第 8 回	929	93	978	904	87	1,009	878	82	1,040	888	111	1,001
第 9 回	951	89	960	976	75	949	1,044	74	882	997	93	910
第 10 回	1,002	96	902	965	119	916	941	103	956	957	102	941
第 11 回	1,319	72	609	1,228	63	709	1,241	83	676	1,334	87	579
第 12 回	650	599	751	665	618	717	706	581	713	624	567	809
第 13 回	894	328	778	939	326	735	823	382	795	915	376	709
第 14 回	679	601	720	692	626	682	704	653	643	637	590	773
第 15 回	766	332	902	757	340	903	816	350	834	759	331	910
第 16 回	750	527	723	750	562	688	750	593	657	700	565	735
第 17 回	723	395	882	788	393	819	773	445	782	750	424	826
第 18 回	706	600	694	687	637	676	665	614	721	693	635	672

表 3.2.9. 審査書類別 文単位 RIBES スコア向上／不変／低下カウント結果

	RIBES											
	拒絶査定			拒絶理由通知書			補正却下の決定			前置報告書		
	向上	不変	低下	向上	不変	低下	向上	不変	低下	向上	不変	低下
第 1 回	1,381	11	608	1,403	9	588	1,500	9	491	1,486	9	505
第 2 回	1,011	77	912	987	75	938	996	75	929	1,028	98	874
第 3 回	967	77	956	1,002	81	917	946	72	982	985	89	926
第 4 回	1,010	98	892	974	95	931	995	94	911	1,000	90	910
第 5 回	936	126	938	896	114	990	960	101	939	921	108	971
第 6 回	952	105	943	904	105	991	920	113	967	965	135	900
第 7 回	880	127	993	985	111	904	936	132	932	888	150	962
第 8 回	966	108	926	936	97	967	927	92	981	914	121	965
第 9 回	907	99	994	933	87	980	940	82	978	910	107	983
第 10 回	975	108	917	930	133	937	951	111	938	985	115	900
第 11 回	1,256	86	658	1,215	78	707	1,249	83	668	1,243	89	668
第 12 回	647	619	734	661	631	708	700	592	708	649	568	783
第 13 回	891	336	773	910	332	758	823	391	786	879	393	728
第 14 回	672	615	713	658	651	691	695	662	643	669	594	737
第 15 回	729	344	927	804	346	850	779	350	871	758	334	908
第 16 回	730	546	724	729	580	691	711	605	684	727	574	699
第 17 回	755	409	836	775	395	830	732	449	819	776	434	790
第 18 回	741	601	658	699	642	659	689	631	680	698	638	664

各回における文単位の自動評価スコア増減のカウント結果は、上二表に示したとおり、全書類、両指標ともほぼ同じ推移を辿っている。第 1 回学習後のカウントでは全体の 70～80% の文でスコアが「向上」し、一方、スコアが「低下」した文は 20～30%にとどまるなど、各書類ともスコア向上に大きく偏った。なお、スコアが「不変」の文はほとんど存在しなかった。

続く第 2 回学習後から第 10 回学習後にかけても「不変」文は依然として全体の 5 %前後とごく少数であり、全体の約 95%の文のスコアが毎回変化し続けた。変動の内訳は各回とも「向上」文と「低下」文がそれぞれ 900～1,000 文（45%～50%）で拮抗している。

この傾向が変化したのが、「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）直後の第 11 回学習後である。第 11 回学習後の集計結果は「向上」が 1,200～1,300 文、「低下」が

600～700 文となり、初回学習後に準ずる「向上」への顕著な偏りが見られた。その後、第 12 回学習後から最終の第 18 回学習後にかけては再び「向上」と「低下」とが均衡する形に戻ったが、それぞれの文数は各回 650～900 文（33%～45%）程度にとどまることが多くなり、その分「不変」文が 300～650 文（15%～33%）と以前に比べて明らかに増加した²⁴。

学習前後の自動評価スコアの変動は、機械翻訳文の変化とほぼ同義である。まれに機械翻訳文が変化してもスコアが同値となる場合はあるが（このため各回において BLEU と RIBES の「不変」文数は異なっている。）、少なくともスコアが変動した機械翻訳文は、学習前後で何らかの変化が生じていると確定できる。

当初の予測では、学習データを段階的に増量するにつれて機械翻訳文で採用される訳語が学習データ中で大勢を占める有力な訳語に徐々に収れんしていき、学習前後で訳語が変化する語が少なくなって「不変」文が増加していくものと考えられた。しかしながら、上記カウント結果はこの予測とは異なっており、「みんなの自動翻訳」には、学習データが増量されて訳語候補間の頻度順位が安定しても、それが必ずしも機械翻訳文の安定につながらない傾向があるものと推察される。

大規模アップデートを経た第 12 回学習後以降は「不変」文の比率が明らかに増加しており、学習前後で大半の機械翻訳文が不安定に変化するという傾向は大規模アップデート以前に比べてやや抑制されたと見られるが²⁵、それでも少なくとも全体の 70% 弱の文が最終学習後まで変化を続けており、上記傾向自体は根本的には変わっていない。

本項に示したカウント結果からは、各回とも何らかの変化が生じた機械翻訳文の半数程度で自動評価スコアが低下しており、これを額面通りに受け取れば、各回 33～45% 程度の文で翻訳品質の劣化が生じていることになる。本来、審査書類対訳コーパスという均質な学習データを段階的に増量していく過程で毎回これほど多数の文に悪影響が及ぶとは考えにくい。が、「みんなの自動翻訳」は上述のとおり学習データの増量が翻訳文の安定に直結しない傾向が見受けられ、この傾向によって多くの文で品質劣化が生じている懸念も皆無では

²⁴ 第 12 回以降、学習前後で同じバージョンのエンジンを用いた回（第 12、14、16、18 回）では学習後の「不変」文数が 600 文前後、中間に定期バージョンアップ（⇒2.2.1.）があり学習前後でエンジンのバージョンが異なる回（第 13、15、17 回）では学習後の「不変」文が 350～450 文前後となっている。後者はエンジンのバージョンアップの影響で機械翻訳文の変化が多く生じているものと推察される。

²⁵ 大規模アップデートは第 11 回学習前に実施されたが、第 11 回学習後は大規模アップデート前後の機械翻訳文の比較となるため、安定化の傾向は大規模アップデート後の機械翻訳文同士を比較した第 12 回学習後以降に表れている。

ない。これについては次項「実例を用いた機械翻訳文の定点観測」において、各回学習前後の機械翻訳文の具体的な変化の内容と、それに伴う自動評価スコアの変動とを比較して、改めて考察する。

3.2.7. 実例を用いた機械翻訳文の定点観測

本項では、審査書類①に属する 4 書類の評価用コーパスから幾つか実例を取り上げ、各回学習後の機械翻訳文の変化の状況を定点観測する。学習の進度に伴う機械翻訳文の変遷を辿ることにより、審査書類対訳コーパスの学習各段階における翻訳品質改善の具体的な内容について把握するとともに、ここまで自動評価の全体平均スコアから看取してきた各種傾向の妥当性について検証する。加えて、各実例における実際の機械翻訳品質とこれに付与された自動評価スコアとを照合し、その妥当性や評価の精度についても検証する。

3.2.7.1. 実例 3-1：「拒絶査定」由来、文番号 A02-#1047 の分析

「拒絶査定（A02）」の評価用コーパスから文番号#1047 を実例に採る。まずは無学習時と第 1 回学習後の機械翻訳結果を以下に示す。

【文番号：A02-#1047】

原文	したがって、請求項 6 には、発明の詳細な説明に記載も示唆もない事項が記載されているから、発明の詳細な説明に記載したものでない。	
正解訳文	Therefore, as the matter which is not described or suggested in the detailed description of the invention is described in Claim 6, it is not described in the detailed description of the invention.	
無学習	Accordingly, the sixth aspect of the present invention is not limited to the detailed description of the present invention, and is not limited thereto.	
	—	BLEU : 0.1503 (－) RIBES : 0.7477 (－)
第 1 回学習後	Therefore, since the matters that are not described or suggested in the detailed description of the invention are described in Claim 6, it is not described in the detailed description of the invention.	
	向上	BLEU : 0.5584 (+0.4081) RIBES : 0.9554 (+0.2077)

上欄一段目が機械翻訳の対象となる「原文」であり、二段目が自動評価の基準となる「正解訳文」、すなわち本事業で作成した人手翻訳文である。この一～二段目の対訳文が審査書類対訳コーパスの 1 レコードに相当する。そして、三段目と四段目が、それぞれ無学習時（デフォルト状態）と審査書類対訳コーパス初回学習後の審査書類①システムで上記原文を機械翻訳した結果である。各機械翻訳文の末行には目視による品質評価（前回より向上／同等／悪化）と、実際に付された BLEU 及び RIBES のスコアを記載した。

三段目に示した無学習時の機械翻訳文では、特許・審査書類用語の誤訳が多発している。まず冒頭部で「請求項 6」という基本的な特許用語が「**sixth aspect** (第 6 態様)」と誤訳されている。

ただし、無学習時の審査書類①で「請求項」が常に **aspect** と誤訳されているわけではなく、他の多くの文では「請求項」は **claim** と適切に訳されていた。とはいえ、本例と同様に **aspect** と誤訳された文も一定数存在しており²⁶、無学習時の「みんなの自動翻訳」では「請求項」という特許用語が **claim** と **aspect** とで訳ゆれしている状況であった。

本例の無学習時の機械翻訳文では、同じく基本的な特許用語である「**発明の詳細な説明**」も **detailed description of the present invention** と、原文にない **present** が付随する不備が発生している（ただし本例では意味は通る）。また、この語は原文中に二度出現するが機械翻訳文では一度しか出現せず、もう一方が欠落している。

「**記載も示唆もない事項**」も、一般語の組み合わせではあるが審査書類で頻出のフレーズであり、特許・審査書類用語に準ずる。このため、審査書類においては定訳語で訳すことが望ましい。無学習時の機械翻訳文では、「**記載** (がない)」の部分のみ **not limited** と訳されているが、「**示唆**」や「**事項**」に相当する訳語は見当たらず、いずれも欠落している。また、**limited** も「**記載**」の訳語として適訳とはいえない。

本例の無学習時の機械翻訳文は、これら特許・審査書類用語に関連する多数の不備によって意味不明な内容となっており、原文の内容を把握することは不可能である。

これに対し、第 1 回学習後、すなわち審査書類コーパス A を約 10 万文対学習した後の機械翻訳文では、上記特許・審査書類用語は三語とも正解訳文と同一、又は実質的に同一²⁷の適訳語で訳されるようになった。文の全体構造も無学習時における「**発明の詳細な説明**」の欠落が解消されてほぼ正解訳文どおりの文の構造となり、原文の内容を正確に伝達する翻訳文に改善されている。

²⁶ 例えば拒絶査定の評価用コーパス 2,000 文中「請求項」を含むものは 702 文存在するが、その機械翻訳文で **claim** を含むものは 615 件、**aspect** を含むものは 88 件であった。また、双方を含むものは 22 件、どちらも含まないものは 21 文であった。どちらも含まない文の場合、「請求項」の訳語自体が欠落していることが多かった。

²⁷ 「**記載も示唆もない事項**」の「**事項**」が、正解訳文では単数形 (**matter which is**) なのに対して機械翻訳文では複数形 (**matters that are**) となっているが、どちらが適切かは文脈・状況次第であり、単独の文としては優劣はつかない。

つまり、本例においては、デフォルト状態の「みんなの自動翻訳」では正しく翻訳されなかった特許・審査書類用語の全てが、審査書類対訳コーパスを学習させたことで正しく翻訳されるようになった。その結果、内容の欠落も解消され、翻訳精度が顕著に改善した。前項までにまとめた自動評価スコアの集計結果から、初回学習後の自動評価の全件平均スコアが無学習時に比べて大幅に向上したこと（⇒3.2.2.）、そして、このスコア向上が評価用コーパスの大多数の文に及んだこと（⇒3.2.6.）が判明していたが、本例からは、その最大の理由が、無学習時には対応できず誤訳されたり欠落したりしていた多数の特許・審査書類用語が、審査書類対訳コーパスの学習によって正解訳文どおりの適訳語に翻訳されるようになったためであることがうかがえる。

なお、各機械翻訳文の最下行には翻訳品質の学習前後の相対評価を付した。第 1 回学習後の機械翻訳文の実際の品質は、上記のとおり無学習時に比べて明らかに「向上」している。これに対し、本例に付された BLEU、RIBES スコアも学習後に大きく向上しており、学習前後の品質の相対評価としては両指標とも実態を正しく反映できている。

続いて、第 2 回学習後の機械翻訳文を示す。

【文番号：A02-#1047】

第 2 回学習後	Therefore, since Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, they are not the matters described in the detailed description of the invention.	
	悪化	BLEU : 0.4557 (-0.1027) RIBES : 0.8107 (-0.1147)

原文に含まれる三語の特許・審査書類用語は前回に引き続き全て適訳語で訳された。「記載も示唆もない事項」は今回 matters that are neither described nor suggested と訳されており、前回時 (matters that are not described or suggested) ととも正解訳文 (matter which is not described or suggested) ととも異なるが、フレーズの要所である「記載」「示唆」「事項」という語句はいずれも suggest、describe、matter(s) で統一されているため、どの訳し方でも読み手は正しい理解に到達できる。無学習時の分析で、このフレーズは「審査書類においては定訳語で訳すことが望ましい」と述べたが、上記三種の訳し方のように同じ意味であることを読み手がスムーズに認識できる形であれば完全一致でなくても問題はない。

このように、第 2 回学習後の機械翻訳文でも特許・審査書類用語は安定して適訳語に訳されたが、その一方で、翻訳文の全体構造は前回から大きく変化した。具体的には、前回時は前半の since 節の主語が「事項：matters」であったのに対し、今回は「請求項 6：Claim 6」が主語となっており、また後半の主語は前回が it (=Claim 6 を指す) であったのに対

し、今回は they (=matters を指す) となっており、文の前半と後半それぞれで主語が変化している。

このような文構造²⁸の変化は、読み手に伝わる内容が同じである限り、原則として優劣はつかない。だが、本例の場合、前半部はどちらの文構造でも同義だが、後半部は、前回の翻訳文が「(請求項 6 は明細書で説明も示唆もされていない事項を記載しているから、) その請求項 (=it) は明細書に記載されたものではない」と解釈でき、請求項の内容が明細書で説明されていないという原文の文意どおりに理解できたのに対し、今回の翻訳文は「(請求項 6 は明細書で説明も示唆もされていない事項を記載しているから、) これらの事項(=they) は明細書に記載された事項ではない」という、同じ内容を繰り返しただけの、原文の意図が伝わらない翻訳文となっている。つまり翻訳品質は前回に比べて「悪化」している。これに対し、第 2 回学習後の翻訳文に付された BLEU、RIBES スコアも前回より低下しており、引き続き実際の品質変化を正しく反映できている。

続いて、第 3 回～第 5 回学習後の各回の機械翻訳結果を示す。

【文番号：A02-#1047】

第 3 回学習後	Therefore, since Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	
	向上	BLEU : 0.5603 (+0.1046) RIBES : 0.8752 (+0.0645)
第 4 回学習後	Therefore, Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, and therefore, it is not described in the detailed description of the invention.	
	悪化	BLEU : 0.5607 (+0.0004) RIBES : 0.9219 (-0.0342)
第 5 回学習後	Therefore, since Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	
	向上	BLEU : 0.5603 (-0.0004) RIBES : 0.8752 (+0.0342)

第 3 回学習後の翻訳文は、前半部は前回時と同じだが、前回 they と誤訳されていた後半部は it is not described in the detailed description of the invention と前々回の正しい解釈に

²⁸ 本例にみられるように、同じ訳語を用いて同じ内容を表した文であっても、各訳語の配置は一意には定まらない。本報告書では、翻訳文におけるこうした各語の配置を「文構造」と称し、上記のように訳語が変化せず各訳語の配置が同義の範囲内で変化している状況を「文構造の変化」と称した。

戻り、全体的には第 1 回学習後と第 2 回学習後を合成したような翻訳文となった。審査書類用語も引き続き安定して適訳語で訳されており、第 1 回学習後と同じく不備のない翻訳文となった。したがって、品質は前回より「向上」している。

第 4 回学習後の翻訳文は、過去 3 回とも採用されていた前半の since 節が採用されず、前後半を and therefore で接続する形に文構造が変化した。主要な用語の訳語は前回と変わっておらず内容理解上は問題ないが、一文中で therefore を繰り返すのは英文として不自然であり、その分僅かに「悪化」している。

続く第 5 回学習後の翻訳文では前回採用された and therefore が再び since 節に戻され、結果、前々回と完全一致の翻訳文となった。品質評価は前回 and therefore を理由に僅かな「悪化」としたため、今回は僅かな「向上」となる。

ここまでの各回の翻訳文の推移を見ると、特許・審査書類用語が最適訳で安定する一方で、文構造の不安定な変化が目立つ。第 3.2.6.項に示したとおり、審査書類①では、全期間を通じて評価用コーパスの大半の文の機械翻訳結果が学習の都度変化し続けたが、第 2 回学習後以降の期間における各文の変化は、本例に見られるような文構造の頻繁な変化がその一因であることがうかがえる。

なお、この期間における本例の自動評価スコアは、BLEU の第 4 回学習後から第 5 回学習後にかけてのスコア低下が実際の品質向上とは逆方向になっている。しかしながら、誤差レベルの僅かな変動であり、実際の品質変化自体も英語表現が自然なものになった程度の微かな改善であることに鑑みると、実際の状況と矛盾するというほどではない。

続いて、第 6 回学習後から第 10 回学習後までの翻訳結果を示す。

【文番号：A02-#1047】

第 6 回学習後	Therefore, since the matters which are neither described nor suggested in the detailed description of the invention are described in Claim 6, they are not described in the detailed description of the invention.	
	悪化	BLEU : 0.5609 (+0.0006) RIBES : 0.9219 (+0.0467)
第 7 回学習後	Therefore, since Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, they are not described in the detailed description of the invention.	
	同等	BLEU : 0.4761 (-0.0848) RIBES : 0.8334 (-0.0885)
第 8 回学習後	Therefore, Claim 6 is not described in the detailed description of the	

	invention because it describes matters that are not described or suggested in the detailed description of the invention.	
	向上	BLEU : 0.5888 (+0.1127) RIBES : 0.6229 (-0.2105)
第 9 回学習後	Therefore, since Claim 6 describes a matter that is not described or suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	
	同等	BLEU : 0.6789 (+0.0901) RIBES : 0.9059 (+0.2830)
第 10 回学習後	Therefore, since Claim 6 describes matters that are not described or suggested in the detailed description of the invention, they are not described in the detailed description of the invention.	
	悪化	BLEU : 0.5603 (-0.1186) RIBES : 0.8567 (-0.0492)

第 6 回学習後の機械翻訳文では、前半部の主語が Claim 6 から matters に変化して 5 回ぶりに初回学習後の文構造に戻ったが、同義の範囲内の変化であり、この点に関して前回時と品質上の優劣はない。ただし、後半部の主語に第 2 回学習後以来の they (=matters を指すと解釈される) が採用され、当時説明した誤訳が再発している。このため、前回より「悪化」と判定した。続く第 7 回学習後は前半の主語が再び Claim 6 に戻ったが、後半の主語は they のままで誤訳は改善せず、品質は前回と「同等」である。

続く第 8 回学習後の翻訳文では、文の骨格構造が大きく変化した。これまでの翻訳文の前半部と後半部の順序が入れ替わり、「請求項 6 は明細書に記載されたものではない」という結論が先に言われ、その理由である「それ (=請求項 6) は明細書に記載も示唆もない事項を記載している」が because で接続されている。過去回でも見られた「みんなの自動翻訳」特有の文構造の不安定な変化の典型といえるが、この新たな文構造でも文意の理解には何ら問題はなく、むしろ過去 2 回で生じていた they の誤訳が解消されたことで品質的には「向上」している。ただし、この文構造が採用されたのは全期間を通じてこの 1 回のみであり、続く第 9 回学習後では再び従来の文構造に復帰している。実質的な翻訳品質に影響しない表層的な変化であり、内容理解上は完全に「同等」である。なお、続く第 10 回学習後も文構造自体は同じであったが、後半で they の誤訳が再発したため「悪化」と判定した。

このように、第 6 回～第 10 回学習後の期間においても、特許・審査書類用語の適訳語での安定、そして文構造の不安定な変化という傾向は引き続き顕著であった。文構造に関しては、一回限りの突発的な骨格構造の変化 (第 8 回) も見られた。こうした学習回ごとの文構造の不安定な変化は、各時点での学習データ全体での出現頻度と直結しているようには見えず、外部の目にはランダム的な変化とも映る。ただし、ここまでに採用された文構造はど

れもおおむね同義の範囲内にあり²⁹、つまり、本例では、文構造はあたかも同義の複数の候補から毎回ランダムに採用されているかのような変化を見せている。

自動評価スコアに関しては、この期間は実際の品質との矛盾が目立つようになってきた。この期間の各回の翻訳文の実質的な優劣は後半部における they の誤訳の有無のみであったが、学習前後でこの箇所に変化がなく、実質的な翻訳品質が前回と「同等」であった回においても、文構造が不安定に変化することによって、正解訳文との表層的な一致度に基づく自動評価指標ではスコアが必要以上に変動してしまっている。特に第 8 回学習後には BLEU は大幅に向上、RIBES は大幅に低下と、両指標間で大きな齟齬も生じている。

本例のここまでの推移からは、コーパス学習後の各回の機械翻訳文の傾向として、特許・審査書類用語は常時適訳語で安定していること、その反面、文構造は不安定に変化を続けていること、ただし採用される文構造はおおむね同義の範囲内で品質上の優劣はないこと、そして自動評価スコアは翻訳文の表層的な変化によって必要以上に変動する場合があること、を指摘できる。第 3.2.6.項では、学習を重ねても毎回全体の半数近い文で自動評価スコアが低下しており、大規模な品質劣化が生じていることが懸念されたが、本例を見る限りは、各回の機械翻訳文の変化には実際の翻訳品質には影響しない表層的なものが多く、自動評価スコアが低下していても必ずしも翻訳品質が劣化しているとは限らないことがわかった。

続いて、第 11 回学習後から、最終学習回である第 18 回学習後までの翻訳文を示す。

第 11 回学習後	Therefore, since Claim 6 describes a matter that is neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	
	向上	BLEU : 0.5712 (+0.0109) RIBES : 0.8782 (+0.0215)
第 12 回学習後	Therefore, since Claim 6 describes a matter that is neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	
	同等	BLEU : 0.5712 (±0) RIBES : 0.8782 (±0)
第 13 回学習後	Therefore, since Claim 6 describes a matter that is neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.	

²⁹ 第 2 回学習時以降の機械翻訳文の優劣は、後半に補足される主語が it であるか they であるかにほぼ集約される。これは文構造の変化ではなく、個々の語における単数複数の変化（ゆれ）であると判断した。

	同等	BLEU : 0.5712 (±0) RIBES : 0.8782 (±0)
第 14 回学習後		Therefore, since Claim 6 describes matters that are neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.
	同等	BLEU : 0.5603 (-0.0109) RIBES : 0.8752 (-0.0030)
第 15 回学習後		Therefore, since Claim 6 recites matters that are neither described nor suggested in the detailed description of the invention, they are not described in the detailed description of the invention.
	悪化	BLEU : 0.4761 (-0.0842) RIBES : 0.8334 (-0.0418)
第 16 回学習後		Therefore, since Claim 6 recites matters that are neither described nor suggested in the detailed description of the invention, they are not described in the detailed description of the invention.
	同等	BLEU : 0.4761 (±0) RIBES : 0.8334 (±0)
第 17 回学習後		Therefore, since Claim 6 recites matters that are neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.
	向上	BLEU : 0.5603 (+0.0842) RIBES : 0.8752 (+0.0418)
第 18 回学習後		Therefore, since Claim 6 recites matters that are neither described nor suggested in the detailed description of the invention, it is not described in the detailed description of the invention.
	同等	BLEU : 0.5603 (±0) RIBES : 0.8752 (±0)

第 11 回学習後の翻訳文では、これまで常に matters と複数形であった「事項」が初めて単数形 (matter) で訳されたが、この点を除けばこれまでと同じ文といえる。日本語文には単数形・複数形の概念がなく、このため前後の文脈を見ないと、単独の文の記載内容のみからはどちらが適切か判断できないことも多い。本例もこれに該当し、正解訳文からこの文が採取された案件においては単数形が適切とわかるが、状況次第では複数形が適切となる場合も十分にありえる。このため、品質判定ではどちらも正解とした。なお、第 11 回後の翻訳文では後半部の主語が前回時の they から it に変化して誤訳が解消されているため、判定結果は前回よりも「向上」としている。この it と they の不安定な変化も単数形・複数形間の変化の一種といえる。

続く第 12 回学習後と第 13 回学習後は、第 11 回学習後と完全一致の翻訳文が 3 回連続で採用された。同じ翻訳文が連続して採用されたのは本例では初であり、文構造が毎回不安定に変化していたこれまでの傾向から一変している。この傾向の変化は、タイミングから考えて、第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデート (⇒2.2.2.) の

影響が大きいと推察される。

第 3.2.6.項に示した文単位のスコア変動の集計結果でも、第 12 回学習後以降、つまり本例と同じタイミングで前回から「不変」となる機械翻訳文が明らかに増加しており、この傾向の変化が評価用コーパス全体に及んでいることを示している。学習データである審査書類対訳コーパス自体は各回同量かつ均質であり、第 11 回時の学習データ自体に、それまでの傾向を突如覆すような要素が存在するとは考えにくい。よって、この傾向の変化は第 11 回時の学習データの内容が直接もたらしたものではなく、大規模アップデートを機にエンジンのロジック自体が見直され、文構造のランダム的な変化を抑制する方向に調整されたと考えるのが妥当である。

その後、第 14 回、第 15 回学習後も文の骨格構造は変わらなかったが、第 14 回学習後には「事項」が再び matters と複数形となり、第 15 回学習後にはさらに「記載」が describe から recite に変化し、また後半の主語も they に戻るなど、単数・複数の変化を中心とした局所的な訳語変化が見られた。続く第 16 回学習後は第 15 回後から変化せず、第 17 回学習後は they の誤訳が解消されて第 14 回後と同じ翻訳文に復帰し、最終第 18 回学習後もこれと完全一致の翻訳文が出力された。

このように、大規模アップデート以降の各回の翻訳文では、それまで頻発していた文の骨格構造の不安定な変化が一切なくなり、同じ骨格構造上で「事項」の訳語 matter(s)の単複の変化、「記載」の describe から recite への変化、そして後半部の主語に it と they のどちらが用いられるか、という三箇所の変化のみが散発的に発生するにとどまっている。つまり、翻訳文の学習前後の変化が明らかに小規模化し、局所的になっている。第 3.2.6.項の集計結果では、大規模アップデート以降も毎回少なくとも全体の 70%近くの機械翻訳文が学習前後で変化していたが、本例の状況からは、各文における変化は本例と同様にアップデート以前よりも小規模化・局所化している可能性が高い。

自動評価の全体平均スコアでは、第 11 回学習前の大規模アップデートを境にスコア水準が顕著に上昇し、以降もその水準が常時維持されていた。これに対し、本例ではこのタイミングでの特段の品質改善は見られない。ただし、本例の場合、アップデート以前の第 10 回学習後の時点で既に完璧な翻訳文が出力されており、さらなる改善の余地はなかった。それでも、アップデート以降は上述のとおり文構造の不安定な変化が解消し、各所における訳語変化も小規模かつ局所的に抑制されるなど、翻訳傾向は如実に変化していた。こうした傾向の変化、すなわち翻訳文の安定化が、他の多くの文では何らかの品質改善をもたらしているものと推察される。この点に関しては後続の実例で改めて検証する。

変化箇所が局所化されたこともあり、第 11 回学習後以降の各回の翻訳文の実質的な優劣は引き続き they の誤訳の有無のみに左右されることとなった。具体的には、この誤訳が発生した第 15 回学習後は「悪化」、解消した第 17 回学習後は「向上」となり、その他の回は全て前回と「同等」となった。自動評価スコア BLEU、RIBES も、第 14 回学習後を除き、おおむねこの評価に沿った変動を示している。

最後に、大規模アップデート以降も引き続き発生して翻訳文の品質を左右した it と they の変化について考察する。この変化は、先述のとおり日本語文における単数・複数の特定の困難性に起因するものと考えられるが、それに加え、同じく日本語文の特徴である主語の省略への対応の困難性にも等しく起因していると考えられる。

日本語文には主語無し文が多用される。このため、英訳時に主語の補足が必要となるのは本例に限らない。しかしながら、本例の場合は特に、補足する主語の特定が難しい状況が存在する。

本例の原文「請求項 6 には、発明の詳細な説明に記載も示唆もない事項が記載されているから、発明の詳細な説明に記載したものではない。」は、後半の主語も「請求項 6」と解釈されることを意図している。しかし、厳密にはこの原文は文法的にねじれており、純文法的には、冒頭の「請求項 6 には」を後半部の主語とは解釈できない（「請求項 6 には、発明の詳細に記載したものではない。」となってしまう。）。文法的にはむしろ「事項」を後半部の主語と解釈するほうが無理がない。ニューラル機械翻訳は翻訳文の生成に文法規則を用いない方式ではあるが、原文がこのようなイレギュラーな書かれ方であると、学習データ中に参考とすべき文例が少なくなる。このため、本例では、後半部の主語が「請求項 6 : Claim 6」であるか「事項 : matter(s)」であるか決定的な判断がつかず、不安定な変化を続けたとも考えられる。実際、各回で採用される it と they は常時「事項 : matter(s)」の単複に対応しているわけではなく、第 10 回以前や第 14 回後、第 17～18 回後は「事項」が複数形 (matters) であっても後半部の主語には it が採用されている。

以上、「拒絶査定」の評価用コーパスから文番号 A02-#1047 を事例に採り、審査書類①システムによる各学習段階の機械翻訳文の変化を定点観測した。一例のみの分析であるが、本例を通じ、自動評価スコアの集計結果から看取された品質改善の実態について、その具体的な状況を把握することができた。以下に列挙する。

- ・ 無学習時の機械翻訳文では特許・審査書類用語の多くが誤訳されるか、訳出されず翻訳文から欠落していた。つまり、デフォルト状態の「みんなの自動翻訳」は特許・審査書類用語に十分に対応できていなかった。

これに対し、審査書類対訳コーパスの学習後は、無学習時に正しく翻訳できなかった全特許・審査書類用語が初回から正解訳文どおりの適訳語に改善され、その後も全期間安定して適訳語が採用され続けた。

自動評価の全体平均スコアが初回学習後に大幅に向上し、以降もその水準が保たれていたことから、他の文においても本例と同じく、文中の特許・審査書類用語の大多数が初回学習によって正解訳文どおりの適訳語に一挙に改善し、そのまま安定したものと推察される。

- ・ 一方、翻訳文の文構造に関しては、特許・審査書類用語と異なり、学習の都度不安定に変化する傾向が顕著であった。ただし、本例を見る限り、こうした文構造の変化はおおむね同義の範囲内にとどまり、実質的な翻訳品質には影響していない。

- ・ 本例では後半部に補足された主語が適訳語の *it* と不適訳語の *they* とで不安定に変化する傾向も見られた。ただし、本例は原文自体の文法的な不備に起因している可能性もあり、「みんなの自動翻訳」の主語補足の精度に問題があるか否かは本例のみからは判断できない。

- ・ 第 3.2.6.項に示した文単位のスコア向上／低下カウント結果では、学習前期（第 10 回学習後まで）において、特許・審査書類用語が適訳語で安定した第 2 回学習後以降も全体の約 95%の機械翻訳文が毎回変化し続けていた。上述した「みんなの自動翻訳」における文構造の不安定な変化が主な原因と考えられる。

ただし、このように多数の文が変化を繰り返したにもかかわらず、自動評価の全体平均スコアの変動は毎回誤差レベルの微かな変動にとどまっていた。本例も第 10 回学習後まで全回で翻訳文に変化が見られたが、その多くは文構造のランダム的な変化によるものであり、翻訳品質を左右しない同義の範囲内のものであった。よって、本例も、原文の不備によるところが大きい *it* と *they* の間の変化を度外視すれば、自動評価の全体平均スコアが示す全体傾向、すなわち第 2 回～第 10 回学習後の期間の翻訳品質はほぼ現状維持という傾向におおむね合致している。

- ・ 第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデート以降、本例ではそれまでの傾向が一変し、文構造の不安定な変化が一切見られなくなった。アップデート以降も翻訳文が変化することはあったが、その変化は文中の特定の箇所のみ小規模化・局所化された。

こうした傾向の変化は、評価用コーパス全体において、大規模アップデートを境に「不変」文が顕著に増加したこと（⇒3.2.6.）と整合する。大規模アップデート以降も過半数の文は学習の都度変化し続けてはいるが、本例の状況からは、各文における変化はアップデ

ート以前よりも小規模・局所的になっている可能性が高い。

- ・ 第 3.2.2.項に示した自動評価の全体平均スコアでは、大規模アップデート直後の第 11 回学習後にスコアが顕著に向上しており、審査書類①全体に及ぶ何らかの品質改善が果たされたことが示されていたが、本例に関しては、大規模アップデート時の顕著な品質改善が確認できなかった。本例の場合、大規模アップデート以前から高品質に達しており、改善の余地がなかったためと考えられる。

ただし、上述のとおり本例においても大規模アップデートを境に翻訳傾向は如実に変化しており、こうした傾向の変化、具体的には翻訳文における文構造の安定化、及び訳語変化の小規模化・局所化が、他の多くの文において何らかの品質改善をもたらしたものと推察される。

- ・ 本例の各回の機械翻訳文に付された自動評価スコアは、おおむね実際の翻訳品質の推移と整合していたが、回によっては実態とかい離していたり、BLEU と RIBES とで矛盾した結果となったりすることもあった。

こうした知見に関しては、次例以降の実例分析で引き続き検証し、補強・修正していく。

3.2.7.2. 実例 3-2 : 「前置報告書」由来、文番号 A913-#314 の分析

実例 3-2 には「前置報告書 (A913)」由来の文を取り上げる。まずは無学習時と第 1 回学習後の機械翻訳結果を示す。

【文番号 : A913-#314】

原文	バリア性フィルム層を積層してから、バリア性フィルム層に脆弱部を形成するか、あるいはその逆とするかは、 当業者が適宜選択し得た事項 である (例えば、 引用文献 3 の段落 0 0 1 0 参照)。	
正解訳文	Forming the weakened portion in the barrier film layer after laminating the barrier film layer, or laminating the barrier film layer after forming the weakened portion in the barrier film layer can be appropriately selected by a person skilled in the art (for example, see paragraph 0010 of Cited Document 3).	
無学習	The barrier film layer may be laminated to form a fragile portion in the barrier film layer or vice versa (see, for example, Paragraph 0010 of Reference document 3).	
	—	BLEU : 0.1905 (—) RIBES : 0.7275 (—)
第 1 回学習後	It is a matter that can be appropriately selected by a person skilled in the art whether the barrier film layer is laminated and then the fragile part is formed on the barrier film layer, or vice versa (for example, refer to paragraph 0010 of Cited Document 3).	
	向上	BLEU : 0.4599 (+0.2694) RIBES : 0.7162 (-0.0114)

本例には特許・審査書類用語として「**当業者**」「**適宜選択し得た事項**」「**引用文献 3**」の 3 語が含まれる。特に「**当業者**」と「**適宜選択し得た事項**」は審査書類特有の用語・表現といえ、その専門性、翻訳難易度は比較的高い。

無学習時の機械翻訳文では、これら専門性の高い「**当業者**」と「**適宜選択し得た事項**」の 2 語は訳出されず、翻訳文から完全に消失してしまっている。「**引用文献 3**」のみは **Reference document 3** と妥当に訳されたが、本事業で定めた統一訳語 (正解訳文で用いている **Cited Document 3**) ではない。

これに対し、審査書類対訳コーパス第 1 回学習後の機械翻訳文では「**当業者**」と「**引用文献 3**」はそれぞれ **a person skilled in the art**、**Cited Document 3** と正解訳文どおりの訳語が採用され、「**適宜選択した事項**」も **a matter that can be appropriately selected** と正解訳文に

準ずる形³⁰で訳出されるようになった。つまり、コーパスの初回学習により、専門性の高いものも含め文中の全ての特許・審査書類用語が適訳語に改善されている。

本例における、無学習時に特許・審査書類用語が適訳語に訳されず、専門性の高い用語・表現はしばしば翻訳文から欠落してしまう傾向、そしてこうした特許・審査書類用語の不備がコーパスの初回学習によって解消され、正解訳文どおりの適訳語に改善する傾向は、いずれも前例（実例 3-1）と共通しており³¹、審査書類①全体にわたる傾向である可能性がさらに高くなった。どちらの実例も、無学習時はこうした審査書類用語関連の不備によって正解訳文との一致率が低く、このため自動評価スコアも低い数値が付されていたが、初回学習によって正解訳文どおりの適訳語が採用されるようになったことで、スコアは大幅に向上している。この点も審査書類①の全体平均スコアの動向と合致しており（⇒3.2.2.）、他の多くの文でも同様の現象が発生していたことを暗示している。

なお、本例には技術用語もいくつか含まれている。具体的には「バリア性フィルム層」「積層する」「脆弱部」の3語である。また、「形成する」「その逆とする」「参照」といった一般用語も存在する。これら技術／一般用語は、上掲のとおり学習前後とも全て妥当な訳語で訳され、かつ多くの語は学習前後で同じ訳語が採用された。ただし、「脆弱部」と「参照」の2語のみは、前者が fragile portion から fragile part へ、後者が see から refer to へと同義の範囲内で訳語が変化している。

文構造に関しては、本例の無学習時の翻訳文は「当業者が適宜選択し得た事項である」に相当する内容が完全に欠落し、原文とかい離した状況になってしまっているため、学習前後の文構造を比較できる状況にない。

初回学習前後の翻訳文の実質的な翻訳品質を判定すると、無学習時は上記のとおり原文の主旨である「当業者が適宜選択し得た事項である」に相当する部分が完全に欠落しており、さらには「積層してから形成する」という作業順序のニュアンス（正解訳文の after に相当）もうまく表現できておらず、原文の内容をほとんど把握できない。これに対して、初回学習後の翻訳文は、全ての特許・審査書類用語が正確に訳出されるようになり、また作業順序のニュアンスも and then と訳出されるようになった。正解訳文の after に比べると（特に vice versa との関係性において）まだ分かりにくく、改善の余地はあるが、全

³⁰ 正解訳文は a matter の訳出を省略して「(当業者であれば) 適宜選択し得る (appropriately selected)」と動詞に意識しており、機械翻訳文は同じ訳語を直訳調で用いている。

³¹ 実例 3-1 でも無学習時には「発明の詳細な説明」が1回欠落し、「記載も示唆もない事項」も「示唆」や「事項」に相当する語句が欠落していた。

体として原文の内容は十分に把握できるレベルといえる。よって、翻訳品質は大幅に「向上」している。

これに対して自動評価スコアは、BLEU スコアはコーパス学習後に大きく向上しており実際の品質改善と整合するが、RIBES スコアは学習後に低下するという、実態とは正反対の評価になってしまっている。学習前後での翻訳文の改善は極めて顕著であり、これを反映できていないことは問題である。

続いて、第2回学習後から第5回学習後までの機械翻訳結果を示す。

【文番号：A913-#314】

第2回学習後	It is a matter that can be appropriately selected by a person skilled in the art that after laminating the barrier film layer, the fragile part is formed on the barrier film layer or vice versa (for example, see paragraph 0010 of Cited Document 3).	
	悪化	BLEU : 0.5314 (+0.0715) RIBES : 0.7270 (+0.0108)
第3回学習後	After laminating the barrier film layer, whether to form a fragile portion on the barrier film layer or vice versa is a matter that could have been appropriately selected by a person skilled in the art (see, for example, paragraph 0010 in Cited Document 3).	
	悪化	BLEU : 0.4024 (-0.1290) RIBES : 0.8535 (+0.1266)
第4回学習後	A person skilled in the art could appropriately select whether to form a fragile part in a barrier film layer after laminating a barrier film layer, or vice versa (for example, refer to paragraph 0010 of Cited Document 3).	
	向上	BLEU : 0.3592 (-0.0432) RIBES : 0.7127 (-0.1409)
第5回学習後	It is a matter that can be appropriately selected by a person skilled in the art whether the barrier film layer is laminated and then a fragile portion is formed on the barrier film layer or vice versa (for example, refer to paragraph 0010 of Cited Document 3).	
	悪化	BLEU : 0.4250 (+0.0658) RIBES : 0.7853 (+0.0727)

第2回から第5回学習後にかけても、特許・審査書類用語の「当業者」「[適宜選択した事項](#)」「[引用文献3](#)」にはそれぞれ初回学習後と同じ適訳語が採用され続け³²、高い安定性を示

³² 「[適宜選択した事項](#)」に関しては「事項：a matter」が訳出されていない回もあるが（第4回）、これは正解訳文と同様の、意識に伴う適切な省略であり、事実上同一と見なした。

している。

技術／一般用語に関しては、初回学習前後で変化しなかった各語（「バリア性フィルム層」「積層する」「形成する」「その逆とする」）は引き続き無変化であり、一方、初回学習時に変化していた「脆弱部（fragile portion → fragile part）」と「参照（see → refer to）」の2語は、その後も上記二種の訳語間での変化を繰り返している。ただし、前述のとおり、どちらも同義内での変化であり、内容理解上の優劣はない。

初回学習後の翻訳文の唯一の改善点として指摘した and then については、第2回から第4回学習後にかけて正解訳文どおりの after に改善されたが、第5回学習後には再び and then に戻っており、こちらも2つの訳語間での不安定な変化が見られる。

文構造に関しては、前例と同じく毎回大きく変化している。第2回と第3回学習後には after 節を先頭に配置する文型が採用されたが、その結果、正解訳文どおり after が採用されているにもかかわらず、vice versa がどこに係るのかが初回学習後よりも不明瞭になってしまっている。このことは after 節が本文の後に配置された第4回学習後の翻訳文の明解さと比較すれば明らかである。このように、文構造的には同義の範囲であっても文の内容や含まれる語句次第では適不適、優劣が生じることもある。続く第5回学習後の翻訳文では前述のとおり and then が再び採用され、初回時とほぼ同一の翻訳文（相違点は「脆弱部」が the fragile part から a fragile portion に変化しているのみ）に戻った。

本例の第1回から第5回学習後までの機械翻訳文の変遷をまとめると、特許・審査書類用語は前例と同じく初回学習時から最適訳語で常時安定している。一方、技術／一般用語においては、訳語が常時安定している語もあるものの、一部の語は複数の訳語間で変化を繰り返した。文構造に関しても、前例と同様に頻繁に変化していた。ただし、こうした技術／一般用語や文構造の変化の大半は、翻訳品質に影響しない同義語の範囲内の変化であった。その結果、本例における各回の翻訳品質の優劣は、「積層してから形成する」という作業順序が採用される文構造や訳語によっては伝わりにくくなるというやや例外的な一点に絞られ、第2回学習後の翻訳文は after 節の配置が悪く第1回学習後よりやや「悪化」、第3回学習後は after 節の配置がさらに悪くなり引き続き「悪化」、第4回学習後は after 節の配置が改善して最良の語順となったため「向上」、そして第5回学習後は初回学習後と同等の品質だが第4回学習後の完璧な翻訳文と比べると若干わかりにくく、軽度の「悪化」とした。

こうした実際の品質変化の推移に対し、各回の自動評価スコアの変動はこれと整合しないことが多い。顕著な品質改善があった初回学習後には RIBES スコアの不当な低下があり、その後も第3回学習後に BLEU スコアが急増、RIBES が急落と両指標間で極端な矛盾が生

じているなど、ここまで本例に付された自動評価スコアから実際の翻訳品質の推移を正しく読み取することは困難であり、個々の文単位での評価精度は高いとはいえない。

続いて、第6回学習後から第10回学習後までの機械翻訳結果を示す。

【文番号：A913-#314】

第6回学習後	It is a matter that can be appropriately selected by a person skilled in the art to form a fragile portion in the barrier film layer after laminating the barrier film layer or vice versa (see, for example, paragraph 0010 of Cited Document 3).	
	向上	BLEU : 0.5188 (+0.0938) RIBES : 0.7175 (-0.0678)
第7回学習後	It is a matter that a person skilled in the art could appropriately select whether a barrier film layer is formed by laminating a barrier film layer or a reversed barrier film layer (for example, refer to paragraph 0010 of Cited Document 3).	
	悪化	BLEU : 0.3126 (-0.2062) RIBES : 0.7257 (+0.0082)
第8回学習後	It is a matter that can be appropriately selected by a person skilled in the art to form a fragile portion on a barrier film layer or vice versa after laminating a barrier film layer (for example, see paragraph [0010] of Cited Document 3).	
	向上	BLEU : 0.3989 (+0.0863) RIBES : 0.7625 (+0.0369)
第9回学習後	It is a matter that could have been appropriately selected by a person skilled in the art whether the barrier film layer is formed by laminating the barrier film layer or the opposite portion thereof (see, for example, paragraph 0010 of Cited Document 3).	
	悪化	BLEU : 0.3969 (-0.0020) RIBES : 0.7773 (+0.0147)
第10回学習後	It is a matter that could have been appropriately selected by a person skilled in the art whether a barrier film layer is laminated, a fragile portion is formed on the barrier film layer, or the barrier film layer is reversed (see, for example, paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.4239 (+0.0270) RIBES : 0.7109 (-0.0663)

第6回から第10回学習後にかけても、特許・審査書類用語は3語とも正解訳文どおりの最適訳語で引き続き安定している。一方、技術／一般用語に関しては、「脆弱部」と「参照」の2語において訳語が頻繁に変化する状況はこれまでどおりだが、さらに「その逆とする：vice versa」というフレーズの誤訳が第7回、第9回、第10回学習後と頻発するようになった。

た。例えば第7回学習後では後半部が **whether a barrier film layer is formed by laminating a barrier film layer or a reversed barrier film layer** (バリア性フィルム層が、バリア製フィルム層、又は裏返したバリア性フィルム層を積層することによって形成される) と、原文の内容 (バリア性フィルム層を積層してから、バリア性フィルム層に脆弱部を形成するか、あるいはその逆とする) を大きく誤解した内容になっている。vice versa という成句は採用されず、代わりに「逆」の意味を持つ **reversed** (裏返した) が採用された。さらには「脆弱部 (fragile potion)」という語も消失し、これに代わって原文には二度しか出現しない「バリア性フィルム層 (barrier film layer)」が三回出力されている。このように原文中の語が翻訳文から欠落し、その埋め合わせとして原文中の別の語が重複して出力されるという現象は多くの文で見られ、当時の「みんなの自動翻訳」の典型的なエラーのひとつである³³。

第9回、第10回学習後も、「その逆とする」に相当する vice versa は採用されず、代わりに第9回では **opposite**、第10回では **reversed** とやはり誤解釈に基づく「逆」の意味の訳語が採用されている。第7回学習後のような語の欠落こそ生じていないものの、こうした「その逆とする」の誤解釈によって文の後半はミスリーディングな内容となっている。

各回で品質劣化の原因となった「その逆とする」というフレーズの誤解釈に関しては、他の回では vice versa が採用されていることから、このフレーズを成句と見なし vice versa と訳す選択肢と、「逆」ほか個々の語句に分解して訳す選択肢の二つがエンジン内に存在し、学習回によってその優先順位が変化しているものと推察される。ただし、外部からはこの優先順位が学習データ全体での出現頻度に基づいて決められているようには見えず、毎回ランダムに選ばれているように映る。

本例や前例の各所に見られる、訳語や文構造が複数の候補間で学習の都度ランダム的な変化が生じるという傾向は、「みんなの自動翻訳」が採用するニューラル翻訳方式での訳語学習のメカニズムに起因すると考えられる。「みんなの自動翻訳」でのメカニズムの詳細は非開示であるが、以下、一般的なニューラル翻訳の訳語選定のメカニズムを簡単に説明する。

ニューラル翻訳は、学習データである大量の対訳文対から翻訳モデルを構築するという点では従来の統計翻訳方式と同様であるが、統計翻訳のようにフレーズテーブル (句単位の対訳辞書) を生成・利用するのではなく、学習データに基づいて個々の単語や構文を多次元のベクトルで表現している³⁴。具体的には、学習データ中の各文例の各用語を、その用語の

³³ ただし、その後の大規模アップデート (⇒2.2.2.) により本エラーは完全に解消されたと見られる。詳しくは第3.3.5.3.6.項を参照のこと。

³⁴ 江原暉将、岡俊行：ニューラル機械翻訳における訳語誤りについての分析，2019。
https://japio.or.jp/00yearbook/files/2019book/19_4_05.pdf

周辺に存在する語句と関連付けて学習し、多次元ベクトル化する。そして、学習データ中の全ての文例、用語のベクトルを集約して「翻訳モデル」を作成する。さらに、作成した翻訳モデルを用いて学習データ中の各原文を仮翻訳し、これと学習データ中で原文と対をなす訳文とを比較する。翻訳モデルは学習データ全体を集約したものであるため、個々の文レベルでは仮翻訳結果と学習データ中の訳文とは異なることが多くなる。この場合、仮翻訳結果が訳文に近づくようベクトルを調整し、翻訳モデルを更新する。この翻訳モデルの更新と仮翻訳とを繰り返して、仮翻訳結果と学習データの訳文の差が収束したら、その時点で最も仮翻訳結果との差の小さい翻訳モデルを採用する。これが一般的なニューラル翻訳における学習の手順である。

こうした学習メカニズムにより、ニューラル機械翻訳では、ごく平易に述べると、ある用語について学習データ中に複数の訳語が存在する場合、それぞれの訳語に対し、その訳語が使われた各文例における周辺情報に基づいた多次元ベクトルが付与される。そして、訳語選定の際は、入力文における用語の周辺情報と、学習データ中の各訳語候補のベクトルとを照合して、最もフィットすると判断された訳語が採用されることになる。

この多次元ベクトルは、各用語間の学習データ全体における共起関係に基づくため、学習データが追加されると、その追加データ中に存在しない訳語であっても、周辺に位置する語のベクトルの変化により、間接的に変化する可能性がある。学習の都度、このように文中の各語のベクトルが変化するため、同じ入力文であっても学習前後で採用される訳語や文構造が変化する場合がある。

特許・審査書類用語の場合、その多くは審査書類対訳コーパスで統一訳語が定められており、事実上、学習データ中に他の訳語候補が存在しない。このため、学習の都度ベクトルが変化しても、ほかに選択の余地がないため、常時安定してコーパスどおりの適訳語が採用されていると考えられる。これに対し、技術／一般用語の多くは学習データ中にさまざまな訳語候補が存在し、かつ、段階的に追加される学習データに新たな訳語や新たな文例が含まれる確率も高いため、学習前後で訳語のベクトルが変化する傾向が強い。特に「みんなの自動翻訳」の場合、このベクトルが学習データ追加の都度大きく変化する傾向があると見られ、これにより訳語間の優先順位が学習の都度頻繁に入れ替わっているものと推察される。これが外部からは、ランダム的な訳語変化と映る。

定点観測に戻り、第6回から第10回学習後にかけての翻訳品質を見ていくと、第6回学習後こそ第5回時の軽微な不備（and then の使用）が解消して第4回学習後と同等の水準に「向上」したが、続く第7回学習後は前述した「その逆とする」の誤解釈により品質は大幅に「悪化」した。第8回学習後にはこの誤解釈が解消して品質は前回よりは「向上」した

が、after 節の配置が悪く、第 1 回～第 5 回学習後に比べると依然低品質である。第 9 回学習後は再度「その逆とする」の誤解釈が発生したため「悪化」、第 10 回学習後もこの誤解釈が解消されず品質は「同等」である。こうした各回の品質の変動に対し、自動評価スコアは BLEU と RIBES で変動方向が矛盾している回（一方が増加、他方が減少）が多く、引き続き評価精度は低い。

最後に、第 11 回学習後から最終第 18 回学習後までの機械翻訳結果を示す。

【文番号：A913-#314】

第 11 回学習後	Whether the fragile portion is formed in the barrier film layer after laminating the barrier film layer or vice versa is a matter that can be appropriately selected by a person skilled in the art (for example, see paragraph 0010 of Cited Document 3).	
	向上	BLEU : 0.5933 (+0.1694) RIBES : 0.9152 (+0.2043)
第 12 回学習後	Whether to form the fragile portion in the barrier film layer after laminating the barrier film layer or vice versa is a matter that can be appropriately selected by a person skilled in the art (for example, see paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.6077 (+0.0144) RIBES : 0.9215 (+0.0063)
第 13 回学習後	It is a matter that could have been appropriately selected by a person skilled in the art whether to form the fragile portion in the barrier film layer after laminating the barrier film layer or vice versa (for example, see paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.5438 (-0.0639) RIBES : 0.7261 (-0.1955)
第 14 回学習後	Whether to form a fragile portion in the barrier film layer after laminating the barrier film layer or vice versa is a matter that would have been appropriately selected by a person skilled in the art (for example, see paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.5685 (+0.0247) RIBES : 0.9063 (+0.1802)
第 15 回学習後	It is a matter that could have been appropriately selected by a person skilled in the art whether to form the fragile portion in the barrier film layer after laminating the barrier film layer or vice versa (see, for example, paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.4865 (-0.0820) RIBES : 0.7195 (-0.1868)
第 16 回学習後	Whether to form a fragile portion in a barrier film layer after laminating the barrier film layer or vice versa is a matter that could have been	

	appropriately selected by a person skilled in the art (for example, see paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.5258 (+0.0393) RIBES : 0.8862 (+0.1667)
第 17 回学習後	Whether to form a fragile portion in the barrier film layer after laminating the barrier film layer or vice versa is a matter that could have been appropriately selected by a person skilled in the art (see, for example, paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.4977 (-0.0281) RIBES : 0.8992 (+0.0130)
第 18 回学習後	Whether to form a fragile portion in the barrier film layer after laminating the barrier film layer or vice versa is a matter that could have been appropriately selected by a person skilled in the art (see, for example, paragraph 0010 of Cited Document 3).	
	同等	BLEU : 0.4977 (± 0) RIBES : 0.8992 (± 0)

第 11 回学習前に「みんなの自動翻訳」の大規模アップデートが実施されているため、上掲の翻訳文は全てアップデート後のエンジンによる機械翻訳結果となる。

これら第 11 回学習後以降の機械翻訳文では、第 7 回時から第 10 回学習後にかけて頻発していた「その逆とする」というフレーズの誤解釈が一切発生しなくなり、常時 vice versa という成句が安定して採用されるようになった。この期間も学習データ中には「その逆とする」を vice versa と訳さないという選択肢を誘発する文例は引き続き存在しているはずだが、大規模アップデートを境に、この不適な選択肢は一切採用されなくなった。

また、これまでランダム的な変化が続いていた技術／一般用語「脆弱部」と「参照」も、第 11 回学習後以降はそれぞれ fragile portion と see に訳語が固定された。これら 2 語にも学習データ中に別の訳語候補（例えば fragile part や refer to）が存在することは、学習初期の機械翻訳結果から明らかである。

前例にも見られたこうした大規模アップデート以降の傾向の変化と、第 3.2.6.項における同じタイミングでの「不変」文の顕著な増加という事実を考え合わせると、大規模アップデートを境に、学習データ中の最有力の訳語候補や文構造候補が以前より安定して採用されるようになった可能性が高い。その結果、各訳語の不安定な変化が抑制されるとともに、本例の「その逆とする」におけるような、学習データ中のマイナーな訳語が突発的に採用されることによる不適訳の発生も減少したものと推察される。大規模アップデート以降、正解訳文との一致率に基づく自動評価スコアも全体的に向上しており、正解訳文で採用されるような適訳語、適切な文構造が機械翻訳文でも順当に採用される頻度が増加したことを示し

ている。

こうした傾向の変化は、前述したニューラル翻訳方式の学習メカニズムに照らせば、「みんなの自動翻訳」において各語に付与する多次元ベクトルの算出ロジックが改良された結果と考えられる。各用語に対し、学習データ全体の内容を反映した精度の高いベクトルが付与されるようになったことで、学習データ追加の都度のベクトルの過剰な変化が抑制され、各用語とも最有力な訳語が順当に採用される確率が高まったものと推察される。本例における大規模アップデート以降の「その逆とする：vice versa」や「脆弱部：fragile portion」、「参照：see」に見られる訳語の安定化も、学習データ全体における訳語の出現頻度等がよりの確に翻訳文に反映されるようになった結果であろう³⁵。

文構造に関しては、大規模アップデート以降、Whether to form…を主語とする構造（第 11、12、14、16、17、18 回学習後）が頻繁に採用されるようになり、形式主語 it を用いる構造（第 13、15 回学習後）がその合間にランダム的に採用されている状況となっている。どちらの文構造も完全に同義であり、少数かつ同義の候補間での限定的な変化に収まっていると見なせる。技術／一般用語に関しても、大規模アップデート以降は原文中の各用語の訳語はほぼ不変となり、barrier layer や fragile portion に付される冠詞が定冠詞 the と不定冠詞 a との間で変化する程度の小規模かつ局所的な変化にとどまっている。つまり文構造、技術／一般用語の双方において、第 10 回学習後までに比べて翻訳文は明らかに安定化している。

第 11 回学習後から最終学習後までの各回の翻訳品質の推移は、「その逆とする」の誤訳が解消された第 11 回学習後に第 4 回、第 6 回と同等の最高水準に再び「向上」し、以降は最終学習時まで全回これと「同等」であった。これも、大規模アップデート以降の翻訳文が学習データに基づく適訳語・適訳文で安定したことを示す結果といえる。これに対し、自動評価スコアは Whether 節を主語とする文構造が採用された回では向上し、It を形式主語とする文構造の回では低下する傾向を示し、引き続き表層的な変化に左右されている。各回の変動幅も非常に大きく、各回とも同品質という実態を反映できていない。

以上、実例 3-2 の各回の機械翻訳結果の推移を分析した。最後に、前例から看取した各種傾向について、本例の状況と照合していく。

³⁵ 本例の正解訳文では「その逆とする」は vice versa と訳されておらず「脆弱部」も fragile portion ではないが、技術／一般用語の場合、審査書類対訳コーパス中でも訳語が統一されているわけではなく、かつ、主要な学習データも「みんなの自動翻訳」の基礎学習データであるため（⇒3.2.4.）、特許・審査書類用語のように常に正解訳文と同じ訳語が最有力候補となるとは限らない。

- ・ 本例も前例と同様に無学習時の機械翻訳文では特許用語・審査書類用語の誤訳や欠落が多発していたが、審査書類対訳コーパス初回学習後には全ての語が適訳語に改善され、以降も全期間にわたり完璧に安定した。学習データである審査書類対訳コーパス中で各用語の訳語が統一されており、他の訳語候補が存在しないためと考えられる。

本例の状況から、大多数の文における大半の特許・審査書類用語は審査書類対訳コーパスの初回学習によって正解訳文どおりの適訳語に一挙に改善し、そのまま安定したという前例での推論がさらに裏付けられた。

- ・ 文構造に関しては、本例でも学習の都度不安定に変化する傾向が顕著であった。本例ではさらに、一部の技術用語や一般用語においても訳語の頻繁な変化が見られた。

こうした文構造や技術／一般用語の不安定な変化は、ニューラル翻訳方式における訳語学習のメカニズムと、学習データ中に複数の訳語候補が存在するという技術／一般用語の性質とが相俟った結果と推察される。ここまでの二例を見る限り、これら技術／一般用語や文構造の変化は同義の範囲内にとどまり、実質的な翻訳品質に影響しない場合が大半であったが、本例の「その逆とする」の誤訳など、重大な品質悪化を招く不適な訳語や文構造が突発的に採用されるケースも存在した。

- ・ こうした不安定な訳語／文構造の変化は、第 11 回学習前の「みんなの自動翻訳」の大規模アップデート以降は明らかに抑制された。本例においても、重大な品質悪化をもたらしていた「その逆とする」の誤解釈を始め、「脆弱部」や「参照」など技術／一般用語の不安定な訳語変化がアップデート以降は一切発生しなくなり、文構造の変化も完全に同義な 2 種の文パターン間に限定されるなど、各回における変化は明らかに小規模化・局所化した。

- ・ 前例では確認できなかった「みんなの自動翻訳」大規模アップデートの品質改善効果だが、本例においては、アップデート以前は度々採用され品質悪化を招いていた「その逆とする」の誤解釈がアップデート以降は一切発生しなくなり、この点において、品質水準はアップデート以前よりも向上した。

この傾向の変化は、「みんなの自動翻訳」自身の訳語学習ロジックの改良により、学習データ中に複数の訳語候補が存在する場合でも、最有力の訳語／文構造候補が安定して採用されるようになったためと考えられる。自動評価の全体平均スコアが大規模アップデート時に顕著に向上し、以降もその水準が維持されたのも、各文において突発的な不適訳語の採用が抑制され、適訳語が安定して採用されるようになったためと考えられる。

- ・ 本例における各回の翻訳品質の推移は、自動評価の全体平均スコアが示す推移とおおむね整合しており、全体平均スコアから看取される品質改善が実体を伴っていることが裏付

けられた。ただし、本例の機械翻訳文に付された自動評価スコア自体は実際の翻訳品質と
かい離した変動を示すことが多く、また BLEU と RIBES とで矛盾した評価がなされるな
ど、個々の文単位での評価精度には限界があることも分かった。本例を見る限り、自動評
価スコアはあくまで一定規模以上の母集合による平均スコアとして利用すべきものとい
える。

3.2.7.3. 実例 3-3 : 「補正却下の決定」由来、文番号 A191-#38 の分析

第三の実例には「補正却下の決定」の評価用コーパスの文番号#38 を取り上げる。無学習時から最終回（第 18 回）学習後まで、全回の機械翻訳結果を一覧で示す。

【文番号：A191-#38】

原文	引用刊行物 1 に記載の衝撃緩和積層体も、二層構造である。	
正解訳文	A cushioning laminate described in Cited Publication 1 also has a double-layer structure.	
無学習	The shock relief laminate described in reference 1 is also a 2 layer structure.	
	—	BLEU : 0.1383 (－) RIBES : 0.8265 (－)
第 1 回学習後	The impact relaxation laminate described in Cited Publication 1 is also a 2 layer structure.	
	向上	BLEU : 0.3583 (+0.2200) RIBES : 0.8666 (+0.0395)
第 2 回学習後	The impact absorbing laminate described in Cited Publication 1 is also a 2 layer structure.	
	同等	BLEU : 0.3583 (±0) RIBES : 0.8666 (±0)
第 3 回学習後	The impact relaxation laminate described in Cited Publication 1 has a 2 layer structure.	
	悪化	BLEU : 0.4002 (+0.0419) RIBES : 0.9036 (+0.0376)
第 4 回学習後	The shock absorbing laminate described in Cited Publication 1 also has a 2 layer structure.	
	向上	BLEU : 0.5555 (+0.1553) RIBES : 0.9106 (+0.0070)
第 5 回学習後	The impact relaxation laminate described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.5555 (±0) RIBES : 0.9106 (±0)
第 6 回学習後	The impact relaxation laminated body described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.4527 (-0.1028) RIBES : 0.8758 (-0.0348)
第 7 回学習後	The impact relaxation laminated body described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.4527 (±0) RIBES : 0.8758 (±0)
第 8 回学習後	The shock absorbing laminate described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.5555 (+0.1028) RIBES : 0.9106 (+0.0348)

第 9 回学習後	The impact relaxation laminated body described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.4527 (-0.1028) RIBES : 0.8758 (-0.0348)
第 10 回学習後	The impact mitigating laminate described in Cited Publication 1 also has a 2 layer structure.	
	同等	BLEU : 0.5555 (+0.1028) RIBES : 0.9106 (+0.0348)
第 11 回学習後	The shock absorbing laminate described in Cited Publication 1 also has a two - layer structure.	
	同等	BLEU : 0.5194 (-0.0361) RIBES : 0.8969 (-0.0137)
第 12 回学習後	The shock absorbing laminate described in Cited Publication 1 also has a two - layer structure.	
	同等	BLEU : 0.5194 (± 0) RIBES : 0.8969 (± 0)
第 13 回学習後	The impact mitigation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (+0.0361) RIBES : 0.9106 (+0.0137)
第 14 回学習後	The impact mitigation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (± 0) RIBES : 0.9106 (± 0)
第 15 回学習後	The impact mitigation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (± 0) RIBES : 0.9106 (± 0)
第 16 回学習後	The impact relaxation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (± 0) RIBES : 0.9106 (± 0)
第 17 回学習後	The impact relaxation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (± 0) RIBES : 0.9106 (± 0)
第 18 回学習後	The impact relaxation laminate described in Cited Publication 1 also has a two layer structure.	
	同等	BLEU : 0.5555 (± 0) RIBES : 0.9106 (± 0)

本例に含まれる特許用語は「引用刊行物 1」のみである。無学習時の機械翻訳文においては reference 1 と訳されている。誤訳ではないが「刊行物」のニュアンスに乏しく、最適訳とも言い難い。これに対し、初回学習後の翻訳文では Cited Publication 1 と正解訳文どおりの適訳語が採用され、以降も最終学習時まで常時この訳語が採用され続けた。つまり、本例においても前二例と同じく、特許・審査書類用語が学習初期から適訳語で安定するという傾向が確認された。

一方、前二例に見られた文構造の不安定性は本例ではほぼ見られず、第 4 回学習後以降は常に「A described in B also has C」という骨格構造が安定的に採用されている。無学習時から第 3 回学習後にかけても、上記構造の has の代わりに is が採用される、あるいは also が訳出されないという局所的な相違にとどまっており、事実上、文構造は全期間を通して不変であったといえる。原文自体が単純な構造であるため、ほぼ変化する余地がなかったものと推察される。

その反面、本例では、前例でも見られた技術用語の訳語のランダム的な変化がさらに顕著であった。以下、各回で採用された「衝撃緩和積層体」の訳語の変遷を一覧で示す。

原文	衝撃緩和積層体
正解訳文	cushioning laminate
無学習	shock relief laminate
第 1 回学習後	impact relaxation laminate
第 2 回学習後	impact absorbing laminate
第 3 回学習後	impact relaxation laminate
第 4 回学習後	shock absorbing laminate
第 5 回学習後	impact relaxation laminate
第 6 回学習後	impact relaxation laminated body
第 7 回学習後	impact relaxation laminated body
第 8 回学習後	shock absorbing laminate
第 9 回学習後	impact relaxation laminated body
第 10 回学習後	impact mitigating laminate
第 11 回学習後	shock absorbing laminate
第 12 回学習後	shock absorbing laminate
第 13 回学習後	impact mitigation laminate
第 14 回学習後	impact mitigation laminate
第 15 回学習後	impact mitigation laminate

第 16 回学習後	impact relaxation laminate
第 17 回学習後	impact relaxation laminate
第 18 回学習後	impact relaxation laminate

上掲のとおり、本例では「衝撃緩和積層体」という一つの用語に対して 7 種類の訳語が頻繁に入れ替わって採用されている。事例 3-2 でも「脆弱部」「参照」など一部の技術／一般用語について同様のランダム的な訳語変化が見られたが、本語で採用された訳語はさらに多数で、変化もより頻繁である。本語は「衝撃」「緩和」「積層体」の三語からなる複合語であるが、それぞれの語が複数の訳語候補を有しており、それらが多様に組み合わせられた結果とも映る。

技術用語に関しては、第 3.2.4.項において、審査書類対訳コーパスでカバーされる技術用語は非網羅的であり、したがって技術用語の翻訳品質は基本的には機械翻訳エンジンのデフォルトの学習内容（基礎学習）に委ねられる旨を述べた。つまり技術用語の訳語は、多くの場合「みんなの自動翻訳」の基礎学習データが主な学習対象となると考えてよい。

この基礎学習データが全期間を通じて同一と仮定すれば³⁶、審査書類対訳コーパス中で使用例のない用語は、学習データ全体における文例の内容や発生頻度が学習全期間を通じて一切変化しない。だが、事例 3-2 の分析時に述べたとおり、ニューラル機械翻訳の訳語学習のメカニズムでは、各用語に対し、学習データ中の各文例においてその用語の周辺に存在する語句と関連付けて多次元ベクトルを付与する。このベクトルは各文例の各用語が相互に影響し合う形で算出されるため、学習データが追加される都度、学習データ全体の全ての用語に対して再計算される。さらに、再計算された各用語のベクトルは、その集合体というべき翻訳モデルの全体最適化の過程でさらに試行錯誤的に調整されていく。このため、審査書類対訳コーパス中で使用例のない技術用語であっても、付与されるベクトルは学習の都度変化することが多いと考えられる。「みんなの自動翻訳」の場合、特に大規模アップデート以前は、こうした一連の過程において各語のベクトルがやや過剰に変化していた可能性がある。このため、基礎学習データを含む全学習データ中に複数の訳語候補が存在する技術用語や一般用語、文構造では、学習前後で最優先となる訳語が不安定に変化していたものと推察される。

本例の「衝撃緩和積層体」という技術用語も、審査書類対訳コーパスでは第 16 回時の学習データ中に 1 回のみ含まれたが（その訳語は impact relaxation laminate であり、以降本例では各回ともこの訳語が採用された）、それ以前のコーパスには含まれていなかった。基

³⁶ 実際には、定期バージョンアップの際に基礎学習データも補充されている可能性がある。

基礎学習データの内容は外部からは知り得ないが、上掲した各回の訳語の変遷を見る限りは、基礎学習データ中にもおそらく存在せず、このため「衝撃」「緩和」「積層体」と単語単位に分解して翻訳し、それぞれの訳語を連結していたと考えられる。これら三語とも技術用語としては一般的であり、学習データ中に複数の訳語候補が存在していることは明らかである。こうした状況と、旧「みんなの自動翻訳」³⁷における各用語のベクトルの過度の変動とが相俟って、上表のような多様なバリエーションが生じたものと考えられる。

なお、前例（実例 3-2）では、「みんなの自動翻訳」のベクトル算出ロジックが第 11 回学習前の大規模アップデート時に改良され、学習データ中の最有力の訳語候補や文構造候補が安定して採用される傾向が強くなったと推察した。本例の「衝撃緩和積層体」も、第 10 回学習後まではほぼ毎回訳語が変化していたが、第 11 回学習後以降は、第 11 回～第 12 回が shock absorbing laminate、第 13 回～第 15 回が impact mitigation laminate、第 16 回～第 18 回が impact relaxation laminate と学習前後で同一の訳語が連続して採用されることが多くなり、訳語のバリエーションも少数となった。前述のとおり、第 16 回時に追加された学習データ中には「衝撃緩和積層体：impact relaxation laminate」という訳例が 1 件存在した。それ以降、この訳語が安定して採用されるようになったこと自体は、学習データ中にこの訳語候補しか存在しないという、特許・審査書類用語と同様の状況になったためと考えられるが、それ以前の第 11 回～第 15 回学習後の翻訳文でも訳語変化は明らかに抑制されている。

ただし、以前より安定化したとはいえ、この用語も第 11～12 回学習後と第 13 回～15 回学習後とで訳語が変化しており、学習データ中に有力な訳語候補が複数存在する一部の用語では、大規模アップデート以降もランダム的な訳語変化が引き続き発生している状況がうかがえる。これは、前述のとおり、ニューラル機械翻訳方式の訳語学習メカニズム自身に起因するものであり、大規模アップデート以降も過半数の文で何らかの訳語変化が生じ続けている状況からも、「みんなの自動翻訳」においても本質的な傾向といえる。

最後に本例の実際の翻訳品質の判定結果をまとめると、学習初期の第 3 回学習後に突発的な also の欠落が発生して一時的に「悪化」となったものの、第 4 回学習後以降は変化箇所が「衝撃緩和積層体」の訳語のみに限定され、かつ、どの訳語もおおむね同義であって明確な優劣はつかない。このため、第 4 回学習後以降は一律「同等」とした。これに対し、自動評価スコアは第 4 回学習後以降も大きな変動を繰り返している。変化箇所が一語のみ、かつ同義語の範囲内での変化であるにしてはスコアの変動が極端であり、前例に引き続き、実態を的確に反映しているとはいえない。

³⁷ 大規模アップデート以前のバージョン 3.8 を指す。以下同じ。

以上、本例では主に「みんなの自動翻訳」における技術用語のランダム的な変化の状況と、大規模アップデート以降この傾向がどのように変化したかを中心に各回の翻訳文を分析した。以下、事例 3-1～2 で指摘した審査書類①システムの各種傾向を、本例の状況と照合する。

- ・ 本例に含まれる特許・審査書類用語は「引用刊行物 1」のみであった。この語も無学習時には reference 1 と不適訳であったが、コーパス初回学習以降は常時 Cited Publication 1 と適訳語で訳されるようになった。よって、前二例と同様に、審査書類対訳コーパスの学習によって無学習時には正しく翻訳できなかった特許・審査書類用語が適訳語に改善される状況が本例でも等しく確認された。
- ・ 本例は原文が単純な構造であるせいか文構造には不安定な変化は生じなかったが、技術／一般用語に関しては「衝撃緩和積層体」の訳語が学習の都度不安定に変化し、前例と同じく、学習データ中に複数の訳語候補が存在する場合に学習前後で訳語が変化しやすいという「みんなの自動翻訳」の傾向が顕著に表れていた。本語の各回での訳語変化は前例で見られたものよりも激しいが、採用される各訳語がいずれも同義の範囲内であって翻訳品質には影響しない点や、大規模アップデート以降は訳語変化に一定の抑制が見られた点など、全期間を通じての傾向自体も前例と共通している。これらの傾向は全体平均スコアが示す状況とも整合しており、審査書類①の全体的な傾向であることは確実と見られる。
- ・ 本例は特許・審査書類用語、文構造とも学習初期から安定し、技術用語「衝撃緩和積層体」も各回で採用された訳語がいずれも同義の範囲内であったことから、第 3 回学習後の also の突発的な欠落による悪化を除き、全期間を通して翻訳品質は常時同等であった。これに対し、自動評価スコアは「衝撃緩和積層体」の表層的な訳語変化によって過大な増減を繰り返しており、前二例と同じく、個々の文単位の評価指標としては精度に限界があることを示す結果となった。

3.2.7.4. 実例 3-4 : 「拒絶理由通知書」由来、文番号 A131-#68 の分析

本章最後の実例として「拒絶理由通知書」の評価用コーパスの文番号#68 を取り上げる。まずは無学習時と初回学習後の機械翻訳結果を対比する。

【文番号：A131-#68】

原文	請求項 3 に、「許容基準」なる記載があるが、これが具体的にどのような基準であるのか把握できず、発明を不明確にしている。	
正解訳文	Claim 3 describes "an acceptable standard" but it cannot be grasped what standard it means specifically, and thus invention is made unclear.	
無学習	The description of the third aspect of the present invention is given by way of example, but it is to be understood that the present invention is not limited to this particular reference, and that the invention is unclear.	
	—	BLEU : 0.0360 (－) RIBES : 0.6389 (－)
第 1 回学習後	Although there is a description of "acceptable criteria" in Claim 3, it cannot be grasped what criteria this is, and the invention is made unclear.	
	向上	BLEU : 0.2806 (+0.2446) RIBES : 0.8249 (+0.1859)

無学習時の翻訳文では実例 3-1 と同じく「請求項 3」が third aspect と誤訳されるエラーが発生しているが、それ以前の問題として、原文と機械翻訳文とで文意が全く異なっている。

無学習時の機械翻訳文を日本語に戻すと、「本発明の請求項 3 は例示する形で記載されているが、本発明はこの特定の例のみに限定されないと理解すべきであり、よって発明が不明瞭である」という意味と理解される。これを原文と対比すると、まず、原文で主題となっている「許容基準」という技術用語が完全に消失しており、また、「これが具体的にどのような基準であるのか把握できず」という部分も全く異なる内容に訳されている。このため、文末の「発明を不明確にしている」という結論部分以外、この翻訳文から原文の内容は全く把握できない。

なぜこのような翻訳文が出力されたのかは不明だが、原文と合致しているのが文末の **the invention is unclear** (発明を不明確にしている) の箇所のみであることから、この箇所を手がかりに基礎学習データ中の英文がそのまま採用されたのかもしれない。いずれにせよ、無学習時の翻訳文は、原文内容の大部分が失われ、かつ、無関係の内容が大量に湧き出す³⁸な

³⁸ 原文中に存在しない要素が翻訳文中に混入することを指す。「湧き出しエラー」の詳細は第 3.3.5.3.6.項で後述する。

ど、きわめて不正確である。自動評価スコアは両指標とも非常に低い値となっているが、妥当な評価である。

これに対し、審査書類対訳コーパス初回学習後の機械翻訳文は非常に精度が良くなり、原文の内容を完璧に伝えるレベルになっている。「許容基準」は acceptable criteria と訳されており、正解訳文の acceptable standard とは異なるが、原文でも「どのような基準が書かれていない」と指摘しているようにこの語が具体的に何を指すかは不詳であり、両者の優劣はつけ難い。

初回学習後の翻訳文の実質的な品質は当然ながら無学習時より「向上」となる。自動評価スコアも無学習時と比べて大幅に増加しており、学習前後の相対評価としては妥当な結果となっている。ただし、スコアの絶対値としては、翻訳文が原文の内容を完璧に伝えるものになっているにもかかわらず、用いられている訳語や文構造が正解訳文と異なるため、特に BLEU では不当に低いスコアとなっている。

続いて、第2回～第5回学習後の機械翻訳結果を示す。

【文番号：A131-#68】

第2回学習後	Although there is a description of "allowable criteria" in Claim 3, it is impossible to grasp what kind of criteria this is, and the invention is unclear.	
	同等	BLEU : 0.0513 (-0.2293) RIBES : 0.7282 (-0.0966)
第3回学習後	In Claim 3, there is a description "allowable reference", but it is impossible to grasp what kind of standard this is, and the invention is unclear.	
	同等	BLEU : 0.0564 (+0.0051) RIBES : 0.7657 (+0.0375)
第4回学習後	Claim 3 describes "an allowable criteria", but it cannot be grasped what criteria this is, and the invention is unclear.	
	同等	BLEU : 0.3817 (+0.3253) RIBES : 0.9045 (+0.1388)
第5回学習後	In Claim 3, there is a description of "acceptable criteria", but it cannot be understood what criteria this is, and the invention makes the invention unclear.	
	悪化	BLEU : 0.1221 (-0.2596) RIBES : 0.7825 (-0.1219)

第 2 回学習後の翻訳文は、文の骨格構造は前回時とほぼ同じであるが、「許容基準」が allowable criteria へとさらに変化したのを始め、「把握できず」が cannot be grasped から impossible to grasp へ、「どのような基準」が what criteria から what kind of criteria へ、そして「発明を不明確にしている」が the invention is made unclear から the invention is unclear へと、各所の技術用語や一般用語において同義の範囲内での訳語変化が多発している。

この傾向は続く第 3 回～第 5 回学習後にかけての各翻訳文にも見られ、「許容基準」はこの期間も allowable reference (初出) → allowable criteria (2 度目) → acceptable criteria (2 度目) と複数の訳語候補間でランダム的な変化を続けている。「把握できず」や「どのような基準」といったフレーズも、言い回しの変化に加え、フレーズを構成する一般用語も「把握する」が grasp と understand、「基準」が criteria と standard と、やはり複数の同義語間で不安定な変化が生じている。

先に分析した三実例では、特許・審査書類用語の早期からの適訳語での安定、そして大規模アップデート以前における技術／一般用語や文構造のランダム的な変化という共通の傾向が見られた。これらの傾向は、第 5 回学習後までの機械翻訳文を見る限り、本例にも等しく当てはまる。

また、こうした技術／一般用語や文構造のランダム的な変化がおおむね同義の範囲内に収まり、実質的な翻訳品質への影響は小さいという点においても、本例を含め各例の状況は一致している。本例の第 1 回～第 4 回学習後の翻訳文はそれぞれ異なるが、各回の実質的な翻訳品質は全て「同等」で優劣はない。

ただし、第 5 回学習後の「発明を不明確にしている」の訳文 **the invention makes the invention unclear** のみは、主語無し原文の主語を補うにあたり、形式主語 it でなく「発明：invention」という具体的な名詞を採用した結果、「発明が発明を不明確にしている」という、意味の通らないものとなっている。このため「悪化」と判定した。主語無し原文に補った主語が不適訳になるという不備は、実例 3-1 で繰り返し発生した「they の誤訳」とも類似性があり、日英方向の翻訳における主語の補足が、機械翻訳にとって難易度の高い処理であることが分かる。

本例に付された自動評価スコアは、実際の品質が「同等」であった第 2 回～第 4 回学習後にかけても正解訳文との表層的な一致度の違いで毎回大きく増減しており、学習前後の相対評価として実際の状況を反映しているとは言い難い。特に BLEU では学習前後でスコアが数倍に増減するなど極端な変動が目立ち、絶対評価としても精度は低い。

続いて、第 6 回学習後から第 10 回学習後までの機械翻訳結果を見ていく。

【文番号：A131-#68】

第 6 回学習後	Although there is a description of "acceptable criteria" in Claim 3, it cannot be grasped what kind of criteria it specifically refers to, and the invention is unclear.	
	向上	BLEU : 0.1722 (+0.0501) RIBES : 0.8055 (+0.0229)
第 7 回学習後	In Claim 3, there is a description of "allowable standard", but it cannot be grasped what kind of criteria this is, and the invention makes the invention unclear.	
	悪化	BLEU : 0.2015 (+0.0293) RIBES : 0.8011 (-0.0044)
第 8 回学習後	Although there is a description of "allowable criteria" in Claim 3, it is not possible to grasp what kind of criteria it is, and thus the invention is unclear.	
	向上	BLEU : 0.0749 (-0.1266) RIBES : 0.7364 (-0.0646)
第 9 回学習後	In Claim 3, there is a description of "allowable criteria", but it is not possible to grasp what kind of criteria this is, and the invention is unclear.	
	同等	BLEU : 0.0520 (-0.0229) RIBES : 0.7542 (+0.0178)
第 10 回学習後	Although there is a description of "allowable criteria" in Claim 3, it cannot be understood what kind of criteria this is, and the invention is unclear.	
	同等	BLEU : 0.0864 (+0.0344) RIBES : 0.7888 (+0.0346)

第 6 回～第 10 回学習後の翻訳文も引き続き同傾向であり、技術用語「許容基準」は第 6 回が acceptable criteria(3 度目)、第 7 回が allowable standard(初出)、第 8 回が allowable criteria(3 度目)と同義の範囲での変化が継続している。その後、第 9 回と第 10 回は第 8 回と同じ allowable criteria が 3 回連続で採用されて安定化の兆しが見えたが、大規模アップデート以前であり、他の実例ではこの期間も技術／一般用語は不安定に変化していたことを考えると、偶然の安定という可能性もある。

各所の言い回しや文構造の変化も引き続き多発しており、第 6 回学習後には「どのような基準であるのか」の箇所が what kind of criteria it specifically refers to という初出の表現で訳出された（ただし同義の範囲内である）。また、第 5 回学習後の翻訳文で発生していた **the invention makes the invention unclear** という主語補足の誤りが、第 6 回学習後に一旦解消したものの第 7 回学習後に再発しており、実例 3-1 の they の誤訳と同じくランダム的に出沒している。初回学習後以降、実質的な翻訳品質の変動要因はこの一箇所の良否のみであ

り、この箇所に主語誤りが発生した回が「悪化」、解消した回が「向上」、それ以外の回は「同等」となる。これに対し、自動評価指標は実際の品質推移と逆の方向に変動していることも多く、また、同一回で BLEU と RIBES の変化の方向が矛盾する回も存在している。

最後に、第 11 回学習後から最終回（第 18 回）学習後までの翻訳結果を示す。第 11 回学習前の「みんなの自動翻訳」の大規模アップデートによって、本例にも他の実例と同様の傾向の変化が生じているかが焦点となる。

【文番号：A131-#68】

第 11 回学習後	In Claim 3, there is a description of "allowable criteria", but it is not possible to grasp what criteria this is specifically, and the invention is unclear.	
	同等	BLEU : 0.0809 (-0.0055) RIBES : 0.7765 (-0.0122)
第 12 回学習後	Claim 3 recites "allowable criteria", but it is not possible to grasp what kind of criteria this is specifically, and the invention is unclear.	
	同等	BLEU : 0.0926 (+0.0117) RIBES : 0.8020 (+0.0255)
第 13 回学習後	Although Claim 3 recites an "allowable standard", it is impossible to grasp what kind of standard this is, and the invention is unclear.	
	同等	BLEU : 0.0661 (-0.0265) RIBES : 0.8091 (+0.0071)
第 14 回学習後	In Claim 3, there is a description of "allowable standard", but it cannot be grasped what kind of standard this is, and the invention is unclear.	
	同等	BLEU : 0.2237 (+0.1576) RIBES : 0.8409 (+0.0318)
第 15 回学習後	Claim 3 recites an "allowable standard", but it is not possible to grasp what kind of standard this is specifically, and the invention is unclear.	
	同等	BLEU : 0.0975 (-0.0748) RIBES : 0.8409 (± 0)
第 16 回学習後	Claim 3 recites "allowable criteria", but it is not possible to grasp what kind of criteria this is specifically, and the invention is unclear.	
	同等	BLEU : 0.0926 (-0.0049) RIBES : 0.8020 (-0.0389)
第 17 回学習後	Claim 3 recites an "allowable standard", but it is not possible to grasp what kind of standard this is, and the invention is unclear.	
	同等	BLEU : 0.0677 (-0.0249) RIBES : 0.8336 (+0.0316)
第 18 回学習後	Although there is a recitation of "allowable criteria" in Claim 3, it is not possible to grasp what criteria this is, and the invention is unclear.	
	同等	BLEU : 0.0531 (-0.0146) RIBES : 0.7340 (-0.0996)

上掲した各回の翻訳文の変遷からは、他の事例において大規模アップデート以降に顕著に見られた二つの傾向、すなわち技術／一般用語や文構造の不安定な変化が抑制され、学習データに準拠した有力な訳語候補が安定して採用されるようになったことと、その反面、訳語や文構造のランダム的な変化はニューラル翻訳方式の本質的な性質であってアップデート以降も完全には解消せず、一部の語における局所的かつ少数の訳語候補間での変化は引き続き発生し続けたことの双方が、本例にも等しく当てはまることが見て取れる。

まず、「技術／一般用語や文構造の不安定な変化が抑制され、学習データに準拠した有力な訳語候補が安定して採用されるようになったこと」を端的に示すのが、第 5 回と第 7 回学習後に突発的に発生していた **the invention makes the invention unclear** という主語の補足誤りが、大規模アップデートを経た第 11 回学習後以降は一度も発生しなくなったという事実であろう。訳語学習のロジックの改良によって学習データの内容が機械翻訳文により的確に反映されるようになり、その結果、不適訳語が突発的に採用される確率も低減されたと推察される。

技術用語「許容基準」に関しても、第 11 回学習後以降は allowable criteria（第 11 回、12 回、16 回、18 回）と allowable standard（第 13 回～第 15 回、第 17 回）のいずれかが採用されており、訳語候補がアップデート前の 4 種から 2 種に絞られたことがわかる。これも訳語学習ロジックの改良により、訳語が少数の有力な候補からのみ採用されるようになった結果と考えられる。

ただし、本語に関しては、訳語候補は 2 種に絞られたとはいえ、その 2 種のどちらかに固定されることはなく、両者は最終学習時まで不安定な変化を続けた。また、本例の文構造も、第 12 回、第 15 回～第 17 回学習後の翻訳文で「Claim 3 recites」で始まる文型が頻繁に採用されるなど一定の安定化は見られるものの、学習前後で変化する回も依然として少なくはなく、「訳語や文構造のランダム的な変化はニューラル翻訳方式の本質的な性質であってアップデート以降も完全には解消せず、一部の語における局所的かつ少数の訳語候補間での変化は引き続き発生し続けた」という傾向にも本例は合致する。

なお、本例は「拒絶理由通知書」由来の実例である。先述のとおり、拒絶理由通知書は全期間とも他の 3 種の書類の 7 倍の量のデータが学習されているが、本例の機械翻訳文の品質やその推移を見る限り、前三例として取り上げた他書類由来の実例との特段の差異は認められなかった。この結果は第 3.2.3.項に示した書類別の自動評価スコア集計結果とも整合しており、そこで述べた、審査書類用語や言い回しの多くは各書類間で共通しているため、各書類由来の学習データ間には高い互換性があるという結論を肯定している。

以下、本例から得られた知見により、これまでの考察内容を補足するとともに、新たな考察を追加する。

- ・ 本例で特許・審査書類用語表現といえるのは文末の「発明を不明確にしている」という審査書類特有のフレーズのみである。ただし、このフレーズを構成する「発明」「不明確」とも、特段の学習を要さずとも直訳すればよいレベルの一般語であり、事実、無学習時の「みんなの自動翻訳」でも the invention is unclear と問題なく訳せている。

本フレーズは第 5 回、第 7 回学習後のみ突発的に the invention makes the invention unclear と誤訳されたが、これは特許・審査書類用語の常時安定という傾向に反する不備というよりは、実例 3-1 における they の誤訳と同種の、日英機械翻訳における主語無し原文への主語の補足の難易度の高さによる不備と見なされる。日本語文では前後の文脈から明らかな場合は主語を省略すること多いが、英文では原則として主語の省略は許容されない。このため、英訳に際しては主語を補う必要がある。人間であれば文脈から適切な主語を特定できるが、機械翻訳の場合、処理単位が単文であるため、前後の文脈を参考にすることはできない。このことにより、主語の補足は機械翻訳にとって非常に難易度の高い課題となっている。本例を見る限り、補足する主語は具体的な用語での補足は避け、できるだけ形式主語 it を使用するほうが無難と映る。

- ・ 技術／一般用語の訳語の不安定な変化という傾向は本例においても「許容基準」の訳語において顕著であった。また、この傾向が大規模アップデート以降はある程度抑制され、訳語変化が少数かつ同義の候補間の変化に限定された点も、他の各実例と同様であった。
- ・ 自動評価スコアに関しては、本例でも個々の文レベルのスコア変動は実質的な品質の動向とかい離しており、学習前後の相対評価、すなわちスコアの増減方向のみに限定しても精度の低い回が散見された。

3.2.7.5. 実例を用いた定点観測のまとめ

ここまで 4 件の実例について、各学習段階での機械翻訳文の内容を詳細に観測し、自動評価の全体平均スコア等から看取した品質改善状況の実態について確認・検証してきた。その結果、以下のことが確認された。

- ・ デフォルト状態の「みんなの自動翻訳」では、特許・審査書類用語にはほとんど対応できておらず、大半の実例で誤訳や欠落が発生していた。これに対し、審査書類対訳コーパスの学習後は、各実例とも初回学習後から全ての特許・審査書類用語が正訳文どおりの適訳語で安定して訳されるようになり、誤訳は一切発生しなくなった。

こうした初回学習時の特許・審査書類用語の訳質改善により各実例の翻訳品質は大幅に改善され、その後も特許・審査書類用語に関しては常時高い翻訳品質が維持された。この一連の推移は、自動評価の全体平均スコアが示す品質改善の状況と一致しており、審査書類①システム全体において、各実例と同等の特許・審査書類用語の訳質改善が果たされたと考えられる。

- ・ 一方、技術／一般用語や文構造に関しては、学習データ（デフォルト含む）中に複数の訳語候補が存在するものが多く、こうした状況とニューラル機械翻訳の訳語学習メカニズムとが相俟って、各実例とも学習の都度不安定な訳語変化を見せる語が多数見られた。第 3.2.6.項に示したカウント結果では、第 2 回学習後以降も大多数の機械翻訳文が学習前後で変化を続けていたが、特許・審査書類用語の訳語が学習初期から非常に安定していたことから考えて、各文における変化は、本項の各実例と同様、その大半が技術／一般用語や文構造の変化であるものと考えられる。

- ・ 観測した各実例においては、技術／一般用語や文構造の学習前後での変化の多くは、実質的な翻訳品質には影響しない同義の範囲内の変化に収まっていた。ただし、不適訳語が突発的に採用されて翻訳品質を低下させるケースも皆無ではなかった。

こうした技術／一般用語や文構造の不安定な変化は大規模アップデート以降はかなり抑制され、各実例とも、学習データ中の最有力の訳語候補や文構造候補がより安定して採用されるようになった。この傾向の変化は、「みんなの自動翻訳」自身の訳語学習ロジックが見直され、学習データ全体の状況がより的確に各訳語の優先順位に反映されるようになったためと推察される。その結果、各実例とも訳語変化がアップデート以前よりも小規模、局所的になり、突発的な不適訳語の採用も明らかに減少した。

- ・ 大規模アップデート以降、不適訳語の突発的な採用による品質低下のリスクが低減したことで各実例の翻訳品質は常時「同等」で安定し、アップデート以前のように学習後に「悪化」するケースがほぼ見られなくなった。自動評価の全体平均スコアも、アップデート以

前は現状維持状態が続いていたが、大規模アップデート直後に再び向上し、以降も向上した水準が常時保たれた。

こうした全体平均スコアの推移を各実例の状況と照合すると、大規模アップデート時の訳語学習ロジックの改良により、評価用コーパス全体で最有力の訳語候補が常時安定して採用される傾向となり、それまで一定数発生していた突発的な不適訳語の採用による品質低下が根本的に抑制されるようになった結果、評価用コーパス全体の翻訳品質が底上げされたという状況が見て取れる。

第 3.2.6.項に示した文単位の自動評価スコアの向上／低下カウント結果でも、大規模アップデート直後の第 11 回学習後は「不変」文の数は変化せず、「向上」と「低下」のバランスが前者に大きく偏っていた。これは、多数の文で採用される訳語が最有力のものになった結果、実際に品質が「改善」する文が多かったことを暗示している。そして、第 12 回学習後以降は「不変」文が顕著に増加しており、これは、アップデート以降、最有力の訳語が安定して採用されるようになった結果、学習前後で同じ翻訳文となるケースが増加したためと解釈できる。

- ・ このように、自動評価の全体平均スコアが示した審査書類①の翻訳品質の改善状況と、本項で観測した各実例の実際の翻訳品質の推移とは、各学習段階において矛盾なく整合した。このことから、自動評価 BLEU 及び RIBES は、一定規模の文集合の学習前後の相対的な翻訳品質を測る指標としては十分な信頼性があると認められる。

ただし、その一方で、各実例に付された自動評価スコアは、BLEU、RIBES とも精度が低く、実際の翻訳品質の変動とかい離するケースが多かった。この結果から、自動評価指標は個々の文レベルでの評価精度には限界があり、あくまで一定規模の平均スコアを取得して全体的な品質の変動を測る用途に用いるべきものであることが判明した。

これらの考察内容は、次項に示す人手評価結果とも照合して更なる検証を行う。

3.3. 審査書類①システムの人手評価結果

本項では、審査書類①システムに対して実施した人手評価の結果を記す。

3.3.1. 人手評価の概要

審査書類①システムの人手評価は、審査書類①に属する4種の書類（拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書）の評価用コーパスからそれぞれ100文、計400文を対象に実施した。また、自動評価は審査書類対訳コーパスを約10万文対学習させる都度、全19回（無学習時の1回を含む）実施したのに対し、人手評価は約50万文対の学習ごと、つまり第5回学習後、第10回学習後、第15回学習後の3回に、無学習時と最終学習時（第18回学習後）を加えた計5回のみ実施した。

人手評価は1文あたりの作業負荷が大きいため大量の文を評価するには適さないが、正解訳文との表層的な一致度を根拠とする自動評価に比べ、機械翻訳文の品質をより正確かつ精密に評価することができる。本調査分析ではこの利点を活かし、自動評価結果から看取した審査書類①における機械翻訳品質の改善状況や各種傾向について、人手評価結果と照合することで、その妥当性を検証した。

3.3.2. 人手評価観点

人手評価では、評価対象各文について下記3つの観点から評価を実施した。

- ① 内容伝達レベルの評価
- ② 学習前後の機械翻訳文の相対評価
- ③ エラーのカテゴリ別カウント

各観点における評価の基準及び手法の詳細は、それぞれの評価結果の項で説明する。

3.3.3. 人手評価者

人手評価者には、本事業の翻訳内容校閲者を起用した。文ごとに担当評価者をあらかじめ定め、原則全ての回の評価を同じ者に担当させた。

3.3.4. 人手評価の対象文

人手評価の対象文は、審査書類①に属する 4 種の審査書類の評価用コーパス各 2,000 文から 100 文ずつ、計 400 文を選定した。

これら 4 書類の評価用コーパスでは、それぞれ 2,000 文のうちの 800 文について「認定」「対比」「判断」「その他」という文種のラベリングを行っている（⇒3.2.5.）。人手評価対象文の選定にあたっては、各コーパスとも 4 文種から均等に 25 文ずつを選定して計 100 文とした。

3.3.5. 各評価項目の集計結果

全 5 回にわたり実施した人手評価の集計結果を本項にまとめる。第 3.3.2.項に示した三つの評価観点（①内容伝達レベルの評価、②学習前後の機械翻訳文の相対評価、③エラーのカテゴリ別カウント）ごとに各回の集計結果を示し、分析・考察する。

3.3.5.1. 内容伝達レベルの評価

機械翻訳文の翻訳品質の絶対評価として、評価対象文 400 文それぞれに対し、特許庁が提案する「特許文献機械翻訳の品質評価手順（ver.1.0）」の主要な評価基準である「内容伝達レベル」の評価を行った。

内容伝達レベルの評価では、原文の実質的内容が機械翻訳文からどの程度把握できるかを、以下の 5 段階の評価基準で主観的に判定した。

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

本調査分析では、4 種の審査書類それぞれの人手評価対象文 100 文、全 400 文について、各評価回の機械翻訳文の内容伝達レベルを上記評価基準により評価した。下表に、審査書類別に集計した各回の内容伝達レベルの全件平均スコアと、100 文におけるスコア別分布を示す。

表 3.3.1. 各回学習前後の審査書類①各書類の「内容伝達レベル」集計結果

書類種別	学習	平均	5	4	3	2	1
拒絶査定	無学習時	3.08	17	16	34	24	9
	第 5 回学習後	4.13	54	19	16	8	3
	第 10 回学習後	4.16	53	22	16	6	3
	第 15 回学習後	4.61	72	17	11	0	0
	第 18 回学習後	4.69	77	15	8	0	0
	向上度／増減	+1.61	+60	-1	-26	-24	-9
拒絶理由通知書	無学習時	3.44	22	27	28	19	4
	第 5 回学習後	4.28	53	30	10	6	1
	第 10 回学習後	4.45	64	22	10	3	1
	第 15 回学習後	4.85	88	9	3	0	0
	第 18 回学習後	4.86	88	10	2	0	0
	向上度／増減	+1.42	+66	-17	-26	-19	-4
補正却下の決定	無学習時	3.20	9	29	38	21	3
	第 5 回学習後	4.09	43	32	18	5	2
	第 10 回学習後	4.18	49	26	20	4	1
	第 15 回学習後	4.63	77	12	8	3	0
	第 18 回学習後	4.62	75	15	7	3	0
	向上度／増減	+1.42	+66	-14	-31	-18	-3
前置報告書	無学習時	2.84	8	25	24	29	14
	第 5 回学習後	3.85	41	23	19	14	3
	第 10 回学習後	4.01	47	23	15	14	1
	第 15 回学習後	4.58	72	18	6	4	0
	第 18 回学習後	4.59	71	21	4	4	0
	向上度／増減	+1.75	+63	-4	-20	-25	-14
審査書類① 合計	無学習時	3.14	56	97	124	93	30
	第 5 回学習後	4.09	191	104	63	33	9
	第 10 回学習後	4.20	213	93	61	27	6
	第 15 回学習後	4.67	309	56	28	7	0
	第 18 回学習後	4.69	311	61	21	7	0
	向上度／増減	+1.55	+255	-36	-103	-86	-30

審査書類①全体の内容伝達レベルは、最下段「審査書類①合計」欄に示したとおり、無学習時の 3.14 から最終学習後（第 18 回）には 4.69 へと+1.55 向上した。各学習段階におけ

る伸び幅は、第 5 回学習後に+0.95、第 10 回学習後に+0.11、第 15 回学習後に+0.47、第 18 回学習後に+0.02 であり、無学習時から第 5 回学習後にかけて最も顕著に向上した。第 10 回～第 15 回学習後の期間がそれに次ぎ、第 5 回～第 10 回学習後の期間と第 15 回～第 18 回学習後の期間は微増にとどまった。4 種の審査書類ごとの内容伝達レベルの水準及び推移も、これとほぼ合致している。

4 書類間の相対順位は、無学習の時点では「拒絶理由通知書」>「補正却下の決定」>「拒絶査定」>「前置報告書」の順であったが、その後も最終学習後に「補正却下の決定」と「拒絶査定」が僅差で入れ替わった程度で、ほぼ全期間にわたり同じ順位が維持された。ただし、書類間のレベル差を見ると、無学習時は最上位の拒絶理由通知書と最下位の前置報告書との差が 0.60 と開いていたのに対し、最終学習後には 0.27 と差が縮まっており、コーパスの学習によって各書類間の均質化が進んでいることがわかる。

第 3.2.3.項に示した自動評価では、4 書類のうち「補正却下の決定」のスコアが全期間を通じて有意に低く、このため翻訳難易度が他の書類より高い可能性があることを指摘していた。だが、人手による内容伝達レベルの評価では「補正却下の決定」は無学習時からほぼ全期間にわたり第 2 位とむしろ上位（最終学習後のみ第 3 位）であり、他の 3 書類に比べて特に品質が低いことを示す結果とはならなかった。集計対象が自動評価の 2,000 文に対し人手評価は 100 文のみと異なるため両者の評価結果が常に整合するとは限らないが、より精度の高い人手評価の結果からは、「補正却下の決定」の翻訳品質が他の 3 書類に劣るといふ事実は認められない。

審査書類①全体、全 400 文のレベル別の分布を見ると、無学習時はレベル 3 をピークにレベル 5 から 1 まで分布が広がっていたのに対し、最終学習後は全体の約 78%に相当する 311 文がレベル 5 に集中してピークを形成しており、レベル 3 以下の文は全体の 7%（28 文。レベル 3：21 文、レベル 2：7 文）とごく少数となった。

内容伝達レベルの集計結果は、自動評価の対象とした評価用コーパスの一部に対し、人手で精密かつ具体的な評価を行った結果である。これを自動評価指標 BLEU 及び RIBES の全体平均スコアから看取された審査書類①の機械翻訳品質の動向(⇒3.2)と照合することで、自動評価結果の妥当性を検証できる。また、自動評価では困難であった機械翻訳品質の絶対評価という側面も有しており、審査書類対訳コーパスの学習により達成された審査書類①の翻訳品質を具体的に把握できる。

上掲の内容伝達レベルの平均値の推移を見ると、各書類とも無学習時の平均値はレベル 3 前後であったのに対し、学習初期にあたる第 5 回学習後にはレベル 4 前後へと+1.00 程度

レベルが向上している。表中「増減」欄に示したとおり、各書類とも全期間を通じて+1.50ほどレベルが向上したが、そのうちの3分の2相当が第5回までの初期学習で達成されていたことになる。この状況は、BLEUとRIBESそれぞれの初回学習後の全体平均スコアの向上度が全期間の75%、66%を占めた自動評価結果と一致している。

続く第5回学習後と第10回学習後の比較では、内容伝達レベルは審査書類①全体で+0.12と微増にとどまった。この期間は自動評価スコアも停滞期にあたり、両評価の動向はおおむね整合している。ただし、自動評価では第5回後から第10回後にかけてBLEUが-0.0008、RIBESが-0.0003とスコアが僅かに低下していたのに対し、内容伝達レベルは微増ながら各書類とも向上している。自動評価スコアからは、この期間は学習効果が完全に飽和しているようにも映ったが、より精密な評価である内容伝達レベルの結果から、この期間も一定の品質改善があったことが示された。

第10回学習後と第15回学習後の比較では、第11回学習前に実施された「みんなの自動翻訳」の大規模アップデート(⇒2.2.2.)の効果もあり、内容伝達レベルは全体平均で+0.46と、学習初期に次ぐ顕著な向上を果たした。その結果、審査書類①全体の内容伝達レベルは5点満点で平均4.67と非常に高い水準に達し、続く最終学習後(第18回学習後)の評価においても前回比+0.02の4.69と、この高い水準がそのまま維持された。こうした学習後期における一連の推移も、自動評価の全件平均スコアが示した動向と完全に一致している。

このように、人手評価対象文400文における内容伝達レベルの推移と、自動評価スコアから看取された評価用コーパス全件(8,000文)の機械翻訳品質の推移とは、全期間を通じてほぼ完全に一致している。したがって、審査書類①の自動評価結果は、実際の機械翻訳品質の全体傾向を正確に表していると認められ、この自動評価結果に基づいた前項までの分析・考察の内容も十分な妥当性を有すると認められる。

なお、自動評価結果から把握できるのは学習前後の機械翻訳の優劣、つまり相対的な評価のみであって、BLEUやRIBESのスコアから、機械翻訳結果の品質が具体的にどの程度であるかを把握することは難しかった。その点、内容伝達レベルは翻訳品質の絶対評価であり、自動評価の不足を補える。

審査書類①の無学習時の内容伝達レベルは全書類平均で3.15であった。つまり、審査書類対訳コーパス学習前のデフォルト状態の「みんなの自動翻訳」で審査書類を機械翻訳した場合、得られる翻訳文の品質は内容伝達レベル3相当の「半分以上の重要情報は正確に伝達されている(50%～)」程度であると見なせる。これに対し、コーパスを全件学習させた最終学習後の審査書類①の内容伝達レベルは4.69まで向上した。内容伝達レベルの定義に

照らせば「レベル 4：ほとんどの重要情報は正確に伝達されている（80%～）」と「レベル 5：すべての重要情報が正確に伝達されている（100%）」の中間に位置し、機械翻訳としては極めて高精度といえる。

審査書類①全 400 文のレベル別の分布状況を見ても、無学習時はレベル 3 をピークにレベル 5 から 1 まで分布が広がり、文ごとの品質のばらつきが大きかったのに対し、最終学習後は全体の約 78%の文がレベル 5 に集中し、レベル 3 以下の文は 7%とごく少数にとどまるなど、個々の文単位の品質も非常に高いレベルで均質化されている。審査書類対訳コーパスの学習効果が各書類の各文に均等かつ十分に及んでいることを示す結果である。

これらは人手評価対象 400 文のみにおける状況であるが、前記したとおり、これら 400 文における内容伝達レベルの推移は、評価用コーパス全件の自動評価スコアの推移と完全に一致している。このため、機械翻訳品質の絶対評価としても、これら 400 文の内容伝達レベルの評価結果は、評価用コーパス全体の状況を端的に表したものと見なせる。さらに敷衍すれば、本評価における審査書類①の最終学習後の内容伝達レベルが、本事業で作成した審査書類対訳コーパス及び「みんなの自動翻訳」を利用して審査書類用の日英機械翻訳システムを実用化した際に期待できる翻訳品質であるといえる。

3.3.5.2. 学習前後の機械翻訳文の相対評価

各審査書類の人手評価対象文 100 文について、学習前後の機械翻訳文のどちらが優れているかを文ごとに比較し、学習後の機械翻訳品質が学習前に比べて「向上」したか「劣化」したか、又は「同等」であるかを判定した。下表に各書類・各回の集計結果を示す。

表 3.3.2. 各回学習前後の翻訳品質の「相対評価」結果

書類種別	比較対象		学習後に		
	学習前	学習後	向上	同等	劣化
拒絶査定	無学習時	第 5 回学習後	65 文	28 文	7 文
	第 5 回学習後	第 10 回学習後	16 文	71 文	13 文
	第 10 回学習後	第 15 回学習後	36 文	63 文	1 文
	第 15 回学習後	第 18 回学習後	11 文	86 文	3 文
拒絶理由通知書	無学習時	第 5 回学習後	70 文	18 文	12 文
	第 5 回学習後	第 10 回学習後	26 文	56 文	18 文
	第 10 回学習後	第 15 回学習後	36 文	60 文	4 文
	第 15 回学習後	第 18 回学習後	6 文	89 文	5 文
補正却下の決定	無学習時	第 5 回学習後	82 文	6 文	12 文
	第 5 回学習後	第 10 回学習後	34 文	42 文	24 文
	第 10 回学習後	第 15 回学習後	41 文	56 文	3 文
	第 15 回学習後	第 18 回学習後	6 文	89 文	5 文
前置報告書	無学習時	第 5 回学習後	61 文	32 文	7 文
	第 5 回学習後	第 10 回学習後	22 文	67 文	11 文
	第 10 回学習後	第 15 回学習後	37 文	62 文	1 文
	第 15 回学習後	第 18 回学習後	4 文	92 文	4 文
審査書類① 合計	無学習時	第 5 回学習後	278 文	84 文	38 文
	第 5 回学習後	第 10 回学習後	98 文	236 文	66 文
	第 10 回学習後	第 15 回学習後	150 文	241 文	9 文
	第 15 回学習後	第 18 回学習後	27 文	356 文	17 文

最下段の審査書類①全体の集計結果を見ると、まず無学習時と第 5 回学習後の相対評価では、400 文中の約 70%にあたる 278 文において第 5 回学習後に翻訳品質が「向上」したと判定された。

続く第 5 回学習後と第 10 回学習後の相対評価では、品質が「同等」の文が 236 文と過半数を占め、残り 164 文を「向上」と「劣化」が 3 : 2 の割合（98 文 : 66 文）で分け合った。

第 10 回学習後と第 15 回学習後の比較では、「同等」は前回とほぼ同じ 241 文であったが、それ以外はほぼ全て「向上」文が占め（150 文）、「低下」文は 9 文と僅かであった。さらに、第 15 回学習後と最終第 18 回学習後の比較では「同等」が 356 文と全体の 89%を占め、残る 44 文は「向上」と「劣化」が再び 3：2 の割合（27 文：17 文）で分け合った。4 種の書類ごとの集計結果も、各回これと同傾向であった。

文単位の相対評価の集計値は、自動評価スコアや内容伝達レベルの全体平均が示した審査書類①全体の翻訳品質の動向が、個々の文単位でどの程度共通の傾向であるのか、その内訳を示すものである。自動評価でもスコアの向上／低下を文単位で集計してはいるが（⇒ 3.2.6.）、その後の各実例の定点観測で明らかになったように、自動評価スコアは個々の文レベルでの精度は低く、このため学習前後での機械翻訳結果の変化の有無を把握する以上の分析は難しかった。その点、本評価は機械翻訳文の実質的な翻訳品質を目視で正確に判定したうえでの相対評価であり、各学習段階における文単位の品質改善状況をより正確に把握できる。

この観点から改めて各回の評価結果を見ていくと、まず無学習時と第 5 回学習後の相対評価は、実質的には、審査書類対訳コーパスの初期学習前後の評価という位置づけになる。コーパス初回学習後に顕著な品質改善があったことは自動評価や内容伝達レベルの評価によって既に示されているが、本評価で全体の約 70%にあたる 278 文の翻訳品質が学習後に「向上」と判定された³⁹ことにより、この時期の品質改善が一部特定の文のみでなく、評価用コーパスの大半の文で広範囲に果たされていたことが改めて確認できた。実例の定点観測では、各実例とも学習初期に特許・審査書類用語の大部分が適訳語に改善されており、この時期の品質改善の最大の要因となっていたが、本評価結果から、同様の改善が他の文でも広く生じていたことは確実といえる。

続く第 5 回学習後と第 10 回学習後の相対評価は、自動評価スコア、内容伝達レベルとも誤差レベルの微動にとどまっていたことから、現状維持期間内での比較となる。評価結果は前掲のとおり、前回時とは傾向が一変し、審査書類①全体の半数強の 236 文を「同等」文が占め、残りを「向上」と「劣化」が 3：2 の比率で分け合った。この結果により、この期間の現状維持状態の理由が、学習前後で品質に変化のない文が全体の過半数を占め、かつ、品質が変化した文では向上文と劣化文がほぼ均衡していたためであったことが明らかとなった。この時期は各学習回で全体の約 95%の機械翻訳文が変化していたが、このうちの半数

³⁹ 人手評価対象 400 文には、無学習の時点で既に内容伝達レベルが 5、つまり事実上改善の余地がない文も 56 文（約 14%）存在しており（⇒ 3.3.5.1.）、これらを対象外とすれば「向上」文の実質的な比率はさらに高くなる。

程度は、品質上優劣のない同義の範囲の変化であったということが本評価結果により示された。

この結果は、実例の定点観測から看取されたこの期間の状況と整合する。定点観測では、この期間、各実例とも特許・審査書類用語は既に適訳語で安定していたが、技術／一般用語や文構造は学習の都度不安定に変化していた。これにより、機械翻訳文は毎回変化した。ただし、こうした技術／一般用語の変化の多くは同義の範囲内にとどまっており、不適訳語が突発的に採用されて一時的な品質劣化を招く場合もあるものの、品質上の優劣は生じていないことも多かった。本カウント結果は、こうした状況が、この時期の審査書類①全体の状況であることを示唆している。

第 10 回学習後と第 15 回学習後との相対評価は、第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデート前後の比較となる。審査書類①全体では、「同等」文は 241 文で前回（236 文）と同程度であったが、「向上」と「劣化」のバランスは前者が 150 文、後者は 9 文のみと「向上」に大きく偏った。この期間、自動評価スコア、内容伝達レベルとも、大規模アップデート直後に顕著な品質改善があったことを示していたが、本評価結果からは、この品質改善が評価用コーパス全体に一律に発生したものではなく、全体の半数弱の文でそれまで拮抗して発生していた品質の「向上」と「劣化」のうち「劣化」がほとんど発生しなくなる一方、「向上」の発生数が倍増したことで、全体品質が底上げされた結果であるとわかる。

前項までの分析では、大規模アップデート以後の品質改善の主な理由として、訳語学習ロジックの改良により学習データを正確に反映した適訳語が採用される確率が高まったことを挙げていた。本評価結果も、アップデート前の第 10 回学習後においては、一定数発生する突発的な不適訳語の採用とその改善によって多数の「劣化」文とこれと拮抗する「向上」文が発生していたが、大規模アップデートを経た第 15 回学習後は、ロジックの改良によって突発的な不適訳語の採用が抑制されたため「劣化」がほとんど発生しなくなり、その一方で前回時に採用された不適訳語の適訳語への改善（すなわち「向上」）はこれまで以上に進んだ、という状況を表しており、上記分析内容と整合している。

そして、続く第 15 回学習後と最終第 18 回学習後の比較では、前回と「同等」となった文が 356 文と大勢を占め、「向上」文（27 文）、「劣化」文（17 文）ともごく少数となった。この期間は、第 5 回学習後と第 10 回学習後の相対評価時と同じく自動評価スコア、内容伝達レベルとも誤差レベルの微動にとどまっており、再び現状維持期間内での比較となるが、当時に比べて「同等」の文数をはるかに多くなり、「向上」文と「劣化」文は、バランスこそ 3：2 で同じだが、該当文数は大きく減少した。つまり、当時と今回とでは、全体的に現

状維持傾向という点では同じだが、今回は当時と比べて個々の文レベルでも品質が同等という文が大幅に増加したことになる。第 15 回後と第 18 回後はどちらも大規模アップデート後の機械翻訳文であるため、各文において最有力の訳語候補が毎回安定して採用されるようになった結果、個々の文レベルでも大半の文が「同等」となったものと考えられる。つまり、この回の評価結果も、先述の分析内容を裏付けている。

なお、第 3.2.6.項に示した自動評価スコアの文単位のカウント結果では、大規模アップデート以降も最終学習後に至るまで全体の約 66～80%の機械翻訳文が学習の都度変化しており、そのうちの半数で評価スコアが低下していることが判明した。このことから、学習後期においても学習の都度多数の文で翻訳品質が劣化している懸念が生じていた。その後の実例の定点観測結果により、自動評価スコアのこうした変動の大半は、少なくとも学習後期においては同義の範囲内の表層的な変化が招いたものであり、実際の翻訳品質は個々の文レベルでも常時ほぼ同等に保たれている蓋然性が高いと推察した。本評価において、第 15 回後と第 18 回後の比較で全体の 89%の文が学習前後で実質的に同等であり、品質が劣化した文は 5%に満たないことが示されたことで、この推察が妥当であることが実証された。これにより、学習データの追加の都度その副作用で多数の文で翻訳品質が劣化するという懸念は完全に払拭された。

3.3.5.3. エラーのカテゴリ別カウント

人手評価の一環として、機械翻訳の特徴的なエラーを 10 種類カテゴライズし、各文における発生の有無⁴⁰を個別にカウントする「エラーのカテゴリ別カウント」を実施した。本項では各種エラーの各回でのカウント結果を示すとともに、具体的なエラーの発生状況や、学習前後での改善の有無等について考察する。

3.3.5.3.1. 「特許誤訳」のカウント結果

特許・審査書類用語や表現に関する誤訳、不適訳は「特許誤訳」としてカウントした。各回学習前後のカウント結果は下表のとおりである。

表 3.3.3. 各回学習後の「特許誤訳」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	71	16	12	5	4
	—	-55	-4	-7	-1
拒絶理由通知書	63	11	4	4	2
	—	-52	-7	±0	-2
補正却下の決定	68	14	5	2	2
	—	-54	-9	-3	±0
前置報告書	85	13	13	5	3
	—	-72	±0	-8	-2
審査書類①	287	54	34	16	11
合計	—	-233	-20	-18	-5

自動評価結果における実例の定点観測（⇒3.2.7.）では、各実例とも特許・審査書類用語が学習初期から正解訳文どおりの定番訳で安定する傾向が顕著であった。上記カウント結果により、この傾向が全体的なものであったことが明確に示された。

まず、無学習時においては、各書類とも過半数の文（100 文中 63～85 文）で何らかの「特許誤訳」が発生していた。デフォルト状態の「みんなの自動翻訳」が特許・審査書類用語に十分に対応できていなかったことが分かる。

これに対し、審査書類対訳コーパスを学習させた第 5 回学習後には、「特許誤訳」は各書類とも 11～16 文と大幅に減少した。つまりコーパス初期学習の結果、全体の 8 割強の文に

⁴⁰ 同一カテゴリのエラーを一文中で都度カウントする方式ではないため、カウントは各書類とも最大で 100（文）、審査書類①合計で最大 400（文）となる。

において、文中の全ての特許・審査関連用語が適訳語で訳されるようになったことになる。この結果は、第 3.3.5.2.項で述べた考察内容、すなわち学習初期において各文の特許・審査書類用語が一斉に適訳語に改善され、この時期の顕著な品質改善の最大の要因となっていたことを裏付けている。その後も学習の進行に伴い「特許誤訳」は順調に減少し、最終学習後には各書類とも「特許誤訳」の発生率は全体の 5%未満とごく少数となった。

機械翻訳システムを審査書類用語・表現に対応させるという本事業の目的に照らせば、全種のエラーの中でも「特許誤訳」の解消こそが審査書類対訳コーパス学習の主眼であるといえる。本項に示したカウント結果の推移は、本事業で作成したコーパスがその役割を十分に果たしたことを示している。

以下、「特許誤訳」の典型的な改善事例として、前置報告書の評価用コーパスから、文番号 A913-#13 の各回評価時の機械翻訳結果の推移を示す。まずは原文と正解訳文を以下に示す。

A913-#13

原文	請求項 1 についての補正は、実質的に平成 23 年 2 月 14 日付け手続補正書の請求項 1 に係る発明に平成 23 年 2 月 14 日付け手続補正書の請求項 2 の発明特定事項を追加したものであり、限定的減縮を目的としている。(なお、請求項 1 を引用する請求項 2-6 については、平成 23 年 2 月 14 日付け手続補正書の請求項 3-7 に記載されたとおりの内容である。)
正解訳文	The amendment in respect of Claim 1 is substantially the addition of the matter specifying the invention of Claim 2 of the written amendment dated February 14, 2011, to the invention according to Claim 1 of the written amendment dated February 14, 2011, and intends for restriction in a limited way. (Additionally, Claims 2 to 6 depending on Claim 1 is the contents as described in Claims 3 to 7 of the written amendment dated February 14, 2011.)

本例の原文は文頭から「請求項 n」「補正」「手続補正書」「発明特定事項」「限定的減縮」と 5 種の特許・審査書類用語を含んでいる。審査書類対訳コーパスでは、各語についてそれぞれ正解訳文のとおり「Claim n」「amendment」「written amendment」「matter specifying the invention」「restriction in a limited way」が統一訳語として用いられている。これを踏まえて、無学習時の機械翻訳文を見る。

機械翻訳結果 【無学習時】 内容伝達レベル：1 特許誤訳：有	The correction of claim 1 is obtained by adding the inventive specification of the procedure correction program of Feb. 14, 2011 to the invention according to claim 1 of the process correction program of February 14, 2011, and is intended for limiting reduction , and it is intended for the purpose of limiting the restriction (note that, for example, as described in claim 3-7 of the process correction program of Feb. 14, 2011 for claim 2-6 which refers to claim 1). 2
	(BLEU : 0.2031、RIBES : 0.7114)

無学習時の機械翻訳文では、5 語の特許・審査書類用語のうち適訳語で訳されているのは「請求項：claim」のみであり、残る 4 語は「補正」が **correction**、「手続補正書」が **procedure correction program**、「発明特定事項」が **inventive specification**、「限定的減縮」が **limiting reduction** と誤訳されている。これら 4 語は全て「特許誤訳」に該当する。特に専門性が高い後半 3 語に関しては、本来の語義とかい離れた訳語になっており、デフォルト状態の「みんなの自動翻訳」ではこうした専門的な審査書類用語を正しく訳出することは難しいとわかる。

文の要所となる審査書類用語が多数誤訳されたことで、無学習時の機械翻訳文はほぼ意味不明である。加えて、文末のカッコ補記も、それ自体が審査書類特有表現といえるものだが、無学習時の翻訳文は文章として成立しておらず意味をなさない。結果、内容伝達レベルは最低レベルのレベル 1 が付されている。

これに対し、審査書類対訳コーパス学習後（第 5 回学習後）の機械翻訳文は以下のように変化した。

機械翻訳結果 【第 5 回学習後】 内容伝達レベル：4 特許誤訳：有	The amendment to Claim 1 adds the matters specifying the invention of Claim 2 of the written amendment dated February 14, 2011 to the invention according to Claim 1 of the written amendment dated February 14, 2011, and is made for the purpose of restriction in a limited way . (In addition, regarding Claims 2-6 to dependent on Claim 1, it is a content as described in Claims 3-7 to of the written amendment dated February 14, 2011.)
	(BLEU : 0.5730、RIBES : 0.8888)

上掲のとおり、第 5 回学習後の機械翻訳文では 5 語の特許・審査書類用語全てが正解訳

文どおりの適訳語⁴¹に改善された。

これら審査書類用語の適訳語への改善により、第 5 回学習後の機械翻訳文は原文の内容をほぼ正確に伝達するものとなった。唯一の不備はカッコ補記部分において「請求項 2 - 6」が **Claims 2-6 to**、続く「請求項 3 - 7」も **Claims 3-7 to** と、それぞれに不要な **to** が付随されている点であり、評価者はこれらを「特許誤訳」と判定した。この不備により内容伝達レベルは 4 にとどまり、また「特許誤訳」もカウントされたままとなったが、無学習時に比べて特許・審査書類用語の翻訳品質が大幅に改善されたことは明らかである。

なお、上掲したのは第 5 回学習後の機械翻訳文であるが、人手評価対象外である第 1 回～第 4 回学習後の機械翻訳文でも、上記 5 語のうち 4 語は適訳語が採用されていた。唯一「**発明特定事項**」が初回学習後のみ **invention specific matter** と不適訳であったが、この語も第 2 回学習後以降は定番訳 **matter(s) specifying the invention** に改善されている。したがって、本例も、第 3.2.7.項で定点観測した各実例から看取された、特許・審査書類用語は学習初期から定訳語で安定するという傾向に等しく当てはまる。

続いて第 10 回、第 15 回、第 18 回学習後の機械翻訳文を示す。

機械翻訳結果 【第 10 回学習後】 内容伝達レベル：5 特許誤訳：無	The amendment to Claim 1 is to add the matter specifying the invention of Claim 2 in the written amendment dated February 14, 2011 to the invention according to Claim 1 of the written amendment dated February 14, 2011, and is made for the purpose of restriction in a limited way . (In addition, Claims 2-6 dependent on Claim 1 are the contents as described in Claims 3-7 of the written amendment dated February 14, 2011.)
	(BLEU : 0.6295、RIBES : 0.8920)
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：5 特許誤訳：無	The amendment to Claim 1 substantially adds the matter specifying the invention of Claim 2 of the written amendment dated February 14, 2011 to the invention according to Claim 1 of the written amendment dated February 14, 2011, and is made for the purpose of restriction in a limited way . (In addition, Claims 2-6 dependent on Claim 1 are the contents as recited in Claims 3-7 of the written

⁴¹ 「**発明特定事項**」は正解訳文では **matter specifying the invention** と単数形であるのに対し第 5 回時は **matters specifying the invention** と複数形であるが、どちらが適切かはケースバイケースで優劣はないため、同一訳語として扱った。以下同じ。

機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 特許誤訳：無	amendment dated February 14, 2011.)
	(BLEU : 0.6211、RIBES : 0.9217)
	The amendment to Claim 1 substantially adds the matters specifying the invention of Claim 2 of the written amendment dated February 14, 2011 to the invention according to Claim 1 of the written amendment dated February 14, 2011, and is made for the purpose of restriction in a limited way. (In addition, Claims 2-6 dependent on Claim 1 have the contents as recited in Claims 3-7 of the written amendment dated February 14, 2011.)
	(BLEU : 0.4324、RIBES : 0.8934)

各回とも特許・審査書類用語 5 語はいずれも正解訳文どおりの適訳語で常時安定している。第 5 回学習後に見られたカッコ補記中の「特許誤訳」(Claims 2-6 to、Claims 3-7 to)も第 10 回学習後以降は発生しなくなり、文中の全ての「特許誤訳」が解消された。各回の文構造は異なっているが、全て同義の範囲内であり、内容伝達レベルは等しく最高レベルのレベル 5 が付されている。

この期間の人手評価対象外の各回（つまり第 6 回～第 17 回学習後）においても、これら 5 語は常に正解訳文どおりの統一訳語で訳されており、特許・審査書類用語が全期間にわたって安定して適訳語で訳されている状況が如実に見て取れる。冒頭に示した「特許誤訳」のカウンタ結果の動向からも、特許・審査書類用語に関しては大多数の文が本例と同様の状況にあると考えられる。

こうした中、最終学習後も残留した少数の「特許誤訳」は、翻訳難易度の高いものが主となっている。以下、典型的な実例を示す。

A913-#214

原文	相違点 2： 請求項 1 に係る発明では集音部を頭部運動に基づき動かすのに対して、文献 1 に記載された発明ではテレビカメラを頭部運動に基づき動かすことについては明記されているが、マイクロホンを動かすことについては明確な規定がない点。
正解訳文	Difference 2: In the invention according to Claim 1, the sound collection unit is moved based on head movement, whereas the invention described in Cited Document 1 specifies that the TV camera moves based on head movement, but does not clearly define

	the moving of the microphone.
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：4 特許誤訳：有	Difference 2 : In the invention according to Claim 1, the sound collecting unit is moved based on the head movement, whereas in the invention described in Cited Document 1, it is specified that the television camera is moved based on the head movement, but there is no clear provision for moving the microphone.
	(BLEU : 0.4830、RIBES : 0.9074)

本例原文の「明確な規定がない」の「規定」は、技術的な特定、限定といった意味であり、正解訳文のように **define** や **specify** が適訳である。しかしながら、「規定」という語は、審査書類においては「特許法に規定する」という文脈で使用されることが多く、このため本例の機械翻訳文でもそのニュアンスで **provision**（「法律としての規程」の意味）という訳語が採用され続けた。最終学習後も上掲のとおり **provision** が採用され、「特許誤訳」と判定されている。審査書類対訳コーパスでは「規定」の最有力訳語が **provision** であることは確実であり、高度な訳し分けが求められる難易度の高いケースである。

最終学習後に残存した他の「特許誤訳」も、このように難易度が高い訳し分けが求められるものが多かった。例えば、「(発明の) 構成 : configuration」は文脈に鑑みると「特徴」を意図しており **feature** と訳すべきという指摘 (A02-#33) や、「特許要件 : patent requirements」の場合の「特許」は **patentability** とすべきという指摘 (A131-#70、同#442)、「参考文献 : cited document」は「引用文献」と区別して **reference document** と訳すべきという指摘 (A191-#26) などが挙げられる。いずれも指摘内容は正しいが、現状の翻訳文でもおおむね意味は類推可能であり、致命的な誤訳ではない。

審査書類対訳コーパスの学習の主眼は、「みんなの自動翻訳」がデフォルトでカバーできていない特許・審査書類用語を正しく翻訳できるようにすることにある。本項に示した無学習時から最終学習後までの「特許誤訳」のカウント数の推移や、最終学習後に残存した少数のエラーが翻訳難易度の高いものに絞られていることから考えて、審査書類①に関しては、上記目的は十分に達成されたと結論される。なお、現時点でも高度な訳し分けを要する少数の「特許誤訳」は残存しているが、例えば上掲の例においても、人間の目には本例の「規定」が法律上の規定のニュアンスではないことは明らかであり、今後の AI 技術の進展により、より人間に近い判断が可能となるにつれ、将来的にはこうした高度な訳し分けも的確になされるようになるものと考えられる。

3.3.5.3.2. 「技術／一般誤訳」のカウント結果

このカテゴリには、用語単位の誤訳のうち「特許誤訳」に該当しないもの、具体的には技術用語や一般用語の不適訳が含まれる文をカウントした。集計結果を下表に示す。

表 3.3.4. 各回学習後の「技術／一般誤訳」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	44	21	17	3	2
	—	-23	-4	-14	-1
拒絶理由通知書	10	5	8	4	5
	—	-5	+3	-4	+1
補正却下の決定	19	8	13	7	10
	—	-11	+5	-6	+3
前置報告書	35	21	18	6	8
	—	-14	-3	-12	+2
審査書類①	108	55	56	20	25
合計	—	-53	+1	-36	+5

審査書類対訳コーパスの主用途は、デフォルト状態の機械翻訳エンジンがカバーしていない審査書類特有の用語や表現、文型を学習させることである。技術用語や一般用語も多数含んではいるが、あらゆる語彙を広範に網羅する性質ではなく、したがって技術／一般用語の翻訳品質はエンジンの基礎学習に委ねられる部分が大きい。このため、「技術／一般誤訳」のカウント結果は、多分に「みんなの自動翻訳」それ自身の技術／一般用語の翻訳品質を測る意味合いをもつ。

上掲の集計結果を見ると、「技術／一般誤訳」は無学習の時点から「特許誤訳」に比べてカウント数は明らかに少数である。技術／一般用語はデフォルト時点で一定の学習（基礎学習）がなされていることを考えれば、これは妥当な結果である。ただし、それでも審査書類①全体の約 4 分の 1 の文（400 文中 108 文）には何らかの「技術／一般誤訳」が発生している。

その後、審査書類対訳コーパスを学習させた第 5 回学習後には、各書類でカウント数がほぼ半減した。ただし、これはデフォルトの「みんなの自動翻訳」がカバーしていなかった専門性の高い技術／一般用語がコーパスを通じて学習された結果ではなく、主に審査書類特有の言い回しに含まれる一般用語・表現が改善されたことによる。例えば以下のようなケースである。

原文	出願人は意見書で、引例 7 には、ホメオパシー化合物を含む H A S A が先ず製造される点が記載されていないと主張する。
正解訳文	Applicant argues in the written opinion that the cited example does not mention that HASA containing a homeopathic compound is first produced.
機械翻訳結果 【無学習時】 内容伝達レベル：2 技術一般誤訳：有	Applicant, in the opinion, claims that no HASA containing homeopathic compounds is first produced in the literature 7. (BLEU : 0.1055、RIBES : 0.8014)
機械翻訳結果 【第 5 回学習後】 内容伝達レベル：5 技術一般誤訳：無	In the written opinion, the applicant argues that Cited Document 7 does not disclose that HASA containing the homeopathic compound is produced first. (BLEU : 0.2207、RIBES : 0.8670)

本例では、無学習時に「主張する」という一般用語が **claims** と訳され、これが一般誤訳と判定されていた。判定の意図は、通常の文であれば「主張する」は claim でも大きな問題はないが、審査書類では claim は権利請求のニュアンスで使用されるため、本例の文脈での「主張」には argue など別の語を用いて区別を図るべき、というものである。いわば一般用語の審査書類での訳し方に関する不備であり、「特許誤訳」に準ずるものともいえる。こうしたタイプの一般誤訳は審査書類対訳コーパスの学習による改善が期待でき、本例も上掲のとおり第 5 回学習後にエラーは解消している。第 5 回学習後に改善した「技術／一般誤訳」の多くは、このように審査書類特有の用法による一般用語の改善が大半であった。

続く第 10 回学習後の「技術／一般誤訳」のカウント結果は、ほぼ第 5 回学習後から横ばいであり、書類によっては増加しているものもある。これは、上例のように「特許誤訳」に準ずる一般誤訳が審査書類対訳コーパスの初期学習によって早期に解消された結果、第 5 回学習後に残留していた「技術／一般誤訳」の大半が、デフォルトの学習データでカバーされていない難訳語の誤訳と、実例の定点観測（⇒3.2.7.）で多数確認されたような技術／一般用語の不安定な訳語変化による突発的な誤訳の二種に絞られていたためと推察される。

このうち後者のタイプの「技術／一般誤訳」に関しては、第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデートによってエンジンのロジックが改良された結果、訳語の不安定な変化が抑制され、不適訳語の突発的な採用も顕著に減少したことが、

実例の定点観測（⇒3.2.7.）によって確認されている。第15回学習後に「技術／一般誤訳」のカウンタ数が大幅に減少（前回比-36）したのも、これが主な理由であると考えられる。

一方、大規模アップデート以降も残存している「技術／一般誤訳」は、学習データ中に適訳が存在しないような特殊な語や、特殊な文脈に組み込まれたことで翻訳難易度が高くなってしまっている語が大半となった。以下、一例を示す。

A191-#47

原文	2) 参考資料2に記載の参考例1及び2のごとく、カルシウム分及びカリウム分が本願発明の規定する範囲外の自然塩ないし天然塩を使用した可塑性油脂組成物は何れも、塩味のとげとげしさのマスクングが不十分で、えぐ味が目立つものであり、本願発明の効果を奏するものではない。
正解訳文	2) As in Reference Examples 1 and 2 described in Reference material 2, a plastic oil composition using the crude salt or natural salt of which the calcium and potassium components do not exceed the scope as defined in the invention of the present application has an inadequate masking of salty roughness and has a conspicuous astringent taste and does not exert the effect of the present invention.
機械翻訳結果 【第18回無学習時】 内容伝達レベル：3 技術一般誤訳：有	2) As in Reference Examples 1 and 2 described in Reference Material 2, in any of the plastic oil and fat compositions using natural salts or natural salts whose calcium content and potassium content are outside the range defined by the present invention, the masking of the spiciness of the salty taste is insufficient, the astringent taste is conspicuous, and the effect of the present invention is not exerted.
	(BLEU : 0.3265、RIBES : 0.8356)

本例の原文では、「自然塩」と「天然塩」を異なるタイプの塩として使い分けており、正解訳文もこれに合わせて crude salt、natural salt と異なる訳語を当てている。一方、最終学習後の機械翻訳文では両者を natural salt と訳されており、双方の区別がつかない。このため技術誤訳と判定された。この誤訳は無学習時から最終学習後まで常時発生している。「自然塩」と「天然塩」は単独ではどちらも natural salt の意味で使われることが多く、学習データ中にもこの訳例しか存在していない可能性が高い。通常の文であれば「自然塩」を

natural salt と訳しても誤訳とはならず、特殊な文脈に組み込まれたことで翻訳難易度が高くなった事例である。

技術／一般用語の翻訳品質は、審査書類特有の言い回しに属するものを除き、「みんなの自動翻訳」自体のデフォルト性能、基礎学習内容に委ねられるところが多い。本調査分析の「技術／一般誤訳」のカウンtr結果では、最終学習後も残存した「技術／一般誤訳」は少数の難易度の高いものに限られており、「みんなの自動翻訳」における技術／一般用語の翻訳品質はおおむね良好と結論される。

3.3.5.3.3. 「未知語」のカウンtr結果

原文の日本語（又はその一部）が英訳されずにそのまま日本語で出力されている場合、「未知語」エラーとしてカウンtrした。集計結果は以下のとおりである。

表 3.3.5. 各回学習後の「未知語」該当文数カウンtr結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	0	2	0	0	0
	—	+2	-2	±0	±0
拒絶理由通知書	2	1	0	0	0
	—	-1	-1	±0	±0
補正却下の決定	1	0	0	0	0
	—	-1	±0	±0	±0
前置報告書	1	0	0	0	0
	—	-1	±0	±0	±0
審査書類① 合計	4	3	0	0	0
	—	-1	-3	±0	±0

カウンtr結果を見る限り、「みんなの自動翻訳」は「未知語」がほとんど発生しないタイプのシステムと見なせる。各書類とも無学習の時点からカウンtr数はごく少数であり、第 10 回学習後以降は一切発生しなくなった。以下、学習初期に発生していた未知語の実例を示す。

A131-#24

原文	本願求項 2 - 4 に係る発明についても同様である。
正解訳文	The same applies to the inventions according to Claims 2 to 4 of the present application.

機械翻訳結果 【無学習時】 内容伝達レベル：2 未知語：有	The same applies to the invention according to the present 願求 section 2-4. (BLEU : 0.2962、RIBES : 0.8998)
機械翻訳結果 【第 5 回学習後】 内容伝達レベル：4 未知語：有	The same applies to the inventions according to Claims 2-4 of the present application. 願求. (BLEU : 0.7332、RIBES : 0.9418)
機械翻訳結果 【第 10 回学習後】 内容伝達レベル：5 未知語：無	The same applies to the inventions according to Claims 2-4 of the present application. (BLEU : 0.8282、RIBES : 0.9699)
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：5 未知語：無	The same applies to the inventions according to Claims 2-4 of the present application. (BLEU : 0.8282、RIBES : 0.9699)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 未知語：無	The same applies to the inventions according to Claims 2-4 of the present application. (BLEU : 0.8282、RIBES : 0.9699)

本例は、原文に「本願求項」という誤記がある。人間にはこれが「本願請求項」の書き損じであることは明らかであり、正解訳文のように **Claims 2 to 4 of the present application** と正しく訳すことも容易である。だが、機械翻訳の場合、誤記への対処は融通が利かないことも多く、「みんなの自動翻訳」も無学習時には「本／願求／項」と分解して直訳しようとした結果、語の意味を成さない「願求」が「未知語」となり、「present 願求 section」という意味不明な文言になっている。

これに対し、審査書類対訳コーパスを学習した第 5 回学習後は、「願求」自体は引き続き文末に「未知語」として出力されているが、翻訳文の該当箇所は「本願請求項」という正しい解釈に基づく訳語 **Claims 2-4 of the present application** で訳出されるようになった。つまり、「未知語」は引き続き発生しているが、英文自体は原文本来の文意を正しく反映できており、実質的にはエラーは解消したとも見なせる。

さらに、続く第 10 回学習後の機械翻訳文では文末の「願求」も出力されなくなり、以降、

第 15 回、第 18 回学習後にも再発せず、本例の「未知語」は完全に解消した⁴²。上掲のカウント集計結果に示したとおり、第 5 回学習後に残存していた他の文の「未知語」も同じタイミングで全て解消している。

本例における「未知語」解消の経緯は、「みんなの自動翻訳」が採用するニューラル機械翻訳方式の利点を端的に表している。各回の機械翻訳文の変遷をまとめると、無学習時には対処不可能であった「本願求項」という誤記が、審査書類対訳コーパスを学習した結果、本来の意図に沿った解釈で翻訳できるようになった、ということになる。この改善は、人手評価対象外の各学習回を調査した結果、実質的には初回学習後に既に達成されていた（つまり [Claims 2-4 of the present application](#) と訳出された）ことが判明した。一方、初回学習時のコーパス中には、「請求項」が「求項」と誤記されているデータは存在しない⁴³。したがって、「みんなの自動翻訳」は「求項」という誤記パターンをコーパスから直接学習したわけではなく、本例に類似する（誤記のない）文パターン、文脈を大量に学習することにより、人間と同様にこれを「請求項」と同義であると認識したと考えられる。ニューラル機械翻訳方式は人間の思考を模した機械学習をその根幹とするが、本例はその利点が表われた事例といえる。

なお、冒頭で述べたとおり「みんなの自動翻訳」はそもそも「未知語」が発生しにくい構造と見受けられる。「未知語」は上例のように原文中に解釈不可能な語句が存在した際に発生する性質であるが、「みんなの自動翻訳」の場合、解釈不可能な語句は未知語として出力されるのではなく、翻訳文から欠落してしまう傾向が強い。第 3.2.7.項で定点観測した各実例においても、無学習時において専門性の高い特許・審査書類用語がしばしば翻訳文から欠落しており、カテゴリ別エラーカウントでも、第 3.3.5.3.7.項で後述するとおり、無学習時には「未知語」を大幅に上回る数の「欠落エラー」が発生していた。したがって、「みんなの自動翻訳」においては、「未知語」が少ないことが必ずしも翻訳不能語が少なかったことを意味しない点に注意が必要である。

⁴² 人手評価対象外の各学習回においても、第 1 回から第 8 回学習後までは第 5 回学習後と同様に「願求」が出力されていたが、第 9 回学習後以降は一切出力されなくなった。

⁴³ 審査書類対訳コーパス全体では「請求項」を「求項」と誤記している文は本例を含め 3 文存在する。

3.3.5.3.4. 「係り受けエラー」のカウント結果

文中の要素間の関係性（係り受け）が原文と異なっている場合、「係り受けエラー」と見なした。カウント結果は以下のとおりである。

表 3.3.6. 各回学習後の「係り受けエラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	18	17	21	9	10
	—	-1	+4	-12	+1
拒絶理由通知書	18	14	12	6	2
	—	-4	-2	-6	-4
補正却下の決定	22	13	13	10	10
	—	-9	±0	-3	±0
前置報告書	26	26	27	17	17
	—	±0	+1	-10	±0
審査書類①	84	70	73	42	39
合計	—	-14	+3	-31	-3

「係り受けエラー」は、無学習時には審査書類①全体の 21%に相当する 84 文で発生していた。審査書類対訳コーパスの学習を経た第 5 回、第 10 回学習後後もそれぞれ 70 文、72 文とさほど改善しなかったが、大規模アップデートを経た第 15 回学習後には 42 文と顕著な減少（-31 文）が見られ、第 18 回学習後も 39 文（-3 文）とその水準が維持された。

大規模アップデートでは「みんなの自動翻訳」の訳語学習ロジックが改良され、その結果、学習データに準拠した適訳な訳語や文構造が安定して選定される確率が高まった。「係り受けエラー」がコーパス学習後もさほど改善せず、大規模アップデートを経た第 15 回学習後のみ顕著に改善したという事実は、無学習時に発生していた「係り受けエラー」の多くは、審査書類由来の学習データの不足によるものではなく、むしろ当時の「みんなの自動翻訳」の訳語学習の不安定性に起因しており、エンジン本体のロジックの調整が不可欠であったことを示している。

以下、大規模アップデートを機に「係り受けエラー」が解消された実例を示す。

A131-#153

原文	引用文献 8 として示した出願(本分割出願の親出願)の請求項 1 に係る発明は、本願請求項 6（請求項 1 を引用する請求項 2 を引用する請求項 3 を引用する請求項 6）に係る発明と同一である。
----	---

正解訳文	The invention concerning claim 1 of the application (parent application of the divisional application) shown as cited document 8 is identical to the invention concerning claim 6 of the present application (claim 6 depending on claim 3 which depending on claim 2 which depending on claim 1).
機械翻訳結果 【無学習時】 内容伝達レベル：2 係り受けエラー：有	The invention according to claim 1 of the application shown as reference 8 (parent application of the present application) is identical to the invention according to the invention according to claim 6 (claim 1 in which claim 2 is cited in claim 3) and the invention in accordance with claim 6. (BLEU : 0.2975、 RIBES : 0.7438)
機械翻訳結果 【第 5 回学習後】 内容伝達レベル：2 係り受けエラー：有	The invention according to Claim 3 of the application (the parent application of the present divisional application) shown as Cited Document 8 is the same as the invention according to Claim 1 of the present application (Claim 6 citing Claim 2 citing Claim 6 citing Claim 1). (BLEU : 0.4798、 RIBES : 0.7085)
機械翻訳結果 【第 10 回学習後】 内容伝達レベル：2 係り受けエラー：有	The invention according to Claim 8 of the application (the parent application of this divisional application) shown as Cited Document 1 is the same as the invention according to Claim 6 of the present application (Claim 6 dependent on Claim 3 citing Claim 1 citing Claim 2 citing Claim 1). (BLEU : 0.4197、 RIBES : 0.8646)
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：5 係り受けエラー：無	The invention according to Claim 1 of the application (parent application of the divisional application) shown as Cited Document 8 is the same as the invention according to Claim 6 of the present application (Claim 6 depending on Claim 3 depending on Claim 2 depending on Claim 1). (BLEU : 0.7227、 RIBES : 0.9435)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 係り受けエラー：無	The invention according to Claim 1 of the application shown as Cited Document 8 (parent application of this divisional application) is the same as the invention according to Claim 6 of the present application (Claim 6 citing Claim 3 citing Claim 2 citing Claim 1). (BLEU : 0.4422、 RIBES : 0.8574)

本例は、無学習時、第 5 回学習後、第 10 回学習後と、いずれも原文中の青字部分、「請求項 1 を引用する請求項 2 を引用する請求項 3 を引用する請求項 6」の箇所が「係り受けエラー」となっていた。一方、大規模アップデートを経た第 15 回学習後、第 18 回学習後は、それぞれ Claim 6 depending on Claim 3 depending on Claim 2 depending on Claim 1、Claim 6 citing Claim 3 citing Claim 2 citing Claim 1 と、「引用する」の訳語は変化したものの、いずれも正しい係り受けで訳出された。

本例の場合、大規模アップデート以前においても、人手評価対象外の各回では正しい引用関係（請求項 6→3→2→1）で翻訳できている回も少なくない（第 2～4 回、7～9 回学習後）。つまり、この箇所の引用関係を正しく解釈するに足る学習データは当初から存在していたことになる。それでもたびたび誤解釈が生じたのは、学習データから導出される有力な解釈が必ずしも安定して採用されないという旧「みんなの自動翻訳」の性質によるものと考えられる。

これに対し、大規模アップデートを経た第 11 回学習後以降は、この箇所は人手評価対象外の回を含めた全ての回で正しい解釈で訳されるようになった。この状況は、実質的には、大規模アップデート前は散発的に採用されていた不適な文構造が、アップデート以降は採用されなくなったとも説明できる。つまり、「みんなの自動翻訳」における「係り受けエラー」の改善とは、事実上、不適な訳語・文構造の突発的な採用の抑制の結果であると見なせる。本例と同じタイミングで「係り受けエラー」のカウント数が大幅に減少していることも、これを裏付けている。

裏を返せば、最終学習後も残存している「係り受けエラー」は、大規模アップデート時のロジックの改善では解消できず、その後の学習データの増量でも改善できなかったものとなる。したがって、今後さらに学習データを増量したとしても、確実に改善されるとはいえない。

ただし、審査書類①における最終学習後の「係り受けエラー」の発生率は 10%弱であり、機械翻訳としてはさほど多発していない。もともと原文自体が曖昧な書かれ方であったり、人間が読めば意味的・文脈的に係り受けを特定できても文法的には複数の係り受け解釈が可能であったりするものも少なからず存在することを考えると、若干の「係り受けエラー」は、機械翻訳を使用する以上は不可避と考えるべきであろう。

3.3.5.3.5. 「英語表現エラー」のカウント結果

本カテゴリには、誤訳とはいえないが英語として不自然である、正しい意味にも解釈できるが曖昧、ミスリーディングであるなど、英文として改善の余地があるケースをカウントした。以下、カウント結果を示す。

表 3.3.7. 各回学習後の「英語表現エラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	44	43	45	39	34
	—	-1	+2	-6	-5
拒絶理由通知書	14	11	8	14	15
	—	-3	-3	+6	+1
補正却下の決定	31	24	14	19	15
	—	-7	-10	+5	-4
前置報告書	21	18	21	23	27
	—	-3	+3	+2	+4
審査書類①	110	96	88	95	91
合計	—	-14	-8	+7	-4

審査書類①における「英語表現エラー」の発生数は無学習時には合計 110 文と、全体の約 4 分の 1 の文で発生していたが、審査書類対訳コーパスの学習後もカウント数はさほど変わらず、劇的な改善は見られなかった。この点は「係り受けエラー」と類似している。ただし「英語表現エラー」は、「係り受けエラー」が顕著に減少した大規模アップデート以降も、最終学習後までほとんどカウント数が変わらなかった。

「英語表現エラー」には多様なケースが含まれ、典型例というべきものは存在しないが、例えば以下のような事例が該当する。

A913-#15

原文	引用文献 1 には上記各点にかかる記載はない。
正解訳文	There is no description regarding each of the above points in Cited Document 1.
機械翻訳結果 【無学習時】 内容伝達レベル：1 英語表現エラー：無	There are No. 1 cited above. (BLEU : 0.0360、RIBES : 0.6481)

機械翻訳結果 【第 5 回学習後】 内容伝達レベル：4 英語表現エラー：有	There is no description in Cited document 1 related to each of the above points.
	(BLEU : 0.5473、RIBES : 0.7546)
機械翻訳結果 【第 10 回学習後】 内容伝達レベル：5 英語表現エラー：無	Cited Document 1 does not have the description relating to the above points.
	(BLEU : 0.1779、RIBES : 0.4933)
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：5 英語表現エラー：無	Cited Document 1 does not describe each of the above matters.
	(BLEU : 0.2469、RIBES : 0.5036)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 英語表現エラー：無	Cited Document 1 does not describe the above points.
	(BLEU : 0.1656、RIBES : 0.4972)

本例は、第 5 回学習後の翻訳文のみが「英語表現エラー」と判定された。原文の「[上記各点にかかる](#)」に相当する **related to each of the above points** が、本来係るべき「記載：description」の直後ではなく「引用文献 1：Cited document 1」の直後に配置されていることで、「引用文献 1」に係っていると誤解しやすく、ミスリーディングであると見なされた。これは、第 10 回学習後以降の各翻訳文と見比べてみるとよく分かる。

とはいえ、この翻訳文でも **related to...** が description に係っていないとは断定できず、誤訳とまではいえない。読みづらく、誤解釈を招くおそれがあるという程度の不備であるため、「係り受けエラー」ではなく「英語表現エラー」と判定された。なお、本エラーは上掲のとおり第 10 回学習後以降は解消している。

本例は「英語表現エラー」の一例にすぎず、他にも様々なタイプが存在する。だが、概していえるのは、「英語表現エラー」と判定されるような英文の不備・不自然さは、審査書類特有の表現や文体に起因する不備というよりは、英文のレトリックを機械処理で達成することの難しさに起因する問題であるということである。審査書類対訳コーパスの学習後もさほどエラーが改善しなかったのも、これが一因であろう。レトリックには文法のような明確な基準がなく、このため翻訳処理の中でも難易度は特に高い。このため、審査書類に限らず、どのような文書であっても、機械翻訳を利用する以上ある程度の「英語表現エラー」の発生は不可避と考えられる。

なお、「英語表現エラー」にはもう一つの特徴がある。それは、より重大なエラーが発生している箇所は、英語表現以前の問題となるため「英語表現エラー」にはカウントされない、ということである。上例も、無学習時の翻訳文では、第5回時に「英語表現エラー」と判定された箇所は完全に欠落してしまっているため、「欠落エラー」と判定され、「英語表現エラー」とは見なされていない。

「英語表現エラー」はこうした性質を持つため、重大なエラーは解消したが未だ改善の余地があるというようなケースで、他種のエラーと入れ違いでカウントされる傾向が見られる。大規模アップデート直後の第15回学習後にカウント数が増加するという、他のエラーとは正反対の動向を示したのも、この傾向によるものと考えられる。

審査書類①では、「英語表現エラー」は最終学習後においても全体の約4分の1の文において発生しており、全10種のエラーの中でも発生件数は突出している。コーパス学習前後、大規模アップデート前後とも発生頻度はほとんど変わらず、機械翻訳にとって改善が難しいカテゴリのエラーといえる。ただし、「英語表現エラー」は冒頭で述べたとおり「誤訳とはいえない」程度の不備をカウントしたものであり、致命的な誤訳は含まれない。このため、内容理解上の悪影響は他種のエラーに比べれば小さい。

3.3.5.3.6. 「湧き出しエラー」のカウント結果

「湧き出しエラー」とは、原文中に存在しない不要な語や表現が翻訳文に混入してしまうエラーを指す。従来のルールベース方式の機械翻訳では理論上ほぼ発生せず、ニューラル翻訳など学習データを統計的に処理する方式の機械翻訳に特徴的なエラーである。まずは各回のカウント結果を下表に示す。

表 3.3.8. 各回学習後の「湧き出しエラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	18	21	30	1	0
	—	+3	+9	-29	-1
拒絶理由通知書	14	11	16	0	1
	—	-3	+5	-16	+1
補正却下の決定	13	20	21	0	0
	—	+7	+1	-21	±0
前置報告書	30	27	27	0	0
	—	-3	±0	-27	±0
審査書類①	75	79	94	1	1
合計	—	+4	+15	-93	±0

「湧き出しエラー」は、無学習時から第 10 回学習後にかけては審査書類①全体で 76 件、79 件、95 件と漸増し改善は見られなかったが、大規模アップデートを経た第 15 回学習後に一挙に解消され、第 15 回後、最終学習後とも 1 件のみと僅かな発生に抑えられている。

コーパス学習後も発生件数がほとんど減らず、大規模アップデート時のみ顕著に減少したという本エラーの改善過程は、第 3.3.5.3.4.項で分析した「係り受けエラー」のものと類似している。つまり、「湧き出しエラー」も「係り受けエラー」と同じく、学習データの増量のみでは解決できない、本質的にはエンジンのロジックに起因するエラーと見なせる。

大規模アップデート以前の旧「みんなの自動翻訳」では、「湧き出しエラー」は突発的に、かつ、「欠落エラー」と同時に発生することが多かった。原文中の要素が翻訳文から欠落した際に、それを埋め合わせる形で文中の別の文言が重複して湧き出すというもので、例えば第 3.2.7.2.項で定点観測した実例 3-2 では、第 7 回学習後の翻訳文のみ、それまで問題なく訳出されていた原文中の「脆弱部 (fragile potion)」という語が突如消失し、その埋め合わせとして「バリア性フィルム層 (barrier film layer)」という文中の別の語句が重複出力されていた。

本カウント結果が示すとおり、こうした「湧き出しエラー」は「みんなの自動翻訳」の大規模アップデートを機に一挙に解消された。アップデートの際にエンジンのロジックが改良され、「湧き出しエラー」への抜本的な対策が講じられたためと推察される。

ただし、第 15 回学習後、第 18 回学習後とも 1 件カウントされているように、大規模アップデート以降も「湧き出しエラー」の発生は皆無ではない。特に最終学習後に検出された 1 件は、過去に頻発していた典型的な「湧き出しエラー」と同傾向のものであった。以下、実例を示す。

A131-#14

原文	(3) しかしながら、画像をディスクレーベル面に印刷するために加工するものにおいて、ディスクのレーベル面にあたる部分のみ透過するようなマスクを表示することは、例えば文献 i、文献 j に、撮像テンプレートにおいて必要な部分以外を不透過表示とすることも例えば文献 h にあるよう、対象外の部分を表示しないようにすることは何ら格別のことではない。
正解訳文	(3) However, for example, as described in Cited document i and Cited document j, in the processing for printing an image on the disk label surface, only the partially transparent mask in contact with the disk label surface is displayed, and also as described in Cited document h, the unwanted portion in an imaging template is displayed opaquely, and thus it is not special to display the non-target portion.
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：3 湧き出しエラー：無	(3) However, in the case of processing an image for printing on the Kléber Pereira surface, displaying a mask that transmits only the portion corresponding to the label surface of the disc, for example, as described in Cited Documents I and j, and making the non-transmissive display other than the necessary portion in the imaging template, for example, as described in Cited Document h, it is not special to prevent the non-target portion from being displayed. (BLEU : 0.1804、RIBES : 0.5641)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：3 湧き出しエラー：有	(3) However, displaying a mask that transmits only the portion corresponding to the label surface of the surface of the disc when processing an image to print the image on the surface of the disc is not exceptional at all, as described in, for example, Cited Documents I and j, making the portion other than the necessary

	portion in the imaging template non-transparent, and for example, Cited Document h. Kleber Pereira
	(BLEU : 0.1804、RIBES : 0.5641)

本例⁴⁴の第 18 回学習後の機械翻訳文では、原文の「画像」が processing an image to print the image、「面」が the label surface of the surface of the disc と、文中の二箇所において同じ訳語が不必要に繰り返されている。過去に多発した「湧き出しエラー」は、「欠落エラー」と併発し、欠落語を埋め合わせるために翻訳文中の別の訳語が重複出力される形で生じる傾向が強かったが、本例でも今回、「不透過表示」の「表示」の「欠落エラー」が発生している。湧き出しの発生箇所は欠落箇所とは異なるため典型的な埋め合わせパターンではないが、両種のエラーが併発し、文中の用語が重複出力されるという点で同傾向といえる。

本例が示すように、大規模アップデート以降の「みんなの自動翻訳」においても「湧き出しエラー」が発生する可能性は引き続き皆無ではない。とはいえ、大規模アップデート以前に比べればその発生頻度は激減しており、今後「みんなの自動翻訳」で「湧き出しエラー」が以前のように多発する懸念はほぼ無いと結論できる。

⁴⁴ 本例の機械翻訳文では第 15 回後、第 18 回後とも原文内容に無関係な Kleber Pereira という人名が混入しているが、これは原文の「あるよう」という誤記（正しくは「あるように」）に起因していると考えられる。

3.3.5.3.7. 「欠落エラー」のカウント結果

「湧き出しエラー」とは逆に、原文中に存在する語や表現が翻訳文から消失してしまっている場合は「欠落エラー」としてカウントした。以下、カウント結果を示す。

表 3.3.9. 各回学習後の「欠落エラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	50	17	24	4	7
	—	-33	+7	-20	+3
拒絶理由通知書	36	15	11	1	2
	—	-21	-4	-10	+1
補正却下の決定	44	17	23	4	7
	—	-27	+6	-19	+3
前置報告書	56	23	35	7	7
	—	-33	+12	-28	±0
審査書類①	186	72	93	16	23
合計	—	-114	+21	-77	+7

上掲のとおり、無学習時の審査書類①では全体の半数近い文で「欠落エラー」が発生していた。「みんなの自動翻訳」には、第 3.2.7.項で分析した多くの実例⁴⁵で見られたように、学習データ中に訳語が存在せず翻訳不可能な用語・表現は、「未知語」として出力されるのではなく翻訳文から消失してしまう傾向がある。上記カウント結果からは、審査書類対訳コーパス学習前の「みんなの自動翻訳」では、各文に含まれる審査書類用語・表現に対応できず、その一部が翻訳文から欠落してしまっている状況がうかがえる。このことは、コーパスの学習により特許・審査書類用語が正しく訳されるようになった第 5 回学習後に「欠落エラー」が 60%以上減少していることからわかる。

ただし、同じく審査書類用語・表現への未対応が原因であった「特許誤訳」と比べると、「欠落エラー」のコーパス学習初期における改善の度合いはやや鈍い。これは、「欠落エラー」は必ずしも未学習の審査書類用語・表現のみで発生するわけではなく、それまで問題なく出力されていた技術／一般用語などで突発的に発生する性質も併せ持つためである。以下、一例を示す。

⁴⁵ 実例 3-1 における「記載も示唆もない事項」、実例 3-2 における「当業者」や「適宜選択し得た事項」、実例 3-4 における「許容基準」など。

原文	3) 補正後の請求項 1 に係る発明では、リッジ部上に形成されたマスクを、リッジ部上のマスク上に形成された第 2 絶縁膜とともに <u>リフトオフ</u> により同時に除去するのに対して、引用文献 5 に記載されている G a N 系半導体レーザの製造方法では、リッジ部上のマスクをエッチングにより除去している点（以下、「相違点 3」という。）。
正解訳文	3) In the invention claimed in Claim 1 after the amendment, the mask formed on the ridge part and a second insulation film formed on the mask on the ridge part are <u>removed at the same time through lift-off</u> ; in contrast, in the manufacturing method for the GaN-based semiconductor laser, as described in Cited Document 5, the mask on the ridge part is removed through etching (hereinafter referred to as "Difference 3").
機械翻訳結果 【無学習時】 内容伝達レベル：4 欠落：無	3) In the invention according to claim 1 after correction, while the mask formed on the ridge portion is simultaneously <u>removed by lift-off together with</u> the second insulating film formed on the mask on the ridge portion, in the method for manufacturing the GaN-based semiconductor laser described in Reference Document 5, the mask on the ridge portion is removed by etching (hereinafter, referred to as "difference 3"). (BLEU : 0.4500、RIBES : 0.9182)
機械翻訳結果 【第 5 回学習後】 内容伝達レベル：4 欠落：有	3) In the invention according to Claim 1 after the amendment, the mask formed on the ridge portion is simultaneously <u>removed together with</u> the second insulating film formed on the mask on the ridge portion, whereas in the manufacturing method of the GaN-based semiconductor laser described in Cited Document 5, the mask on the ridge portion is removed by etching (hereinafter, referred to as "Difference 3"). (BLEU : 0.5459、RIBES : 0.9301)
機械翻訳結果 【第 10 回学習後】 内容伝達レベル：5 欠落：無	3) In the invention according to Claim 1 after amendment, the mask formed on the ridge portion is simultaneously <u>removed by lift-off with</u> the second insulating film formed on the mask on the ridge portion, whereas in the manufacturing method of the GaN-based semiconductor laser described in Cited Document 5, the mask on the ridge portion is removed by etching (hereinafter, referred to as

	"Difference 3").
	(BLEU : 0.5124、RIBES : 0.9252)

本例では、無学習時には問題なく訳出されていた「リフトオフ : lift-off」という技術用語が、第 5 回学習後の翻訳文では欠落している。無学習時に訳出できている以上、本例は、学習データ中に存在しない用語・表現が翻訳文から消失するという「欠落エラー」の典型パターンには該当しない。

「リフトオフ」の欠落は、上掲のとおり第 10 回学習後の機械翻訳文では解消し、再び lift-off と訳出されるようになっていく。人手評価対象外の各回に範囲を広げて確認したところ、欠落が発生していたのは第 5 回学習後とその前後の第 4 回、第 6 回学習後の 3 回のみであった。

このように、本例の「リフトオフ : lift-off」の「欠落エラー」は、学習データ中に訳語が存在するにもかかわらず特定の学習回のみで突発的に発生するという状況を呈している。この突発性は、旧「みんなの自動翻訳」で頻発していた技術／一般用語の不適訳の採用や「湧き出しエラー」とも共通する。つまり、「みんなの自動翻訳」に見られる「欠落エラー」には、学習データの不足、すなわち特許・審査書類用語の未学習に起因するタイプと、技術／一般用語の不安定な訳語変化や「湧き出しエラー」と同様にエンジンのロジックに起因し、学習済みの用語であっても突発的に生じるタイプの二種類が混在している。

無学習時から第 5 回学習後にかけての「欠落エラー」の減少は、実際の人手評価データ（⇒3.3.4.）を参照すると、主に前者のタイプが審査書類対訳コーパスの学習によって改善された結果であることがわかる。これに対し、第 5 回学習後も相当数のエラーが残存し、第 10 回学習後も減少しなかったのは、残存した「欠落エラー」の多くが後者のタイプであるためと見られる。実際、「湧き出しエラー」ほかエンジンのロジックが原因と見なされる各種エラーはいずれもこれと同じような推移を辿っている。

その後、これら後者のタイプの「欠落エラー」も、大規模アップデートを経た第 15 回学習後以降は大幅に減少した。「湧き出しエラー」と同様に、アップデート時のロジックの改良によって根本的な対策が講じられたものと推察される。最終学習後のエラーカウント数は 23 件と「湧き出しエラー」に比べて多いが、その大半は「等」(A131-#44、A191-#50 ほか)、「上記」(A913-#84)、「前記」(A913-#215)、「具体的に」(A02-#14) など、欠落しても文意の理解にはさほど大きく影響しない文言であった。いずれも、上掲した「リフトオフ :

lift-off」の事例のようにロジックの問題で文中の語句が消失したというよりは、翻訳時に過度な省略がなされた結果との印象を受ける。

なお、ここまで取り上げてきた「欠落エラー」はいずれも語句が欠落するタイプのエラーであったが、これらとはタイプの異なる「欠落エラー」も存在し、複数の文で最終学習後も残留している。下例のように、翻訳文が文の途中で途絶し、それ以降の部分が完全に欠落するタイプである。

A02-#260

原文	・「前記動作状態は、前記スキャナの走査ベッドの原点座標によって決定される該スキャナの走査ヘッドの定位置である」(請求項3の第1～2行)(何を表しているのかが明確になっていない。特に、動作状態としての「走査ヘッドの定位置」とは、何をさしているのかが不明である。)
正解訳文	The operation state is the fixed position of the scanning head of the scanner to be determined by the origin coordinate of the scanning head of the scanner (rows 1 to 2 of Claim 3) (It is not clear what is meant. In particular, it is not clear what is meant by the "fixed position of the scanning head" as an operation state.)
機械翻訳結果 【無学習時】 内容伝達レベル：3 欠落：有	the operating state is the fixed position of the scanning head of the scanner determined by the origin coordinates of the scanning bed of the scanner (the first ~ 2 row of claim 3), in which case it is unclear what is the "fixed position" of the scanning head. (BLEU : 0.3550、RIBES : 0.8515)
機械翻訳結果 【第5回学習後】 内容伝達レベル：3 欠落：有	"The operation state is a constant position of a scanning head of the scanner determined by the origin coordinate of the scanning bed of the scanner (lines 1 to 2 of Claim 3) (It is unclear what it represents. In particular, it is unclear what " constant (BLEU : 0.5184、RIBES : 0.8753)
機械翻訳結果 【第10回学習後】 内容伝達レベル：3 欠落：有	"The operation state is a constant position of the scanning head of the scanner determined by the origin coordinate of the scanning bed of the scanner" (lines 1 to 2 of Claim 3) (it is not clear what it represents, in particular, it is unclear what the " (BLEU : 0.6089、RIBES : 0.9231)
機械翻訳結果 【第15回学習後】	"The operating state is a fixed position of the scanning head of the scanner determined by the origin coordinates of the scanning bed

内容伝達レベル：3 欠落：有	of the scanner" (lines 1-2 of Claim 3) (It is not clear what it represents. In particular, it is unclear what the "fixed (BLEU : 0.5538、RIBES : 0.9038)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：3 欠落：有	"The operating state is a fixed position of the scanning head of the scanner determined by the origin coordinates of the scanning bed of the scanner" (lines 1-2 of Claim 3) (It is unclear what it represents. In particular, it is unclear what the "fixed p (BLEU : 0.5004、RIBES : 0.8858)

本例は無学習時から最終学習後まで全ての回で「欠落エラー」が検出されている。ただし、無学習時と第 5 回学習後以降とは「欠落エラー」のタイプが異なる。無学習時の「欠落エラー」は原文の「特に、動作状態としての」という語句が訳出されないという、旧「みんなの自動翻訳」の典型的な「欠落エラー」であった。これに対し、第 5 回学習後以降の各回の「欠落エラー」は、いずれも機械翻訳文が文頭から 254～255 文字に達した時点で、たとえば単語の途中であっても唐突に途絶し、それ以降が完全に欠落するという、特殊なタイプのエラーである。

同様のエラーは A191-#6 と A191-#215 でも発生している。上例を含めた 3 文とも第 3 回学習後に同時にエラーが発生し、それ以降は最終学習後まで毎回、機械翻訳文が 254～255 文字に達した時点で必ず途絶している。

これら 3 文の原文を見比べると、いずれも冒頭が中点で始まり、続いてカギカッコの引用文で構成されている点で共通していた。以下、A191-#6 と A191-#215 の原文を示す。

A191-#6 原文	・「これにより、新請求項 1 の発明は、遊技体への給電の確実性を向上させることができるという効果があります。また、カーブ部分の入口側及び出口側のうち少なくとも一方に給電部を配置すればよいので、給電部の設置面積等を削減できるので、コストダウンを図ることができます。」
A191-#215 原文	・「すなわち、本願新請求項 1 の発明は、走行体にカーブを走行させることなく、カーブ部分では減速しなければ走行体は曲がることできないといった特性を利用し給電部を配置することが技術的な特徴です。」

さらに、各書類の評価用コーパスの人手評価対象外の全件を確認したところ、この特徴をもつ原文（例えば A02-#283、A913-#244、A913-#1083 など）は、高確率で同様の「欠落エ

ラー」が全く同じタイミングで発生していた。さらに、冒頭の中点がなく文全体がカギカッコで括られている文（A02-#1180、A913-#1131 など）でも同様のエラーが発生していることも判明した。

このように同じタイミングで一斉に発生し、その後も学習データの増量にかかわらず同一かつ多数の文で常に発生し続けていることから、このエラーが学習データの内容に起因するものとは考えにくく、エンジンのロジックが原因で特定の条件下において発生するエラーであると推察される⁴⁶。本エラーが一斉に発生した第 3 回学習時には、「みんなの自動翻訳」のエンジンが GPMT-3.8_190930_nmt から GPMT-3.8_200131_nmt にバージョンアップされており（⇒2.2.1.）、その際にエンジンのロジックに加えられた調整内容に起因している可能性がある。

審査書類では摘記箇所など文全体がカッコで括られる機会が多く、このため本エラーも広範囲で発生する可能性がある。翻訳文が途絶してしまうため内容理解上の悪影響も大きく、実際、本エラーが発生した 3 文（A02-#260、A191-#6、同#215）の最終学習後の内容伝達レベルはいずれもレベル 2～3 と他に比して低い。つまり、このエラーが解消すれば、審査書類①の翻訳品質はさらに向上する。

⁴⁶ 異なる学習データを使用した審査書類②システム（第 4 章で後述）においても、発生条件に該当する文（A53-#1481）で同種のエラーが発生している。

3.3.5.3.8. 「数値エラー」のカウント結果

機械翻訳の不備のうち、原文中の数値（算用数字、漢数字含む）に起因／関連すると見られるものは「数値エラー」にカウントした。

表 3.3.10. 各回学習後の「数値エラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	7	9	6	0	0
	—	+2	-3	-6	±0
拒絶理由通知書	6	6	4	4	4
	—	±0	-2	±0	±0
補正却下の決定	6	6	8	0	0
	—	±0	+2	-8	±0
前置報告書	13	12	5	1	1
	—	-1	-7	-4	±0
審査書類①	32	33	23	5	5
合計	—	+1	-10	-18	±0

「数値エラー」のカウント結果は、上掲のとおり、コーパス学習後も発生件数がさほど減少せず、大規模アップデートを機に一挙に改善している。これは、「湧き出しエラー」や「欠落エラー」など、エンジンのロジックに起因するタイプのエラーに共通の推移である。

アップデート以前の旧「みんなの自動翻訳」における典型的な「数値エラー」は、下例のように原文中の数値が翻訳文で入れ替わってしまうものであった。

A131-#66

原文	引用文献 1 の図 2 及び図 3 には、積層セラミックコンデンサの幅及びランドにおける幅のいずれも具体的な幅寸法が具体的に記載されているものではない。
正解訳文	Figure 2 and Figure 3 of Cited Document 1 does not concretely describe any concrete width dimension either for the wide of the layered ceramic capacitor and the width in the land.
機械翻訳結果 【第 5 回学習後】 内容伝達レベル：4 数値エラー：有	Figs. 2 and 1 of Cited Document 3 do not specifically describe a specific width dimension of both the width of the multilayer ceramic capacitor and the width of the land. (BLEU : 0.2228、RIBES : 0.8110)

本例では、第 5 回学習後の翻訳文において原文の「引用文献 1 の図 2 及び図 3」の箇所の数字が入れ替わり、Figs. 2 and 1 of Cited Document 3（「引用文献 3 の図 2 及び図 1」）と訳出されている。ただし本例で「数値エラー」が発生したのは、無学習時を含め全 19 回の学習回のうちこの回と第 3 回学習後（Fig. 1 and Fig. 3 of Cited Document 2）の二度のみであり、「湧き出しエラー」や「欠落エラー」と同じく突発的に発生・解消する傾向が見られる。

「数値エラー」は前掲のカウント結果が示すとおり、大規模アップデートを経た第 15 回学習後以降はほぼ完全に解消された。ただし、その一方で、大規模アップデート以降に新たに出現した「数値エラー」も存在する。最終学習後に「数値エラー」と判定された 5 件は全てこの新種のエラーに属していた。以下、事例をいくつか示す。

A131-#48

原文	C) 請求項 1 4, 1 6 に式 (X V I I I) と記載があるが、どのような式であるか不明である。
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：4 数値エラー：有	C) Claim 1416 recites Formula (XVIII), but it is unclear what kind of formula it is.
	(BLEU : 0.3229、RIBES : 0.8993)

A131-#377

原文	したがって、請求項 1 ～ 7, 1 0 ～ 1 2, 1 8 に係る発明以外は、発明の単一性の要件以外の要件についての審査対象としない。
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：4 数値エラー：有	Therefore, the inventions other than the inventions according to Claims 1 to 710 and 1218 are not subject to examination for requirements other than the requirement of unity of invention.
	(BLEU : 0.3874、RIBES : 0.8160)

A131-#10

原文	引用文献 2 には、6 つのブラケット 4 上に載置される回転モジュールが記載されている(図 3, 5, 6 を参照)。図 2 0 には、コア部 3 に対する回転モジュール 2 の相対的な移動に伴う変形を許容した免震機構が図示されている。4. この出願は、特許請求の範囲の記載が下記の点で、特許法第 3 6 条第 6 項第 2 号に規定する要件を満たしていない。
----	---

機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 数値エラー：有	Cited Document 2 describes a rotary module mounted on six brackets 4 (see Fig. 356). Fig. 20 illustrates a seismic isolation mechanism that allows deformation due to relative movement of the rotary module 2 with respect to the core portion 3. 4. In this application, the recitation of the scope of claims does not satisfy the requirements as provided in the Patent Act Article 36 (6) (ii) in the following points.
	(BLEU : 0.4462、RIBES : 0.9240)

各事例とも、原文中のカンマで区切られた数値が翻訳文では連結されてしまっている。例えば最初の例では「請求項 1 4, 1 6」が Claim 1416 となっている。Claim が単数形であることから、1 4 と 1 6 は連続した 1 つの数値 (1 4 1 6) として扱われていることがわかる。

ただし、カンマで区切られた数値はエラーとなった 5 文以外の原文にも多数存在するが、それらには「数値エラー」は発生していない。一例を示す。

A02-#305

原文	本願の補正後の請求項 1, 2 に係る発明について、依然として上記拒絶理由通知に記載した拒絶理由は解消していない。
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 数値エラー：無	Regarding the inventions according to the amended Claims 1 and 2 of the present application, the reasons for refusal described in the above notice of reasons for refusal still have not been resolved.
	(BLEU : 0.4436、RIBES : 0.5056)

エラーが発生している文としていない文を比較した結果、エラーが発生するのは、中間にカンマを挟んだ、合計 3 つ以上となる複数の数値に限られることが判明した。この条件に該当する数値は、第 11 回学習後以降、ことごとく連結されて 1 つの数値として扱われている。

本エラーは、明らかにエンジン内部の数値処理に起因したものである。各事例とも大規模アップデート直後の第 11 回学習後に初めて発生し、それ以降は最終学習時まで毎回必ず発生していることから、アップデート時のエンジンのロジック改良の副作用である可能性がある。

3.3.5.3.9. 「記号エラー」のカウント結果

ピリオドやコロン、カッコやクォーテーションマークなど各種記号に関わるエラーは、この項目にカウントした。

表 3.3.11. 各回学習後の「記号エラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	10	7	6	5	3
	—	-3	-1	-1	-2
拒絶理由通知書	3	2	1	2	2
	—	-1	-1	+1	±0
補正却下の決定	4	2	3	2	3
	—	-2	+1	-1	+1
前置報告書	8	5	5	3	3
	—	-3	±0	-2	±0
審査書類①	25	16	15	12	11
合計	—	-9	-1	-3	-1

「記号エラー」は無学習時から 25 件と比較的少数であったが、その後は最終学習後までさほど減少せず、コーパスの学習や大規模アップデートのタイミングでも特段の改善は見られなかった。

最終学習後に指摘された「記号エラー」の多くは、例えば etc.の直後の不要なカンマの存在（#913-#161）や必要なスペースの不在（A02-#18）、不要なスペースの挿入（A913-#8）など、内容理解上ほとんど影響しない体裁上の不備であった。

そのような中、これらより重大性のある「記号エラー」として、「…」が適切に処理されていないという事例が複数件指摘されている。以下、一例を示す。

A131-#263

原文	請求項 7 - 9 に記載の「循環路に…設けられた…開閉弁」については引用文献 5 の記載（段落【0 0 2 8】、【0 0 3 9】 - 【0 0 4 3】、図 3 など）を参照のこと。
正解訳文	Regarding the feature "a switch valve...provided on...a circulation path" recited in Claims 7-9, reference is made to the disclosure of Cited document 5 (paragraphs [0028] and [0039]-[0043], and Fig. 3).

機械翻訳結果	See the description of Cited Document 5 (paragraphs [0028], [0039] - [0043], Fig. 3, etc.) for the "on-off valve. provided. in the circulation path" recited in Claims 7-9.
【第 18 回学習後】 内容伝達レベル：5 記号エラー：有	(BLEU：0.4537、RIBES：0.5166)

本例原文のカギカッコで括られた摘記箇所「循環路に…設けられた…開閉弁」には、途中を省略したことを表す「…」が使用されている。正解訳文はこれを"a switch valve...provided on...a circulation path"と三連ピリオドで表現しているが、機械翻訳文では「…」を"on-off valve. provided. in the circulation path"とピリオド 1 個に訳出しており、本来の意図が伝わらない。これが「記号エラー」と判定された。同種のエラーは他にも A02-#52、同#60、A191-#19、同#53、A913-#26 と計 6 文で見られ、最終学習後の「記号エラー」の過半数を占めた。

各例を見る限り、「みんなの自動翻訳」では原則「…」はピリオド 1 個に置き換える設定と見受けられる。だが、英文においてピリオド 1 個は原文の「…」の意味とはならない。正解訳文のように三連ピリオド等を使用するべきであろう。

なお、本例に関しては原文の「…」に対応するピリオドの挿入位置自体はおおむね正しく、仮に「みんなの自動翻訳」が「…」を三連ピリオドに置き換える設定であったならば、原文の内容を正確に理解できる。ただし、実際のところ、原文の「…」を翻訳文の適切な位置に反映することは機械翻訳にとってかなり難易度が高い。日本語と英語では、言語的性質の違いにより、同じ内容でも語順が大きく変化するためである。

本例においても、原文と翻訳文とでは「循環路：circulation path」と「開閉弁：switch valve」の出現順序は逆転している。その意味では、正しい位置にピリオドを挿入している「みんなの自動翻訳」は優秀である。だが、「…」を用いた省略のあり方に特段のルールはなく、このため原文における「…」の位置と翻訳文における対応位置との間にも法則性はない。このため、「…」の配置や使われ方によっては、人間であっても翻訳文のどの位置に反映させるべきか判断がつかないことも少なくない。

他文献からの引用が多用される審査書類では、割愛部分の存在を示す「…」が使用される頻度は高く、かつ一定の重要度を有する。だが、これを翻訳文に正確に反映させることは、上述の理由により機械翻訳では非常に難易度が高いと推察される。まずは現在の「みんなの自動翻訳」におけるピリオド 1 個への置換を三連ピリオドに改善することが望まれる。

3.3.5.3.10. 「その他エラー」のカウント結果

ここまでの各カテゴリに該当しない不備が存在する場合、本項目にカウントした。

表 3.3.12. 各回学習後の「その他エラー」該当文数カウント結果

書類種別	無学習	第 5 回後	第 10 回後	第 15 回後	第 18 回後
拒絶査定	6	5	7	1	1
	—	-1	+2	-6	±0
拒絶理由通知書	2	0	1	0	0
	—	-2	+1	-1	±0
補正却下の決定	3	0	2	0	0
	—	-3	+2	-2	±0
前置報告書	2	4	6	0	0
	—	+2	+2	-6	±0
審査書類①	13	9	16	1	1
合計	—	-4	+7	-15	±0

「その他エラー」にカウントされたエラーの大多数は、一文中で同じ用語が複数回出現し、その訳語が不統一であるという「訳語ゆれ」の指摘であった。特に大規模アップデート以前の旧「みんなの自動翻訳」では一部の技術／一般用語の訳語が学習前後で不安定に変化する傾向が顕著であったが、本エラーのカウント結果によって、同一学習回であっても同じ用語に異なる訳語が採用される場合があることが判明した。以下、一例を示す。

A131-#21

原文	(2) 第一流路が、前者においては、遊技球を球流れの勢いを弱めながら可変入賞口に導くのに対し、後者においては、球流れの勢いを弱めているか不明である点。
正解訳文	(2) In the former, the first flow path guides the game ball to the variable winning opening while slowing the momentum of the ball flow, whereas it is unclear whether the first flow path of the latter slows the momentum of the ball flow.
機械翻訳結果 【無学習時】 内容伝達レベル：4 その他エラー：有	(2) In the former, the first flow path guides the game balls to the variable winning port while weakening the momentum of the ball flow, whereas in the latter, it is unclear whether or not the momentum of the spherical flow is weakened. (BLEU : 0.5691、RIBES : 0.8997)

機械翻訳結果 【第 5 回学習後】 内容伝達レベル：5 その他エラー：無	(2) In the former, the first flow path guides the game balls to the variable winning port while weakening the momentum of the ball flow , whereas in the latter, it is unclear whether or not the momentum of the spherical flow is weakened.
	(BLEU : 0.6853、RIBES : 0.9168)
機械翻訳結果 【第 10 回学習後】 内容伝達レベル：4 その他エラー：有	(2) In the former, the first flow passage guides the game ball to the variable winning port while reducing the momentum of the spherical flow , whereas in the latter, it is unclear whether the ball flow is weakened.
	(BLEU : 0.4987、RIBES : 0.8778)
機械翻訳結果 【第 15 回学習後】 内容伝達レベル：5 その他エラー：無	(2) In the former, the first flow path guides the game ball to the variable winning opening while weakening the momentum of the ball flow , whereas in the latter, it is unclear whether the momentum of the ball flow is weakened.
	(BLEU : 0.7237、RIBES : 0.9239)
機械翻訳結果 【第 18 回学習後】 内容伝達レベル：5 その他エラー：無	(2) In the former, the first flow path guides the game ball to the variable winning opening while weakening the momentum of the ball flow , whereas in the latter, it is unclear whether the momentum of the ball flow is weakened.
	(BLEU : 0.7237、RIBES : 0.9239)

本例原文には「球流れ」が2回出現する。こうした場合、特に技術用語においては、双方に同一の訳語を用いないと読み手に無用な混乱を与える。だが、無学習時の翻訳文では、一方は正解訳文どおりの適訳語 **ball flow** を採用しているものの、もう一方には **spherical flow** と別の用語が採用されており、同じ用語が一文中で訳ゆれしてしまっている。これが「その他エラー」にカウントされた。

本例はその後、第 5 回学習後には訳ゆれが解消して双方 **ball flow** が採用されたが、続く第 10 回学習後には再び **spherical flow** と **ball flow** との間での訳ゆれが（無学習時とは入れ替わった形で）再発している。本例では人手評価対象外の第 7 回、第 9 回学習後にも同様の訳ゆれが発生しており、こうした訳ゆれが学習の進捗にかかわらずランダム的に発生する性質であることが見て取れる。ただし、大規模アップデートを経た第 11 回学習後以降は本例では訳ゆれが発生しなくなり、第 15 回、第 18 回学習後の翻訳文とも上掲のとおり **ball flow** で統一されている。

審査書類①全体でのカウント数も、大規模アップデート以降の第 15 回後、第 18 回後はいずれも 1 件のみと大幅に減少しており、アップデート時の訳語学習メカニズムの改良によって、学習前後の訳語の不安定な変化のみならず、同一学習回における用語の訳ゆれも抑制されたことが見て取れる。残存する 1 件も、A02-#38 において「自走」と「自走式」がそれぞれ self-running と self-propelled に訳揺れしていることを指摘したものであり、評価者の言うとおりの両語は訳語を統一することが望ましいが、厳密には「自走」と「自走式」とでは原語自体が異なっており、同じ用語が訳ゆれしているわけではない。

ただし、人手評価の対象文は 400 文と少数であり、同じ技術用語が一文中に複数回出現する文自体が少ない。このため、本カウント結果が実質的にゼロであるからといって、これをもって同一学習回での用語の訳ゆれの懸念なしとは確言できない。「みんなの自動翻訳」では大規模アップデート以降も訳語のランダム的な変化はニューラル翻訳方式の宿命として残存しており、学習前後での技術／一般用語の訳語変化は少なからず発生している（⇒ 3.2.7.4.）。同一学習回での用語の訳ゆれの発生理由もこれと同根であると考えれば、大規模アップデート以降も発生している可能性はむしろ高い。

そもそも、審査書類の機械翻訳を実用する際には、一文中の訳ゆれに限らず、同じ案件の異なる文間でも同じ用語の訳語は統一されるべきである。この観点から審査書類①の評価用コーパスを調査したところ、同一案件（補正却下の決定の第 6112367330 号）に由来し、「内部階段」という同じ技術用語を含む 8 文において、大規模アップデート以降も第 16 回学習後まで文間の訳ゆれが発生していたことが判明した。具体的には、第 16 回学習後の機械翻訳文で「内部階段」を internal staircase と訳した文が 5 文（A191-#1824、#1859、#1908、#1940、#1960）、internal stairs と訳した文が 2 文（#1943、#1967）、inner staircase と訳した文が 1 文（#1988）と、同一案件に属する文間で技術用語の訳ゆれが生じていた。どの訳語も同義の範囲内ではあるが、一つの構成要素に対して複数の呼称が混在しており、読み手の混乱は免れない。

本例に関しては、第 17 回学習後以降は「内部階段」の訳語は internal staircase に統一されるようになった。とはいえ、本例によって、訳語学習ロジックが改良された大規模アップデート以降の「みんなの自動翻訳」でも、同一学習回、同一案件内での訳ゆれが発生する可能性があることが明らかとなった。

審査書類①の評価用コーパスでは、同一の案件から複数の文が選出されているケース自体が少なく、このため最終学習後の同一案件内での訳ゆれの発生は確認できなかった。この点に関しては、同一案件の全文を評価用コーパスに採用する方針としている審決③／審決④の章（第 4.4.項）で改めて調査する。

3.4. 審査書類①システムに関するまとめ

審査書類①システムの機械翻訳の品質及び傾向について、自動評価及び人手評価の結果から以下の点が確認された。

3.4.1. 審査書類対訳コーパスの学習による翻訳品質の改善

審査書類①の機械翻訳品質は、審査書類対訳コーパスの学習により顕著に改善した。

評価用コーパス全件（2,000 文×4 書類）を対象とした自動評価では、併用した二種の自動評価指標 BLEU、RIBES の双方で全件平均スコアが大幅に向上した。無学習時と最終学習後（第 18 回後）の比較では、BLEU は 0.3160 から 0.5082 (+0.1922) へ、RIBES は 0.7558 から 0.8456 (+0.0898) へと、いずれもスコアを大きく伸ばした（⇒3.2.2.）。

評価用コーパスの一部（100 文×4 書類）を対象に精密な評価を実施した人手評価の結果も、自動評価と完全に整合するものであった。翻訳品質の絶対評価である「内容伝達レベル」の評価（5 段階評価）では、人手評価対象文 400 文の平均レベルは無学習時の 3.14 から最終学習後（第 18 回後）には 4.69 (+1.55) へと大幅に向上しており、自動評価結果が示した品質改善が人間の目から見ても同様であることが示された（⇒3.3.5.1.）。

さらに、内容伝達レベルの評価結果から、審査書類対訳コーパス学習前後の「みんなの自動翻訳」の具体的な翻訳品質も把握することができた。内容伝達レベルのレベル定義に照らすと、コーパス学習前（無学習時）の「みんなの自動翻訳」で審査書類を翻訳して得られる平均的な品質はレベル 3、すなわち「半分以上の重要情報は正確に伝達されている（50%～）」程度にとどまっていた。これに対し、コーパス全件学習後の翻訳品質は、評価用コーパス全体の 4 分の 3 の文がレベル 5「すべての重要情報が正確に伝達されている（100%）」と判定され、残りの文も大半がレベル 4「ほとんどの重要情報は正確に伝達されている（80%～）」を付されるなど、機械翻訳として極めて高い翻訳精度に到達したことが具体的に示された。

本事業は、日本の審査書類の機械翻訳品質が現状では不十分であること、そしてニューラル機械翻訳エンジンに審査書類由来の対訳コーパスを大量に学習させることで高品質な機械翻訳が得られることを前提に実施したものである（⇒1.1.）。本調査分析により、これら本事業の前提が的確なものであり、かつ本事業で作成した審査書類対訳コーパスが企図したとおりの訳質改善効果を奏するものであることが実証された。

3.4.2. 特許・審査書類関連用語の早期からの改善と安定

審査書類対訳コーパスの品質改善効果は、デフォルトの機械翻訳エンジンでカバーされていない審査書類特有の用語・表現の翻訳精度に最も顕著に表れた。

第 3.2.7.項に示した実例の定点観測では、無学習時の機械翻訳文において多くの特許・審査書類用語が誤訳されたり翻訳文から欠落したりしていたが、それらの全てが審査書類対訳コーパスの初回学習によってコーパス由来の適訳語に改善され、その後も全期間にわたり常時安定して採用される状況が確認された。

この状況は、人手評価の一環として実施した「エラーのカテゴリ別カウント」の「特許誤訳」のカウント数の推移から、審査書類①全体の傾向であることが示された(⇒3.3.5.3.1.)。無学習時には評価対象 400 文の約 70%に相当する 286 文において特許・審査書類用語の誤訳が発生していたが、これらの「特許誤訳」はコーパス学習初期(第 5 回学習後)の時点で 54 文(13.5%)へと大幅に減少し、最終学習後には 11 文(2.75%)のみと、ほとんど発生しなくなった。

こうした状況から、デフォルトの「みんなの自動翻訳」では特許・審査書類用語が十分にカバーされていないこと、そして本事業で作成した審査書類対訳コーパスによって、こうした特許・審査書類用語をほぼ完璧に「みんなの自動翻訳」に学習させることができることが実証された。エラーのカテゴリ別カウントでは、「特許誤訳」は無学習時の発生数、コーパス学習後の改善度合いとも他のエラーから突出しており、この「特許誤訳」の解消こそが審査書類①の翻訳品質の顕著な改善の最大の理由であることは明白である。つまり、本事業で作成した審査書類対訳コーパスは、その主眼である、機械翻訳システムを審査書類特有の用語・表現に対応させるという役割を十分に果たすものであると結論できる。

なお、特に大規模アップデート以前の「みんなの自動翻訳」では学習前後で訳語が不安定に変化する傾向が強かったが、こと特許・審査書類用語に関しては、こうした傾向にもかかわらず学習初期からコーパス由来の最適訳語で常時安定した。これは、審査書類対訳コーパスの作成にあたり、主要な審査書類用語・表現の訳語を統一していたためと考えられる。学習データ中に訳語候補が一種しか存在しないため、変化の余地が生じなかったものと推察される。

本調査分析により、ニューラル機械翻訳において専門性の高い用語・表現が多用される書類の翻訳精度を向上させるには、その書類に由来する高品質な対訳コーパスを十分な量学習させることが極めて有効であることが実証された。また、こうした学習用コーパスを作成

する際は、主要な専門用語・表現に対して事前に統一訳語を定めておくことで、無用な訳語ゆれを回避できることがわかった。

3.4.3. 特許・審査書類関連用語以外の翻訳品質について

本事業で作成した審査書類対訳コーパスは審査書類関連用語・表現の学習用途に主眼を置いており、技術／一般用語の翻訳品質や網羅性、そして翻訳文の生成などの基幹的な機械翻訳処理の精度は「みんなの自動翻訳」自体のデフォルト性能と基礎学習内容に委ねられる部分が多い。本調査分析により、こうした「みんなの自動翻訳」のデフォルト性能が優秀であり、審査書類の翻訳用途においておおむね不足がないことが確認された。

人手評価の一環として実施したエラーのカテゴリ別カウント（⇒3.3.5.3.）では、「特許誤訳」以外の各カテゴリのエラーは無学習時から比較的少数で、かつコーパス学習後に劇的な改善が見られなかった。この結果は、これらのエラーの多くが審査書類用語・表現の学習量の不足によるものではなく、エンジン自体のロジックや基礎学習データの内容に起因していることを示唆している。

事実、コーパスの学習では改善が乏しかったこれらのエラーは、「みんなの自動翻訳」の大規模アップデート直後の第11回学習後のみ顕著に改善し、以降もその状態が維持された。その結果、翻訳品質の絶対評価である内容伝達レベルの評価において、最終学習後の審査書類①に付されたレベルの平均値は5段階評価で4.69と、機械翻訳としては極めて高い水準に達した。レベル4.69は、内容伝達レベルの定義に照らし、重要情報の80～100%が正確に伝達される品質と見なせる。コーパス学習後の「みんなの自動翻訳」が、特許・審査書類用語のみならず技術／一般用語や文法に関しても正確に翻訳できていることを示すものであり、「みんなの自動翻訳」が優れたデフォルト性能と十分な基礎学習内容を有していることの証左といえる。

また、第3.2.4.項に示した自動評価スコアの技術分野ごとの集計結果では、各技術分野（化学、電気物理、機械の三分野）とも無学習時から最終学習後までの全期間にわたりほぼ同一の水準・推移を示しており、分野間の翻訳品質や傾向に特段のバラつきは見られなかった。この結果から、「みんなの自動翻訳」が各技術分野の用語・表現をおおむね均等にカバーしていることも確認できた。

3.4.4. 技術／一般用語及び文構造のランダム的な変化

審査書類①では、特許・審査書類用語が学習初期から正解訳文どおりの適訳語で安定したにもかかわらず、全体の過半数の機械翻訳文が最終学習時まで変化し続けた（⇒3.2.6.）。各実例の定点観測（⇒3.2.7.）では、技術／一般用語の訳語や翻訳文の文構造が学習前後で不安定な変化を繰り返す状況が確認されており、同様の変化が多数の文で継続的に発生していると考えられる。

こうした技術／一般用語や文構造の不安定性は、第 11 回学習前に実施された「みんなの自動翻訳」の大規模アップデート以降はかなり抑制された。各実例とも学習データ中で有力と見られる適訳語が採用される確率が高まり、変化の頻度や文中の変化箇所が少なくなるとともに、不適訳語の突発的な採用による品質劣化も明らかに減少した。第 3.2.6.項のカウント結果においても、第 12 回学習時以降は「不変」の文が顕著に増加している。

とはいえ、大規模アップデート以降も過半数の機械翻訳文で引き続き何らかの変化が学習の都度生じており、そのうちの半数では自動評価スコアが低下していることも事実である。この状況からは、大規模アップデート後も学習データを追加する都度、多数の文で品質が劣化している懸念が生じる。

だが、第 3.3.5.2.項に示した学習前後の翻訳文の相対評価では、最終学習前後の比較で評価対象全 400 文中 356 文（89%）が、学習前後で翻訳文の品質が「同等」と判定されている。この評価結果により、大規模アップデート以降の自動評価スコアの変動の大部分は同義の範囲内の機械翻訳文の表層的な変化によるものであり、実質的な翻訳品質の変化を伴うものではないことが示された。よって、学習データの追加が多数の文に品質劣化をもたらしているという懸念は否定される。

「みんなの自動翻訳」において技術／一般用語の不安定な訳語変化が生じる理由は、ニューラル翻訳方式における訳語学習のメカニズムと、技術／一般用語の多くが（基礎学習データを含めた）学習データ中に複数の訳語候補を持つことにある。一般的なニューラル翻訳方式では、学習に際し、ある用語について学習データ中に複数の訳語が存在する場合、それぞれの訳語に対し、その訳語の各文例での周辺情報（周辺に存在する語句等の情報）に基づいた多次元ベクトルを付与する。そして、翻訳時には、入力文における用語の周辺情報と、学習データ中の各訳語候補のベクトルとを照合して、最もフィットすると判断された訳語が採用される。

このため、複数の訳語候補を持つ技術／一般用語では、同じ用語であっても入力文の内容によって異なる訳語が採用される可能性がある。つまり、同一用語の訳ゆれは、「みんなの自動翻訳」に限らず現在のニューラル機械翻訳の共通の課題といえる。

なお、用語に付与されるベクトルは、学習データ全体の全ての用語の相互の関連度に基づくため、学習データが追加される都度、全ての用語について再計算される。大規模アップデート以前の旧「みんなの自動翻訳」では、このベクトルが再計算の都度大きく変化する傾向にあったと見られる。このため、入力文が同一である（つまり文中の各語の周辺情報が変化していない）にもかかわらず、多数の技術／一般用語の訳語が学習の前後で不安定に変化したものと推察される。特許・審査書類用語の訳語はそれでも常時安定していたが、これは学習データ中で訳語が統一されていて他の訳語候補が存在しないため、ベクトルが変動しても採否に影響しなかったのであろう。

大規模アップデート以降、それまで頻発していた技術／一般用語の学習前後での訳語変化が抑制され、学習データの内容がよりの確に反映されるようになったのは、「みんなの自動翻訳」内部でこのベクトル付与／再計算のロジックが改良されたためと推察される。各用語に対し、学習データ全体の状況がより正確に反映されたベクトルが付されるようになり、学習前後でベクトルの極端な変動が生じにくくなったことで、各回とも最有力の訳語候補が順当に採用される確率が高まったと考えられる。

ただし、各用語の訳語を文中の周辺語との関連に基づいて決定するというメカニズムがニューラル翻訳方式の本質であることに変わりはない。事実、審査書類①においても、一部の技術／一般用語は大規模アップデート以降も引き続き訳語変化を繰り返した。こうした訳語変化は、ニューラル翻訳方式を採用する「みんなの自動翻訳」の本質的な性質であるといえる。

ここで懸念されるのが、こうした「みんなの自動翻訳」における技術／一般用語の訳語変化は、その発生メカニズムから考えて、学習前後のみならず同一学習回においても発生する可能性が高いということである。

機械翻訳文を実用する際には、学習前後での訳語の変化よりも同一学習回での訳ゆれのほうが内容理解上の悪影響は大きい。特に技術用語は、同じ案件で繰り返し使用される場合、異なる文の間でも訳語が統一されていないと、読み手に混乱を与え、内容理解を阻害する。

第 3.3.5.3.10.項に示した「その他エラー」のカウンタ結果では、当初見られた同一文中の用語の訳ゆれが大規模アップデート以降は解消されたことが示されたが、これはあくまで

同一文中に同じ技術用語が複数回用いられるという限定的なケースでの検証にすぎず、この結果をもって、大規模アップデート以降は同一学習回での技術用語の訳ゆれの懸念なしとはいえない。むしろ、大規模アップデート以降も学習前後の訳語変化は相当数発生している以上、同様の変化が同一学習回における異なる文の間でも発生している可能性は高く、この点で機械翻訳の利用性に悪影響を与える懸念がある。

この懸念については第 3.3.5.3.10.項「その他エラー」でも言及したが、審査書類①の評価用コーパスには同一文献から採取された文が乏しいこともあり、確固とした結論は得られなかった。このため、本項では結論は保留し、文献単位で評価用コーパスを選定した審決③／④システムの章（第 5 章 5.4.5 項）で改めて調査する。

3.4.5. 各種審査書類由来の学習データ間の高い互換性

審査書類①の学習データである「審査書類コーパス A」では、4 種の審査書類（「拒絶査定」「拒絶理由通知書」「補正却下の決定」「前置報告書」）それぞれに由来する対訳文の構成比が 1：7：1：1 となっている。つまり、拒絶理由通知書の学習データ量が他の 3 書類に比して突出して多い。

しかしながら、書類別の自動評価スコア（⇒3.2.3.）や内容伝達レベル（⇒3.3.5.1.）の集計結果からは、学習データ量の圧倒的な違いにもかかわらず、拒絶理由通知書の評価スコアの推移や最終的に到達した品質水準に、他の 3 書類と異なる点は特段見られなかった。拒絶理由通知書の内容伝達レベルは 4 書類中の最上位ではあるが、これは無学習時の順位がそのまま維持された結果であり、学習データの多寡が反映されるべき学習前後のスコアの伸び幅は 4 書類中で最も低かった。

一方、拒絶理由通知書以外の 3 書類は、全学習データ中、自身に由来するデータはそれぞれ 10%にすぎない。つまり、学習データの 90%が他種の書類に由来するデータであったが、このことによる翻訳品質への悪影響は見られず、各書類とも拒絶理由通知書と完全に同等の品質改善が果たされた。

こうした状況から、審査書類①に属する 4 種の審査書類由来の学習データ間には高い互換性があると結論される。4 種の審査書類とも同じ制度に基づく審査手続の過程で特許庁審査官の手により作成される書類であり、使われる専門用語や表現もおおむね共通していることを考えれば、妥当な結果ともいえる。

なお、自動評価では補正却下の決定のみ他の 3 書類よりスコア水準が明らかに低かったが（⇒3.2.3.）、内容伝達レベルの評価では補正却下の決定のレベルは他の 3 書類と同水準であり、エラーのカテゴリ別カウントでも他の書類と異なる傾向は見られなかった。これら全ての評価結果を総合すると、補正却下の決定という書類自体が他の 3 書類よりも翻訳難易度が高いと見なす根拠に乏しく、各書類の翻訳難易度はおおむね同等と見なすべきである。

3.4.6. 文種間の翻訳難易度の違いについて

審査書類文における4種の文種（「認定」「対比」「判断」「その他」）間の自動評価スコアの相対比較（⇒3.2.5.）では、書類によって各文種の平均スコアの相対順位が異なるなどのバラつきが見られ、明確な傾向は見いだせなかった。ただし、各書類とも「その他」に属する文のスコアが下位となる傾向は共通しており、「その他」に属する文が他の文種より翻訳難易度が高い可能性がある。

人手評価対象100文は、各書類とも4文種から25文ずつを選定している（⇒3.3.4.）。このため人手評価結果を文種間で相対比較することが可能である。下表に、各書類の各学習段階における内容伝達レベルを文種ごとに集計して平均レベルを算出した結果を示す。

表 3.4.1. 文種ごとの「内容伝達レベル」平均値の推移

種別	評価対象	無学習	第5回	第10回	第15回	第18回
認定	審査書類①	3.21(2)	3.93(4)	4.10(4)	4.66(3)	4.66(4)
		—	+0.72	+0.17	+0.56	±0
	(拒絶査定)	2.88	4.08	4.12	4.60	4.68
		—	+1.20	+0.04	+0.48	+0.08
	(拒絶理由通知書)	3.88	4.00	4.32	4.88	4.96
		—	+0.12	+0.32	+0.56	+0.08
	(補正却下の決定)	3.12	4.00	4.05	4.62	4.55
		—	+0.88	+0.05	+0.57	-0.07
	(前置報告書)	2.96	3.64	3.92	4.56	4.44
		—	+0.68	+0.28	+0.64	-0.12
対比	審査書類①	3.29(1)	4.20(1)	4.31(1)	4.67(2)	4.73(1)
		—	+0.91	+0.11	+0.36	+0.06
	(拒絶査定)	3.56	4.00	4.20	4.52	4.68
		—	+0.44	+0.20	+0.32	+0.16
	(拒絶理由通知書)	3.36	4.56	4.64	4.96	4.88
		—	+1.20	+0.08	+0.32	-0.08
	(補正却下の決定)	3.32	4.28	4.26	4.65	4.68
		—	+0.96	-0.02	+0.39	+0.03
	(前置報告書)	2.92	3.96	4.12	4.56	4.68
		—	+1.04	+0.16	+0.44	+0.12
判断	審査書類①	3.02(4)	4.10(3)	4.18(3)	4.63(4)	4.68(3)
		—	+1.08	+0.08	+0.45	+0.05

	(拒絶査定)	3.00	4.32	4.12	4.68	4.80
		—	+1.32	-0.20	+0.56	+0.12
	(拒絶理由通知書)	3.32	4.28	4.48	4.72	4.76
		—	+0.96	+0.20	+0.24	+0.04
	(補正却下の決定)	3.12	4.04	4.28	4.68	4.64
		—	+0.92	+0.24	+0.40	-0.04
	(前置報告書)	2.64	3.76	3.84	4.44	4.52
		—	+1.12	+0.08	+0.60	+0.08
その他	審査書類①	3.04(3)	4.12(2)	4.21(2)	4.70(1)	4.69(2)
		—	+1.08	+0.09	+0.49	-0.01
	(拒絶査定)	2.88	4.12	4.20	4.64	4.60
		—	+1.24	+0.08	+0.44	-0.04
	(拒絶理由通知書)	3.20	4.28	4.36	4.84	4.84
		—	+1.08	+0.08	+0.48	±0
	(補正却下の決定)	3.24	4.04	4.13	4.54	4.58
		—	+0.80	+0.09	+0.41	+0.04
	(前置報告書)	2.84	4.04	4.16	4.76	4.72
		—	+1.20	+0.12	+0.60	-0.04

各文種の審査書類①全体の平均レベルには、文種間の相対順位（第 1 位～第 4 位）を補記した。上掲したとおり、「その他」に属する文の内容伝達レベルは無学習時は審査書類①全体で 4 文種中第 3 位（最下位は「判断」と下位であり、「認定」や「対比」に比べて人間の目からも若干品質が劣っていると映ることがわかる。ただし、コーパス学習後は「その他」の相対順位はむしろ上位となっており、また最終学習後には「その他」を含めた 4 文種全てがレベル 4.66～4.73 と同水準の範囲に収まっており、文種間の品質差、翻訳難易度の差は特に見いだせない。むしろ無学習時に比べて各文種間のスコア差は小さくなっており、各文種ともコーパスの学習により同程度の品質に均質化されたといえる。

本集計結果は各文種とも 25 文という少量のサンプルによるものであり、文種間の全般的な傾向を示すものとして十分な信頼性を持つとは言い難い。より多くのサンプル数を用いた自動評価結果のほうが実態に近いという可能性はある。

だが、仮に文種間で多少の翻訳難易度の差異があったとしても、特定の文種のみ学習データを作成するためには、（本調査分析で行ったように）まず個々の文の文種を目視で特定

して選別する必要がある。この選別作業には特許審査に関する知識をもつ作業者が必須であり、文種を意識せずにコーパスを作成するよりも高いコストを要する。少なくとも、上記人手評価結果は、サンプルは少量ながら各文種とも同等の品質に達しており、あえて高いコストを費やして特定の文種の学習データを補充する必然性は薄いと結論される。

3.4.7. 審査書類①に残存する課題

審査書類①システムに残存する機械翻訳品質上の課題として、以下の二点を指摘する。

3.4.7.1. カギカッコで括られた原文において翻訳文が途絶するエラー

原文全体がカギカッコで括られている場合に、機械翻訳文が文頭から 254～255 文字に達した時点で途絶してしまうエラーである。エラーのカテゴリ別カウントで「欠落エラー」の一種として検出された（⇒3.3.5.3.7.）。第 3 回学習後以降、原文が下記のいずれかの条件に該当し、かつ、機械翻訳文が 254 文字以上となる場合に発生し、最終学習後も解消されていない。異なる学習データを用いた審査書類②システム（第 4 章で後述）でも同様に検出されており、「みんなの自動翻訳」自体のロジックに起因するものと推察される。

- ・タイプ A：冒頭に中点（・）が打たれ、直後の本文全体がカギカッコで括られ、その後にカッコ補記が存在する

・「前記動作状態は、…該スキャナの走査ヘッドの定位置である」（請求項 3 の第 1～2 行）（何を表しているのかが明確になっていない。特に、…何をさしているのかが不明である。）

- ・タイプ B：上記タイプ A と同様だが、カッコ補記を伴わない

・「すなわち、本願新請求項 1 の発明は、走行体にカーブを走行させることなく、…給電部を配置することが技術的な特徴です。」

- ・タイプ C：本文全体がカギカッコで括られ、その後にカッコ補記が存在する

「交換機または回線に呼が集中すると過負荷となり、通信網が異常となって、疎通が不可能な輻輳状態になる。…このような通信網の品質低下を抑えるために輻輳時の対策が必要である。」（段落 0 0 0 2）

- ・タイプ D：上記タイプ C と同様だが、カッコ補記を伴わない

「よって、本願発明では、モータの電氣的時定数を転流間隔時間の 1／6 にするための…請求項 1 の従属項として定義したものである。」

審査書類には出願書類や先行技術文献などからの引用文も多く、引用であることを示すためにカギカッコで文を括ることも多用される。このため悪影響は広範囲に及ぶ。翻訳文が途中で断絶するという重大なエラーである。

3.4.7.2. カンマで区切られた複数の数字が合計 3 つ以上のときに連結されるエラー

原文中で複数の数値がカンマで区切られて並び、数字が合計で 3 つ以上である場合に、3 桁以上の 1 つの数値として誤訳されてしまうエラーである。例えば「請求項 1 4, 1 6」であれば、Claims 14, 16 ではなく、カンマで区切られた 4 つの数字 1、4、1、6 が連結されて Claim 1416 と 1 つの数値として連結される。エラーのカテゴリ別カウントで「数値エラー」の一種として検出された (⇒3.3.5.3.8.)。

大規模アップデート直後の第 11 回学習後に突如発生したエラーであり、条件に合致する文では確実に発生していることから、こちらも「みんなの自動翻訳」自体のロジックに起因するエラーと見なされる。

本エラーも最終学習後も解消されていない。カンマで区切られた数値の列挙は審査書類では普通に使用されており、影響範囲が広いエラーである。

4. 審査書類②システムの評価結果とその分析

本章では、審査書類②システムの機械翻訳品質の評価結果についてまとめる。

4.1. 審査書類②システムの評価の概要

審査書類②システムは、審査書類①と同じく「みんなの自動翻訳」の特許 NT エンジンに基づとするが、学習データである「審査書類コーパス B」には審査書類対訳コーパスのうち「意見書(A53)」由来の文のみを用いている。つまり、本システムは意見書専用にチューニングしたシステムである。

システムの学習は、審査書類コーパス B を 1 万文対作成する都度段階的に実施し、計 18 回、最終的には約 18 万文対⁴⁷を学習させた。審査書類①と比較して、学習データ量は各回とも 10 分の 1 となる。

機械翻訳文の評価手法は審査書類①と同一であり、BLEU 及び RIBES による自動評価と、評価者による人手評価とを併用した。評価のタイミングも同時で、自動評価は学習の都度、人手評価は学習 5 回おき（データを 5 万文対学習する都度）に実施した。

審査書類②の評価用コーパスには、意見書由来の審査書類対訳コーパスの初回作成分 10,000 文対から 2,000 対文を選定し、これを各学習段階の審査書類②システムに供して機械翻訳文を取得した。人手評価は評価用コーパスから 100 文対を選定して対象とした。

審査書類②は、「みんなの自動翻訳」の特許 NT エンジンに基づとし、これに審査書類対訳コーパスを段階的に学習させていくという点で審査書類①と共通する。各回の学習・評価のタイミングも同時である。ただし、審査書類①に属する 4 種の書類がいずれも特許庁審査官の作成する書類であったのに対し、審査書類②が対象とする意見書は出願人の作成する書類である。このため、記載内容や文体、使用される語句や表現等に異なる部分があると考えられる。加えて、審査書類②は学習データ量においても審査書類①とは大きな差があり、これらの要因によって機械翻訳文の品質や傾向が審査書類①と異なっている可能性がある。

この観点のもと、本項では審査書類②の品質評価結果について、前章に示した審査書類①システムの評価結果との比較を交えて分析・考察する。

⁴⁷ 初回学習時のみ、1 万文対のうち 2,000 文対を評価用コーパスとして選定しており、これらを除いた 8,000 文対を学習させた。一方、最終学習時には通常より多い 11,830 文対を学習させた。このため学習データ総量はおよそ 18 万文（179,830 文）となった。

4.2. 審査書類②システムの自動評価結果

本項では、審査書類②システムに対して実施した自動評価の結果を記す。

4.2.1. 審査書類②全体の自動評価結果

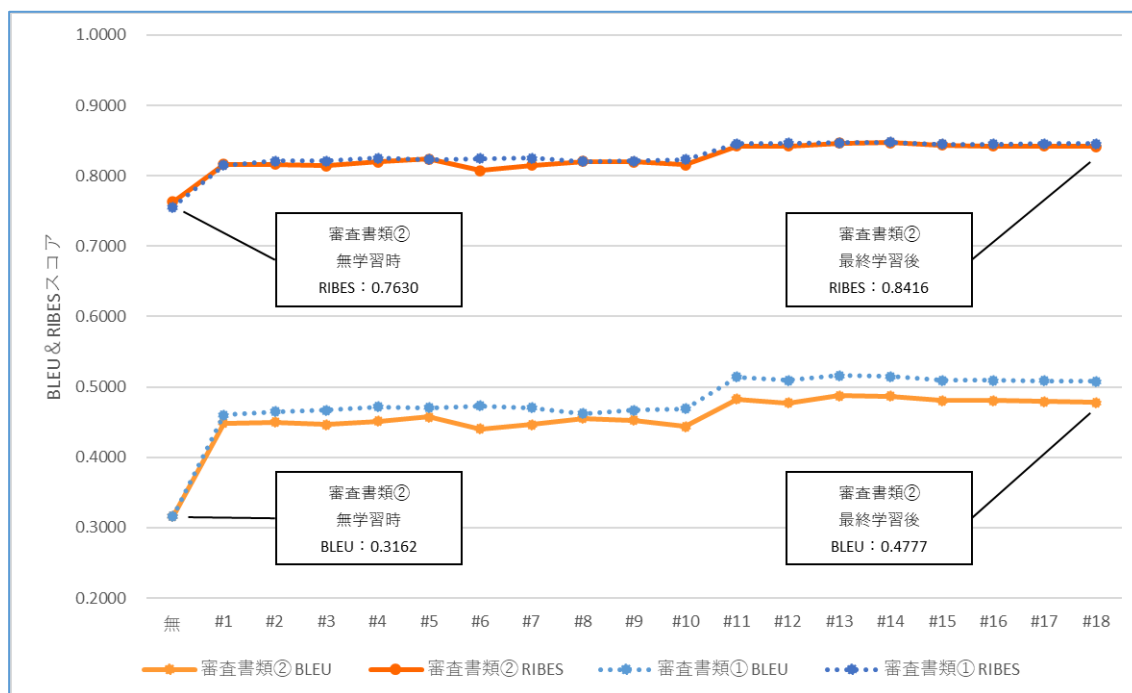
審査書類②の自動評価は、無学習時の1回と、審査書類コーパスBを約1万文対ずつ段階的に学習させる都度の18回、計19回実施した。評価の回数、各回の学習・評価の実施タイミングとも審査書類①と同じである。

下表に、審査書類②の評価用コーパス2,000文の自動評価スコア（BLEU、RIBES）の全件平均値を示す。

表 4.2.1. 審査書類②（2,000 文）の BLEU/RIBES 全件平均スコア

学習回数	学習データ量	バージョン	BLEU		RIBES	
無学習時	0	3.8_190930	0.3162	前回比	0.7630	前回比
第1回学習後	8,000	3.8_190930	0.4483	+0.1321	0.8163	+0.0533
第2回学習後	18,000	3.8_190930	0.4496	+0.0013	0.8162	-0.0001
第3回学習後	28,000	3.8_200131	0.4463	-0.0033	0.8143	-0.0019
第4回学習後	38,000	3.8_200131	0.4515	+0.0052	0.8198	+0.0055
第5回学習後	48,000	3.8_200131	0.4577	+0.0062	0.8238	+0.0040
第6回学習後	58,000	3.8_200131	0.4402	-0.0175	0.8076	-0.0162
第7回学習後	68,000	3.8_200131	0.4465	+0.0063	0.8151	+0.0075
第8回学習後	78,000	3.8_200531	0.4555	+0.0090	0.8203	+0.0052
第9回学習後	88,000	3.8_200531	0.4524	-0.0031	0.8200	-0.0003
第10回学習後	98,000	3.8_200531	0.4434	-0.0090	0.8155	-0.0045
第11回学習後	108,000	3.9_200930	0.4825	+0.0391	0.8426	+0.0271
第12回学習後	118,000	3.9_200930	0.4775	-0.0050	0.8422	-0.0004
第13回学習後	128,000	3.9_210131	0.4874	+0.0099	0.8466	+0.0044
第14回学習後	138,000	3.9_210131	0.4870	-0.0004	0.8468	+0.0002
第15回学習後	148,000	3.9_210531	0.4808	-0.0062	0.8438	-0.0030
第16回学習後	158,000	3.9_210531	0.4809	+0.0001	0.8425	-0.0013
第17回学習後	168,000	3.9_210930	0.4796	-0.0013	0.8426	+0.0001
第18回学習後	179,830	3.9_210930	0.4777	-0.0019	0.8416	-0.0010
スコア向上度（無学習時と第18回後の差）			+0.1615		+0.0786	

図 4.2.1. 審査書類②（2,000 文）の BLEU、RIBES 全件平均スコア推移



審査書類②の自動評価の全件平均スコアは、無学習時と最終学習後の比較で BLEU が +0.1615、RIBES が +0.0786 といずれも大きく向上した。自動評価結果の全体平均スコアの推移が実際の翻訳品質をほぼ正確に反映していることは審査書類①で検証されており、審査書類②においても審査書類①と同様に審査書類対訳コーパスの学習によって翻訳品質が顕著に改善したと見なせる。

学習回ごとのスコアの推移も審査書類②は審査書類①とほぼ同様であった。初回学習後には BLEU が +0.1321、RIBES が +0.0533 と全期間中最大の向上度を示し、それぞれ全期間トータルの向上分の約 82%、68%を占めた。審査書類①はそれぞれ約 75%、66%となっており、両システムはおおむね合致している。

続く第 2 回から第 10 回学習後にかけては BLEU が -0.0049、RIBES が -0.0008 と微減した。この期間、審査書類①のスコア推移は BLEU が +0.0090、RIBES が +0.0082 と微増しており、厳密には両システムの動向は一致しないが、どちらも 9 回の学習を経たことを考えれば誤差レベルの微動にとどまっており、事実上の横ばい状態という点で同傾向ともいえる。

その後、直前に「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）が実施された第 11 回学習後に再び顕著な向上（BLEU：+0.0391、RIBES：+0.0271）を見せ、以降も最終

学習回である第 18 回学習後まで常時そのスコア水準が維持された。この学習後期における審査書類②の一連のスコア推移も審査書類①と同じである。つまり、審査書類②の自動評価スコアは、全学習期間にわたって審査書類①とほぼ同期した推移を辿っている。

審査書類②は、審査書類①に比べて学習データ量が約 10 分の 1 と少なく、このため学習量の不足により翻訳品質が大きく劣る懸念があった。しかしながら、自動評価スコアの水準や動向を見る限り、両システムが得た学習効果に極端な違いはないように見受けられる。下表に、主要な学習回における両システムの評価スコア及び学習データ量の比較結果を示す（なお、表中の「スコア向上度」は各回とも無学習時との比較とした。）。

表 4.2.2. 審査書類②と審査書類①システムのスコア比較

		無学習	第 1 回	第 10 回	第 11 回	第 18 回
BLEU	審査書類②	0.3162	0.4483	0.4434	0.4825	0.4777
		スコア向上度	+0.1321	+0.1272	+0.1663	+0.1615
	審査書類①	0.3160	0.4602	0.4692	0.5142	0.5082
		スコア向上度	+0.1442	+0.1532	+0.1982	+0.1922
	①とのスコア差	+0.0002	-0.0119	-0.0258	-0.0317	-0.0305
RIBES	審査書類②	0.7630	0.8163	0.8155	0.8426	0.8416
		スコア向上度	+0.0533	+0.0525	+0.0796	+0.0786
	審査書類①	0.7558	0.8150	0.8232	0.8458	0.8456
		スコア向上度	+0.0592	+0.0674	+0.0900	+0.0898
	①とのスコア差	+0.0072	+0.0013	-0.0077	-0.0032	-0.0040
学習データ	審査書類②	0	8,000	98,000	108,000	179,830
	審査書類①	0	92,000	992,000	1,092,000	1,810,685
	①との差	—	-84,000	-894,000	-984,000	-1,630,855

※「スコア向上度」は各回とも無学習時との比較を示した。

上表に示したとおり、審査書類②の最終学習後の自動評価スコアは確かに両指標とも審査書類①より低く、特に BLEU は-0.0305 と大きな差が生じている（RIBES は-0.0040）。無学習時からの最終学習後にかけてのスコア向上度も、BLEU では審査書類①の+0.1922 に対して審査書類②は+0.1615 (-0.0307)、RIBES では+0.0898 に対して+0.0786 (-0.0112) と、いずれも審査書類②のほうがやや鈍い。

ただし、前記したとおり、両システムの自動評価スコアの推移自体は全学習期間を通じてほぼ完全に同期していた。つまり、第 2 回～第 10 回後にかけてと第 11 回～第 18 回後に

けては、審査書類①だけでなく審査書類②も自動評価スコアはほぼ横ばいの現状維持状態となっていた。

仮に両システムの最終的なスコア差が学習データ総量の極端な違い、つまり、審査書類②の学習データ量の不足によるものであるならば、審査書類②のスコアはこれらの期間にも一定の向上傾向を示すのが自然である。だが、実際のスコア推移はそうっておらず、このため両システム間のスコア差を学習データ量の差異によるものとは断定できない。むしろ最終学習後までの一連のスコア推移からは、仮に審査書類②の学習データをさらに増量したとしても、審査書類②の評価スコアは引き続き横ばい傾向が続くものと予測される。

そうなると、審査書類①と審査書類②との間の BLEU スコアの差異は学習データの多寡によるものではなく、それぞれが翻訳対象とした書類自体の翻訳難易度の違いによる可能性がある。ただし、もう一方の自動評価である RIBES では両システムのスコア水準に大きな差異は生じておらず（最終学習後での比較で-0.0040）、実際には両システム間に大きな品質差は存在していない可能性もある。

こうした種々の可能性については、自動評価スコアの集計結果からは検証は困難である。よって、ひとまず結論は保留し、後述する実例の定点観測や人手評価結果などの状況を踏まえて改めて考察する。

4.2.2. 技術分野間の差異

審査書類②の評価用コーパス各文の自動評価スコアを「化学」、「電気物理」、「機械」の分野ごとに集計した結果を下表に示す。

表 4.2.3.審査書類②技術分野ごとの BLEU、RIBES 平均スコア

学習回	全件		化学		電気物理		機械	
	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
無学習	0.3162	0.7630	0.3114	0.7661	0.3346	0.7732	0.3069	0.7519
第 1 回	0.4483	0.8163	0.4525	0.8240	0.4741	0.8237	0.4237	0.8025
第 2 回	0.4496	0.8162	0.4522	0.8213	0.4716	0.8283	0.4296	0.8013
第 3 回	0.4463	0.8143	0.4441	0.8180	0.4725	0.8268	0.4283	0.8005
第 4 回	0.4515	0.8198	0.4525	0.8278	0.4716	0.8248	0.4348	0.8075
第 5 回	0.4577	0.8238	0.4530	0.8311	0.4816	0.8319	0.4441	0.8096
第 6 回	0.4402	0.8076	0.4386	0.8115	0.4641	0.8182	0.4233	0.7952
第 7 回	0.4465	0.8151	0.4475	0.8254	0.4719	0.8225	0.4257	0.7985
第 8 回	0.4555	0.8203	0.4524	0.8263	0.4809	0.8340	0.4391	0.8034
第 9 回	0.4524	0.8200	0.4523	0.8314	0.4760	0.8288	0.4342	0.8011
第 10 回	0.4434	0.8155	0.4441	0.8237	0.4676	0.8212	0.4237	0.8024
第 11 回	0.4825	0.8426	0.4806	0.8513	0.5096	0.8511	0.4634	0.8267
第 12 回	0.4775	0.8422	0.4758	0.8507	0.5025	0.8494	0.4599	0.8277
第 13 回	0.4874	0.8466	0.4854	0.8571	0.5137	0.8549	0.4690	0.8290
第 14 回	0.4870	0.8468	0.4898	0.8587	0.5070	0.8519	0.4685	0.8302
第 15 回	0.4808	0.8438	0.4811	0.8532	0.4995	0.8490	0.4658	0.8298
第 16 回	0.4809	0.8425	0.4856	0.8525	0.4982	0.8477	0.4624	0.8278
第 17 回	0.4796	0.8426	0.4804	0.8539	0.4987	0.8486	0.4639	0.8259
第 18 回	0.4777	0.8416	0.4810	0.8536	0.4957	0.8447	0.4603	0.8264
向上	+0.1615	+0.0786	+0.1696	+0.0875	+0.1611	+0.0715	+0.1534	+0.0745

審査書類①で技術分野ごとの集計結果を分析した際、技術用語の翻訳品質は「みんなの自動翻訳」の基礎学習データに委ねられる部分が大きいと述べた（⇒3.2.4.）。さらに、審査書類①では各分野の集計結果に顕著な差が見られなかったことから、「みんなの自動翻訳」における各技術分野の翻訳精度はおおむね同等であると結論した（⇒3.4.3.）。この結論に照らせば、審査書類①と共通の基礎学習データを用いている審査書類②においても、各技術分野の翻訳精度は同等であるはずである。

この観点から上表の集計結果を見ると、無学習時及び最終学習後の各分野の平均スコアは、スコアの水準、そして全期間を通じたスコアの伸び幅（最下行）とも、おおむね全分野の平均値と同等であり、特異な数値を示す分野は存在しない。分野間のスコアのバラつきは審査書類①に比べると若干大きい、これは審査書類①の集計対象が 8,000 文であったのに対して審査書類②は 2,000 文と少数であり、数値のならされ方が粗いためであると考えられる。本集計結果からは、審査書類②においても各技術分野の翻訳品質はおおむね同等であり、翻訳精度が著しく劣る分野は存在しないといえる。

さらに、最終学習後の技術分野間の相対順位を見ると、BLEU が「電気物理」→「化学」→「機械」、RIBES が「化学」→「電気物理」→「機械」と両指標で順位が異なっており、かつ、審査書類①における相対順位（BLEU が「機械」→「電気物理」→「化学」、RIBES が「電気物理」→「化学」→「機械」）とも食い違いが多い。仮に「みんなの自動翻訳」において対応度が著しく劣る技術分野が存在するとすれば、その分野は書類を問わず常に最下位となるであろうことを考えると、このようにシステム間で相対順位がばらついている⁴⁸こと自体、それが「みんなの自動翻訳」自身の各技術分野への対応度の差によるものではなく、各システムの評価用コーパスにおける偶然の偏りの結果であることを示している。

以上、本集計結果の分析からは、審査書類②においても分野間のばらつきは想定される範囲内であった。したがって、審査書類①の分析で得た結論、すなわち「みんなの自動翻訳」における各技術分野の翻訳精度はおおむね同等であり、特に精度が劣る分野は存在しないという結論は審査書類②においても肯定された。

4.2.3. 文の種類による差異

審査書類②システムの対象である意見書は、審査官の手による書類ではないため「認定」「対比」「判断」といった文種に属する文は原則として存在しない。このため意見書では文種の判定を行っておらず、文種別のスコア集計も実施していない。

⁴⁸ 審査書類①及び審査書類②では「機械」分野が最下位となることが多かったが、次章にまとめた審決③／④システムでは「機械」分野が最上位となることが多かった（審決③の RIBES のみ第 2 位）。詳細は第 5.2.2 項参照。

4.2.4. 文単位のスコア向上／低下カウント結果

審査書類②の評価用コーパス 2,000 文を対象に、学習前後で自動評価スコアが向上した文、低下した文、変化しなかった文それぞれのカウント結果を下表に示す。表の右欄には参考値として審査書類①全 8,000 文のカウント結果を、比較を容易にするため 2,000 文に換算して掲載した。

表 4.2.4. 審査書類②文単位スコア向上／不変／低下カウント結果

	審査書類②						審査書類①【2,000 文換算】					
	BLEU			RIBES			BLEU			RIBES		
	向上	不変	低下	向上	不変	低下	向上	不変	低下	向上	不変	低下
第 1 回	1,636	5	359	1,411	6	583	1,683	7	310	1,443	10	548
第 2 回	1,019	43	938	966	56	978	998	71	931	1,006	81	913
第 3 回	948	62	990	947	73	980	988	73	939	975	80	945
第 4 回	998	74	928	1,000	79	921	990	80	930	995	94	911
第 5 回	1,028	75	897	1,034	77	889	940	101	959	928	112	960
第 6 回	855	62	1,083	842	73	1,085	971	106	923	935	115	950
第 7 回	1,010	84	906	993	98	909	912	121	967	922	130	948
第 8 回	1,032	88	880	1,038	95	867	900	93	1,007	936	105	960
第 9 回	913	85	1,002	894	88	1,018	992	83	925	923	94	984
第 10 回	909	91	1,000	931	94	975	966	105	929	960	117	923
第 11 回	1,246	44	710	1,231	47	722	1,281	76	643	1,241	84	675
第 12 回	697	529	774	680	544	776	661	591	748	664	603	733
第 13 回	966	257	777	948	268	784	893	353	754	876	363	761
第 14 回	733	557	710	713	563	724	678	618	705	674	631	696
第 15 回	845	264	891	820	280	900	775	338	887	768	344	889
第 16 回	704	536	760	714	551	735	738	562	701	724	576	700
第 17 回	806	326	868	826	336	838	759	414	827	760	422	819
第 18 回	692	601	707	677	609	714	688	622	691	707	628	665

上表に示したとおり、審査書類②における各回学習前後での文単位の変化の状況は、全学習期間にわたり審査書類①とほぼ同様であった。初回学習後（第 1 回）及び大規模アップデート直後（第 11 回）を除き「向上」と「低下」がほぼ均衡していることや、大規模アップデート前の各回では全体の 95%以上の機械翻訳文が学習の都度変化し、大規模アップデートを経た第 12 回学習後以降も全体の 3 分の 2 程度の文が引き続き変化し続けたことなど、審査書類①に見られた主な傾向は全て審査書類②にも当てはまる。審査書類①の調査分析

では、学習の進捗にかかわらず多数の機械翻訳文が毎回変化する原因は、学習データそのものではなく「みんなの自動翻訳」の訳語学習ロジックにあると論じたが、学習データの量や内容が大きく異なる審査書類②でも全く同じ状況が見られたことで、この考察が妥当であることが裏付けられた。

なお、上掲の集計結果により、審査書類②においても、最終学習後に至るまで毎回少なくとも全体の 3 分の 2 の機械翻訳文が変化し、そのうちの半数程度はスコアが低下していることが示された。これを額面通りに捉えると、学習の都度少なくとも全体の 3 分の 1 の文で翻訳品質が低下していることになる。この懸念は審査書類①においては人手評価結果との照合により払拭されたが（⇒3.3.5.2.）、審査書類②も同じく人手評価結果を用いて検証しておく必要がある。この検証は第 4.3.2.2.項で行う。

4.2.5. 実例を用いた機械翻訳結果の定点観測

審査書類②システムによる各回学習後の機械翻訳文の変化の状況を、実例を用いて定点観測する。

4.2.5.1. 実例 4-1 : 「意見書」由来、文番号 A53-#94 の分析

審査書類②の評価用コーパスは全件「意見書 (A53)」由来の文である。最初の実例にはここから文番号#94 をピックアップした。まずは原文と正解訳文、そして無学習時と初回学習後の機械翻訳文を示す。

【文番号 : A53-#94】

原文	当該補正によって、引用文献 3 に基づく拒絶理由 1 も解消したものと 思料致します。	
正解訳文	The applicant believes that the reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
無学習	It is assumed that the rejection reason 1 based on the reference document 3 is eliminated by the correction.	
	—	BLEU : 0.1206 (－) RIBES : 0.8569 (－)
第 1 回学習後	It is also considered that the reason for refusal 1 based on Cited Document 3 has been eliminated by the amendment.	
	向上	BLEU : 0.6751 (+0.5545) RIBES : 0.9376 (+0.0807)

審査書類①で対象とした「拒絶査定」等の 4 種の審査書類が全て特許庁審査官により作成されるものであるのに対し、「意見書」は書き手が出願人である。このため「ですます調」で書かれることが多く⁴⁹、丁寧語や謙譲語も多用される傾向にある。本例文末の「思料致します」という表現もその一例であり、意見書では多用されるが、他の審査書類では減多に使用されない。

英語にもフォーマルな文体はあるが、日本語の「ですます調」や各種丁寧語・謙譲語のように通常語と全く表記が異なるような丁寧表現は存在しない。このため、機械翻訳精度の観点からは、こうした日本語の丁寧表現を（通常の表現に読み替えて）正しく解釈できるかどうか焦点となる。審査書類②の翻訳品質は、特にこの点において審査書類①よりも難易度が高く、機械翻訳品質上の差異が生じる可能性がある。

⁴⁹ 審査官が作成した書類でも「ですます調」が使用されることもある（例：A02-#36、#265、#286 など）。また、「ですます調」で書かれた意見書の文がそのまま引用される場合もある。

したがって、本例でまず着目すべき箇所も前述の「**思料致します**」となる。単なる「です
ます調」の文であれば日本語を取り扱う機械翻訳エンジンなら対応できて当然だが、「**思料**」
という表現はさほど一般的な語ではなく、必ずしも十分な文例が「みんなの自動翻訳」の基
礎学習データ中に存在しているとは限らない。

正解訳文ではこの箇所を、主語を補い **The applicant believes** としている。原文は主語無
し文であるが、「**思料致します**」を訳出するうえでは主語を配すほうが自然な英文となる。
文の内容から主語は出願人と特定できるので、この補足は適切であり、学習データとしても
問題ない。

一方、無学習時の機械翻訳文ではこの箇所は **It is assumed** と訳されている。**assumed** は
ニュアンスとしては「～と推定される」となり、原文が出願人自身の主張であることを考え
るとやや当事者感を欠く印象はあるが、一般的な語義での「思料」の訳語としては妥当であ
る。よって、この無学習時の結果から、「みんなの自動翻訳」はデフォルト状態でも「です
ます調」を問題なく処理でき、かつ、「思料する」という語句も解釈可能であることがわか
る。

なお、無学習時の機械翻訳文では主語に形式主語 **It** を使用している。この点に関しては、
この文単独で考えれば正解訳文のように **applicant** と具体的な主語を補足するほうが分かり
易いのは事実である。ただし、意見書全文を翻訳する場合、主語無し文にその都度 **applicant**
と具体的な主語を補足するのは過剰であり、文脈から主語が理解できる場合は形式主語 **it**
を用いるほうが自然なことも多い。本例も、意見書であることを前提として読めば形式主語
it でも意味は十分に通じる。

続いて、意見書由来の審査書類対訳コーパスの初回学習を経た機械翻訳文を見ると、「**思
料致します**」の箇所は **It is also considered** に変化した。「**思料**」にあたる **consider** は「～と
考える」の意味であり、無学習時の **assume** よりは当事者感がある。正解訳文の **believe** ほ
どの確信は感じさせないが、出願人の主張・意見であって最終的な判断は審査官が行うこと
を考えると、このように穏当な表現を用いるのも一手であろう。ただし、**also** に関しては、
「拒絶理由1も」に相当するものであり、この配置では誤りである。正しくは **the reason for
refusal 1...has also been eliminated** となる。

なお、本例には「補正」「引用文献3」「拒絶理由1」と審査書類用語が三語含まれている
が、それぞれ **correction** → **amendment**、**reference document 3** → **Cited Document 3**、
rejection reason1 → **reason for refusal 1** と、意見書由来のコーパスの学習により適訳語に
改善された。いずれも審査書類では頻出の語であるが、少なくともこうした頻出語であれば、

意見書に限定され、かつ、8,000 文と比較的少量の学習データでも十分にカバーされることが示された。

こうした各所の改善により、初回学習後の翻訳文は、also の配置誤りという軽微な不備を除けば、原文の内容がほぼ正確に反映されるものとなった。無学習時の翻訳文でも文意はほぼ伝わるが、前述のとおり「思料」のニュアンスがやや逸れており、各特許用語も適訳語となっていない点で劣る。このため、初回学習後の翻訳文は学習前より品質が「向上」していると判定される。

以下、第 2 回学習後から最終学習回である第 18 回学習後まで各回の機械翻訳文を順に見ていく。

【文番号：A53-#94】

第 2 回学習後	By the amendment, it is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved.	
	向上	BLEU : 0.5853 (-0.0898) RIBES : 0.6824 (-0.2552)

「思料致します」は引き続き it is considered と適切に訳されている。前回時に誤っていた also の配置も正され、「解消した」の訳語も前回の eliminated から正解訳文どおりの最適訳 resolved となって、翻訳精度はさらに向上した。機械翻訳文としてほぼ上限のレベルと考えられる。なお、「当該補正によって」にあたる By the amendment が文末から文頭に移動し、当時の「みんなの自動翻訳」の特徴である文構造の不安定な変化が本例でも発生しているが、内容理解上は完全に同義であり優劣はない。

【文番号：A53-#94】

第 3 回学習後	We consider that the Reason for Refusal 1 based on Cited Document 3 has also been resolved by this amendment.	
	同等	BLEU : 0.6364 (+0.0511) RIBES : 0.9485 (+0.2661)
第 4 回学習後	It is considered that the Reasons for refusal 1 based on Cited Document 3 have also been resolved by the amendment.	
	同等	BLEU : 0.5607 (+0.0004) RIBES : 0.9219 (-0.0342)
第 5 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
	悪化	BLEU : 0.8456 (+0.2849) RIBES : 0.9622 (+0.0403)

第 3 回学習後の翻訳文は、主語に形式主語 it ではなく we が初めて採用された。意見書の

書き手（本例では出願人）が単数であるか複数であるかは案件により異なるが、それは正解訳文の applicant にもいえることである。複数形概念のない日本語を英訳する場合、背景情報なしで単数複数を確定することは難しく、機械翻訳で的確な訳し分けを行うことは容易ではない。この単複の相違を不問とすれば、形式主語 **it** よりも出願人自身を指す **we** のほうがこなれた英文といえるが、品質上の判定は優劣無しとした。

続く第 4 回～第 5 回学習後の翻訳文では、再び形式主語 **it** を用いた文構造に復帰した。主語無し原文に補足される主語が学習前後で変化するという状況は審査書類①で分析したいくつかの実例、例えば実例 3-1 における **it** と **they** の間の変化（⇒3.2.7.1.）や実例 3-4 における **it** と **invention** の間の変化（⇒3.2.7.4.）などでも見られ、「みんなの自動翻訳」の傾向の一つといえる。ただし、本例に関しては、これまでのところ主語として補われる語の変化による品質の優劣は生じていない。なお、第 5 回学習後の翻訳文では **also** の欠落が発生しており、この点で品質はやや「悪化」した。

続いて、第 6 回から第 10 回学習後までの機械翻訳文を示す。

【文番号：A53-#94】

第 6 回学習後	We consider that the Reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	向上	BLEU : 0.7618 (-0.0838) RIBES : 0.9622 (±0)
第 7 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
	悪化	BLEU : 0.8456 (+0.0838) RIBES : 0.9622 (±0)
第 8 回学習後	It is considered that the Reason for Refusal 1 based on Cited Document 3 has also been resolved by this amendment.	
	向上	BLEU : 0.6053 (-0.2403) RIBES : 0.9376 (-0.0246)
第 9 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	同等	BLEU : 0.7246 (+0.1193) RIBES : 0.9511 (+0.0135)
第 10 回学習後	The applicant believes that the reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
	同等	BLEU : 1.0000 (+0.2754) RIBES : 1.0000 (+0.0489)

上掲のうち第 6 回から第 9 回学習後までの翻訳文では、これまでと同じような変化が繰り返された。第 6 回学習後には再び主語に **We** が採用され、前回欠落していた **also** も再び訳出されている。第 7 回から第 9 回学習後にかけては再び主語が **It** に戻り、第 7 回では **also**

の欠落も再発している。また、第 8 回のみ amendment に付随する冠詞が this に変化した。いずれも、審査書類①の各実例で見られたのと同種の、当時の「みんなの自動翻訳」における訳語の不安定性や、突発的な「欠落エラー」の発生／解消がもたらした変化である。

これに対し、第 10 回学習後の翻訳文では、それまでの変遷から逸脱した大きな変化が見られた。主語に突如 The applicant が採用され、かつ、「思料致します」の訳語も初回学習後から採用され続けていた consider から初めて believe に変化して、正解訳文と完全に同一の翻訳文となった。

本事業で作成する審査書類対訳コーパスでは同一文は事前に排除しており（⇒1.2.1.）、学習データ中に正解訳文と完全一致するデータは存在しない。したがって、この翻訳文は「みんなの自動翻訳」が学習データから間接的に導き出したものとなる。英文は主語が必須であるため「みんなの自動翻訳」でも主語無し原文に主語を補足して翻訳しているケースはよく見られるが、多くの場合、補足される主語は形式主語や文中の文言の繰り返しであった。だが、本例はそのいずれでもなく、applicant（出願人）という、原文中に存在しない具体的な名詞が補足されている。状況的には、意見書由来の学習データでは applicant を主語として補足している訳例も存在するため、「みんなの自動翻訳」がこれを学習したものと推察される。

ただし、この変化は一回限りの単発的なものであった。以下、第 11 回学習後以降の各回の翻訳文を列挙する。

【文番号：A53-#94】

第 11 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	向上	BLEU : 0.7246 (-0.2754) RIBES : 0.9511 (-0.0489)
第 12 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	同等	BLEU : 0.7246 (±0) RIBES : 0.9511 (±0)
第 13 回学習後	It is considered that Reason 1 for refusal based on Cited Document 3 has also been resolved by this amendment.	
	悪化	BLEU : 0.3502 (-0.3744) RIBES : 0.9187 (-0.0324)
第 14 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
	同等	BLEU : 0.8456 (+0.4954) RIBES : 0.9622 (+0.0435)

第 15 回学習後	It is considered that reason for refusal 1 based on Cited Document 3 has been resolved by the amendment.	
	同等	BLEU : 0.7561 (-0.0895) RIBES : 0.9554 (-0.0068)
第 16 回学習後	It is considered that reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	向上	BLEU : 0.6678 (-0.0883) RIBES : 0.9485 (-0.0069)
第 17 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	同等	BLEU : 0.7246 (+0.0568) RIBES : 0.9511 (+0.0026)
第 18 回学習後	It is considered that the reason for refusal 1 based on Cited Document 3 has also been resolved by the amendment.	
	同等	BLEU : 0.7246 (±0) RIBES : 0.9511 (±0)

第 11 回学習後以降の機械翻訳文は、大規模アップデートを経た後の「みんなの自動翻訳」による翻訳文となる。審査書類①で定点観測した各実例では、大規模アップデートを境に機械翻訳文の安定化が進み、学習前後で変化が生じたとしても小規模・局所的、かつ、実質的な品質変化を伴わない同義内の変化にとどまる傾向がうかがえた。本例の第 11 回学習後以降の変遷もこの傾向に完全に合致している。

まず各回で採用される形式主語は **it** で安定し、第 3、6 回学習後の **we** や第 10 回学習後の **applicant** が採用されることはなくなった。具体的な主語の補足は、適切に行われれば文意理解を助けるが、機械翻訳には難易度が高く、不適切な主語を採用して読み手をミスリードするリスクも高い。それを考えると、形式主語 **it** が安定的に採用されるようになったことは、より安全な方向への変化と評価できる。

第 10 回学習後に **believe** に変化した「思料致します」の訳語も再び **consider** に戻り、この訳語で完全に安定した。こちらも本例に限れば **believe** は適訳だが、本来の意味は「信じる、確信する」であり、やや意識でもある。一般的には「思料」の訳語としては **consider** のほうが穏当であり、本例においても **consider** で特段の問題はない。

文の骨格構造に関しても、第 11 回学習後以降は各回とも **It is considered** から始まる形で完全に安定し、例えば第 2 回学習後のように By the amendment が文頭に配置されるようなこともなくなった。結果、各回の翻訳文の変化箇所は、also が訳出されているか否かと、reason for refusal 1 に冠詞 the が付されるか否かの二箇所のみに局所化された。具体的には、第 11、12、17、18 回学習後が「also あり、the あり」、第 14 回が「also なし、the あり」、

第 15 回が「also なし、the なし」、第 16 回が「also あり、the なし」となっている。なお、第 13 回学習後のみは「拒絶理由 1」が reason for refusal 1 ではなく **Reason 1 for refusal** と突発的に変化した。この付番位置でも意味は通るが、審査書類対訳コーパス中でこの位置に付番されることはない。審査書類では単に「理由 n」とする記載も多く、これに対応する Reason n という訳語がこの回に限り採用されてしまったものと考えられる。こうした突発的な訳語変化や、各回における the や also の出没状況を見ると、こうした訳語の不安定性は、かなり抑制されたとはいえ、「みんなの自動翻訳」本来の性質として大規模アップデート以降も残存していることがわかる。

以上、本例では、意見書特有といえる「ですます調」の文体や「**思料**」という丁寧語への対応度を中心に分析を行ったが、前者は無学習時から、後者もコーパス初回学習後から適切に翻訳されており、翻訳品質上の問題は一切見られなかった。その他の部分における変化の状況はおおむね審査書類①と共通しており、特許・審査書類用語の早期からの安定と一般用語や文構造における不安定な訳語変化、そして大規模アップデート以降のランダム性の小規模化・局所化とニューラル翻訳方式の本質としての訳語のランダム性の残存など、全ての面で両システムの傾向は同期していた。実際の機械翻訳品質もコーパス学習初期から高いレベルで安定しており、本例においては、審査書類②の翻訳品質が審査書類①に比べて劣ることや、学習データが審査書類①に比べて不足していることを示す事象は存在しなかった。

4.2.5.2. 実例 4-2 : 「意見書」由来、文番号 A53-#55 の分析

続いての実例も、意見書特有の丁寧表現に着目して選定した。まずは原文、正解訳文及び初回学習前後の機械翻訳文を示す。

【文番号：A53-#55】

原文	つきましては、本願を再応ご審査の上、本発明は、特許すべきとのご査定を賜わりたくお願い申し上げます次第である。	
正解訳文	Accordingly, it is respectfully requested that the application be re-examined, and the invention be granted a patent.	
無学習	Indeed, the present invention is based on the review of the present application, and the present invention is based on 賜わり's desire to assess the patent to be claimed. 申し上げます.	
	—	BLEU : 0.0544 (－) RIBES : 0.6287 (－)
第 1 回学習後	In response to the above, it is respectfully requested that the present application be reexamined and the examiner's decision to grant a patent be received.	
	向上	BLEU : 0.2723 (+0.2179) RIBES : 0.8409 (+0.2122)

本例の原文は全体にわたって非常に丁寧な表現が用いられている。その結果、無学習時の「みんなの自動翻訳」では正しい構文解析ができず、「賜わり」と「申し上げます」が「未知語」となっている。「賜わる⁵⁰」や「申し上げます」という語句自体は尊敬語・謙譲語としてはさほど特殊な語句とはいえ、単独であれば「みんなの自動翻訳」の基礎学習データ中にも存在している可能性が高いが、「賜わりたくお願い申し上げます次第である」という極端に丁寧な言い回しに組み込まれた文例、つまり周辺の語句が類似する文例が存在せず、学習に支障をきたした結果、「未知語」扱いとなったものと推察される。

これに対し、初回学習後の翻訳文では、この「賜わりたくお願い申し上げます次第である」という一連の丁寧表現を正解訳文と同じく respectfully requested と二語でシンプルに訳出している。かつ、「再応ご審査」も reexamine と正しく解釈して一語で訳すなど、意見書由来のコーパスを学習した効果が十分に発揮されている。文末の the examiner's decision to grant a patent be received は「本発明は、特許すべきとのご査定」に相当し、正解訳文とは全く異なる訳し方でやや回りくどい印象も受けるが、少なくとも文意は類推可能である。

⁵⁰ 「賜わる」の送りがなは「賜る」が正しいが、「賜る」でも「未知語」扱いされている文が存在する。詳しくは第 4.2.5.3. 項を参照のこと。

無学習時に対応できなかった丁寧表現が学習後に正しく訳出されるようになった以上、この改善は審査書類対訳コーパスの学習の結果と見なせる。また、本例ほどの丁寧な表現は審査官の手による審査書類ではまず使われることはないため、コーパスの作成対象として意見書を含めていたことが奏功したといえる。

続いて、第2回から第5回学習後までの機械翻訳結果を列挙する。

【文番号：A53-#55】

第2回学習後	Therefore, it is respectfully requested that the present application be granted a patent after reexamination of the present application.	
	同等	BLEU : 0.4298 (+0.1575) RIBES : 0.8870 (+0.0461)
第3回学習後	Accordingly, it would be respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.1764 (-0.2534) RIBES : 0.8492 (-0.0378)
第4回学習後	Therefore, it is respectfully requested that the application be reexamined and the present application be reexamined and an examiner's decision to grant a patent be received.	
	悪化	BLEU : 0.3242 (+0.1478) RIBES : 0.8336 (-0.0156)
第5回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	向上	BLEU : 0.3061 (-0.0181) RIBES : 0.8579 (+0.0243)

第2回学習後の翻訳文は前回から構造が大きく変化し、原文の前半と後半とを逆の順序で訳出するとともに、「本願を再応ご審査の上」のくだりを after reexamination of the present application とやや意識している。それでも意味は類推できるが、前回の翻訳文のほうが再審査を依頼していることがわかり易く、この箇所に関してはやや悪化している。その反面、前回時に若干回りくどいと評した「特許すべきとのご査定を賜わり」の箇所は present application be granted a patent と明解な訳文に改善された。品質判定は、両者の差し引きで「同等」とした。

続く第3回学習後は、前半部分で Accordingly→Therefore, it would be→ it is など同義の範囲内の表層的な変化があり、後半部分は初回学習後とほぼ同じ翻訳文に戻った（相違点は examiner's decision の冠詞が the から an になったのみ）。事実上、初回学習後と同じ内容であり、品質は引き続き「同等」である。

第4回学習後の翻訳文も初回学習後の翻訳文を踏襲しているが、「本願を再応ご審査の上」の箇所が **the application be reexamined and the present application be reexamined** と、同じ内容を二度繰り返す形になってしまっている。審査書類①でも大規模アップデート以前に頻発していた、文中の語句を繰り返すタイプの突発的な「湧き出しエラー」(⇒3.3.5.3.6.)に該当し、このエラーによって品質は「悪化」している。

第5回学習後の翻訳文では前回の「湧き出しエラー」は解消し、冒頭の接続詞と examiner's decision の冠詞の違いを除いて初回学習後と同一の翻訳文となった。品質判定は、前回時のエラーが解消して初回～第3回学習後と同レベルに復帰したことにより「向上」である。

極端な丁寧表現で翻訳難易度が高いと予想された「**賜わりたくお願い申し上げる次第である**」と「再応ご審査」は、第2回から第5回学習後にかけても respectfully requested、reexamined と初回学習後と同じ訳語が安定して採用され続けた。こうした極度の丁寧表現は、基礎学習ではカバーされておらず、意見書由来の審査書類対訳コーパスで初めて学習されるものが多いと考えられる。その意味では多くの審査書類用語と同じ条件であり、他の訳語候補が存在しないため、コーパス学習後は適訳語で安定する傾向を有しているものと推察される。

続いて、第6回から第10回学習後までの翻訳結果を見ていく。

【文番号：A53-#55】

第6回学習後	Therefore, it is respectfully requested that the present application be reexamined and the examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3208 (+0.0147) RIBES : 0.8739 (+0.0160)
第7回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3357 (+0.0149) RIBES : 0.8739 (±0)
第8回学習後	Therefore, it is respectfully requested that the present application be reexamined and the examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3208 (-0.0149) RIBES : 0.8739 (±0)
第9回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (-0.0147) RIBES : 0.8579 (-0.0160)

第 10 回学習後	Therefore, the present invention is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	悪化	BLEU : 0.1970 (-0.1091) RIBES : 0.7243 (-0.1336)

第 6 回学習後の機械翻訳文は examiner's decision の冠詞が前回時の an から the に変化したのみである。この語の冠詞は頻繁に変化しているが、本例に関してはどちらであっても内容理解上の優劣はない。続く第 7 回～第 9 回学習後にかけても、変化箇所はこの冠詞と、第 7 回時に reexamined が re-examined とハイフン入りに変化したことのみであり、各回とも実質的な翻訳品質は「同等」である。

これに対し、第 10 回学習後は、これまで形式主語 it で安定していた主語が「本発明：present invention」に変化した。もともと原文には「本発明は」という文言が「本願：present application」とは別に存在しており、これが初めて訳出されたことになる。しかしながら、実際の機械翻訳文を読めば明らかなように、こうして原文どおり「本発明」を訳出すると英文として非常に不自然で、the present application との関係性も把握しにくくなる。このため、品質判定は「悪化」とした。本例の場合、「本発明」を無理に訳出しようとする英文が非常に冗長となるため、前回までの翻訳文のように省略するほうが適切であり、正解訳文もそうになっている。

このように第 10 回学習後のみ大きな変化があったが、それ以外の各回における翻訳文の傾向は第 5 回時までの延長線上にあり、文構造や言い回しには微細な変化が常時発生している反面、丁寧表現に関しては引き続き高い安定を見せている。

続いて、大規模アップデートを経た第 11 回学習後以降の翻訳文を示す。

【文番号：A53-#55】

第 11 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	向上	BLEU : 0.3061 (+0.1091) RIBES : 0.8579 (+0.1336)
第 12 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (± 0) RIBES : 0.8579 (± 0)
第 13 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (± 0) RIBES : 0.8579 (± 0)

第 14 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (±0) RIBES : 0.8579 (±0)
第 15 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (±0) RIBES : 0.8579 (±0)
第 16 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (±0) RIBES : 0.8579 (±0)
第 17 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (±0) RIBES : 0.8579 (±0)
第 18 回学習後	Therefore, it is respectfully requested that the present application be reexamined and an examiner's decision to grant a patent be received.	
	同等	BLEU : 0.3061 (±0) RIBES : 0.8579 (±0)

大規模アップデートを経た第 11 回学習後の機械翻訳文は、前回時の突発的な変化から再び前々回時の翻訳文に復帰し、以降、最終学習後（第 18 回）まで 8 回連続で同一の機械翻訳文が出力された。それまで学習の都度何らかの変化が生じていたことを考えると、アップデートを境に傾向が一変したといえる。懸案の丁寧表現「賜わりたくお願い申し上げる次第である」「再応ご審査」は大規模アップデート以降も適訳で安定し、また、第 4 回や第 10 回学習後のような突発的なエラーや不適訳の採用も一切生じなくなった。こうした翻訳文の安定化は審査書類①でも見られたが、学習データの規模、内容ともに異なる審査書類②でも全く同じタイミングで同様の安定化がなされていたことで、これが主に大規模アップデート時のエンジン本体の改良、具体的には訳語学習ロジックの改良によってもたらされたものであることが確実となった。

以上、意見書特有の高度な丁寧表現が多用された実例を用いて、審査書類対訳コーパス学習前後の機械翻訳文の変遷を観測した。本例においては、無学習時の機械翻訳文では丁寧表現にうまく対応できておらず、一部の尊敬語・謙譲語が「未知語」扱いされていたが、コーパス学習後は全ての丁寧表現が安定して適訳語で翻訳されるようになった。本例の観測結果により、「みんなの自動翻訳」に意見書由来のコーパスを学習させることで、難易度の高い丁寧表現にも十分に対応可能となることが示された。

なお、本例の丁寧語以外の部分での各回の機械翻訳文の内容や品質、変化の状況はいずれ

も審査書類①と共通の傾向・推移を示しており、学習データ量の不足を感じさせるような事象も見られなかった。

4.2.5.3. 無学習時の機械翻訳文で「未知語」が検出された実例の分析

前例（実例 4-2）では、無学習時に丁寧表現に起因する「未知語」（「賜わり」「申し上げる」）が発生していた。「みんなの自動翻訳」の場合、解析不可能な語は翻訳文から欠落してしまうことも多く、必ずしも「未知語」として出力されるとは限らないが（⇒3.3.5.3.7.）、少なくとも「未知語」となった語句はその時点では翻訳不能語である可能性が高い⁵¹。

そこで、審査書類②で使用した意見書由来の評価用コーパス 2,000 文の無学習時の機械翻訳文において「未知語」となっていた語句を調査し、そこから意見書特有の丁寧表現と見なしたものをピックアップした。以下、一覧で示す。

表 4.2.5. 審査書類②で無学習時に「未知語」扱いとなった丁寧表現

未知語	発生文	解消状況
したがいまして	#232 ほか	この語を含む 17 文全てで「未知語」扱いとなったが、初回学習後に全て解消し、in other words, therefore など妥当に訳出されるようになった。なお、別表記「従いまして(#714 ほか)」は「未知語」扱いされず。
～におきまして	#748 ほか	この語を含む 13 文全てで「未知語」扱いとなったが、初回学習後に全て解消。in など妥当に訳出されるようになった。
賜る／賜わる	#1708 ほか	この語を含む一部の文で「未知語」扱い。初回学習後に全て解消した。丁寧表現においては訳出不要な語である。
～ざるを得ません	#620 ほか	この語を含む 3 文全てで「未知語」扱いとなったが、初回学習後に全て解消。訳出が省略された回もあるが、おおむね must、should be を用いた妥当な訳出がされるようになった。なお「ざるを得ず (#312)」は「未知語」扱いされず。
～に過ぎません	#217 ほか	この語を含む 4 文全てで「未知語」扱い。初回学習後に全て解消したが、merely／only と訳出されたのは第 3 回学習以降であった。
何とぞ	#270	1 文のみで使用され、第 2 回学習後に「未知語」扱いが解消された。別表記「何卒 (#34)」も無学習時は訳出されず翻訳文から欠落していたが、初回学習後に解消。特段の訳出は不要な語

⁵¹ 他の文では訳出できているが、その文における使われ方や周辺の語などの状況によって「未知語」となるケースも考えられる。

		である。
決してござい ません	#1244	「ございません」の使用例は他にも 3 例（#165 など）があるが、「決してございません」という形で用いられた本例のみ「てございません」が「未知語」扱いされた。第 8 回学習後に解消。
適示する	#591	「示す」の丁寧表現。初回学習後に解消し indicate と訳出されるようになった。
参りました 参ります	#511 #680	いずれも最終学習後まで「未知語」扱いが解消せず。

上表に示した状況から、意見書特有の丁寧表現の一部がデフォルトの「みんなの自動翻訳」では翻訳不能であったことや、それが意見書由来のコーパスの学習によって改善されていたことが見て取れる。つまり、審査書類②には、意見書特有の丁寧表現という、デフォルトの「みんなの自動翻訳」では対応できず、したがって、審査書類①よりも翻訳難易度が高くなる要因が存在していたことになる。また、本調査結果から、その大半は意見書由来のコーパスの学習によって最終学習後までに順次解決されたこともわかる。

なお、上表に示した中で最終学習後まで「未知語」扱いが解消されなかった唯一の丁寧表現が最下段の「参りました／参ります」であった。この語を含む A53-#511 と #680 の最終学習後の機械翻訳文を示す。

【文番号：A53-#511】

原文	以上述べて参りましたように、請求項 1－6 は、特許を受けることができる状態であると思慮致しますので、再度ご審査のうえ、特許査定を賜りたくお願い申し上げます。	
正解訳文	As stated above, it is believed that the invention of Claim 1-6 is in a patentable state; therefore, it is respectfully requested to re-examine the amended application to grant a patent.	
第 18 回学習後	As described above and in 参り, it is respectfully requested that the application be reexamined and an examiner's decision to grant a patent be received since it is considered that Claims 1-6 are patentable.	
	—	BLEU : 0.2167 RIBES : 0.6550

【文番号：A53-#680】

原文	以下各請求項について具体的に述べて参ります。	
正解訳文	I would like to describe the following claims specifically.	
第 18 回学習後	The following is a specific description of each claim and 参り.	
	—	BLEU : 0.0750 RIBES : 0.7071

両例とも「未知語」となっているのは「**て参り**」の部分であり、これに付随して前者は「and in **て参り**」、後者は「and **て参り**」という余計な英語も出力されてしまっている。「**て参り**」という丁寧語自体は特段の意味をもたないため、上二例もこの部分を見れば原文どおりの内容を読み取れるのだが、「**て参り**」という語句を訳文中に無理に組み込もうとした結果、余計な接続詞や前置詞が用いられ、内容理解をさらに阻害する結果となっている。

こうしたごく一部の例外はあるが、本調査により、無学習時に「未知語」扱いされていた丁寧表現の大部分は、意見書由来のコーパスの学習によって早期に改善されていたことが確認できた。ただし、今回「未知語」として検出された丁寧表現の多くは比較的一般的な丁寧語であり、「みんなの自動翻訳」の性質を考えると、さらに特殊・高度な丁寧表現は「未知語」とならず訳文から欠落している可能性も否定できない。したがって、意見書特有の丁寧表現への対応が万全であるか否かは、人手評価結果の分析において引き続き検証する必要がある。

4.3. 審査書類②システムの人手評価結果

本項では、審査書類②システムに対して実施した人手評価の結果を記す。

4.3.1. 人手評価の概要

審査書類②の人手評価は、評価用コーパス 2,000 文のうち 100 文のみを対象とした。自動評価が審査書類対訳コーパス（審査書類コーパス B）を約 1 万文対学習させる都度実施したのに対し、人手評価は約 5 万文対の学習ごと、つまり、第 5 回学習後、第 10 回学習後、第 15 回学習後の 3 回と、これに無学習時及び最終学習時（第 18 回学習後）の 2 回を加えた計 5 回のみ実施した。

人手評価の観点及び人手評価者の人選は審査書類①と同様である（⇒3.3.2.、3.3.3.）。

4.3.2. 各評価項目の集計結果

計 5 回にわたり実施した審査書類②の人手評価の結果を本項にまとめる。審査書類①と同様に内容伝達レベルの評価、学習前後の翻訳文の相対評価、エラーのカテゴリ別カウントの三つの評価観点で評価を行っており、本項ではこれらの観点ごとに各回の集計結果を示し、分析・考察していく。

4.3.2.1. 「内容伝達レベル」の評価

内容伝達レベルの評価では、審査書類②の人手評価対象 100 文それぞれの各学習段階における機械翻訳文の翻訳品質を、以下の 5 段階で絶対評価した。

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

下表に、各学習段階の機械翻訳文 100 文に付された内容伝達レベルの平均値と、レベルごとの分布を示す。

表 4.3.1. 各回学習前後の審査書類②の「内容伝達レベル」集計結果

書類種別	学習	平均	5	4	3	2	1
審査書類②	無学習時	2.99	7	30	31	19	13
	第 5 回学習後	3.89	28	47	15	6	4
	第 10 回学習後	4.10	51	23	14	9	3
	第 15 回学習後	4.68	78	14	6	2	0
	第 18 回学習後	4.70	79	14	5	2	0
	向上度／増減	+1.71	+72	-16	-26	-17	-13
[参考] 審査書類① ※100 文換算	無学習時	3.14	14	24	31	23	8
	第 5 回学習後	4.09	48	26	16	8	2
	第 10 回学習後	4.20	53	23	15	7	2
	第 15 回学習後	4.67	77	14	7	2	0
	第 18 回学習後	4.69	78	15	5	2	0
	向上度／増減	+1.55	+64	-9	-26	-21	-8

審査書類②の内容伝達レベルは、無学習時の 2.99 から最終学習後（第 18 回）には 4.70 へと+1.71 ポイント向上した。表下段に示した審査書類①と比較すると、審査書類②は無学習時から第 10 回学習後まではやや低かったが、第 15 回学習後以降は同レベルとなり、最終学習後（第 18 回学習後）も審査書類①の 4.69 に対し審査書類②は 4.70 と完全に同等の内容伝達レベルが付されている。自動評価では、審査書類②の BLEU スコアが審査書類①よりも有意に低く、このため審査書類②の翻訳精度が劣っている懸念があったが(⇒4.2.1.)、より正確と見なせる本評価結果により、最終学習後の審査書類②の機械翻訳品質は審査書類①と同等と考えるのが妥当である。

なお、審査書類②の無学習時の自動評価スコアは審査書類①とほぼ同等（BLEU）か若干上回る（RIBES）数値となっていたが、内容伝達レベルの評価では審査書類①の 3.14 に対して審査書類②は 2.99 と、やや低くなった。自動評価が評価用コーパス全体（2,000 文）を対象としたのに対して人手評価は 100 文のみを対象としており、必ずしも双方の評価結果が一致するとは限らないが、無学習時の審査書類②は、意見書で多用される丁寧表現の一部に対応できていなかった(⇒4.2.5.3.)。このことから、審査書類②の評価用コーパスの翻訳難易度は審査書類①よりもやや高かったと考えられる。したがって、無学習時の内容伝達レベルが審査書類①に比して低くなっている本評価結果が正しい可能性が高い。

意見書特有の丁寧表現のうち、無学習時に「未知語」として出力されていた丁寧語の大多数は審査書類対訳コーパスの学習によって改善されたことが確認されたが、より難易度の高い丁寧表現に関しては、そもそも「未知語」としても出力されず、翻訳文から欠落してし

まっている懸念もあった。だが、審査書類②の内容伝達レベルは学習後期には審査書類①と同等に迫いついており、難易度の高い丁寧表現がコーパス全件学習後も多数欠落し、それが全体品質を損ねているという懸念は小さくなった。

つまり、本評価結果からは、審査書類②の翻訳対象である意見書は、丁寧表現の多用という特徴により他の審査書類に比べて翻訳難易度はやや高いが、意見書由来のコーパスを十分に学習させることで他の審査書類と同等の高精度な機械翻訳文が得られると結論できる。

また、審査書類②の各学習段階における内容伝達レベルの推移と、第 4.2.1.項に示した自動評価の全体平均スコアの推移とはおおむね同期した。これは審査書類①と同じ状況であり、本調査分析の基礎とした自動評価結果が十分な量のサンプルの平均を採った相対評価という用途において十分な信頼性を有するということが、審査書類②の評価結果から改めて実証された。

なお、審査書類②における自動評価スコアと内容伝達レベルそれぞれの推移における唯一の相違点として、第 5 回学習後から第 10 回学習後にかけて自動評価スコアはやや低下（BLEU：-0.0143、RIBES：-0.0083）していたのに対し、内容伝達レベルは+0.21 と向上しており、実際にはこの期間にも一定の品質改善効果が生じていたことが示された。この不一致は、審査書類①でも同じく発生している。

4.3.2.2. 学習前後の翻訳文の相対評価

審査書類②の評価対象文 100 文について、学習前後の機械翻訳文のどちらの品質が優れているかを文ごとに比較し、学習後の機械翻訳品質が学習前に比べて「向上」したか「劣化」したか、又は「同等」であるかを判定した。下表に各回の集計結果を示す。

表 4.3.2. 審査書類②各回学習前後の翻訳品質の「相対評価」結果

書類種別	比較対象		学習後に		
	学習前	学習後	向上	同等	劣化
審査書類②	無学習時	第 5 回学習後	71 文	19 文	10 文
	第 5 回学習後	第 10 回学習後	23 文	53 文	24 文
	第 10 回学習後	第 15 回学習後	45 文	52 文	3 文
	第 15 回学習後	第 18 回学習後	9 文	83 文	8 文
[参考] 審査書類① ※100 文換算	無学習時	第 5 回学習後	70 文	20 文	10 文
	第 5 回学習後	第 10 回学習後	25 文	58 文	17 文
	第 10 回学習後	第 15 回学習後	38 文	60 文	2 文
	第 15 回学習後	第 18 回学習後	7 文	89 文	4 文

審査書類②における学習前後の相対評価結果は、各回とも表下段に示した審査書類①の評価結果と同様の比率となり、最終学習後においては学習前後で品質が「同等」と判定された文が 83 文と大勢を占めた。

「みんなの自動翻訳」は学習前後で機械翻訳文が変化する傾向が強く、審査書類②においても、最終学習時まで毎回少なくとも全体の 3 分の 2 の文が学習前後で変化し、その半数の文では自動評価スコアが低下していた(⇒4.2.4.)。このため、学習の副作用として翻訳品質が低下する文が多数発生している懸念が生じていた。

同様の状況であった審査書類①の分析では、本評価で学習前後の翻訳品質が「同等」と判定された文が全体の 89%と大勢を占めたことを根拠に、自動評価スコアの変動は機械翻訳文の表層的な変化によるものが主であり、スコアが示すような大量の文の品質低下は実際には生じていないと結論した(⇒3.3.5.2.)。審査書類②に関しても、学習前後の品質が「同等」である文が最終的には 100 文中 83 文と大勢を占め、実際に品質が「劣化」した文は 8 文と僅かであることから、審査書類①に関する上記結論が等しく当てはまる。つまり、審査書類②においても、学習後期まで発生し続けた機械翻訳文の変化の大半は、実質的な翻訳品質に影響しない同義の範囲内での変化であったと考えてよい。

4.3.2.3. エラーのカテゴリ別カウント

人手評価の一環として、評価対象文 100 文それぞれについて、各回の機械翻訳文に含まれるエラーを 10 種に分類してカウントした (⇒3.3.5.3.)。以下、全 10 種のエラー別の各回のカウント結果を示す。

表 4.3.3.審査書類②学習前後のカテゴリ別エラー該当文カウント結果一覧

書類種別	無学習	第 5 回	第 10 回	第 15 回	第 18 回
特許誤訳	61	13	8	4	5
	—	-48	-5	-4	+1
(参考) 審査書類①	72	14	9	4	3
技術／一般誤訳	31	20	25	5	7
	—	-11	+5	-20	+2
(参考) 審査書類①	27	14	14	5	6
未知語	4	0	0	0	0
	—	-4	±0	±0	±0
(参考) 審査書類①	1	1	0	0	0
係り受けエラー	9	5	9	10	8
	—	-4	+4	+1	-2
(参考) 審査書類①	21	18	18	11	10
英語表現エラー	42	39	33	24	23
	—	-3	-6	-9	-1
(参考) 審査書類①	28	24	22	24	23
湧き出しエラー	18	26	22	2	2
	—	+8	-4	-20	±0
(参考) 審査書類①	19	20	24	1	1
欠落エラー	46	25	30	6	4
	—	-21	+5	-24	-2
(参考) 審査書類①	47	18	23	4	6
数値エラー	7	6	4	1	1
	—	-1	-2	-3	±0
(参考) 審査書類①	8	8	6	1	1
記号エラー	7	5	5	3	3
	—	-2	±0	-2	±0
(参考) 審査書類①	6	4	4	3	3

その他エラー	0	2	0	0	0
	—	+2	-2	±0	±0
(参考) 審査書類①	3	2	4	1	1

上表は各種エラーについて3段構成の表示とした。1段目には各回評価時のエラー発生文数、2段目には前回からの増減、そして3段目には参考として審査書類①の各評価回のエラー発生文数（100文換算）を掲載した。

審査書類②は、対象書類が意見書である点、そして学習データ量が約10分の1と少量である点において審査書類①と条件が異なる。こうした条件の違いによりエラーの発生状況に有意な差異が生じているか、両システムのカウント結果を対比することで把握することができる。結果は上表のとおり、審査書類②の各カテゴリのエラーの発生率及び改善状況は各回ともおおむね審査書類①と同様であり、両システム間のエラーの発生／改善状況に特段の差異は見られなかった。

以下、主なカテゴリのエラーの集計結果を両システムで比較する。

4.3.2.3.1. 「特許誤訳」

特許・審査書類用語の誤訳をカウントした「特許誤訳」は、審査書類対訳コーパスの学習効果が最も顕著に反映されるカテゴリといえる。したがって、審査書類②の学習データ総量が審査書類①に比べて大幅に少ないことが、エラーの発生文数や改善状況に影響している可能性がある。また、書類種別が意見書に変化したことによる学習効果の違いが表れている可能性もある。

しかしながら、両システムの「特許誤訳」の発生数は、上表に示したとおり、無学習時から最終学習後まで常時ほぼ同水準で推移した。審査書類②で最終学習後に残留した「特許誤訳」は5件のみであり、その内容も、「(発明の)構成要件」は直訳調の constituent requirement ではなく constituent element 又は constituent feature とすべきという指摘（A53-#42）や、「符号」は図面の符号と解釈される場合は symbols ではなく reference numerals とすべきという指摘（A53-#86）など、審査書類①と同じく高度な訳し分けが求められるケースが目立った。つまり、「特許誤訳」に関しては、発生数、改善の度合い、そして誤訳の傾向のいずれも両システム間で特段の違いはなく、審査書類②の対象書類が意見書であることや、学習データ量が少量であることによる悪影響は全く見られなかった。

この結果から、コーパスの学習を介して機械翻訳システムを審査書類特有の用語・表現に対応させるという本事業の主目的は、審査書類②においても十分に達成されていると結論

できる。

4.3.2.3.2. 丁寧表現に関係する各種エラー

両システム間の顕著な条件の違いとしては、意見書における丁寧表現の多用も挙げられる。事実、デフォルトの「みんなの自動翻訳」では「未知語」扱いされた丁寧表現も少なからず存在した(⇒4.2.5.3.)。審査書類②の各学習段階における丁寧表現への対応に不足があれば、それは「技術／一般誤訳」、「未知語」、「係り受けエラー」、「英語表現エラー」、「欠落エラー」などが審査書類①よりも多発するという形で表れると考えられる。

この観点から各エラーを見ていくと、まず「未知語」が100文中4文とやや多く(審査書類①は100文あたり1文程度)、このうち3文では丁寧表現が「未知語」となっていた(「したがいまして」×2、「に過ぎません」×1)。この結果は、意見書特有の性質に起因したエラーの多発と見なせる。これら3文の「未知語」はいずれもコーパス学習により早期に解消されたが、評価用コーパス全体では最終学習後も「未知語」として残存する丁寧表現も存在する(第4.2.5.3.項に示した「参りました」)。ただし、丁寧表現に関する「未知語」の発生率は無学習の時点でも3%(100文あたり3文)程度であり、かつ、その大半はコーパス学習の過程で解消されている。このためコーパス全件学習後の審査書類②において「未知語」が多数発生し、全体的な翻訳品質に重大な悪影響を及ぼす懸念は小さい。

丁寧表現がうまく解析できないことで、「係り受けエラー」や「英語表現エラー」が頻発する可能性も考えられる。だが、審査書類②の無学習時から第10回学習後にかけて「係り受けエラー」は審査書類①よりも少なく、丁寧表現による悪影響は見られない。一方、「英語表現エラー」はこの期間、審査書類①よりも明らかに多かったが、こちらも実際の該当文を見ると、丁寧表現に関連するものはほとんど見られなかった。こうした状況から、両エラーの多寡は丁寧表現とは無関係の、偶然の偏りであると考えられる。若しくは、「係り受けエラー」と「英語表現エラー」はどちらも判定できるケースも多いため、審査書類②ではそうしたケースが後者と判定される傾向が強かったという可能性もある。いずれにせよ、学習後期には「係り受けエラー」、「英語表現エラー」とも審査書類①と同等のカウント数となっており、両エラーとも丁寧表現の多用による特段の悪影響は見られない。

丁寧表現の多用が影響する可能性がある他種のエラー、具体的には「技術／一般誤訳」(丁寧語は一般用語に属する)や「欠落エラー」(「みんなの自動翻訳」では翻訳不能語が欠落しやすい)に関しても、全期間を通じて審査書類①と同程度のカウント数であり、丁寧表現の多用による悪影響は見られなかった。第4.2.5.3.項では、高度・特殊な丁寧表現が「未知語」ではなく「欠落エラー」として多発している懸念を述べたが、審査書類②で最終学習後に検

出された「欠落エラー」4 件中に丁寧表現に関わるものは存在せず、この懸念は払しょくされた。

これら審査書類②における各種エラーのカウント結果の推移から、意見書特有の丁寧表現の多用は、無学習時こそ「未知語」の多発など翻訳品質を下げる一因となるが、意見書由来の審査書類対訳コーパスの学習によりその大半は早期に改善され、以降は機械翻訳品質上、特に不利な条件とはなっていないことが確認された。

4.3.2.3.3. 上記以外の各種エラー

上記した以外の、学習データ量の多寡や丁寧表現の多用に直接関連しないと考えられる各種エラー、具体的には「湧き出しエラー」や「数値エラー」、「記号エラー」、「その他エラー」に関しても両システム間に差異は見られず、カウント数は全期間ほぼ完全に同等であった。いずれもエンジンのロジックに起因するタイプのエラーと見なされるものであり、その意味では、同じエンジンを使用した両システムで傾向が一致するのは順当といえる。

なお、審査書類①で最終学習後も残存していた特殊なタイプのエラー、具体的には「カギカッコで括られた原文において翻訳文が途絶するエラー」(⇒3.4.7.1.)と「カンマで区切られた複数の数字が合計 3 つ以上のときに連結されるエラー」(⇒3.4.7.2.)については、審査書類②の人手評価対象文 100 文には発生条件に該当するものが存在せず、検出されなかった。ただし、評価用コーパス全体ではそれぞれ A53-#834、A53-#62 などと同タイプのエラーが発生しており、これら 2 種のエラーが、学習データの内容に関わりなく発生する「みんなの自動翻訳」のロジック由来のエラーであることが確実となった。

4.4. 審査書類②システムに関するまとめ

審査書類②システムの機械翻訳の品質及び傾向について、自動評価及び人手評価の結果から以下の点が確認された。

4.4.1. 意見書で多用される丁寧表現の影響

審査書類①に属する 4 種の審査書類はいずれも特許庁審査官が作成する書類であったのに対し、審査書類②で翻訳対象とした意見書は出願人が審査官に宛てて作成する書類であり、このため丁寧表現が多用される傾向にある。こうした丁寧表現への対応が必要である点において、意見書は他の審査書類よりも機械翻訳難易度が高く、翻訳品質で劣る可能性があり得た。

実例を調査したところ、無学習時の審査書類②の機械翻訳文では丁寧表現に起因する「未知語」が多数検出され（⇒4.2.5.3.）、内容伝達レベルも審査書類①よりやや低かった。こうした状況から、デフォルトの「みんなの自動翻訳」は意見書特有の丁寧表現への対応が不十分であり、このため無学習時の審査書類②の機械翻訳品質は審査書類①にやや劣ることが示された。

ただし、その後の意見書由来の審査書類対訳コーパスの学習により、審査書類②で当初「未知語」扱いされていた丁寧表現の大部分は正しく訳出されるようになり、内容伝達レベル、各種エラーの発生件数とも学習後期には審査書類①と完全に同等となった⁵²。

こうした本調査分析の結果から、意見書で多用される丁寧表現は機械翻訳の難易度を高める一因ではあるものの、意見書由来の学習データを機械翻訳エンジンに十分に学習させることで対応可能であり、本事業で作成したコーパス全件を学習させることで、審査書類②においても審査書類①と同等の極めて高品質な機械翻訳文を得られることが確認された。

4.4.2. 学習データ量が比較的少量であることの翻訳品質への影響

審査書類②の学習に用いた意見書由来の学習データ（審査書類コーパス B）の総文数は 179,830 文であり、審査書類①の学習データ（同 A. 1,810,685 文）に比べ約 10 分の 1 と大幅に少ない。この学習データ量の多寡により、審査書類②の翻訳品質が審査書類①ほど改善されない懸念があった。

⁵² 自動評価 BLEU の全件平均スコアのみは学習初期以降も審査書類①とは有意な差が生じたが（最終学習後で-0.0287）、RIBES のスコアは両システムほぼ同等であり、人手評価の各観点においても審査書類②の品質が劣ることを示す状況は一切見られなかった。このため BLEU における差異は自動評価の精度に起因する誤差であると判断した。

審査書類対訳コーパスの学習量に不足があれば、それは主に審査書類用語の誤訳、すなわち「特許誤訳」の発生件数として表れるはずである。また、審査書類②の場合、丁寧表現への対応の不十分さとして顕出する可能性もある。そして、こうした不備が多発すれば、それは当然、内容伝達レベルにも反映される。

しかしながら、前項でも述べたとおり、審査書類②の内容伝達レベルや各種エラーの発生件数は学習後期には審査書類①と完全に同等となっており、「特許誤訳」の発生数も両システム間で多寡は生じなかった。丁寧表現への対応の不足によるエラーも学習後期にはほぼ見られなくなり、あらゆる面において、審査書類②の学習データ量の顕著な不足を示す事象は見られなかった。

学習データの効果は、その書類において頻出の用語であるほど早期から得られる。そして頻出の用語が一斉に改善されるため、学習初期ほど全体的な品質改善の度合いも大きくなる。これに対し、学習が進み、主要な用語が網羅的にカバーされるにつれ、学習データを増量させても、改善される語は出現頻度の少ないマイナーなものに限られるようになる。このため、全体に占める改善の度合いは極めて小さくなり、巨視的には現状維持と映る状態に移行していくことは避けられない。

審査書類②も、大規模アップデートが実施された第 11 回以降は学習データを増量しても自動評価スコア、内容伝達レベルとも現状維持状態となっており、審査書類①と同様、意見書で用いられる主要な特許・審査書類用語はおおむねカバーされた状態に達していることが見て取れる。

ただし、一般に学習データの総量が多くなるほど、そこに含まれる用語や表現のバリエーションは増加する。その意味では、第 11 回以降の学習データにも、それ以前の学習データではカバーされていなかったマイナーな用語や表現が一定量含まれていると考えられる。こうしたマイナーな用語・表現は無数に存在し、評価用コーパス中に含まれるのはそのごく一部にすぎないため、学習した内容が必ずしも本調査分析の評価結果に直結するとは限らないが、学習データを増量するほどマイナーな用語・表現に幅広く対応できるようになることが期待される。

4.4.3. 審査書類②の機械翻訳品質

本調査分析の結果から、審査書類対訳コーパス学習後の審査書類②の機械翻訳品質は審査書類①と同水準であると結論される。最終学習後に到達した内容伝達レベル 4.70 は審査書類①の 4.69 と完全に同等であり、かつ、機械翻訳として極めて高水準である。

一方、審査書類①において指摘した「カギカッコで括られた原文において翻訳文が途絶するエラー」(⇒3.4.7.1.)、「カンマで区切られた複数の数字が合計3つ以上のときに連結されるエラー」(⇒3.4.7.2.)は審査書類②においても同様に発生しており、両システム共通の課題となる。

5. 審決③／審決④システムの評価結果とその分析

本章では、審決③システム及び審決④システムの機械翻訳品質の評価結果についてまとめる。

5.1. 審決③／④システムの評価の概要

審査書類①／②システムと異なり、審決③／④システムは「審決」由来のコーパスを学習させたシステムである。具体的には、本事業で全文翻訳した審決（全 300 案件）に由来する「審決対訳コーパス（⇒1.2.2.）」及び過去事業で作成された既存の「作成済み審決コーパス」（特許庁貸与物）を「みんなの自動翻訳」の特許 NT エンジンに学習させている。なお、審決④システムには更に約 250 万文対の「審決逆翻訳コーパス」を追加学習させた（第 5.1.1. 項で後述）。

審決③／④の学習は、審決 100 案件分の審決対訳コーパスが完成する都度、全 3 回実施した。下表に、各学習回におけるコーパス別の学習データ量をまとめた。

表 5.1.1. 審決③／④システム各回の学習データ量

システム	学習	審決コーパス※	作成済みコーパス	審決逆翻訳コーパス ⁵³	総計
審決③	第 1 回学習	32,601 文対	56,964 文対	—	86,565 文対
	第 2 回学習	+35,089 文対			121,654 文対
	第 3 回学習	+31,498 文対			156,152 文対
審決④	第 1 回学習	32,601 文対	56,964 文対	2,521,792 文対	2,611,357 文対
	第 2 回学習	+35,089 文対		2,521,790 文対	2,646,444 文対
	第 3 回学習	+31,498 文対		2,521,826 文対	2,677,978 文対

※各回の新規データ数のみを示した。各回の学習データには過去回の分も含まれる。

審決③／④システムの品質評価は、自動評価・人手評価とも無学習時と上記 3 回の学習の都度、計 4 回実施した。評価用コーパスには、初回作成分（100 案件）の審決対訳コーパスから案件単位に 2,024 文を選定し、これを両システムで共用した。評価用コーパスを案件単位に選定した理由は、同一案件の一部の文のみを評価用コーパスに選定すると、残りの文

⁵³ 各回学習に用いる審決逆翻訳コーパスは、原文となる英文は全回同じものを用いたが、その原文を英日機械翻訳するために用いる「みんなの自動翻訳」は学習データ量やエンジンのバージョンが各回異なり（第 5.1.2. 項で後述）、このため各英文に対応する日本語翻訳文は各回同一とは限らない。また、各回とも英日機械翻訳処理時に若干数の翻訳不能文が発生したため、各回の文数にも僅かな相違がある。

は全て学習データに含まれることとなり、同じ案件由来の文が学習データ中に大量に存在するという、実用時に比して不当に有利な条件となってしまうためである。

なお、審査書類①／②では原出願が特許の審査書類のみを学習・翻訳の対象としていたのに対し、審決③／④では実用新案、意匠、商標の審決も対象としている⁵⁴。また、審査書類対訳コーパスが 40 文字～250 文字の非重複文で構成されるのに対し、審決対訳コーパスは審決全文を対象に作成しており、こうした限定はない。審決③／④の分析・考察にあたっては、こうした条件の違いも考慮する必要がある。

5.1.1. 審決逆翻訳コーパス

審決④システムの学習には、審決③と共通の学習データ（本事業で作成した審決コーパス及び過去に作成済みの審決コーパス）に加え、審決逆翻訳コーパスも併用した。両システムの品質を比較することで、逆翻訳を用いた機械学習（Back-Translation）の品質改善効果を検証できる。

審決逆翻訳コーパスは、海外特許庁（USPTO）の英語審決 pdf データから抽出した約 250 万文（2,527,476 文）の英文データを利用して作成した。具体的には、各回学習のタイミング、すなわち審決対訳コーパスを 100 案件分作成する都度、以下の手順で作成した。

- ① 審決③システムに対して学習を施す。
- ② 学習後の審決③システムを逆方向（英→日）に使用して、約 250 万文の英文データの機械翻訳文（日本語文）を得る。
- ③ 英文データ約 250 万文と、上記②で得た日本語文約 250 万文を対訳文対にアライメントして審決逆翻訳コーパスを作成する。

つまり審決逆翻訳コーパスは、各回の学習と同じタイミングで、計 3 回作成した。原文となる約 250 万文の英文データは毎回同一だが、上記②で逆方向に使用する審決③システムの学習量が毎回異なり、かつ、エンジンのバージョンもその時点の最新のものとなるため、各英文に対応する日本語翻訳文は同一とは限らない。また、各回とも上記②の英日機械翻訳処理時に若干数の翻訳不能文が発生したため、総文数も僅かに相違する。

⁵⁴ ただし、本事業でコーパスを作成した審決 300 案件には実用新案は含まれなかった。詳しくは第 1.3.4. 項を参照されたい。

5.2. 審決③／④システムの自動評価結果

本項では、審決③及び審決④システムの自動評価の結果を記す。

5.2.1. 審決③／④の自動評価結果

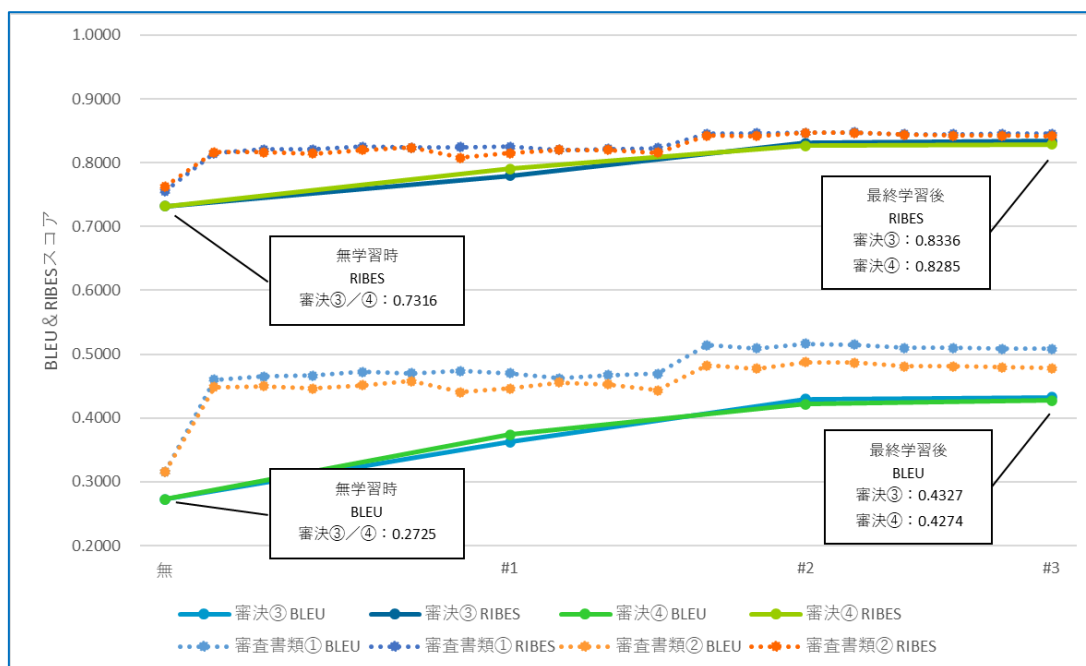
まずは審決③及び審決④それぞれについて、無学習時及び第 1 回～第 3 回学習後における BLEU と RIBES の全件平均スコアを示す。

表 5.2.1. 審決③／④（各 2,024 文）の BLEU／RIBES 全件平均スコア

学習回	審決対訳 コーパス数*	バージョン	審決③		審決④	
			BLEU	RIBES	BLEU	RIBES
無学習時	—	3.8_200131	0.2725	0.7316	0.2725	0.7316
第 1 回 学習後	86,565 (2,611,792)	3.8_200131	0.3627 +0.0902	0.7793 +0.0477	0.3740 +0.1015	0.7904 +0.0588
第 2 回 学習後	124,654 (2,646,444)	3.9_210131	0.4297 +0.0670	0.8310 +0.0518	0.4217 +0.0477	0.8269 +0.0364
第 3 回 学習後	156,152 (2,677,978)	3.9_210930	0.4327 +0.0030	0.8336 +0.0026	0.4234 +0.0017	0.8285 +0.0016
無学習時からのスコア向上			+0.1602	+0.1020	+0.1509	+0.0969

* 各回とも「作成済みコーパス」56,964 文を含む。なお、カッコ内の数字は審決④に使用した「審決逆翻訳コーパス」を含めた文数である。

図 5.2.1. 審決③／④（各 2,024 文）の BLEU、RIBES 全件平均スコア推移



5.2.1.1. 各回学習後のスコア変動

上表では各学習回の上段に評価用コーパス 2,024 文の全件平均スコア、下段に前回時からのスコア変動を示した。

無学習時の自動評価スコアは審決③と審決④で同値となっている。無学習の時点では、両システムともコーパスを学習させる前のデフォルト状態の「みんなの自動翻訳(特許 NT)」で同一の評価用コーパスを翻訳しており、得られる機械翻訳文も同一であるためである。

審決③の無学習時における BLEU スコア 0.2725、RIBES スコア 0.7316 は、いずれも審査書類①／②のものより有意に低い。例えば審査書類①との比較では BLEU が-0.0435、RIBES が-0.0242 とかなりの差が生じている。各システムとも「みんなの自動翻訳」をデフォルトで使用している状態での品質差であり、それぞれの評価用コーパス自体の翻訳難易度の差と見なせる。

審決対訳コーパスの初回学習を経た第 1 回学習後は、審決③、④とも評価スコアは大幅に向上した。無学習時との比較で、審決③では BLEU が+0.0902、RIBES が+0.0477、審決④も BLEU が+0.1015、RIBES が+0.0588 と、両者とも全学習回を通じて最大の向上を遂げている。ただし、審査書類①／②の初回学習後のスコアの向上度に比べると、伸びはやや鈍い（審査書類①の初回学習後の向上度は BLEU が+0.1442、RIBES が+0.0592。審査書類②は BLEU が+0.1321、RIBES が+0.0533）。

続く第 2 回学習後においても、審決③が BLEU で前回比+0.0670、RIBES で+0.0518、審決④も BLEU で+0.0477、RIBES で+0.0364 と、引き続き顕著なスコア向上を見せた。これは、中間に実施された「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）の影響が大きいと考えられる。

審決③／④は学習・評価の回数やタイミングが審査書類①／②と大きく異なっており、審決③／④の第 2 回学習のタイミングは、審査書類①／②の第 13 回学習とほぼ同時期となる。よって、審決③／④の第 2 回学習時の「みんなの自動翻訳」は、審査書類①／②の第 13 回学習時と同じく、既に大規模アップデートを経ている（⇒2.2.1.）。つまり、審決③／④では第 2 回学習前後の比較が、大規模アップデート前後の機械翻訳文の比較となる。

続く第 3 回学習が、審決③／④の最終学習回となる。最終学習後のスコアは審決③の BLEU が前回比+0.0030、RIBES が+0.0026、審決④では BLEU が+0.0017、RIBES が+0.0016 と、両システム、両指標とも過去 2 回と比べて微増にとどまった。大規模アップデート後の機械翻訳文同士の比較となるこの期間は、審査書類①／②でも現状維持状態が続いていた

が、審決③／④においても同様の状況、すなわち大規模アップデートを機に改善された翻訳品質がその後も維持された状態であることが見て取れる。

以上、審決③／④の各回学習後の自動評価スコアの推移を概観した。コーパスの初回学習後に最大のスコア向上が果たされ、大規模アップデート直後にもこれに準ずる顕著な向上があり、以降は最終学習後までその水準が維持された、という各学習段階におけるスコア推移は審査書類①／②と一致しており、審決③／④においても審査書類①／②と同様の品質改善がなされたことを示唆している。

ただし、審決③／④の全体平均スコアは無学習時から最終学習後に至るまで審査書類①／②よりも有意に低く、両者の翻訳品質には一定の差異があることが示された。また、審決③と審決④の比較においては、第 2 回学習後以降は審決逆翻訳コーパスを追加学習させていない審決③のほうが、学習させた審決④よりも若干高いスコアを得る結果となった。

以下、上述した審決③／④の自動評価結果について、「審決③と審決④のスコア比較」及び「審査書類①／②とのスコア差」の二観点から詳しく考察する。

5.2.1.2. 審決③と審決④のスコア比較

審決③と審決④との各回のスコアを比較すると、第 1 回学習後の時点では審決逆翻訳コーパスを追加学習させた審決④のスコアが有意に高く、審決③は審決④よりも BLEU では **-0.0113**、RIBES では **-0.0111** と両指標とも大きく下回っていた。だが、続く第 2 回学習後の自動評価では両者のスコアは逆転し、逆翻訳コーパスを学習させていない審決③のほうが審決④よりも BLEU では **+0.0080**、RIBES では **+0.0041** と、若干ながら上回った。最終学習回となる第 3 回学習後もこの状況は変わらず、審決③のほうが審決④よりも BLEU では **+0.0093**、RIBES では **+0.0051** と、引き続きやや上回る結果となった。

第 1 回学習後に審決④のスコアが審決③を上回っていたことから、逆翻訳コーパスにも一定の品質改善効果はあると考えられる。ただし、第 2 回学習後以降の評価結果から、それは審決対訳コーパスを十分に学習すれば同様に得られる性質のものであると見なせる。つまり、審決対訳コーパスを十分に学習させる前提であれば、審決逆翻訳コーパスの追加学習には特段の品質改善効果はないと結論される。

審決③／④において審決逆翻訳コーパスによる品質改善効果が生じなかった理由については、第 5.4.2 項で詳しく考察する。

5.2.1.3. 審査書類①／②とのスコア差

審決③／④の自動評価スコアは、無学習時から最終学習後に至るまで常時、審査書類①／②のスコアを下回った。各システムの無学習時及び最終学習後の自動評価スコアの比較表を以下に示す。

表 5.2.2. 審決③／④と審査書類①／②の自動評価スコア水準比較表

学習回	学習 コーパス数	自動 評価	無学習時		最終学習後	
			スコア	差異	スコア（向上度）	差異
審決③	156,152	BLEU	0.2725	—	0.4327(+0.1602)	—
		RIBES	0.7316	—	0.8336(+0.1020)	—
審決④	2,677,978	BLEU	0.2725	±0	0.4274(+0.1549)	+0.0053
		RIBES	0.7316	±0	0.8285(+0.0969)	+0.0051
審査 書類①	1,810,685	BLEU	0.3160	-0.0435	0.5082(+0.1922)	-0.0755
		RIBES	0.7558	-0.0242	0.8456(+0.0898)	-0.0120
審査 書類②	179,830	BLEU	0.3162	-0.0437	0.4777(+0.1615)	-0.0450
		RIBES	0.7630	-0.0314	0.8416(+0.0786)	-0.0080

表中の「差異」欄は、審決③の無学習時及び最終学習後のスコアを各システムのスコアと比較した結果を示した。最終学習後の比較では、審決③の自動評価スコアは審査書類①に対して BLEU で-0.0755、RIBES で-0.0120、審査書類②に対しても BLEU で-0.0450、RIBES で-0.0080 と低くなっている。つまり、最終学習後の審決③の翻訳品質は、審査書類①／②に比べて劣っている。もちろん、審決③よりもスコアが低かった審決④も同様である。

その主な理由としては、先述したとおり、審決③／④の評価用コーパスの翻訳難易度そのものが審査書類①／②よりも有意に高かったことが挙げられる。無学習時の「差異」欄に示したとおり、この時点で既に審決③／④と審査書類①／②の間には大きなスコア差があり、最終学習後の両者のスコア差も、無学習時のスコア差を埋めきれなかった結果と見なせる。

こうした翻訳難易度の差は、偶然の偏りの可能性もあるが、審決③／④と審査書類①／②との条件の違いに起因する必然的なものである可能性もある。そこで、まずは物理的条件、具体的には平均文字数の観点で双方のコーパスを比較した。

審決③／④の評価用コーパスの採取元である審決対訳コーパスは、審決 300 案件の全文を対訳コーパス化したものである。このため、審査書類対訳コーパスでは作成対象外とした 250 文字以上の長文も含んでいる。事実、審決③／④の評価用コーパス 2,024 文中にも 250

文字以上の文が 67 文存在していた。こうした長文の混入が、評価用コーパス全体の翻訳難易度を高めていた可能性がある。だが、評価用コーパス全体の平均文字数を比較すると、審決③／④の平均文字数は 95.7 文字であり、審査書類①／②の平均文字数 (101.7 文字) よりもむしろ短かった。この結果からは、審決③／④の評価用コーパスが文字数の面で不利であったとは言い難い。

そうである以上、審決③／④の評価用コーパスは、各文の記載内容そのものが、審査書類①／②の評価用コーパスよりも翻訳難易度が高いということになる。

概して、審決の記載内容は、個々の審査書類の記載内容よりも広範囲に及んでいる。各審査書類の記載内容が特定の手続に関する内容に限定されるのに対し、審決には審査の各過程での経緯が記されるとともに、審理の主旨や審決に至る判断の内容なども詳述されるためである。加えて、審査書類対訳コーパスの作成対象が特許のみであるのに対し、審決対訳コーパスの作成対象案件には意匠や商標の案件も含まれる⁵⁵。このため、同じ 2,000 文規模の評価用コーパスであっても、審決③／④のものは、そこに含まれる特許・審査書類関連の用語・表現・文体等のバリエーションが審査書類①／②のものよりも広いと考えられる。そして、こうしたバリエーションの中に、コーパスの学習ではカバーし切れなかった要素が存在するものと推察される。それが具体的にどのような要素であるかは、後続の各項で考察していく。

⁵⁵ 審決③／④の評価用コーパス 2,024 文のうち 195 文 (9.6%) は原出願が商標の案件 (審判番号 2016-300561) に由来している。なお、意匠案件は評価用コーパスには含まれていない。

5.2.2. 技術分野間の差異

続いて審決③／④の評価用コーパス各文の自動評価スコアを「化学」、「電気物理」、「機械」の三分野と、審決③／④のみに存在する「商標」とに区分して集計した結果を示す。

表 5.2.3. 審決③／④の技術分野ごとの BLEU、RIBES 平均スコア

学習	システム	化学		電気物理		機械		商標 ⁵⁶	
		BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
無学	審③	0.2629	0.7196	0.3165	0.7620	0.2808	0.7335	0.1338	0.6641
	審④								
#1	審③	0.3648	0.7695	0.3767	0.7957	0.3943	0.7901	0.2224	0.7236
	審④	0.3735	0.7815	0.3860	0.8039	0.4091	0.8037	0.2311	0.7348
#2	審③	0.4317	0.8268	0.4495	0.8441	0.4518	0.8386	0.2929	0.7789
	審④	0.4117	0.8168	0.4403	0.8391	0.4592	0.8431	0.2797	0.7695
#3	審③	0.4249	0.8258	0.4551	0.8483	0.4600	0.8422	0.3107	0.7847
	審④	0.4174	0.8199	0.4437	0.8423	0.4641	0.8395	0.2950	0.7767
向上	審③	+0.1620	+0.1062	+0.1386	+0.0863	+0.1792	+0.1087	+0.1769	+0.1206
	審④	+0.1545	+0.1003	+0.1272	+0.0803	+0.1833	+0.1060	-0.1612	+0.1126

技術用語の訳質改善は本事業の主眼ではなく、「みんなの自動翻訳」のデフォルト性能、すなわち基礎学習データに委ねるところが大きい。審査書類①／②の技術分野間の比較（⇒3.2.4.、4.2.2.）では、翻訳品質が他より顕著に劣る分野は存在しないと結論したが、審決③／④の集計結果もこの結論を否定するものではなかった。審決③／④では上表のとおり、全期間を通じて「化学」のスコアが他の二分野に比べて低いが、審査書類①／②では必ずしも化学分野のスコアが他に比して低いわけではなく⁵⁷、各システムにおける技術分野間の相対順位は一致しない。このように、同じ基礎学習データを用いた各システムで相対順位が変化していること自体が、各技術分野の間に顕著な基礎学習量の差が存在しないことを間接的に示唆している。

⁵⁶ 審決対訳コーパスには商標に属する案件も含まれており、評価用コーパスにも商標案件を1案件選定している。商標案件は技術分野情報を有さないため「商標」扱いとした。なお、意匠に属する文は評価用コーパスには含めていない。

⁵⁷ 各システムの最終学習後での分野間の比較においては、審査書類①の BLEU では「化学」が最下位であったが RIBES では第2位、審査書類②では BLEU が第2位、RIBES では第1位であった。

本集計結果ではむしろ「商標」、すなわち商標審決由来の文の評価スコアが極めて低い点に注目すべきである。前項において、審決③／④の翻訳難易度が高い理由の一つとして商標や意匠由来の文の存在を挙げたが、上記集計結果はこれを肯定している。

商標に属する文は評価用コーパス 2,024 文のうち 195 文存在する。また、これら 195 文中には、正解訳文（英文）に日本語カタカナ表記やローマ字表記が使用された、やや特殊な文が複数文（19 文）含まれている。こうした文の存在が審決③／④全体の評価スコアを低下させている可能性がある。以下、一例を示す。

【文番号：#15】

原文	上記のとおり、「TOP-SIDER」,「トップサイダー」の靴は、発売から数十年を経てもなお、復刻版が販売されるなど、根強い人気を維持してきた。	
正解訳文	As described above, "TOP-SIDER," and "トップサイダー (toppusaida)" shoes have maintained enduring popularity as shown by the fact that the former models are reproduced and sold even after several decades since their launch.	
無学習	As described above, the shoes of "TOP-SIDER" and "Tops-sider" have maintained the popular popularity such that the 復刻 version is sold even after several 10 of years have passed from the release.	
	—	BLEU : 0.1777 (－) RIBES : 0.8228 (－)

本例は、原文に商標名として「TOP-SIDER」と「トップサイダー」の二種が記載されている。後者も英訳すれば top sider となるが、この審決では商標を英語表記とカタカナ表記とで区別しており、後者を英訳してしまうとこの点が伝わらない。このため、人手翻訳では正解訳文のとおり、カタカナ表記にローマ字表記を付記して対応した⁵⁸。一方、機械翻訳文（無学習時）では「トップサイダー」は Tops-sider と英訳されており、「トップ」が Tops と複数形で訳されたため結果的に異なる表記にはなったが、英語表記か日本語表記か、という原文本来の意味にはなっていない。

⁵⁸ 審査書類や審決には本例のように英訳すると主旨が不明となるような文や、誤表記の指摘など英訳不可能な文も少なからず存在する。審査書類対訳コーパスは対象文選定の際にこのような文は極力除外しているが、審決対訳コーパスには審決全文の翻訳結果を用いたため、本例のように翻訳文中に日本語表記を使用せざるを得なかったような文も含まれる。

本調査分析の自動評価で商標案件由来の文のスコアが低くなった一因には、正解訳文にこの「**トップサイダー**」のカナ表記を含む文が存在していることが挙げられる。事実、該当する 19 文の無学習時の BLEU スコアの平均は 0.1207、RIBES は 0.6214 と非常に低い。

ただし、これら 19 文を除外しても、残る商標由来の 176 文の無学習時の BLEU スコアは 0.1352、RIBES は 0.6688 と依然として特許由来の文より極端に低い。つまり、商標由来の文には他にも翻訳難易度を高めている何らかの要因が潜んでいる。まず考えられるのは、商標審決特有の商標用語の多用である。

商標案件由来の文には商標専門用語が多数含まれる。評価用コーパスに選定した各文にも、例えば「指定商品：designated goods」「本件使用商標：the used trademark」「称呼：pronunciation」「類否：similarity」「専用使用権：exclusive use right」「通常使用権：non-exclusive use right」など、特許案件由来の文では使用されない専門用語が数多く使用されている。だが、こうした商標用語は、無学習時の機械翻訳文では正しく翻訳されなかったものの、審決対訳コーパス学習の過程で全て正解訳文どおりに訳されるようになった。つまり、特許・審査書類用語とほぼ同じ改善状況であり、商標用語の存在は、商標案件由来の文の評価スコアが全期間を通じて特許案件由来の文よりも著しく低くなった主要因とは考えにくい。

そこで、改めて商標案件由来の各文を概観すると、特許案件由来の文に比べて固有名詞の多用が目立つ。商標案件由来の文のスコアが低くなった最大の理由は、むしろこの点にあるものと推察される。上記 195 文においても、「トップサイダー」を始め、「スペリ トップーサイダー インコーポレーテッド」「株式会社ビイエムブランニング」「ケネディー家」「プレッピーハンドブック」「デニス・コナー」「株式会社丸井」など、何らかの固有名詞を含む文が全 90 文に及んでいる。そして、こうした固有名詞の多くが、無学習時から最終学習後まで全期間を通じて誤訳又は「未知語」扱いされていた。

概して固有名詞は案件ごとに全く違うものが使われる。このため、コーパスの学習効果がほとんど得られない、翻訳難易度の極めて高いジャンルといえる。固有名詞が使われるのは商標審決に限らないが、商標審決の場合、主題である商標自体が固有名詞である場合が多く、このため多くの文で繰り返し使用される傾向が強い。つまり、商標審決由来の文は、コーパスを学習させても改善しにくい固有名詞を含む文が多いため、全学習期間を通じて特許審決由来の文ほど品質が向上しなかったと考えられる。そして、こうした商標審決由来の文が評価用コーパスに多数含まれていたことが、審決③／④の自動評価スコアが審査書類①／②に比して低かった原因の一つであると推察される。

5.2.3. 文の種類による差異

審決③／④の評価にあたっては、評価用コーパスの一部（759 文）に対して審査書類①と同様に「認定」「対比」「判断」「その他」という文の種類の特定（ラベリング）を実施した。このため文種ごとのスコア集計・比較が可能である。

下二表に、審決③／④の文種ごとの平均スコアの推移を BLEU、RIBES 別に示す。1 行目に無学習時、2 行目に最終回である第 3 回学習後の平均スコアを示し、3 行目には無学習時から第 3 回学習後にかけてのスコアの向上度（±）を示した。なお、表の見出し行には 759 文に占める各文種の文数を表示している。

表 5.2.4. 審決③／④ 文種ごとの自動評価結果の推移（BLEU）

書類		全件	認定	対比	判断	その他
		2,024 文	60 文	67 文	172 文	460 文
審決③	無	0.2725	0.3403(1)	0.2647(3)	0.2059(4)	0.2738(2)
	03	0.4327	0.5067(1)	0.4747(3)	0.4783(2)	0.4171(4)
	±	+0.1602	+0.1664	+0.2100	+0.2724	+0.1433
審決④	無	0.2725	0.3403(1)	0.2647(3)	0.2059(4)	0.2738(2)
	03	0.4274	0.4976(1)	0.4475(3)	0.4923(2)	0.4103(4)
	±	+0.1549	+0.1573	+0.1828	+0.2864	+0.1365

表 5.2.5. 審決③／④ 文種ごとの自動評価結果の推移（RIBES）

書類		全件	認定	対比	判断	その他
		2,024 文	60 文	67 文	172 文	460 文
審決③	無	0.7316	0.7640(1)	0.7186(3)	0.6822(4)	0.7348(2)
	03	0.8336	0.8561(1)	0.8377(3)	0.8380(2)	0.8339(4)
	±	+0.1020	+0.0921	+0.1191	+0.1558	+0.0991
審決④	無	0.7316	0.7640(1)	0.7186(3)	0.6822(4)	0.7348(2)
	03	0.8285	0.8424(1)	0.8277(3)	0.8421(2)	0.8238(4)
	±	+0.0969	+0.0784	+0.1091	+0.1599	+0.0890

文種間の相対順位はシステム、評価指標、学習前後で若干の違いはあるが、最終学習後は両システムとも「認定」が首位で、以下「判断」→「対比」→「その他」という相対順位となった。審査書類①の文種間の相対順位（⇒3.2.5.）と比較すると、双方で一致する部分もあるが、不一致も多く、同傾向とは言い難い。

ただし、審査書類①で見られた、無学習時の相対順位が低い文種ほど最終学習後までのスコアの伸び幅が大きく、文種間のスコア差が縮まるという傾向は審決③／④にも等しく当てはまる。例えば審決③では無学習時の最上位文種（認定）と最下位文種（判断）との BLEU スコア差は 0.1344 であったが、最終学習後は最上位（認定）と最下位（その他）のスコア差は 0.0896 まで縮まった。RIBES ではスコア差はさらに縮まっており (0.0818→0.0222)、審決④の両指標も同様（BLEU が 0.1344→0.0873、RIBES が 0.0818→0.0186）である。この結果は、コーパスの学習によって文種間の均質化が進んだことを示しており、特定の文種の学習データが不足しているという状況はうかがえない。

第 3.4.6.項で述べたとおり、特定の文種のための学習データを作成するには、文へのラベリングという事前作業が必要となり、余分なコストが発生する。本調査結果からは、追加コストを費やして対処する必要性が生じるほど品質が著しく劣る文種は存在しないと判断される。

5.2.4. 文単位のスコア向上／低下カウント結果

審決③／④の評価用コーパス 2,024 文における、各回学習前後で自動評価スコアが向上した文、低下した文、そして変化しなかった文それぞれのカウント結果を下表に示す。

表 5.2.6.審決③／④文単位スコア向上／不変／低下カウント結果

	審決③						審決④					
	BLEU			RIBES			BLEU			RIBES		
	向上	不変	低下	向上	不変	低下	向上	不変	低下	向上	不変	低下
第 1 回	1,401	33	590	1,293	38	693	1,426	26	572	1,358	38	628
第 2 回	1,394	49	581	1,422	55	547	1,314	47	663	1,301	53	670
第 3 回	927	227	870	906	248	870	904	241	879	908	247	869

審決③／④は、期間中の学習回数が全 3 回と審査書類①／②に比べて少なく、各回の間隔が開いている。このため、第 1 回は審査書類①／②と同様にコーパス初回学習前後の比較となるが、続く第 2 回は「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）前後の比較となり、審査書類①／②の第 11 回学習後に相当する⁵⁹。第 3 回は大規模アップデート以降の機械翻訳文同士の比較であるが、第 2 回から間隔が開いているため、中間に複数回のエンジンのバージョンアップを経ている（⇒2.2.1.）。

⁵⁹ 審決③／④の第 2 回学習のタイミングは審査書類①／②の第 13 回学習と同時期だが、審査書類①／②で大規模アップデート前後の比較となるのはアップデート直後の第 11 回学習後である。

このように各学習回とも中間に大きな変動要素が存在するため、上表に示したとおり、審決③／④では毎回、評価用コーパス 2,024 文の大多数において学習前後で機械翻訳文が変化している。これは審査書類①／②の同時期の変動状況から予測されたとおりの結果であり、各システム間に特段の傾向の違いは見られない。

審決③と審決④との比較においても、両者のカウント結果は各回ともおおむね一致しており、審決④のみに大量の審決逆翻訳コーパスを追加学習させたことによる顕著な傾向の差異は生じていない。

ただし、本カウント結果が両システム同様であっても、それは単に変化した文の量が同程度であったというにすぎず、各回で変化した文や、その変化の内容が両システム同じとは限らない。逆翻訳コーパスの学習の有無により、双方の翻訳傾向に有意な違いが存在している可能性もある。この点については、次項以降に示す機械翻訳実例の定点観測や人手評価結果によりさらに分析していく。

5.2.5. 実例を用いた機械翻訳結果の定点観測

審決③／④システムによる各回学習後の機械翻訳文の変化の状況を、実例を用いて定点観測する。審決③／④の場合、各システムの学習前後の変化に加え、同じ文の機械翻訳結果が審決逆翻訳コーパスの追加学習の有無でどのように変わるかも分析の焦点となる。

5.2.5.1. 実例 5-1：特許審決由来、文番号#1736 の分析

最初の実例には特許審決由来の文番号#1736 を選んだ。まずは原文と正解訳文、そして無学習時と初回学習後の両システムの機械翻訳文を示す。なお、無学習時の機械翻訳文は両システム同一である。

【文番号：#1736】

原文	本件出願は、特許請求の範囲の記載が特許法 3 6 条 6 項 1 号に規定する要件を満たしていない。	
正解訳文	The description in the Claim in the application does not comply with the requirement set forth in Article 36(6)(i) of the Patent Act.	
審決③／④ 無学習時	The present application does not satisfy the requirement that the description in the claims 36 and 6 meets the patent law 1 .	
	—	BLEU : 0.1447 (－) RIBES : 0.6685 (－)
審決③ 第 1 回学習後	In the application, the statements of the scope of claims do not meet the requirement stipulated in Article 36 (6) (1) of the Patent Act .	
	向上	BLEU : 0.4063 (+0.2616) RIBES : 0.8848 (+0.2163)
審決④ 第 1 回学習後	In the application, the recitation of the scope of claims does not meet the requirement stipulated in Article 36 (6) (1) of the Patent Act .	
	向上	BLEU : 0.4182 (+0.2735) RIBES : 0.8962 (+0.2277)

本例は特許・審査書類用語とそれに付随する一般用語のみで構成された文である。無学習時の機械翻訳文では、このうち「本件出願：present application」、「要件：requirement」、「記載：description」などは適切に訳されているが、「特許請求の範囲」と「特許法 3 6 条 6 項 1 号」に関しては、それぞれ **claims 36 and 6**、**patent law 1** と両者が混同され、条文番号の一部が請求項番号に誤解釈されている。このため、後半部分は「請求項 36 及び 6 の記載が**特許法 1** に適合するという要件を満たしていない」という、原文と全く異なる意味になってしまっている。

これに対し、審決③の初回学習後の機械翻訳文では、まず「特許請求の範囲」が本事業の統一訳語である **scope of claims** に改善された。また「特許法 3 6 条 6 項 1 号」も、**Article 36 (6) (1) of the Patent Act** と条文番号が正しい位置に配置されるようになり、「1 号」がロ

ーマ数字(i)となっていない点を除けば、正解訳文どおりのフォーマットで翻訳されている。つまり本例では、審決対訳コーパスの初回学習によって、無学習時に対応できなかった全ての特許・審査書類用語が正しく翻訳されるようになった。「記載」が description から statement に変化するなど正解訳文どおりでない訳語も見られるものの、原文の内容を正確に把握できる翻訳文となり、無学習時に比べて格段に改善されている。

こうした状況は審決④においても同様であり、審決③との相違点は「記載」の訳語に statement ではなく recitation が採用されていることのみであった。

第 1 回学習後の時点では、審決③の学習データが審決対訳コーパス 86,565 文のみであったのに対し、審決④はさらに 2,521,792 文の審決逆翻訳コーパスを追加学習している。このため学習データ総量には約 28 倍の差がある。しかしながら、両者の機械翻訳文の差異は上記のとおりごく局所的であり、逆翻訳コーパスの規模に見合うような顕著な違いは見られない。ただし、審決④で「記載」の訳語に採用された recitation は、特許請求の範囲の記載という本例の文脈では statement よりも適訳である。このため、この時点での翻訳品質は審決④のほうがやや優れている。この優劣は、第 5.2.1.項で示した自動評価の全体平均スコアでの優劣とも一致する。

続いて、第 2 回学習後の両システムの機械翻訳結果を示す。

審決③ 第 2 回学習後	Regarding the present application, the recitation of the scope of claims does not meet the requirement stipulated in Article 36 (6) (I) of the Patent Act.	
	向上	BLEU : 0.4740 (+0.0677) RIBES : 0.8778 (-0.0070)
審決④ 第 2 回学習後	In the present application, the statement of the scope of claims does not meet the requirement stipulated in Article 36 (6) (I) of the Patent Act.	
	劣化	BLEU : 0.4871 (+0.0689) RIBES : 0.8778 (-0.0184)

審決③／④では、第 1 回学習と第 2 回学習の間に「みんなの自動翻訳」の大規模アップデート（⇒2.2.2.）が実施されている。したがって、第 2 回学習後の機械翻訳文は、大規模アップデート後の機械翻訳文となる。

審決③の機械翻訳文では、まず冒頭部が前回の In the application から Regarding the present application に変化した。ただし、どちらの訳し方でも文意は変わらず、審査書類①／②でもよく見られた、同義の範囲内での文構造の変化といえる。また、「記載」の訳語が前回の statement から recitation へと変化した。前回時に審決④で採用されていた訳語であ

る。前述のとおり、本例では statement より **recitation** のほうが適訳であり、この点において、審決③の第2回学習後の翻訳品質は前回よりも「向上」している。

一方、審決④の機械翻訳文では、こちらも冒頭部が In the application から In the **present** application へと変化しているが、審決③と同じく、同義の範囲内の変化である。そして、「記載」の訳語が前回の recitation から今回は **statement** へと、審決③と逆の方向に変化した。

前回時の審決③と審決④の実質的な相違点は、この「記載」の訳語に審決③は statement を採用したのに対し、審決④はより適切な recitation を採用したことであった。つまり、この訳語 (recitation) の採用は逆翻訳コーパスの追加学習がもたらしたものである可能性があった。しかしながら、今回、審決③でも recitation が採用され、かつ、審決④では recitation が不採用となったことで、前回時の審決④での recitation の採用も、逆翻訳コーパスがもたらしたのではなく、「みんなの自動翻訳」の特徴である同義の範囲内のランダム的な訳語変化によるものである蓋然性が高くなった。そうすると、本例では今のところ逆翻訳コーパスの追加学習は審決④の機械翻訳文に実質的な影響を及ぼしていないことになる。

なお、上記 recitation から **statement** への変化により、審決④の翻訳品質は前回よりもやや「劣化」し、相対的には審決③のほうが優位となった。

最後に、最終学習回である第3回学習後の両システムの機械翻訳文を示す。

審決③ 第3回学習後	In the present application, the statement of the scope of claims does not meet the requirement stipulated in Article 36 (6) (I) of the Patent Act.	
	劣化	BLEU : 0.4871 (+0.0131) RIBES : 0.8778 (±0)
審決④ 第3回学習後	In the Application , the statement of the scope of claims does not meet the requirement stipulated in Article 36 (6) (I) of the Patent Act.	
	同等	BLEU : 0.5216 (+0.0345) RIBES : 0.9072 (+0.0294)

審決③の第3回学習後の機械翻訳文では、冒頭部が前回の Regarding から前々回の In へ再び変化した。同義の範囲内であり優劣はない。また、前回時に recitation に変化した「記載」の訳語も **statement** に戻っており、やはりこの訳語変化が「みんなの自動翻訳」のランダム性によるものであることを示唆している。本例では recitation のほうが適訳であるため、審決③の翻訳品質は前回よりもやや「劣化」となる。なお、審決③の今回の翻訳文は、審決④の前回の翻訳文と完全一致している。

一方、審決④も冒頭部の present application が Application に変化した。こちらも、先頭文字がキャピタライズされたことを除けば前々回時に訳語に戻った形となるが、内容理解上の優劣はない。「記載」の訳語は引き続き statement で変化しておらず、このため翻訳品質は前回と完全に「同等」である。なお、「記載」の訳語が両システムとも statement となったことで、システム間の品質にも優劣はなくなった。

以上、文番号#1736 の無学習時から最終学習後までの両システムの機械翻訳文を定点観測した。本例においては、審決③、④のいずれの翻訳文とも、無学習時には対応できていなかった特許・審査書類用語がコーパスの学習により適訳語で安定し、その一方で文構造や一般用語は学習の都度不安定に変化するという、審査書類①／②で観測した各実例と同様の傾向を示した。

本項の焦点である審決③／④間の比較では、各学習回とも双方の機械翻訳文は異なっていたが、例えば「記載」の訳語として両システムで recitation と statement の 2 語が採用されていたように、訳語のバリエーション自体は両システム共通であり、それらが「みんなの自動翻訳」特有のランダム的な訳語選定によってそれぞれで不規則に採用されているだけの違いと見受けられた。つまり、本例を見る限りにおいては、逆翻訳コーパスの有無による機械翻訳文の内容や品質への影響は見いだせなかった。この結果は、自動評価の全体平均スコアから導出した第 5.2.1.2.項の考察内容、すなわち「審決対訳コーパスを十分に学習させる前提であれば、審決逆翻訳コーパスの追加学習には特段の品質改善効果はない」という結論と整合している。

5.2.5.2. 実例 5-2：特許審決由来、文番号#1277 の分析

続いて、同じく特許審決由来の文番号#1277 の機械翻訳文の推移を示す。本例は審査書類用語に加え、前例には含まれなかった技術用語も含んでいる。

【文番号：#1277】

原文	本件訂正請求により、請求項 1 に「シリコーンゴムからなる第 1 被覆層及び第 2 被覆層とを有し、前記シリコーンゴムの塗布量は 1 0 0 ～ 1 1 0 g / m 2 であり」との事項が付加された。	
正解訳文	By the correction request of the case , the matter of "comprising a first coating layer and a second coating layer composed of silicone rubber, wherein an application quantity of the silicone rubber is 100-110 g/m2" was added to Claim 1 .	
審決③／④ 無学習時	The present correction claims added the matters "the silicone rubber has a first coating layer and a second coating layer made of silicone rubber, and the amount of the silicone rubber is 100 g / m2 to 110g / m 1 ".	
	—	BLEU : 0.3546 (－) RIBES : 0.8119 (－)
	向上	BLEU : 0.2185 (+0.2735) RIBES : 0.8217 (+0.2277)

本例で使用されている審査書類用語は「本件訂正請求：[correction request of the case](#)」と「請求項 1：[Claim 1](#)」の 2 語、そして「事項：[matter](#)」も一般用語ではあるが審査書類特有の用法であり審査書類用語に準ずる。無学習時の機械翻訳文では、「事項」には正解訳文と同じ訳語 [matters](#) が採用されたが（ただし複数形は不自然である）、「本件訂正請求」には [present correction claims](#) という意味不明の訳語が採用され、また、「請求項 1」は消失して [1](#) のみが文末に出力されている⁶⁰。つまり、前例と同じく、無学習時の審決③／④では審査書類用語への対応がほとんどできていない。

原文ではカギカッコで括られている技術内容の摘記部分は、文の一部分のみが切り取られており、文が完結していないため翻訳しにくい形になっている。無学習時の機械翻訳文は、これを文として翻訳すべく主語を補っているが、不適切な the silicone rubber を補ったことで「シリコーンゴムがシリコーンゴムからなる被覆層を有し」という誤った内容となってしまった。このように原文に存在しない主語を補足する際に不適切な語が選択される事象は、審査書類①の実例 3-4 (the invention makes the invention unclear) でも見られた。主語の省略が可能な日本語と主語が必須な英語との言語としての性質の違いに起因する不可避の

⁶⁰ [present correction claims](#) の [claims](#) が「本件訂正請求」ではなく「[請求項 1](#)」に対応している可能性はあるが、文法的には考えにくい。

課題であるが、「みんなの自動翻訳」の場合、本例の silicone rubber や実例 3-4 の the invention のように具体的な名詞を主語として補足する傾向があると見受けられる。本例も形式主語 it を採用していれば、曖昧ではあっても誤訳にはならなかった。

この不適切な主語の補足と、末尾の数値単位「g/m²」が「g/m」となっていて不完全であることを除けば、摘記箇所の各技術／一般用語の訳語は正確である。

続いて、第 1 回学習後の両システムの機械翻訳文を示す。

審決③ 第 1 回学習後	According to the Correction request, the matter that "a first coat layer made of silicone rubber and a first coat layer and a second coat layer are included, and the amount of the silicone rubber is 100 to 110g / m²" is added to Claim 1.	
	向上	BLEU : 0.3081 (+0.2616) RIBES : 0.8138 (+0.2163)
審決④ 第 1 回学習後	The correction request added the matter "comprising a second coat layer and a first coat layer each consisting of a silicone rubber, wherein the amount of the silicone rubber to be applied was 100 to 110g / m²" in Claim 1.	
	向上	BLEU : 0.2185 (+0.2735) RIBES : 0.8217 (+0.2277)

審決対訳コーパスの初回学習により、審決③の機械翻訳文では「請求項 1」は **Claim 1** と正しく訳されるようになり、「本件訂正請求」も the **Correction request** と、本事業の統一訳語ではないものの意味の通る訳語に改善された。「事項」も **matter** と単数形になっている。ただし技術内容の摘記部分は、文法的に不完全な原文を英文として成立させようとした結果、今回は「第 1 被覆層」にあたる a first coat layer が重複出力されて「シリコンゴムからなる第 1 被覆層及び第 1 被覆層及び第 2 被覆層が含まれ」と、引き続き誤った内容になっている。

これに対し、審決④はこの摘記部分を正解訳文と同じく ing 節を用いて comprising～と適切に処理している。なぜか first coat layer と second coat layer の順序が逆になっているが、内容理解への影響はない。「本件訂正請求」は審決③と同様に the **correction request** と訳しており改善の余地はあるが、原文の内容は問題なく理解できる。したがって、前例と同様、第 1 回学習後の時点では審決④のほうが翻訳品質は優れている。ただし、審決④におけるこの ing 節の採用が審決逆翻訳コーパスの効果であるのか、「みんなの自動翻訳」特有の文構造のランダム的な変化によるものかは、現時点では判断できない。

技術用語に関しては、両システムとも「第 1 / 2 被膜層」の訳語が無学習時の first/second coating layer から first/second coat layer に変化した (coating→coat)。意味はどちらも同じだが、無学習時の訳語のほうが特許文献での使用頻度は高く、マイナーな訳語に悪化したとの印象も受ける。審査書類①／②においても、大規模アップデート以前の「みんなの自動翻訳」では技術／一般用語のランダム的な訳語変化が顕著であったが、審決③／④にも同様の傾向がうかがえる。

続いて、第 2 回学習後の機械翻訳結果を示す。

審決③ 第 2 回学習後	By the request for correction of the case, the matter of "having a first coating layer and a second coating layer made of silicone rubber, wherein a coating amount of the silicone rubber is 100 to 110g / m ² " was added to Claim 1.	
	向上	BLEU : 0.6305 (+0.0677) RIBES : 0.9493 (-0.0070)
審決 第 2 回学習後	By the request for correction of the case, the matter "comprising a first coating layer and a second coating layer made of silicone rubber, wherein an application amount of the silicone rubber is 100 to 110g / m ² " was added to Claim 1.	
	向上	BLEU : 0.8217 (+0.0689) RIBES : 0.9660 (-0.0184)

第 2 回学習後の機械翻訳文では、両システム間の差異がかなり少なくなった。「本件訂正請求」はどちらも request for correction of the case と正解訳文どおりの適訳語が採用されるようになり、かつ摘記部分も審決③は having、審決④は comprising と両者で ing 節が採用され、訳語にも優劣はない。「塗布量」の訳語も審決③は coating amount、「審決④」は application amount と異なっているがやはり優劣はなく、翻訳品質は優秀かつ完全に同等である。

なお、今回、審決③でも摘記部分に ing 節が採用されたことから、前回時の審決④での ing 節の採用は、審決逆翻訳コーパスの追加学習の効果ではなく、文構造のランダム的な変化による偶然の結果である可能性が高くなった。

第 2 回学習後の機械翻訳文は、大規模アップデート後の「みんなの自動翻訳」による機械翻訳文である。審査書類①／②では、大規模アップデート以降、学習前後で訳語や文構造のランダム的な変化が抑制される傾向が見られた。今回、審決③／④の翻訳文の類似度が向上したのも、この傾向の変化が関係していると考えられる。審決③と審決④は、それぞれ逆翻訳コーパスを学習させる前と学習させた後、という見方もできるためである。

ただし、この傾向はこれまで、同種のコーパスの学習前後という条件で見られていたものである。同種のコーパスであれば、増量しても学習前後でコーパス全体の訳語の優先順位はさほど変化しないであろうから、訳語選定の精度が向上することで、各回の訳語が安定するのは理に適う。今回、逆翻訳コーパスの学習前後で同じ傾向が生じたということは、裏を返せば、審決対訳コーパスと逆翻訳コーパスが学習データとしてほぼ同質であることを意味する。

また、第2回学習後の両システムの「第1／2被覆層」の訳語は、前回時の first/second coat layer から、再び無学習時の first/second coating layer に戻った (coat→coating)。訳語としては前述したとおり後者のほうが一般的であり、より適切な訳語に戻されたことになる。これも、大規模アップデート以降に審査書類①／②で見られた傾向のとおり、学習データ中で頻度上位であろう訳語が順当に採用されるようになった結果と見なせる。

続いて、最終回である第3回学習後の機械翻訳文を比較する。

審決③ 第3回学習後	By the request for correction of the case, the matter of " comprising a first coating layer and a second coating layer made of silicone rubber, wherein an application amount of the silicone rubber is 100 to 110g / m ² " was added to Claim 1.	
	同等	BLEU : 0.7318 (+0.1013) RIBES : 0.9668 (+0.0175)
審決④ 第3回学習後	By the request for correction , the matter "comprising a first coating layer and a second coating layer made of silicone rubber, wherein an application amount of the silicone rubber is 100 to 110g / m ² " was added to Claim 1.	
	同等	BLEU : 0.6083 (-0.0657) RIBES : 0.9593 (-0.0067)

審決③の第3回学習後の機械翻訳文で前回と変化した箇所は2箇所であり、「有し」が having から comprising へ、「塗布量」が coating amount から **application amount** へと変化した。その結果、今回の翻訳文は前回時の審決④の翻訳文とほぼ同じとなり、matter の直後に of が付されるか否かのみ的小差となっている。

一方、審決④の前回時からの変化は、冒頭の「本件訂正請求」が前回の the request for correction of the case から今回は **the request for correction** と後半の of the case 省略された一点のみであった。したがって、今回の審決③と④との差はこの「本件訂正請求」での of the case の有無と、上記 matter の直後の of の有無の二箇所のみであり、前回時から更に類

似度を増した。

なお、審決④の「本件訂正請求」における of the case（「本件」に相当）の省略も、定冠詞 the が存在するので「本件」のニュアンス自体は読み手には伝わる。このため内容理解上の優劣はない。したがって、第 3 回時の両システムの機械翻訳文の品質はそれぞれ前回時と「同等」であり、両システム間にも優劣はない。

以上、本例においても、コーパスの学習によって特許・審査書類用語が適訳語に改善される一方、技術／一般用語や文構造は同義の範囲内で不安定に変化するという、各システムの各実例に見られたのと共通の傾向が顕著であった。

審決③と審決④との比較では、各学習回とも両者の相違点は「みんなの自動翻訳」の訳語選定のランダム性によるものと見られ、かつ、学習を重ねるにつれ、双方の機械翻訳文は類似度を増していった。採用される訳語にも有意な差異はなく、前例と同じく、逆翻訳コーパスの追加学習による機械翻訳文への影響は全く見られなかった。

5.2.5.3. 実例 5-3：商標審決由来、文番号#122 の分析

審決対訳コーパスは意匠審決や商標審決も作成対象としており、評価用コーパスにも商標審決由来の文が含まれる⁶¹。第 5.2.2.項に示した技術分野ごとの自動評価スコアの集計結果では、商標審決由来の文は特許審決由来の文に比べて明らかに低スコアであり、商標審決由来の文には、翻訳難易度を他より高めている何らかの特徴を有していると考えられる。第 5.2.2.項ではその最大の要因として固有名詞の多用を挙げた。一方、商標用語に関してはコーパスの学習によって他の審査書類用語と同等に改善され、このため低スコアの原因とはならないと分析した。本項では評価用コーパスから商標審決由来の文番号#122 を実例に採り、実際の機械翻訳結果の変化を観測して、これらの推論を検証するとともに、他の特徴の有無を調査する。

【文番号：#122】

原文	水甚社による本件商標の使用は、被請求人が、同社に対して、本件商標の本件譲渡契約時の口頭による「セイルボート図形」を使用するようにとの請求人側の依頼によるものであり、当該図形の使用を含めて使用を許諾したものである。	
正解訳文	The use of the Trademark by Mizujin was carried out in accordance with the request by Demandee to use the "sailboat figure" as verbally requested on the occasion of conclusion of the Transfer Agreement for the Trademark from Demandant side, and Demandee has granted a license to use the Trademark including the use of the figure.	
審決③／④ 無学習時	The use of the book by the water 甚社 is based on the request of the person skilled in the art to use the spoken "sailboat figure" at the time of the formal transfer agreement of the present book to the company, and the use uses the use including the use of the figure.	
	—	BLEU：0.2417（－） RIBES：0.6966（－）

本例には「水甚社：Mizujin」という固有名詞（社名）が含まれている（この固有名詞は本案件中で他にも 11 文で使用されている）。上掲した無学習時の機械翻訳文では、「water 甚社」と「未知語」扱いされ、一部日本語のまま出力されている。

また、本例には審査書類用語も多数含まれており、冒頭から順に「本件商標：the Trademark」「被請求人：Demandee」「本件譲渡契約：Transfer Agreement for the Trademark」「請求人：Demandant」「図形の使用：use of the figure」の 5 語が該当する。特に「本件商標」「本件

⁶¹ 審判番号 2016-300561 の商標審決の全文に相当する 195 文。

譲渡契約」「図形の使用」は商標審決特有の商標用語である。

無学習時の機械翻訳文では、「本件商標」は **book** と訳されており、「本」のみが翻訳されて残りの部分は欠落したと見られる。「被請求人」は **person skilled in the art**（当業者）と全く意味の違う語に誤訳され、「本件譲渡契約」も **formal transfer agreement of the present book** と、「譲渡契約 (transfer agreement)」は正しく訳されたものの「本件」が **present book** と誤訳されており、**formal** の意味も不明である。さらに「請求人」も翻訳文から欠落してしまっており、唯一「図形の使用」のみが **use of the figure** と正しく訳されている。つまり、本例も、「みんなの自動翻訳」がデフォルトでは審査書類用語（商標用語を含む）にほとんど対応できないことを示している。

無学習時の機械翻訳文はこれら審査書類用語の誤訳や欠落により既に原文の意味を把握することは不可能な状態であるが、その他の部分も非常に不正確であり、例えば原文末尾の「当該図形の使用を含めて使用を許諾したものである」という箇所は the use uses the use including the **use of the figure** と use が多発している。原文で「使用」が2回出現するのに対して翻訳文では use が4回も出現しているが、このうち1回は「許諾した」の欠落の埋め合わせとして湧き出したもの、もう1回は主語無し原文に主語を補う際に文中の the use が採用されたためと推察される。どちらも「みんなの自動翻訳」の典型的なエラーパターンである。

続いて、第1回学習後の両システムの機械翻訳文を示す。

【文番号：#122】

審決③ 第1回学習後	The use of the Trademark by water 甚社 is made by the demandee's request to use the oral "sailboat figure" by the oral cavity at the time of the transfer agreement of the Trademark with respect to the same company, and the use of the figure including the use of the figure is approved.	
	向上	BLEU : 0.2868 (+0.0451) RIBES : 0.7813 (+0.0847)
審決④ 第1回学習後	Use of the Trademark by water 甚社 is made by the demandee on the basis of the request made by the demandee so as to use the oral "sailboat figure" at the time of the transfer agreement of the Trademark, and is approved for use including the use of the figure.	
	向上	BLEU : 0.2969 (+0.0552) RIBES : 0.7928 (+0.0962)

固有名詞「水甚社」は審決③、審決④とも引き続き「water 甚社」と「未知語」扱いされ

ており、コーパス学習後も改善されていない。一方、審査書類用語に関しては、「本件商標：the Trademark」「被請求人：demandee」「本件譲渡契約：transfer Agreement of the Trademark」「図形の使用：use of the figure」と5語中の4語が両システムで同じ訳語となり、「本件譲渡契約」の前置詞が for ではなく of であるという微細な違いを除けばいずれも正解訳文どおりとなった。ただし、「請求人：Demandant」のみは両システムとも引き続き欠落しており、審決④ではこれを埋め合わせるように demandee（被請求人）が重複出力されている。

審査書類用語以外の部分では、審決③で「口頭による：oral」が「請求人の依頼」ではなく「セイルボート図形」に係ると誤解釈されている。ただし、この係り受けは、意味的にはナンセンスだが、原文自体が文法的にはこう解釈するほうがむしろ自然な書きぶりになっており、やむを得ない面はある。なお、審決③では別の箇所にも oral cavity（口腔）という語が重複出力されており、この語が前述の「請求人：Demandant」の欠落を埋め合わせているものと見られる。また、「当該図形の使用を含めて使用を許諾したものである」という文末部分も、the use of the figure including the use of the figure is approved と「使用：use」のみでよい部分に「図形の使用：use of the figure」が重複出力され、意味不明な文になってしまっている。

一方、審決④では、「口頭による：oral」の係り受け誤りが発生している点は審決③と同じだが、文末部分は is approved for use including the use of the figure と正しい文意を類推できる訳文になっている。この点において審決④のほうが翻訳品質は優れている。つまり、本例も前二例と同じく、第1回学習後の時点では審決④のほうが品質が優れており、自動評価の全体平均スコアが示した両システム間の優劣と合致している。

続いて、第2回学習後の両システムの機械翻訳文を示す。

審決③ 第2回学習後	The use of the Trademark by Mizuho 甚社 was requested by the demandee on the side of the demandant so as to use the oral "Sailboat figure" of the Trademark at the time of the agreement for transfer of the case, and the use of the Trademark including the use of the figure was permitted.	
	向上	BLEU : 0.3119 (+0.0251) RIBES : 0.7694 (-0.0119)
審決 第2回学習後	The use of the Trademark by 甚社 was requested by the demandant for the demandee to use the oral "sailboat figure" of the Trademark at the time of the contract for the transfer of the case, and the demandee allowed the use of the Trademark including the use of the figure.	
	向上	BLEU : 0.3406 (+0.0437) RIBES : 0.8164 (+0.0236)

固有名詞「水甚社」は審決③では「Mizuho 甚社」、審決④では「甚社」と引き続き「未知語」扱いされ、改善の兆しは見られない。一方、審査書類用語については、前回時まで欠落していた「請求人」が **demandant** と両システムで正しく訳出されるようになった。その反面、「本件譲渡契約：Transfer Agreement for the Trademark」が審決③では **agreement for transfer of the case**、審決④では **contract for the transfer of the case** と、ともに「本件」が **Trademark** ではなく **the case** と訳されるようになり、誤訳ではないものの（原文には「商標」とは書かれていない）、やや分かりにくくなった。

今回、「請求人」の欠落が解消したことで原文中の文言は（「水甚社」を除けば）全て訳出されるようになったが、各語の係り受けは依然として不正確である。本例が由来する商標審決では、被請求人と請求人との間に商標の譲渡契約が結ばれており、被請求人はその契約に則り水甚社にその商標を使用した商品を発注したが、その際、譲渡契約に含まれない「セイルボート図形」が使用されたことについて争われている（被請求人は請求人から口頭で使用を許諾されたと主張している）。審決③の機械翻訳文は係り受けが不正確なため、こうした状況を把握することは困難である。審決④の翻訳文も、審決③よりは係り受けが整理されているが、それでも「水甚社による本件商標の使用」を「請求人が依頼」したことになっている点が誤っている（実際は「被請求人」が「許諾」している）。ただし、本例は日本語原文自体が文法的に不明瞭で、きわめて難解であることも事実である。

前回時の「口頭による」の「係り受けエラー」は引き続き両システムで発生しているが、文末の **use of the figure** 周辺箇所は両システムとも正しい解釈となり、かつ英文の品質もそれぞれ違う形ながら適切に語を補足して分かり易く訳している。

第2回学習後の両システムの翻訳品質は、どちらも「請求人」の欠落と不要な語によるその埋め合わせという重大な不備が解消し、かつ **use of the figure** 周辺の解釈も正しくなったことで、顕著な「向上」と判定される。システム間の優劣は、前半部の係り受け誤りの程度が軽い分、審決④のほうが若干優れている。

最後に第3回学習後の機械翻訳文を示す。

審決③ 第3回学習後	The use of the Trademark in 甚社 is based on the demandant's request to the demandee to use the oral "sailboat figure" at the time of the agreement for the transfer of the Trademark, and the demandee permits the use of the Trademark including the use of the figure.	
	向上	BLEU : 0.3376 (+0.0257) RIBES : 0.7637 (-0.0057)

審決 第3回学習後	The use of the Trademark by 甚社 is made by the demandant's request that the demandee use the oral "sail boat figure" at the time of contract for the transfer of the Trademark to the company, and the use of the Trademark including the use of the figure is approved.	
	向上	BLEU : 0.2629 (-0.0777) RIBES : 0.7821 (-0.0343)

最終学習後においても、各審査書類用語は引き続き適訳語が安定して採用されている。前回やや分かりにくくなった「本件譲渡契約」も、今回は審決③が **agreement for the transfer of the Trademark**、審決④が **contract for the transfer of the Trademark** と、それぞれで **Trademark** が再び訳出されるようになった。

一方、固有名詞「水甚社」は今回も両システムで「甚社」と「未知語」扱いされており、最後まで改善は見られなかった。「口頭による：oral」の「係り受けエラー」も引き続き発生している。前半部の係り受けも、原文自体の記載が不明瞭なこともあり依然として不正確であるが、今回は審決③も前回時の審決④と同程度には改善しており、このため審決③は前回に比べて若干の「向上」とした。審決④も Trademark が再度訳出されるようになった点で前回時よりも良化しており、こちらも「向上」とした。なお、システム間の品質差はなくなり、両者同等となった。

以上、本例では、商標用語、固有名詞という、審決③／④の翻訳品質に大きく影響すると考えられるジャンルの語の改善状況を確認することができた。観測の結果、商標用語は審決対訳コーパス学習が進むにつれて適訳語に改善されていき、最終学習後まで誤訳されている用語は存在しなかった。係り受けに関しても、文末の **use of the figure** 周辺箇所などにおいて一定の改善が見られた。最終学習後においても不正確な箇所はあったが、本例は原文自体が難解な書きぶりであり、仕方のない面もある。

一方、固有名詞である「水甚社」は常時「未知語」扱いされ、審決対訳コーパスの学習後も改善は見られなかった。この結果は、第 5.2.2 項で述べた、審決由来の文の機械翻訳の難易度が各審査書類よりも高くなる主要因は商標用語ではなく固有名詞の多用にある、という考察内容と整合している。

また、本例においても、審決③と審決④の最終学習後の翻訳品質は完全に同等であり、逆翻訳コーパスの追加学習の効果を示す特段の差異は見られなかった。

5.2.5.4. 実例 5-4：商標審決由来、文番号#159 の分析

商標審決由来の文をもう一例分析する。文番号#159 は、固有名詞は含まないが商標審決特有の判断内容が書かれた文である。

【文番号：#159】

原文	したがって、引用商標と本件使用商標とは、外観において酷似し、称呼及び観念を同じくする類似の商標であって、その類似性の程度は、極めて高いものと認められる。	
正解訳文	Accordingly, it can be recognized that the Cited Trademark and the Used Trademark are very similar to each other in their appearances, and have common pronunciation and meaning, and the degree of their similarity is quite high.	
審決③／④ 無学習時	Therefore, it is understood that the quoted trademarks and books are similar in appearance and are similar trademarks, and the degree of similarity is very high.	
	－	BLEU：0.1322（－） RIBES：0.7517（－）

本例に含まれる商標用語は「引用商標」「本件使用商標」程度であるが、一般用語に属する「外観」「酷似」「称呼」「類似性」なども商標審決では多用されており、商標用語に準ずる語といえる。

無学習時の機械翻訳文では、これら 6 語のうち正解訳文どおりに訳せていたのは「外観：appearance」と「類似性：degree of similarity」の 2 語のみであり、残る 4 語は「引用商標」が quoted trademarks、「本件使用商標」が books、「酷似」が similar（ニュアンスが弱く、正解訳文のように very similar 等と強調すべきである）と 3 語が誤訳され、「称呼」は翻訳文から欠落している。つまり、本例も前三例と同じく、無学習時には審査書類用語（商標用語を含む）にほとんど対応できていない。

続いて、第 1 回学習後の機械翻訳文を示す。

審決③ 第 1 回学習後	Therefore, it is recognized that the Cited Trademark and the Used trademark are similar in appearance, and are similar to each other in terms of similarity and meaning, and the degree of similarity is extremely high.	
	向上	BLEU：0.4933（+0.3611） RIBES：0.9124（+0.1607）

審決 第1回学習後	Therefore, Cited Trademark and the Used Trademark are similar to each other in terms of appearance and similar to each other, and it can be recognized that the degree of similarity is extremely high.	
	向上	BLEU : 0.4047 (+0.2725) RIBES : 0.6562 (-0.0955)

審決対訳コーパス初回学習後の機械翻訳文では、審決③、④とも「引用商標」が Cited Trademark、「本件使用商標」が Used Trademark と正解訳文どおりの適訳語に改善された。無学習時から適訳語に訳されていた「外観」(appearance)と「類似性」(degree of similarity)は今回も正しく訳されている。一方、「酷似」は引き続き両システムとも similar のみでニュアンスが弱く、「称呼」は依然として欠落している(審決③では埋め合わせ的に similarity が湧き出している)。

こうした、初回学習の時点では審査書類用語への対応がまだ完全ではなく、不適訳語や欠落が残留するという状況は、本例に限らず、前例(実例 5-3)で初回学習後も「請求人: Demandant」が引き続き欠落したり、実例 5-2 で「本件訂正請求」が定番訳 correction request of the case ではなく Correction request と訳されたりと、審決③／④ではよく見られる。審査書類①／②では初回学習後に大半の審査書類用語が正解訳文どおりの適訳語に改善されていたことを考えると、学習の進みはやや遅い。これは、第 5.2.1.3 項で考察したとおり、審決は個々の審査書類に比べて記載される範囲が広く、使用される審査書類用語の種類も多いことが影響していると考えられる。カバーすべき範囲が広いため、初回学習時の審決対訳コーパスの規模では、主要な用語を全て網羅することができなかったものと推察される。その意味では、約 250 万文と大規模な審決逆翻訳コーパスの追加学習には、用語の網羅性を増強させるという効果はほとんど無いといえる。

続いて、第2回学習後の両システムの機械翻訳文を示す。

審決③ 第2回学習後	Therefore, the Cited Trademark and the Used Trademark are similar trademarks that resemble each other in appearance and have the same pronunciation and meaning, and it is recognized that the degree of similarity is extremely high.	
	向上	BLEU : 0.3798 (-0.1135) RIBES : 0.7509 (-0.1615)
審決 第2回学習後	Therefore, the Cited Trademark and the used trademark of the case are similar trademarks which are very similar in appearance and have the same pronunciation and meaning, and it is recognized that the degree of similarity is extremely high.	
	向上	BLEU : 0.3411 (-0.0636) RIBES : 0.7315 (+0.0753)

両システムとも、前回時まで欠落していた「称呼」が **pronunciation** と正しく訳出されるようになった。「酷似」に関しては、審決③は訳語が **resemble** に変化したが、ニュアンス的には依然として弱く、前回の **similar** と大差ない。一方、審決④は **very similar** と強調された適切なニュアンスに改善された。なお、審決④は今回「本件使用商標」を **Used Trademark** ではなく **used trademark of the case** と訳しているが、意味的には同義であり内容理解上の不足はない。

第 2 回学習後の両システムの機械翻訳文では、前回時に対応できていなかった商標用語「称呼」が **pronunciation** と正しく訳されるようになるなど、審決対訳コーパスの増量によって審査書類用語の網羅性が向上したことを示す改善が見られた。これも前例で「請求人」の欠落が解消したタイミングと同じであり、先に述べた「カバーすべき範囲が広いため、初回学習時の審決対訳コーパスの規模では、主要な用語を全て網羅することができなかった」という考察と整合する。

最後に、第 3 回学習後の機械翻訳文を示す。

審決③ 第 3 回学習後	Therefore, Cited Trademark and the Used Trademark are similar trademarks that resemble each other in appearance and have the same pronunciation and meaning, and it is recognized that the degree of similarity is extremely high.	
	同等	BLEU : 0.3528 (-0.0270) RIBES : 0.7441 (-0.0068)
審決④ 第 3 回学習後	Therefore, the Cited Trademark and the Used Trademark are similar trademarks that resemble each other in appearance and have the same pronunciation and meaning, and it is recognized that the degree of similarity is extremely high.	
	劣化	BLEU : 0.3798 (+0.0387) RIBES : 0.7509 (+0.0194)

審決③の第 3 回学習後の翻訳文は、冒頭の **Cited Trademark** に定冠詞 the が付されなくなったこと以外、前回時と完全に同一である。この定冠詞はあるべきだが、内容理解上の優劣をつけるべきかは微妙である。

一方、審決④の翻訳文では、前回 the used trademark of the case と訳されていた「本件使用商標」が **the Used Trademark** へ、そして「酷似し」が very similar から **resemble each other** へと、いずれも審決③と同じ訳語に変化した。その結果、今回の審決④の翻訳文は前回の審決③の翻訳文と完全一致となった。ここまで見てきた各例とも、このように同じ訳語

が時間差で両システムで採用されるケースが目立つ。これも、両システム間の訳語の相違が審決逆翻訳コーパスの有無ではなく「みんなの自動翻訳」のランダム的な訳語変化によるものであることを示唆する現象といえる。

以上、本例は商標審決特有の判断内容を記した文の翻訳品質について観測した。本例においては、コーパス初回学習後の時点では商標用語への対応はまだ完全ではなかったが、その後のコーパス増量により、第 2 回学習後からは正確な翻訳文が出力されるようになった。つまり、商標審決由来の文である前例と本例とも、最終学習後の時点までに全ての商標用語が適訳語に改善されていた。これら二例の状況は、第 5.2.2.項で述べた「商標用語の存在は、商標案件由来の文の評価スコアが全期間を通じて特許案件由来の文よりも著しく低くなった主要因ではない」という考察のとおりであり、商標審決特有の商標用語・表現やこれに準ずる一般用語の存在は、最終学習後においては、翻訳品質への悪影響はほとんど及ぼしていないと結論できる。

なお、審決逆翻訳コーパスの追加学習が翻訳品質に特段の効果を奏していないという点においても、本例も他の実例と同様であった。

5.2.5.5. 実例の定点観測のまとめ

ここまで審決③／④の評価用コーパスから 4 文を対象に、各学習段階の機械翻訳文を定点観測してきた。各実例とも、審査書類①／②と同様、審査書類用語は審決対訳コーパスの学習によって適訳語に改善され、その結果、無学習時に比べて大幅な品質向上が果たされていた。商標審決で多用される商標用語・表現に関しても、第 5.2.2.項で考察したとおり、コーパスの学習につれて他の審査書類用語と同等に改善されていき、最終学習後の時点ではほぼ正しく訳されるようになっていた。

一方、同じく商標審決で多用される傾向が強い固有名詞に関しては、実例 5-3 で「水甚社」という固有名詞の状況を確認したが、無学習時から最終学習後まで常に「未知語」となっており、コーパス学習後にも改善は見られなかった。こちらも第 5.2.2.項で述べたとおり、固有名詞がコーパスの学習効果をほとんど得られない、翻訳難易度の極めて高いジャンルであることが示された形となる。商標審決由来の文は固有名詞の含有率が高く、こうした文が評価用コーパスに多数含まれていたことが、審決③／④の全体品質を低下させた一因である可能性が高い。

本項の分析の焦点であった審決逆翻訳コーパスの有無による翻訳品質や傾向の違いに関しては、各実例とも、審決③と④の機械翻訳文の間に逆翻訳コーパスの影響がうかがえるような品質差や傾向の違いは見られなかった。両システムの機械翻訳文には異なる部分はあ

ったが、共通の訳語が異なるタイミングで採用されるようなケースが大半であり、審決逆翻訳コーパスがもたらした変化というよりは、「みんなの自動翻訳」自体の傾向である学習前後の訳語の不安定な変化によるものと見られる。事実、「みんなの自動翻訳」で訳語変化が抑制された大規模アップデート以降（第 2 回後以降）は両システムの機械翻訳文は類似度が非常に高くなり、各例とも最終学習後には両システムの機械翻訳文の差異は誤差レベルの僅かな違いのみに収れんした。こうした各実例の状況からは、審決逆翻訳コーパスの有無は、自動評価スコアが示したとおり、実質的な機械翻訳品質にはほぼ影響していないと結論される。

5.3. 審決③／④システムの人手評価結果

本項では、審決③／④システムに対して実施した人手評価の結果を記す。

5.3.1. 人手評価の概要

審決③／④の人手評価は、評価用コーパス 2,024 文のうちの 100 文を対象とした。両システムとも同じ 100 文を使用し、各学習回におけるシステム間の相対比較も実施した。なお審決③／④に関しては、人手評価の実施回数及びタイミングは自動評価と同一であり、無学習時及び第 1 回～第 3 回学習後の全 4 回実施した。

5.3.2. 人手評価観点

審決③／④の人手評価は、審査書類①／②と共通の 3 観点到下記④を加えた 4 つの観点から行った。

- ① 「内容伝達レベル」の評価
- ② 学習前後の翻訳文の相対評価
- ③ エラーのカテゴリ別カウント
- ④ 審決③と審決④の相対評価

5.3.3. 人手評価の対象文

人手評価対象文は審決③／④の評価用コーパス 2,024 文から 100 文を選定した。

5.3.4. 各評価項目の集計結果

全4回にわたり実施した審決③及び審決④システムの人手評価の結果を本項にまとめる。
第5.3.2.項に挙げた4つの評価観点ごとに各回の集計結果を示し、分析・考察する。

5.3.4.1. 「内容伝達レベル」の評価

人手評価対象 100 文について、各学習段階の審決③／④システムによる機械翻訳文の内容伝達レベルの平均値と、レベル別の分布を下表に示す。

表 5.3.1. 各回学習前後の審決③／④の「内容伝達レベル」集計結果

書類種別	学習	平均	5	4	3	2	1
審決③	無学習時	2.52	8	16	24	24	28
	第1回学習後	3.31	26	19	28	14	13
	第2回学習後	4.37	60	25	8	6	1
	第3回学習後	4.42	58	31	7	3	1
	増減	+1.90	+50	+15	-17	-21	-27
審決④	無学習時	2.52	8	16	24	24	28
	第1回学習後	3.32	24	29	15	19	13
	第2回学習後	4.24	48	36	9	6	1
	第3回学習後	4.31	52	33	10	4	1
	増減	+1.79	+44	+17	-14	-20	-27
[参考] 審査書類① ※100 文換算	無学習時	3.14	14	24	31	23	8
	第5回学習後	4.09	48	26	16	8	2
	第10回学習後	4.20	53	23	15	7	2
	第15回学習後	4.67	77	14	7	2	0
	第18回学習後	4.69	78	15	5	2	0
	増減	+1.55	+64	-9	-26	-21	-8

無学習時の審決③／④の内容伝達レベルの平均値は 2.52 であり、審査書類①の無学習時のレベル (3.15) に比べて-0.63 と顕著に低かった。この状況は、自動評価スコアが示した両者の優劣と一致している。内容伝達レベル 2.52 は、レベル 2 「いくつかの重要情報は正確に伝達されている (20%～)」とレベル 3 「半分以上の重要情報は正確に伝達されている (50%～)」の中間に相当する。

これに対し、審決対訳コーパスを学習させた第 1 回学習後の内容伝達レベルは、審決③が 3.31 (前回比+0.79)、審決④が 3.32 (前回比+0.80) と大幅に、かつ両システムほぼ同

程度に向上した。ただし、審査書類①の初期学習時（第 5 回学習後）の向上度（+0.95）と比べると伸びはやや鈍く、その結果、両者の差は無学習時よりもさらに広がっている。

ただし、これに続く第 2 回学習後においては、審決③の内容伝達レベルは 4.37 (+1.06)、審決④は 4.24 (+0.92) と、両システムとも初回学習後を上回る伸びを見せた。第 2 回学習前には「みんなの自動翻訳」の大規模アップデートが実施されており、今回の顕著なスコア向上の主要因であることは確実であるが、審査書類①の同時期（第 5 回学習後～第 15 回学習後）の伸びは+0.58 にとどまっており、ここで学習初期の改善の遅れをかなり挽回した形となる。

続く第 3 回学習後は、審決③が 4.42 (+0.05)、審決④が 4.31 (+0.07) と両システムとも微増にとどまった。ただし、審決③／④では第 3 回学習が最終学習回であり、この期間は大规模アップデート以後の期間にあたる。この時期は審査書類①／②も品質は現状維持傾向となっており、審決③／④もこれと同様、大規模アップデート直後に向上した品質水準がその後も維持されていることが示された形となる。なお、審決③／④の最終学習後の内容伝達レベルはそれぞれ審査書類①と比べて-0.27、-0.38 と依然低いが、その差は無学習時の-0.63 に比べるとかなり縮まっている。

なお、各学習段階での審決③と審決④の内容伝達レベルを比較すると、初回学習後はほぼ同値、第 2 回学習後は審決③が審決④をやや上回る結果となった。

これら審決③／④の各回の内容伝達レベルの状況は、各回における推移、審査書類①／②との比較、そして審決③／④間の相対比較のいずれにおいても、第 5.2.項に示した自動評価スコアの推移とほぼ完全に合致している⁶²。この結果により、自動評価結果に基づいて導出し、実例の定点観測で検証した各種推論の妥当性が裏付けられた。以下、具体的に示す。

まず、無学習時の内容伝達レベルが審査書類①に比べて明らかに低かったことにより、審決③／④の評価用コーパス自体の翻訳難易度が他のシステムよりも有意に高かったことが改めて示された。

無学習時の審査書類①との内容伝達レベルの差は、審決対訳コーパス全件を学習したことで明らかに縮まったが、それでも完全に同等とはならず、最終学習後も審決③／④の内容伝達レベルは審査書類①よりもやや低かった。これは、審決③／④の評価用コーパスの難易

⁶² 唯一、第 1 回学習後の審決③／④間の相対比較では、自動評価では審決④が審決③を有意に上回っていたが、内容伝達レベルではほぼ同水準であった。第 5.2.5 項に示した実例定点観測からは、どちらが正しいとも言い切れない。

度を高めていた要素の一部は審決対訳コーパスで学習可能だが、一部は学習困難であることを示している。自動評価結果の分析では、審決③／④の翻訳難易度が高い理由として、審決がカバーする審査書類用語（商標用語）のバリエーションは各審査書類よりも広いことと、特に商標審決において固有名詞が多用されることを挙げた。前者は「審決対訳コーパスで学習可能」なもの、後者は「（コーパスでは）学習困難である」ものの代表例といえ、内容伝達レベルの示した状況は、上記分析内容と整合している。

また、審決③／④と審査書類①との内容伝達レベルの差は、コーパス初回学習後の第1回学習後には無学習時よりもさらに広がったが、その後は差が縮まっていき、最終学習後は無学習時よりも明らかに差が詰まった。この一連の経緯も、審決に用いられる審査書類用語のバリエーションが広いため初回学習では完全にはカバーされず、このため初回学習時点で大半の審査書類用語が適訳語に改善された審査書類①との差は一時的に広がったが、その後コーパスを増量していくにつれて初回時に改善できなかった用語（例えば実例 5-3 の「請求人：demandee」や 5-4 の「称呼：pronunciation」など）も徐々に改善されていき、最終的にはおおむね正しく訳されるようになった、という定点観測の各実例の状況と合致する。

さらに、全学習期間を通じて審決④の内容伝達レベルが審決③と同値か、あるいは若干下回る結果であったことから、自動評価スコアや各実例の定点観測から看取されたとおり、審決逆翻訳コーパスの有無が審決③／④の翻訳品質にほぼ影響を及ぼしていないことも改めて確認された。

5.3.4.2. 学習前後の翻訳文の相対評価

審決③／④の評価対象文 100 文について、学習前後の機械翻訳文のどちらが翻訳文として優れているかを文ごとに比較し、各回学習後の機械翻訳品質が学習前に比べて「向上」したか「劣化」したか、それとも「同等」であるかを判定した。下表に各書類・各回の集計結果を示す。

表 5.3.2. 審決③／④各回学習前後の翻訳品質の「相対評価」結果

書類種別	比較対象		学習後に		
	学習前	学習後	向上	同等	劣化
審決③	無学習時	第 1 回学習後	66 文	20 文	14 文
	第 1 回学習後	第 2 回学習後	61 文	30 文	9 文
	第 2 回学習後	第 3 回学習後	15 文	74 文	11 文
審決④	無学習時	第 1 回学習後	65 文	18 文	17 文
	第 1 回学習後	第 2 回学習後	46 文	44 文	10 文
	第 2 回学習後	第 3 回学習後	21 文	69 文	10 文
[参考] 審査書類① ※100 文換算	無学習時	第 5 回学習後	70 文	20 文	10 文
	第 5 回学習後	第 10 回学習後	25 文	58 文	17 文
	第 10 回学習後	第 15 回学習後	38 文	60 文	2 文
	第 15 回学習後	第 18 回学習後	7 文	89 文	4 文

無学習時と第 1 回学習後の比較では、審決③／④とも全体の約 65%の文が学習後に「向上」、約 20%の文が「同等」、約 15%の文が「劣化」という評価となった。これは [参考] として示した審査書類①の無学習時と第 5 回学習後の相対評価結果とほぼ一致する。各実例の定点観測では初回学習後の機械翻訳文では審査書類用語を中心に品質が顕著に向上していたが、本評価結果は、他の多くの文でも同様の改善が果たされたことを示唆している。

続く第 1 回学習後と第 2 回学習後との比較は、「みんなの自動翻訳」の大規模アップデート前後の比較となり、審決③では 61 文、審決④では 46 文と、全体の半数前後の文で引き続き品質が「向上」していると判定された。審査書類①の同時期（第 10 回と第 15 回の比較）の評価では「同等」が 60 文と過半数を占め、「向上」は 38 文にとどまったが、審決③／④の場合、前項で述べたとおり、商標用語をはじめとする審査書類用語の改善が審査書類①／②に比べると遅れていた傾向が見られており、第 2 回学習時点での改善の余地が大きかったものと推察される。

第 2 回学習後と最終学習回である第 3 回学習後の相対評価では、過去 2 回とは傾向が一変し、学習前後で「同等」と評価された文が審決③で 74 文、審決④で 69 文と 70%前後を

占めた。この結果は、審査書類①／②と同様、大規模アップデート以降は個々の文においても翻訳品質はほぼ現状維持状態に移行したことを示している。それでも審査書類①の 89 文に比べると「同等」文のカウント数はやや少ないが、これは前記したとおり「審査書類用語の改善が審査書類①／②に比べると遅れていた」こと、及び審決③／④の第 2 回後と第 3 回後との間隔が審査書類①の第 15 回後と第 18 回後の間隔よりも広がったという条件の違いによる差異と推察される。

本評価結果により、審決③／④においても、審査書類①／②と同じく、大規模アップデート以降は個々の文レベルでも翻訳品質が現状維持状態に移行していたことが確認された。この結果はすなわち、審決③／④においても、審決由来のコーパスで学習可能な審査書類用語（商標用語を含む）は最終学習後の時点ではおおむねカバーされており、さらにコーパスを増量したとしても改善はごく一部の文に限られるであろうことを暗示している。これは見地を変えれば、本事業で作成した審決対訳コーパスが、審決の全体的な翻訳品質を効果的に向上させるための学習データとしては、ほぼ十分な量に達していたことを意味しているといえる。

5.3.4.3. エラーのカテゴリ別カウント

審決③審決④それぞれの各学習時点における機械翻訳文 100 文について、カテゴリ別にエラーの有無を文単位でカウントした。全 10 種のエラーの各回のカウント結果を下表に示す。

表 5.3.3. 審決③／④各学習回時のカテゴリ別エラー該当文カウント結果

カテゴリ	システム	無学習	第 1 回	第 2 回	第 3 回	向上度
特許誤訳	審決③	60	17(-43)	5(-12)	4(-1)	-56
	審決④		12(-48)	6(-6)	5(-1)	-55
技術／ 一般誤訳	審決③	25	24(-1)	8(-16)	10(+2)	-15
	審決④		18(-7)	10(-8)	9(-1)	-16
未知語	審決③	2	0(-2)	0(±0)	0(±0)	-2
	審決④		0(-2)	0(±0)	0(±0)	-2
係り受け エラー	審決③	34	30(-4)	6(-24)	8(+2)	-26
	審決④		37(+3)	9(-28)	9(±0)	-25
英語表現 エラー	審決③	39	43(+4)	37(-6)	39(+2)	±0
	審決④		39(±0)	41(+2)	35(-6)	-4
湧き出し	審決③	41	34(-7)	6(-28)	2(-4)	-39
	審決④		35(-6)	6(-29)	2(-4)	-39
欠落	審決③	57	35(-22)	20(-15)	13(-7)	-44
	審決④		41(-16)	16(-25)	11(-5)	-46
数値 エラー	審決③	18	20(+2)	6(-14)	6(±0)	-12
	審決④		20(+2)	9(-11)	7(-2)	-11
記号 エラー	審決③	1	1(±0)	1(±0)	1(±0)	±0
	審決④		1(±0)	1(±0)	0(-1)	-1
その他の エラー	審決③	0	0(-1)	0(±0)	0(±0)	±0
	審決④		1(±0)	0(-1)	0(±0)	±0

以下、主要なエラーカテゴリについて考察する。

5.3.4.3.1. 「特許誤訳」

特許誤訳は、審決対訳コーパス学習の主眼である特許・審査書類用語の誤訳や不適訳をカウントするカテゴリである。審決③／④においては、商標用語の誤訳もここに含まれる。

無学習時点（デフォルト）の「みんなの自動翻訳」では、人手評価対象 100 文中 60 文と過半数の文で何らかの「特許誤訳」が発生していた。審査書類①（平均 72 文⁶³）、審査書類②（61 文）の無学習時とおおむね同じ状況といえる。

これに対し、最終回となる第 3 回学習後の「特許誤訳」の発生文数は審決③で 4 文、審決④で 5 文と大幅に減少した。これも審査書類①（平均 4 文）、審査書類②（4 文）と同水準であり、この結果から、審決対訳コーパスが、特許・審査書類用語（商標用語も含む）の翻訳品質において、審査書類対訳コーパスと同等の高い改善効果を奏したことが見て取れる。

なお、審決③／④で最終学習後に指摘された「特許誤訳」には「甲 n」の訳し方に関するものが多く、両システムで同じ 3 文（#1174、#1187、#1946）がカウントされている。以下、一例を示す。

【文番号：#1946】

原文	甲第 1 号証発明に、上記「1（2）ア」の設計変更をすれば、甲第 1 号証発明は、[相違点 2] に係る構成と同様の構成を有することは明らかである。
正解訳文	It is obvious that, if the design change of the above-mentioned "1(2)A" is applied to Evidence A No. 1 Invention, Evidence A No. 1 Invention has a constitution similar to the constitution concerning [Different Feature 2].
審決③ 第 3 回学習後 特許誤訳：有	If the design change of the above-mentioned "1 (2) A" is made to Invention a- 1, it is obvious that Invention a- 1 has a configuration similar to the configuration relating to [Different Feature 2].

原文の「甲第 1 号証発明」が、審決③では Invention a- 1 と訳されている。評価者はこれに対し、正しいフォーマットは正解訳文で採用された Evidence A No. 1 Invention であるとして、「特許誤訳」と判定した。なお、#1174 と #1187 でも「甲 1 発明」が Invention a- 1 と訳されており、同じ理由で「特許誤訳」と判定されている。な審決④も同じ状況である。

⁶³ 審査書類①は 4 種の審査書類それぞれについて 100 文ずつ全 400 文をカウント対象としたが、比較を容易にするためここでは 4 書類の平均値を示す。以下同じ。

「甲 1」を Evidence A No.1 と訳す正解訳文のフォーマットは、過去事業から継続的に採用されている審決翻訳の定型フォーマットであり、上記「甲 n 発明」や「甲第 n 号証発明」のように「甲 n」を証拠文献の付番と確定できる場合は、具体的で非常に分かりやすい。しかしながら、「甲」という語自体は Evidence（証拠）という意味は持たず、このため使われ方によっては誤訳となるリスクもある。その点、現状の「甲 1 : a-1」という訳し方は無難であり、機械翻訳された案件内で統一されていれば、読み手にも特に不都合はないであろう。

この「甲 n」以外の最終学習後の審決③／④における「特許誤訳」は、審決③で 1 文、審決④で 2 文のみであった。以下、各例を示す。

【文番号：#259】

原文	ところで、平成 1 5 年特許法 1 2 3 条 2 項では、「何人も請求することができる」とされ、利害関係人に限らず、利害関係を有さない者も請求可能であったのであるから、請求人の当該主張は根拠がない。
正解訳文	However, Article 123(2) of the 2003 Patent Act specifies that "any person may file a request for a trial for patent invalidation" with no restriction on the interested person, and those who do not have an interest are also allowed to file a request. Hence, said demandant's allegation lacks basis.
審決③ 第 3 回学習後 特許誤訳：有	Incidentally, in Article 123 (2) of the Patent Act in 2003, it was said that "anyone can request", and not only interested persons but also a person who does not have an interest could request, so that the allegation of the demandant is groundless.

上掲した文番号#259 では、「(特許法では) …とされ」という箇所が it was said that…と訳されていることについて、原文のニュアンスと異なっており、訳語及び時制に問題ありと指摘された。「される」は一般語であるが、特許法の文脈で使われているため評価者は「特許誤訳」と判定している。

機械翻訳文の it was said that は「…と言われていた」というニュアンスであり、本例の文脈では確かに適訳とは言い難い。正解訳文のように specify を用いるか、評価者の提案どおり it is stipulated that…等とするのが最善であろう。

ただし、「～とされる」という語はあらゆる文脈で多用される一般語であり、本例は、この語が特許法の文脈で使われている場合のみ specify/stipulate と訳すという、極めて高度

な訳し分けが求められるケースである。このため、機械翻訳にとっては非常に難易度が高く、審決④でも同じ誤訳が発生しているのをはじめ、両システムとも無学習時から最終学習後に至るまで常に「特許誤訳」が発生している⁶⁴。

残る 1 文は文番号#290 である。こちらは審決④のみで検出された。

【文番号：#290】

原文	訂正が、明瞭でない記載の釈明を目的としたものとして認められるには、 特許がされた 明細書又は特許請求の範囲のそれ自体意味の不明瞭な記載、又は、 特許がされた 明細書又は特許請求の範囲の他の記載との関係で不合理を生じているために不明瞭となっている記載を正し、その本来の意味を明らかにするものであることが必要であるといえる。
正解訳文	When a correction is approved as a correction made for the purpose of clarification of an ambiguous statement, it is necessary to correct a statement in the patented specification or the claims whose meaning per se is ambiguous or a statement that became ambiguous because of unreasonableness caused in relation to other statements in the patented specification or the claims, and then clarify the original meaning of the statement.
審決③ 第 3 回学習後 特許誤訳：無	In order for the correction to be recognized as aiming at clarification of an ambiguous statement, it is necessary to correct an ambiguous statement of the meaning of the description or the scope of claims for which a patent has been granted , or a statement that has become ambiguous due to unreasonableness in relation to other statements of the description or the scope of claims for which a patent has been granted , and to clarify its original meaning.
審決④ 第 3 回学習後 特許誤訳：有	In order for correction to be approved for the purpose of clarification of an ambiguous statement, it can be said that it is necessary to correct an ambiguous statement of the meaning of itself of the specification or the scope of claims for which a patent has been made , or a statement that is ambiguous because of unreasonableness in relation to other statements in the specification or the scope of claims for which a patent has been made , and clarify the original meaning thereof.

⁶⁴ ただし、採用されている訳語は it was said that とは限らず、it is described と訳されていたり、the demandant alleges と原文中に存在しない文言が湧き出したりしている。

本例では、「特許がされた（明細書）」という箇所の審決④の訳語 **for which a patent has been made** が「特許誤訳」と判定された。つまり、前例と同じく、一般語「された」を特許の文脈でどう訳すかという訳し分けに関する不備である。本例の「特許がされた」は「特許が付与された」の意味であり、正解訳文のように **patented** とするか、審決③のように **a patent has been granted** とするのが適切である。

審決④では第 2 回学習後にも同じく **made** を採用しており、審決④のみに使用した審決逆翻訳コーパスの内容が関与している可能性もある。ただし、海外においても、少なくとも「特許が付与された」の意味で **a patent is made** と言うことはなく、誤訳判定は妥当である。

以上、審決③／④の最終学習後の「特許誤訳」の状況を確認した。実際に誤訳された語句は「甲_n」「（特許法では）…とされ」「特許がされた（明細書）」の 3 語のみであり、いずれも一般的な用語を審決特有の文脈でどう訳すかという、高度な訳し分けに関する不備といえる。ニューラル機械翻訳は文脈に応じた柔軟な訳し分けを得意とする方式ではあるが、上記各例のようにさほど頻出とは言えないような用語・表現においては、必ずしも十分な学習データが確保できるとは限らず、正しく対応できないケースもあり得る。とはいえ、最終学習後に検出された「特許誤訳」はこうした少数の特殊なケースに限られ、全体的に見れば、審決文に用いられる主な審査書類用語は、審決対訳コーパス全件の学習により、審決③／④ではほぼ完璧に翻訳されるようになったと結論できる。

なお、最終的な審決③／④の翻訳品質への悪影響は小さい（⇒5.2.2.）と論じていた商標用語に関しては、最終学習後に「特許誤訳」と判定された語は存在しなかった。この結果からは、当時の推論どおり、商標用語も他の審査書類用語と同じく、審決対訳コーパス全件の学習によっておおむねカバーされたと考えられる。

5.3.4.3.2. 「未知語」

審決③／④の人手評価対象文 100 文中、「未知語」が発生したのは無学習時の 2 文のみであった。具体的には文番号#637 において「乙第 21 号証」という審査書類用語に対応できず「乙第」が「未知語」となり、文番号#962 においても審査書類用語「委任省令要件」を正しく解釈できず「令要」が「未知語」扱いとなった。ただし、これら 2 語もコーパス初回学習後にはそれぞれ Evidence B No.21、Ministerial Ordinance Requirement と正しい訳語に改善され、以降は最終学習時まで「未知語」は発生していない。

ただし、評価用コーパス全体を見ると、最終学習後も固有名詞を中心に「未知語」はいくつか残留している。第 5.2.5.3.項に示した実例 5-3 の「水甚社」（文番号#46 など 12 文）を始め、文番号#9 の「プレッピーハンドブック」、#72 の「織研新聞」、#1444 の「カムバックミニメ」の 4 語、15 文が「未知語」となっている。

なお、人手評価対象文にも文番号#52 に「株式会社ビイエムプランニング (BM Planning Co., Ltd.)」という固有名詞を含む文が存在する。この語は「未知語」とはならなかったが、無学習時は the company yemm Planning と誤訳されたため「技術／一般誤訳」、第 1 回学習後は両システムとも翻訳文から消失したため「欠落エラー」、第 2 回学習後は審決③では Bye M-planning Co., Ltd.、審決④では BII M-planning Co., Ltd と誤訳されており再び「技術／一般誤訳」、そして第 3 回学習後も審決③が BII M-planning Co., Ltd、審決④では BUI M-planning Co., Ltd と誤訳されたため「技術／一般誤訳」と、最終学習後まで毎回エラーが発生していた。このように、固有名詞はたとえ「未知語」扱いされていなくても、誤訳や欠落など正しく訳されていないことが多いと推察される。

5.3.4.3.3. 「英語表現エラー」

「英語表現エラー」のカウント数は無学習時が 39 文であったのに対し、最終学習後も審決③で 39 文、審決④で 35 文と他のエラーに比べて改善が鈍かった。これは審査書類①／②にも見られた各システム共通の現象であるが、エラーの性質上、重大視する必要はないと考える。この現象が生じる理由に関する考察は第 3.3.5.3.5.項を参照されたい。

なお、「英語表現エラー」は、ネイティブの英文に由来している審決逆翻訳コーパスの追加学習の効果が比較的表れやすいカテゴリと推察されたが、実際のカウント結果は両システムほぼ同等であり、顕著な効果は見られなかった。

5.3.4.3.4. 「湧き出しエラー」

審決③／④と審査書類①／②のカウント結果の比較で最も差が大きかったのが、無学習時の「湧き出しエラー」のカウント数である。審査書類①／②がそれぞれ 18 文、19 文であったのに対し、審決③／④では 41 文と倍以上の発生率であった。

「みんなの自動翻訳」では、「湧き出しエラー」は対処不能な語が翻訳文から欠落した際に、その埋め合わせとして文中の別の語が重複出力されるという形で発生することが多い。例えば実例 5-3 (⇒5.2.5.3.) の第 1 回学習後の機械翻訳文では、審決③／④とも「請求人」に相当する訳語が欠落し、その埋め合わせとして審決③では oral cavity (口頭)、審決④では demandee(被請求人)という文中の別の語が重複出力されていた。

また、実例 5-2 (⇒5.2.5.2.) では、無学習時に「被請求人」が a person skilled in the art (当業者)と誤訳されたが、両者は全く意味が異なり、単なる不適訳の域を超えている。このような場合も、単なる誤訳ではなく、「被請求人」の「欠落エラー」と、その埋め合わせとして a person skilled in the art の「湧き出しエラー」が併発したと見るべきである。

このように、「湧き出しエラー」は、「欠落エラー」の埋め合わせとして併発するケースが多い。人手評価対象文 100 文においても、例えば文番号#22 で「本件商標」の欠落と present invention (本発明) の湧き出しが併発していたり、#69 で「請求人」の欠落と claims (請求項) の湧き出しが併発していたりと、「湧き出しエラー」が発生した 41 文のうち 29 文が、「欠落エラー」との併発であった。

こうした状況から、審決③／④における無学習時の「湧き出しエラー」の多発は、審決③／④の評価用コーパス中に「みんなの自動翻訳」では解析不能で訳出できなかった語句が審査書類①／②よりも多く含まれていたためである可能性が高い。ここまで、第 5.2.1.項では審決が各審査書類に比べて審査書類用語のバリエーションが広いこと、第 5.2.2 項では商標審決で商標用語や固有名詞が多用されていることを指摘し、これらが審決③／④の翻訳難易度を高めていると考察したが、本カウント結果で判明した無学習時の「湧き出しエラー」の頻発も、これらの指摘と整合する。なお、前掲した「特許誤訳」のカウント結果では、審決③／④と審査書類①／②との間とで差がなかったが、審決③／④では審査書類用語が欠落してしまうことが審査書類①／②より多かったためとも考えられる。

審決③／④の「湧き出しエラー」発生数は学習が進むにつれて順調に減少し、最終学習後には審決③、審決④ともに 2 件と、審査書類①／② (1 件、2 件) と同水準となった。最終学習後に「湧き出しエラー」としてカウントされた事例も、大規模アップデート以前に多発していた典型的な「湧き出しエラー」とは性質が異なるものが大半であった。以下、一例を

示す。

文番号-#1731

原文	そうすると、前記アで述べた構成の変更を行った「前槽 1 2」における膨潤水 1 4 の可塑剤の濃度の上限値や、「後槽 1 3」における膨潤水 1 4 の可塑剤の濃度の上限値を、どのような値に設定するのかは、製造される偏光フィルムの品質や安定性と、必要とされるコスト等を総合的に勘案して、当業者が適宜決定すれば足りる設計上の事項というべきである。
正解訳文	Then, to what value the upper limit value for the concentration of plasticizer in swelling water 14 in the "fore tank 12" that has a configuration modified as described in above A, and the upper limit value for the concentration of plasticizer in swelling water 14 in the "aft tank 13" should be set should be deemed as design choice that may be appropriately determined by a person skilled in the art comprehensively taking into consideration quality and stability of manufactured polarizing films and required cost, etc.
審決③ 第 3 回学習後 湧き出し：有	Then, it should be said that it is a design matter that is sufficient for a person skilled in the art to appropriately determine the upper limit value of the concentration of the plasticizer of the swelling water 14 in in the "front tank 12" and the upper limit value of the concentration of the plasticizer of the swelling water 14 in the "rear tank 13" to which the change of the configuration described in the above A has been made, in comprehensive consideration of the quality and stability of polarizing film to be produced, required cost, and the like.

本例は、原文の『「前槽 1 2」における膨潤水 1 4 の可塑剤』に該当する部分が he plasticizer of the swelling water 14 in in the "front tank 12"となっており、in が重複している。これにより、不要な in の「湧き出しエラー」と判定された。

本例も、広い意味では翻訳文中の語句が不必要に繰り返されていると言える。ただし、学習初期の典型的な「湧き出しエラー」では、文中の語句、主に具体的な名詞やフレーズが、（しばしば欠落した語句の埋め合わせとして）本来無関係な位置に重複出力されていたのに対し、本例は前置詞 in が、正しい位置で単に重複しているのみである。このため、印象が大きく異なる。エンジン内でどのような事象が起こっているかは外部からは知り得ないが、例えば「A における B の C」というフレーズの学習の際、例えば「A における」が in A、

そして「における B」が B in と双方が in を伴ったまま翻訳文に採用された等、翻訳文の組み立て時の誤りである可能性が高いと推察される。

このように、最終学習後の「湧き出シエラー」のカウンタ数はごく少数であり、かつ、旧「みんなの自動翻訳」で頻発していた典型的な「湧き出シエラー」は検出されなくなった。この状況は審査書類①／②と同様であり、審決③／④においても、エンジンのロジックに起因する典型的な「湧き出シエラー」は、「みんなの自動翻訳」の大規模アップデート以降は根本的に解消されたと結論できる。

このように、審決③／④においても、「みんなの自動翻訳」の大規模アップデートを経た第2回以降は、アップデート以前に見られた典型的な「湧き出シエラー」は根本的に解消されたと見なせる。ただし、上例のような具体的な主語の補足は、ここまで見てきた各実例では裏目に出ていることが多く、「湧き出シエラー」ではないにせよ、リスクの高い処理であることは確かである。

5.3.4.3.5. 「欠落エラー」

審決③／④では、「湧き出シエラー」と同じく、「欠落エラー」も全期間にわたって審査書類①／②より多く発生している。原文中の語句が解析できなかったことに直結するエラーであり、本エラーの多発により、審決③／④の評価用コーパスに解析不能な語句がより多く含まれていたことがわかる。

「欠落エラー」も学習の進行により改善が進み、無学習時の 57 文から最終学習後は審決③が 12 文、審決④が 10 文まで減少した。とはいえ、審査書類①／②（4 文、6 文）に比べると残留数はやはり多い。

ただし、その大多数は、例えば「必要であるといえる」(#290、審決③)、「図 2 にも」(#458、審決③)、「そして、前記端縁部…」(#575、審決④)、「おおむね次の記載…」(#886、審決③)、「Q E - 0 6 0」(#908、審決③)、「発明の詳細な説明の記載は…」(#962、審決④)など、やや過度な省略の結果とも捉えられるものであり、難易度の高い語句が消失したり、それまで訳せていた重要な語句が突発的に消失するといった典型的な「欠落エラー」とはタイプが異なる。つまり、典型的な「欠落エラー」に関しては、審決③／④でも根本的に解消されたものと見なせる。

なお、審査書類①／②においては、最終学習後に残存した「欠落エラー」は主に第 3.4.7.1. 項で述べた「カギカッコで括られた原文において翻訳文が途絶するエラー」であった。審決③／④の人手評価対象 100 文中にはこの条件に該当するものが存在しなかったが、評価用

コーパス全体では文番号#340 や#342 など多数の文で最終学習後まで同エラーが発生しており、本エラーは審決③／④でも等しく発生することがわかった。

加えて、審決③／④の人手評価対象文#1174 では、上記エラーの発生条件には該当していないが、同様に翻訳文が文頭から 254 文字に達した箇所で途絶するという、新たなタイプの途絶エラーが検出された。以下、実例を示す。

【文番号：#1174】

原文	甲 1 発明における「強化繊維を長さ 5 mm にカットして得たチョップド強化繊維」は、本件特許発明 1 における「強化繊維」に相当し、以下、同様に、「熱可塑性樹脂 (A)」及び「熱可塑性樹脂 (B)」は「熱可塑性樹脂」に、「繊維強化樹脂シート」は「繊維強化プラスチック成形体用シート」に、それぞれ相当する。
正解訳文	The "chopped reinforcing fibers obtained by cutting reinforcing fibers with a length of 5 mm" in the A-1 Invention corresponds to "reinforcing fibers" in Patent Invention 1. Hereinafter, similarly, "thermoplastic resin (A)" and "thermoplastic resin (B)" correspond to "thermoplastic resin", and "a fiber reinforced resin sheet" corresponds to "a sheet for fiber reinforced plastic mold".
審決③ 第 3 回学習後 欠落：有	"Chopped reinforcing fibers obtained by cutting reinforcing fibers into a length of 5mm" in Invention a- 1 corresponds to "reinforcing fibers" in Patent Invention 1, and, in a similar fashion, "thermoplastics (A)" and "thermoplastics (B)" respectively co
審決④ 第 3 回学習後 欠落：無	The "chopped reinforcing fibers obtained by cutting reinforcing fibers to a length of 5mm" in Invention a- 1 corresponds to the "reinforcing fibers" in Patent Invention 1. Similarly, the "thermoplastics (A)" and "thermoplastics (B)" correspond to the "thermoplastics", and the "fiber-reinforced resin sheet" corresponds to the "sheet for fiber-reinforced plastic moldings".

本例の第 3 回学習後の翻訳文は、文頭から 254 文字の時点で翻訳文が途絶しており、「カギカッコで括られた原文において翻訳文が途絶するエラー」と同じ様相を呈している。しかしながら、本例の原文は、カギカッコは多用されているものの、本文全体がカギカッコで括られているという、第 3.4.7.1.項に示したエラーの発生条件には該当していない。かつ、同

じ原文であるにもかかわらず、審決④ではエラーが発生していない⁶⁵。

評価用コーパス全体を調査したところ、上例と同様の途絶エラーは本例以外にも例えば文番号#1929 や#1978 などでも発生している。ただし、文中にカギカッコを多数含む文は他にも存在し、どのような条件で本エラーが発生するかは不明である。また、上記#1929 や#1978 は審決③、④の双方でエラーが発生したが、#910 や#913 など審決③のみで発生している文もある。さらに、#910 のように、第 3 回学習後は審決③のみでエラーが発生したが、第 2 回学習後は審決④のみで発生していたなど、同じ文でも必ず発生するとは限らず、学習回やシステムによって発生する場合としない場合がある。

このように発生原因、発生条件とも不明なエラーであるが、翻訳文が 254 文字で途絶するという状況自体は第 3.4.7.1.項で報告した翻訳文途絶エラーと共通しており、発生メカニズムは同根である可能性が高い。

5.3.4.3.6. 「数値エラー」

審決③／④の最終学習後の「数値エラー」はそれぞれ 6 文、7 文と、こちらも審査書類①／②（各 1 文）に比べて明らかに多い。これは、和暦→西暦の変換誤りという、審決③／④のみで複数件が検出された「数値エラー」の存在による。以下、一例を示す。

【文番号：#626】

原文	その後、平成 2 7 年 5 月 1 1 日付けで審理事項通知がされ、平成 2 7 年 5 月 2 4 日付け（同年 5 月 2 5 日差出し）で被請求人より上申書が提出され、平成 2 7 年 6 月 1 5 日付けで被請求人より口頭審理陳述要領書が提出され、平成 2 7 年 6 月 2 2 日差出しで請求人より口頭審理陳述要領書が提出された。
正解訳文	Notification of matters to be examined were notified as of May 11, 2015. A written statement was submitted by the demandee as of May 24, 2015 (sent on May 25, 2015). An oral proceedings statement brief was submitted by the demandee as of June 15, 2015. An oral proceeding statement brief was sent and submitted by the demandant on June 22, 2015.

⁶⁵ 審決④も、「熱可塑性樹脂」が thermoplastics と訳されていることについて「樹脂：resin」が欠落しているとされ、実際は欠落エラーにカウントされているが、本項の論点ではないため、上欄では混乱を避けるべく「欠落：無」とした。

審決③ 第3回学習後 数値エラー：有	Thereafter, a notice of matters to be examined was issued on May 11, 2015, a written statement was submitted by the demandee on May 24, 2015 (submitted on May 25, 2015), an oral proceedings statement brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2019.
審決④ 第3回学習後 数値エラー：有	After that, a notification of matters to be examined was issued on May 11, 2015, a written statement was submitted by the demandee on May 24, 2015 (delivered on May 25, 2016), an oral proceedings statement brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2015.

「みんなの自動翻訳」では、原文中の和暦は翻訳文では西暦に変換される。本例の原文には「平成27年」が4回、「同年」が1回出現する。平成27年は西暦2015年だが、審決③の機械翻訳文では文末の1回のみ2019と誤った西暦年に変換されている。一方、審決④もカッコ補記内の1回（「同年」に相当）のみ2016と、やはり誤った西暦年となっている。

両システムとも西暦誤りは5箇所中1箇所のみであり、それ以外は正しく2015と変換されている。また、審決③と④とでエラーの発生箇所や誤って採用された西暦年も異なっており、この一例の状況からは、現象が判然としない。

そこで、もう一例、同じく和暦→西暦変換エラーが生じているケースを、評価用コーパス中からピックアップして示す。

【文番号：#1233】

原文	特許第6153560号の請求項1～3に係る特許についての出願は、平成24年8月6日（優先権主張平成23年8月12日）に出願した特願2012-174425号の一部を平成27年4月10日に新たな特許出願としたものであって、平成29年6月9日にその特許権の設定登録がされ、平成29年6月28日に特許掲載公報が発行された。
正解訳文	The patent application concerning claims 1-3 of Japanese Patent No. 6153560 is a divisional application filed on Apr. 10, 2015 from Japanese Patent Application No. 2012-174425 filed on Aug. 6, 2012 (Priority Claim, Aug. 12, 2011), the establishment of the patent right was

	registered on Jun. 9, 2017, and a gazette containing the patent was issued on Jun. 28, 2017 .
審決③ 第3回学習後 数値エラー：有	The application for the patent according to Claims 1 to 3 of Japanese Patent No. 6153560 is a divisional application filed on April 10, 2017 from Japanese Patent Application No. 2012 - 174425 filed on August 6, 2012 (priority claim : August 12, 2011). The establishment of the patent right was registered on June 9, 2017, and a gazette containing the patent was issued on June 28, 2015 .
審決④ 第3回学習後 数値エラー：無	The application for the patent according to Claims 1 to 3 of Japanese Patent No. 6153560 is a divisional application filed on April 10, 2015 from Japanese Patent Application No. 2012 - 174425 filed on August 6, 2012 (priority claim : August 12, 2011). The establishment of the patent right was registered on June 9, 2017, and a gazette containing the patent was issued on June 28, 2017 .

本例は、同じ原文を機械翻訳した結果、審決④は「平成27年4月10日」を [April 10, 2015](#)、「平成29年6月28日」を [Jun. 28, 2017](#) とそれぞれ正しく西暦変換できているのに対し、審決③はそれぞれ [April 10, 2017](#)、[June 28, 2015](#) と誤った西暦変換がなされている。

ただし、審決③の西暦誤りは、それぞれが誤った西暦に変換されたのではなく、2015 と 2017 が入れ替わった結果であるとも見なせる。他の事例を見ても、例えば#198 などでも同様に 2 つの西暦が入れ替わっており、西暦が単独で誤っているというケースは見られなかった。

つまり、本エラーは基本的には、文中の数値が入れ替わってしまうという、ニューラル機械翻訳に散見される数値エラーのバリエーションである可能性が高い。冒頭の例はこれには当てはまらないが、これは、「同年」という語句の処理に問題があると考えられる。

冒頭例に見られるように、「みんなの自動翻訳」では、「同年」を具体的な西暦で訳出するように設定されているか、若しくはそう学習されているものと見受けられる。「同年」は、英語に直訳すると of the same year 等となり、具体的な西暦年よりも大幅に冗長となってしまう。このため、英文では、既出の年を再度述べる文脈でも具体的な西暦年を用いるのが一般的である。よって、こうしたデータを学習した機械翻訳エンジンも、「同年」を具体的な西暦で翻訳する傾向が強くなる。

しかしながら、この場合、機械翻訳エンジンは、「同年」が具体的に指す西暦年を、文ご

とに正確に選定しなければならない。現状においては、この選定を誤る場合があり、このため、西暦の誤変換が散見されるものと推察される。

冒頭例もこれに該当する。冒頭例の審決④では、西暦が誤っているのは「同年」の箇所ではないが、これは、「同年」の誤変換と、第二例に見られた数値の文中での入れ替わりとが併発した結果である可能性が高い。

つまり、この和暦西暦変換の誤りは、ニューラル機械翻訳において突発的に発生する文中の数値の入れ替わりと、「同年」は英文では具体的な西暦年で訳されるのが一般的であるという、言語的性質の差異とが複合した結果であるといえる。このため、現状においては、一定量の発生は不可避であろう。

なお、上記「カンマで区切られた複数の数値が3桁以上のときに連結されるエラー」に関しては、審決③／④の人手評価対象文では発生条件に合致する文が存在せず検出されなかったが、評価用コーパス全体では文番号#354 や#1725 など条件に該当する文では同様のエラーが発生していた。

5.3.4.3.7. 審決③と審決④のエラー数比較

前掲の集計結果一覧表に示したとおり、審決③と審決④のエラー発生文数は、各学習段階、全てのカテゴリにおいてほぼ同等であった。つまり、審決④のみに使用した審決逆翻訳コーパスは、各種エラーの発生／解消に関して特段の影響を及ぼさなかったと結論される。

この結果は、ここまで各種評価から看取されていた、審決逆翻訳コーパスは審決④の翻訳品質に特段の効果も副作用ももたっていないという考察内容を肯定している。

5.3.4.4. 審決③と審決④の相対評価

人手評価の最後の観点として、審決③と審決④の相対評価の結果を示す。

審決③と審決④は、学習データは異なる（審決④のみ審決逆翻訳コーパスを追加学習させている）が、評価用コーパスは同一のものを使用している。つまり、同じ文を、学習内容の異なる2つのシステムでそれぞれ機械翻訳していることになる。

ここまで自動評価、内容伝達レベル、そしてエラーのカテゴリ別カウントのいずれにおいても、審決③と審決④の評価結果に有意な差異は見られなかった。ただし、これらの評価結果はあくまで評価用コーパス 2,024 文若しくは人手評価対象文 100 文全体を比較した結果であり、個々の文単位でも両システムの翻訳品質が常に同等とは限らない。

そこで、審決③／④の人手評価では、追加の観点として、両システムの各回学習後の機械翻訳文を文単位で比較し、その優劣を文ごとに判定する相対評価を実施した。以下、各回の集計結果を示す。

表 5.3.4. 審決③／④各回学習後のシステム間相対評価

	第 1 回学習後	第 2 回学習後	第 3 回学習後
審決③が優位	42	24	14
両システム同等	24	64	70
審決④が優位	34	12	16

無学習時は両システムとも「みんなの自動翻訳」をデフォルト状態で使用しているため、機械翻訳文も完全に同一となり、優劣はつかない。このため、相対評価は第 1 回学習後から実施した。

第 1 回学習後の相対評価では、全 100 文中 24 文が「両システム同等」、残る 76 文のうち 42 文が「審決③が優位」、34 文が「審決④が優位」と判定された。この結果から、当時の審決③／④の翻訳品質は、全体としては同等であったが、個々の文単位では品質に優劣がある場合が多く、こうした文単位の優劣がトータルでは両システムで拮抗していたにすぎないことがわかる。つまり、当時の審決③／④は、どちらも翻訳品質が不安定であり、このため文ごとに優劣のばらつきが激しかったといえる。

これに対し、第 2 回学習後、第 3 回学習後と学習が進むにつれて、「両システム同等」と判定される文が 64 文、70 文と大幅に増加し、過半数を超えるようになった。つまり、第 2 回学習後以降の審決③／④では、評価用コーパス全体としておおむね同等の品質である点

は第 1 回学習後と同じだが、個々の文においても、それまで激しかった文ごとの品質のばらつきが解消され、大多数の文が、どちらのシステムでも同等の翻訳文が得られるようになったということになる。こうした傾向の変化は第 5.2.5.項で定点観測した各実例からも見て取れたが、本評価結果により、それが審決③／④の全体傾向であることが改めて示された。

そして、この評価結果は、審決対訳コーパスを全件学習させた後であれば、審決③と審決④は、個々の文単位でも翻訳品質はおおむね同等であることを意味する。第 5.2.4.項の項末では、「変化した文の量が同程度であっても、逆翻訳コーパスの学習の有無により、双方の翻訳傾向に有意な違いが存在している可能性もある」として判断を保留していたが、本評価結果により、少なくとも第 2 回学習後以降、大多数の文において両システム間に実質的な差異が存在していないことが明確に示された。これはすなわち、審決④のみに追加学習させた審決逆翻訳コーパスは、個々の文単位でも、翻訳品質にほとんど影響を及ぼしていないことを意味している。

5.4. 審決③／④システムに関するまとめ

審決③／④システムの機械翻訳の品質及び傾向について、自動評価及び人手評価の結果から以下の点が確認された。

5.4.1. 審決③／④の機械翻訳品質

審決③／④の無学習時の内容伝達レベルは 2.52 であった。レベル 2.52 は、内容伝達レベルの定義に照らせばレベル 2「いくつかの重要情報は正確に伝達されている（20%～）」とレベル 3「半分以上の重要情報は正確に伝達されている（50%～）」の中間値にあたる。第 5.2.5.項で観測した各実例においても、無学習時の機械翻訳文では特許・審査書類用語にほとんど対応できていなかった。審決文ではこうした用語が正しく訳されていない限り文意の理解は困難であり、この低評価も妥当であった。

これに対し、審決対訳コーパス全件学習後の審決③／④の内容伝達レベルは、審決③が 4.42（無学習時から+1.90）、審決④が 4.31（同+1.79）と、両システムとも顕著な向上を示した。最終学習後に到達した両システムの品質水準はレベル 4（「ほとんどの重要情報は正確に伝達されている（80%～）」）を大きく上回っており、機械翻訳として十分に高い品質水準といえる。

なお、審決③と審決④との比較では、自動評価、内容伝達レベルとも、第 2 回学習後以降は、審決逆翻訳コーパスを追加学習させていない審決③のほうが逆翻訳コーパスを学習させた審決④よりも若干高い評価となった。しかしながら、定点観測した各実例において、学習後期の審決③と審決④の品質差は皆無といえ、また、各種エラーの発生状況にも両システム間に差異はなかった。こうした結果からは、審決逆翻訳コーパスは、審決対訳コーパスを学習させたシステムに対しては、特段の品質改善効果も副作用も及ぼさないと結論される。

5.4.2. 審決逆翻訳コーパスについて

本調査分析では、審決逆翻訳コーパスは特段の翻訳改善効果を奏さなかった。

審決逆翻訳コーパスは、英文審決から採取した約 250 万文の英文を、その時点の審決③システムを逆方向（英→日）で使用して日本語翻訳文を取得し、対訳コーパスに加工したものである。

つまり、審決逆翻訳コーパスの日本語文における全ての訳語は、その時点の審決③の学習データで学習可能な語彙となる。審決逆翻訳コーパスの元となる英文約 250 万文には、審決③の学習データ中に存在しない英語も多々含まれているはずだが、審決③の学習データ中に存在しない英語は、当然ながら、それに対応する日本語も学習データ中に存在していな

い。このため、審決③を用いて作成した逆翻訳コーパスの日本語文では、こうした英語に相当する箇所は、「未知語」となるか欠落していることが多いと考えられる。このような状態の逆翻訳コーパスを追加学習しても、有益な学習効果が得られるとは考えられない。

そうすると、審決逆翻訳コーパスに期待できる主な品質改善効果は、審決③の学習データ、すなわち審決対訳コーパスにおいて特定の用語の統一訳として用いられた英語と、海外審決文で多用される英語とが異なっている場合に得られるものであると考えられる。

以下、架空の例で説明する。審決対訳コーパスでは「明細書」の訳語を *description* と定め、これに統一しているが、英語審決文においては *subscription* が明細書の意味で用いられる最有力の語であったとする。そして、審決③の基礎学習データ中には「明細書」を *subscription* と訳している文例が存在し、これにより審決逆翻訳コーパス中の *subscription* は全て「明細書」と訳されていたと仮定する。

このような場合、「明細書」の訳語は、審決③では審決対訳コーパスで統一訳語となっている *description* が採用されてしまうが、審決④では、審決逆翻訳コーパス中の「明細書：*subscription*」という大量の訳例によって *subscription* が採用されるようになる。つまり、逆翻訳コーパスには、審決対訳コーパスで統一訳語とした英語が海外ではあまり使われていないような場合に、これをネイティブがよく使う英語に正す働きがある。

仮に審決③と審決④が実際に上例のような状況であったならば、それは特に自動評価スコアに顕著な差として顕れるはずである。自動評価スコアは、正解訳文との表層的な一致度に大きく左右されるためである。ただし、このような場合、両システムの差は、審決④のほうが有意に低くなる形で現れる。これは、自動評価の正解訳文も審決対訳コーパスの統一訳である *description* が用いられているためである。*subscription* が採用された審決④の機械翻訳文は正解訳文と一致しなくなるため、審決③よりもスコアが低くなる。

だが、本調査分析で取得した審決③／④の自動評価スコアは、各回とも両システムほぼ同水準であり、上記「審決④のほうが有意に低くなる」という状況は見られない。これはすなわち、審決対訳コーパスのみを学習させた審決③の機械翻訳文と、海外審決文由来の審決逆翻訳コーパスを追加学習させた審決④の機械翻訳文とで、主要な特許・審査書類用語の訳語となる英語がほぼ一致していたことを意味する。つまり、審決対訳コーパスにおける主な特許・審査書類用語の訳語は、海外審決文でも一般的に用いられる英語、海外で通用する英語であると見なせる。このような場合、審決逆翻訳コーパスを追加学習させても大きな改善効果は生じない。

審決逆翻訳コーパスは、その規模の大きさにより、追加学習させると機械翻訳文が学習の中心となってしまふ。前記したとおり、機械翻訳で作成した日本語文には、語の欠落や「未知語」が発生する可能性もある。本調査分析では審決逆翻訳コーパスの追加学習による品質への顕著な悪影響は見られなかったが、一定のリスクが存在することは確かである。また、約 250 万文という大規模なデータの作成や学習には多くの時間やリソースを消費する。こうした諸条件に鑑みると、審決対訳コーパス全件を学習データに使用する前提であれば、逆翻訳コーパスの追加学習は不要と判断される。

5.4.3. 審決の翻訳難易度について

審決③／④が審決対訳コーパス全件学習後に到達した内容伝達レベルはそれぞれ 4.42、4.31 と、機械翻訳としては十分に高い水準であった。しかしながら、審査書類①／②の最終学習後の内容伝達レベル(4.69、4.70)と比べるとやや低い。つまり、審決③／④の最終的な翻訳品質は、審査書類①／②よりも若干劣ることになる。

この品質差は、無学習の時点から存在していた。審決③／④の無学習時の内容伝達レベルは 2.52 であり、審査書類①／②の 3.14、2.99 に比べて明らかに低い。また、審決③／④の全学習期間の内容伝達レベルの向上度はそれぞれ+1.87、+1.76 であり、審査書類①の+1.55、審査書類②の+1.71 を上回っている。つまり審決③／④も、コーパスの学習による品質改善効果は審査書類①／②と同等以上に得ている。

こうした状況から、審決③／④と審査書類①／②との品質差は、学習データ量の不足や内容の不備ではなく、審決由来の文の難易度が各種審査書類由来の文よりも高いことが主な原因であると考えられる。

概して、審決の記載内容は、個々の審査書類の記載内容よりも広範囲に及ぶ。各審査書類の記載内容が特定の手続に関する内容に限定されるのに対し、審決には審査の各過程での経緯が記されるとともに、審理の主旨や審決に至る判断の内容なども詳述されるためである。加えて、審査書類対訳コーパスの作成対象が特許のみであるのに対し、審決対訳コーパスの作成対象案件には意匠や商標の案件も含まれる⁶⁶。このため、審決文で用いられる特許・審査書類関連の用語・表現・文体等のバリエーションもより広いと考えられる。こうしたバリエーションの中に、コーパスの学習ではカバーし切れなかった要素が存在するものと推察される。

⁶⁶ 審決③／④の評価用コーパス 2,024 文のうち 195 文 (9.6%) は原出願が商標の案件に由来している (JP_1978020807_2016300561)。なお、意匠案件は評価用コーパスには含まれていない。

本調査分析では、この「コーパスの学習ではカバーし切れなかった要素」の一つとして、固有名詞を挙げた。固有名詞は、特に商標審決では多用される傾向にあり、審決③／④の評価用コーパスでも、商標審決由来の 195 文中、約 46%にあたる 90 文に何らかの固有名詞が含まれていた。商標の場合、それ自体が固有名詞であることが多く、審決ではこれが争点となって多用される傾向が強い。

そして固有名詞は、審査書類用語等とは異なり、無限に存在し、かつ、案件ごとに全く異なる語が使用される。このため、学習データで十分に網羅することが事実上不可能なジャンルの語である。これは「みんなの自動翻訳」やニューラル翻訳方式のみに限らず機械翻訳全体の課題であり、現時点では画期的な解決策は存在しない。

審決③／④の品質評価結果が全期間を通じて審査書類①／②より若干劣っていたのも、商標審決由来の文でこうした固有名詞が多用されていたことが一因である可能性が高い。事実、第 5.2.2.項に示した分野ごとの自動評価スコアの集計でも、商標審決由来の文のスコアは特許審決由来の各分野よりも極端に低く、かつ、全期間を通じてのスコアの向上度も同様に低い。その原因が固有名詞の多用のみにあるとは限らないが、少なくとも商標審決に由来する文が審決③／④の全体品質を低下させている大きな要因であることは確実といえる。

なお、審決③／④の第 2 回学習後から第 3 回学習後にかけての自動評価スコア及び内容伝達レベルは、審査書類①／②と同じく、ほぼ現状維持状態に移行していた。このことは、審決対訳コーパスでカバーすべき主要な審査書類用語・表現は、第 2 回学習後の時点でおおむねカバーされていたことを意味する。もちろん、それ以降の学習でも改善される用語・表現は存在するが、比較的マイナーなものに限られるため、全体スコア上はさほど顕著な変化は生じない。審決③／④も、審査書類①／②と同じく、既にこの段階に達しているものと考えられる。したがって、仮に審決対訳コーパスをさらに増量したとしても、審決③／④の品質が直ちに審査書類①／②と同水準に達するとは考えにくい。つまり、審決③／④の品質水準が審査書類①／②よりも低いことは、学習データ量が不足していたためではない。むしろ審決③／④の翻訳品質は、現時点においては、審決由来の対訳コーパスを学習させることで到達可能なほぼ上限に達していると結論できる。

5.4.4. 固有名詞の翻訳品質について

審決由来文の翻訳難易度を高めている要因として、商標審決における固有名詞の多用を指摘した。本項では、人手評価対象文に含まれる固有名詞について、各学習段階における機械翻訳結果を示し、その翻訳品質を分析する。

審決③／④の人手評価対象 100 文中には、固有名詞を含む文は 7 文存在した。このうち商標審決由来の文は 2 文であった。人手評価対象 100 文中に商標審決由来の文は 7 文しか選ばれておらず、このうちの 2 文なので商標審決由来文の固有名詞含有率は約 29%とやはり高い（特許審決由来の文での含有率は 93 文中の 5 文なので約 5.4%である）。

固有名詞を含む 7 文を順に挙げると、文番号#52（「ビイエムプランニング」）、#165（「TOP-SIDER」、「SPERRY」）、#908（「三井化学」「宇部興産」「アドマー」）、#1234（「河野亜美」）、#1341、#1372、#1376（いずれも「ドラゴンコレクション」）となる。ただし、このうち#165 の「TOP-SIDER」と「SPERRY」は原文の時点で英字表記であり、また、#1341～#1376 の「ドラゴンコレクション」も一般英語のカタカナ表記にすぎず、どちらも翻訳難易度は低い。実際、これらは無学習時からおおむね問題なく訳出されていた⁶⁷。

よって、本項では、これら以外の固有名詞、すなわち#52 の「ビイエムプランニング」、#908 の「三井化学」、「宇部興産」、「アドマー」、そして#1234 の「河野亜美」について、各学習段階での訳され方を見ていく。

まず、文番号#52 の「ビイエムプランニング」は、BM Planning と訳するのが正しいが、第 5.3.4.3.2 項で既示したとおり、無学習時は yemm Planning と誤訳され、続く第 1 回学習後の翻訳文では審決③／④ともに欠落していた。その後も第 2 回学習後は審決③では Bye M-planning、審決④では BII M-planning とそれぞれ無学習時とは異なる訳語で誤訳され、最終学習後も審決③では BII M-planning、審決④が BUI M-planning と引き続き誤訳されて、改善は見られなかった。

文番号#908 では、著名企業名である「三井化学」と「宇部興産」は全期間を通じてほぼ正しく訳されていた⁶⁸が、登録商標である「アドマー」は無学習時には「欠落エラー」となっていた。その後、第 1 回学習後の翻訳文では、審決③では引き続き欠落したものの審決④では admer と正しく訳されるようになり、第 2 回～第 3 回学習後は双方のシステムで Admer

⁶⁷ 「SPERRY」のみ、第 1 回学習後に両システムで SPERRY と訳出され、表記上の軽微な不自然さが生じていた。

⁶⁸ ただし、第 2 回学習後の審決③のみ「宇部興産」が欠落した。

と正しい訳語が採用された。つまり本例は、審決対訳コーパスの学習の過程で固有名詞が正しく訳されるようになったケースである。ただし、「アドマー」は接着性ポリオルフェンの商品名（登録商標）として4千件以上の日本特許文献で使用されており、実質的には準技術用語である。

文番号#1234に含まれる個人名「河野亜美」も、無学習時は翻訳文から完全に欠落していたが、第1回学習後には審決③で Kawano Sub-Aesthetic とひとまず訳出されるようになった。ただし、苗字は正しいが名前が「亜：Sub」と「美：Aesthetic」とに分解して訳されており、読み手には意味不明である。一方、審決④では文末に苗字のみ Kawano と出力されたが、名前が訳出されておらず、本来あるべき位置でもないため、やはり読み手には意味不明である。

その後、第2回学習後には両システムとも Kono Ami と人名で訳出されるようになり、誤訳が解消した⁶⁹。続く第3回学習後も、両システムとも訳語が Ami Kawano に変化したが、引き続き人名として妥当に訳出されている。つまり、この語も、審決対訳コーパスの学習の過程で固有名詞が正しく訳されるようになったケースに該当する。

以上、人手評価対象文中の各固有名詞の翻訳品質について調査した。人手評価対象文中に含まれた全8種の固有名詞は、学習初期には正しく翻訳されないものも多かったが、学習の進行につれて改善されていき、最終学習後に正しく訳出されなかったのは#52の「ビイエムプランニング」のみであった。

ただし、人手評価対象文中に含まれていた固有名詞8語は、英字表記のもの（2語）や英語をカタカナで表記したもの（1語）、著名企業（2語）、特許文献で多用されている商品名（1語）など、その大部分は翻訳難易度が低かった。審決で使用される固有名詞は必ずしもこうした難易度の低いものとは限らず、実際には「ビイエムプランニング」や実例5-3の「水甚社」のように、コーパスを大量に学習させても誤訳が改善されないものも多いはずである。

固有名詞は、必ずしも正確でなくとも、案件内で訳語が統一されていれば、審決内容の理解にはさほど影響しないことも多い。しかしながら、「ビイエムプランニング」や#1234の「河野亜美」のように、学習前後やシステム間で異なる訳語が使用されているものも見られた。このような語は、一案件中においても同様の訳ゆれが生じている可能性がある。商標審決のように同一案件内で一つの固有名詞が繰り返し出現するような場合、こうした訳ゆれ

⁶⁹ 「河野」が Kawano か Kono かは字面からは判断できず、どちらでも不問とした。

は内容理解を大きく損なう懸念がある。

機械翻訳において、無限に存在する固有名詞を網羅的にカバーすることは事実上不可能といえる。このため、固有名詞の機械翻訳品質はどうしても低くなる。審決、とりわけ商標審決は、その影響を受ける度合いが他の審査書類に比べて大きい点に留意が必要である。

5.4.5. 同一案件内の訳語ゆれの発生について

審査書類①の分析において、同一学習回における同一案件内の訳語ゆれの懸念について言及した(⇒3.3.5.3.10.)。しかしながら、審査書類①／②の評価用コーパスは同一案件から複数の文が選定されることが少なく、十分な調査はできなかった。このため、案件単位で選定されている審決③／④の評価用コーパスを用いて、改めて調査を行う。

本調査の焦点は、最終学習後の最新の「みんなの自動翻訳」において、同一学習回における同一案件内の訳語ゆれが発生する可能性があるか否かである。この観点から審決③の最終学習後の機械翻訳文を確認した結果、審判番号 2017-800128 号(評価用コーパス文番号 #402～#617)の審決における「突出部」(56 文中、projecting portion が 27 文、protrusion portion が 17 文、protrusion が 11 文、protrusion part が 1 文)や 2017-006367 号(同#746～#938)における「タイヤ骨格体」(tire frame が 14 文、tire skeleton が 4 文)、2018-700001 号(同#1230～#1313)における「被覆層」(covering layer が 15 文、coating layer が 13 文)、2017-012323 号(同#1563～#1738)における「泡立ち」(foaming が 20 文、bubbles が 5 文、bubbling が 4 文)など、同一学習回・同一案件内での技術用語の訳ゆれが複数件検出された。なお、上記各語とも、審決④でも同様の訳ゆれが発生していた。

上記 4 例で採用された各訳語を見ると、(直訳しただけで意味が通じない「タイヤ骨格体」の tire skeleton を除けば)いずれも同義の範囲内かつ頻用の訳語と見なせる。つまり、上記各語に見られる同一学習回での訳ゆれの状況は、審査書類①／②を始め、大規模アップデート以降の各システムで共通して見られた、学習前後の技術／一般用語の訳語変化の状況と一致している⁷⁰。

これは、審査書類①の分析において第 3.4.4.項で考察したとおり、「みんなの自動翻訳」における同一学習回・同一案件内の訳語ゆれが、大規模アップデート以降も引き続き発生していた学習前後の訳語変化と同じく、ニューラル機械翻訳の訳語決定方式に根差した宿命的なものであることを示唆している。したがって、理論的には学習データの量や内容にかか

⁷⁰ 学習前後で「タイヤ骨格体：tire skeleton」のような不適訳語が突発的に採用されるケースは、大規模アップデート以降は顕著に減少したものの、引き続き検出されている。

ならず、ニューラル翻訳方式のあらゆるシステムで発生する可能性がある。本調査分析では、最終学習後の審査書類①／②では同一案件内の訳ゆれは検出されなかったが、案件単位での翻訳が基本となる実用時においては、審決③／④と同様、一定量の訳ゆれの発生が想定される。

5.4.6. 最終学習後に検出された主なエラー

審査書類①／②では、最終学習後に検出された主なエラーとして、「カギカッコで括られた原文において翻訳文が途絶するエラー（⇒3.4.7.1.）」、「カンマで区切られた複数の数字が合計3つ以上のときに連結されるエラー（⇒3.4.7.2.）」の二種を指摘した。これらのエラーは、いずれも審決③／④でも発生が確認された。

審決③／④では、これらに加えてさらに二種、報告すべき新たなエラーが検出された。以下に記す。

5.4.6.1. 同一文でランダム的に発生する途絶エラー

上記「カギカッコで括られた原文において翻訳文が途絶するエラー（以下、「カギカッコ途絶エラー」という。）」と同様、翻訳文が254文字に達した時点で途絶してしまうエラーである。第5.3.4.3.5項で詳述したとおり、審決③／④の複数の文で見られた。

本エラーが発生する文もカギカッコは多用されているが、本文全体がカギカッコで括られているわけではなく、発生条件が定かでない。また、同じ文であっても審決③／④の双方で発生する場合と、審決③のみで発生する場合とがあり⁷¹、文によっては、同じシステムでも発生する回と発生しない回とがある。このように、同じ文でも発生がランダム的である点が、特定の文で毎回必ず発生していたカギカッコ途絶エラーと異なっている。

これら2種のエラーの発生状況からは、「みんなの自動翻訳」において翻訳文が254文字に達した時点で途絶するという現象は、種々の理由によって機械翻訳処理が完了できなかった場合にやむなく内容が出力される作業用バッファの容量が254文字相当であるということではないかと推察される。つまり、本エラーとカギカッコ途絶エラーは、機械翻訳処理が正常に完了しなかった理由はそれぞれ異なるが、機械翻訳処理が完了できなかったために作業用バッファの内容がそのまま出力されたという点では共通しているため、エラーの状況としては同じ様相を呈しているものと考えられる。

これら2種のエラーの最終学習後の発生件数は、審決③では評価用コーパス2,024文中、

⁷¹ 今回の調査分析では、審決④のみで発生したケースは検出されなかった。

22 文、審決④では同 17 文と、いずれも全体の 1%前後であった。

5.4.6.2. 和暦→西暦の誤変換

「数値エラー」の一種として審決③／④で新たに検出された不備である（⇒5.3.4.3.6.）。

「みんなの自動翻訳」では原文中の和暦を西暦に変換して翻訳文に反映しているが、まれに誤った西暦年が採用されていることが判明した。エラーの実例分析結果からは、この変換誤りは、実際には、同一文中において数値（すなわち西暦年）の配置が入れ替わった結果である蓋然性が高い。つまり、和暦と西暦の変換自体が誤っているのではなく、文中に複数の年が存在する場合に、それぞれの配置を誤ってしまっているということになる。こうした数値の入れ替わり自体は、ニューラル機械翻訳では比較的良好に見られる。

ただし、こうした数値の入れ替わりが特に西暦年で多く発生した背景には、「同年」という語句の存在がある。「同年」という語は、英語に直訳すると of the same year 等なるが、この訳語を「同年 X 月 X 日」といった日本語文での用法で用いると、具体的な西暦年で訳すよりも大幅に冗長となってしまう。このため、英文では、「同年 X 月 X 日」のように既出の年を再度述べる文脈でも、具体的な西暦年を用いるのが一般的である。

その結果、こうしたデータを学習した機械翻訳エンジンでも、「同年」は具体的な西暦で訳される傾向が強くなる。しかしながら、この場合、「同年」が具体的にどの西暦年を指しているかを文ごとに正確に判断しなければならない。現状においては、この判断が誤る場合もあり、このため、他の数値と比べて、西暦に誤りが多くなっているものと推察される。

つまり、この和暦西暦変換の誤りは、ニューラル機械翻訳において突発的に発生する文中の数値の入れ替わりと、「同年」は英文では具体的な西暦年で訳されるのが一般的であるという、言語的性質の差異の双方の理由による。いずれも学習データでの対応は困難であり、ニューラル機械翻訳技術のさらなる発展が待たれる。

6. 各システムの統合の可能性について

本調査分析では、審査書類対訳コーパス及び審決対訳コーパスを使い分け、審査書類①、審査書類②、そして審決③／④と、それぞれ特定の書類を対象としたシステムを個別に構築した。だが、書類ごとに個々のシステムを構築するよりも、各システムの学習データを統合し、単独のシステムで全ての書類に対応できるならば、より効率的である。

本調査分析では、全学習データを統合したシステムは評価対象としていない。このため、全学習データを統合したシステムにおいて、各書類が個別のシステムと同等の品質で翻訳されるかは未調査である。

ただし、審査書類①では、4種の異なる審査書類（拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書）に由来する対訳コーパスを一括で学習させている。その結果、各書類とも同等の翻訳品質が得られており、異種の書類に由来する学習データを併用したことによる悪影響は見られなかった。これにより、これら4種の審査書類に由来するコーパスは学習データとして互換性が高いと結論した（⇒3.4.5.）。

異なる書類であるにもかかわらず互換性が高い理由は、各書類で使われる審査書類用語や表現がおおむね共通しており、意味や用法も書類間で統一されているためと考えられる。そして、この点に関しては、審査書類②で取り扱った意見書や、審決③／④で扱った審決も同様である。したがって、これらに由来する学習データも高い互換性を有している蓋然性が高い。

これを簡易的に調査するため、本項では、審査書類①、審査書類②、そして審決③の全学習データを一括で学習させた全書類統合システムを構築した⁷²。この統合システムを用いて個々のシステムの評価用コーパスを機械翻訳し、得られた翻訳文の自動評価スコアを、個々のシステムの最終学習後の自動評価スコアと比較した。こうすることで、全書類統合システムで得られる機械翻訳文の翻訳品質が、個別のシステムで得られた品質と比較してどの程度であるかを把握できる。結果を下表に示す。

⁷² 審決④システムで追加の学習データとした審決逆翻訳コーパスは、本調査分析では特段の品質改善効果が認められなかったため（⇒5.4.2.）、使用しなかった。

表 6.1.1 全書類統合システムによる各評価用コーパスの機械翻訳文の自動評価スコア一覧

評価用コーパス	指標	全書類統合	個別（最終学習後）	スコア差
審査書類①	BLEU	0.5059	0.5082	-0.0023
	RIBES	0.8447	0.8456	-0.0009
審査書類②	BLEU	0.4770	0.4777	-0.0007
	RIBES	0.8423	0.8416	+0.0007
審決③	BLEU	0.4207	0.4327	-0.0120
	RIBES	0.8280	0.8336	-0.0056

表の最上段は、審査書類①の評価用コーパスを全書類統合システムで機械翻訳して得た翻訳文の自動評価スコアである。BLEU が 0.5059、RIBES が 0.8447 と、それぞれ、個別システム（審査書類①）の最終学習後のスコアと比較すると **-0.0023**、**-0.0009** と僅かに劣るものの、誤差レベルの差異にとどまっている。実質的には同水準と見なせる。

二段目に示した審査書類②の評価用コーパスにおける状況もこれと同様であり、BLEU で **-0.0007**、RIBES では **+0.0007** とやはり同水準の品質が得られた。

一方、最下段に示した審決③の評価用コーパスのみは、全書類統合システムによる機械翻訳文に対する自動評価スコアは、個別システムに比べて BLEU では **-0.0120**、RIBES でも **-0.0056** に低下した。誤差レベルと見なせる差異ではなく、この結果からは、審決に関しては、全書類統合システムの機械翻訳品質は、審決単独のシステム（審決③）に比べて劣っていると見なされる。

この結果を受け、審決③の評価用コーパス 2,024 文において特にスコアが下落していた数文について、最終学習後の機械翻訳文と全書類統合システムの機械翻訳文とを比較し、スコア下落の要因を調査した。以下、BLEU と RIBES それぞれで、スコアが最も下落した文を 2 文ずつ示す。

【文番号：審決-#1552】 BLEU 下落第 1 位 (**-0.6744**)

原文	必要とあらば、以下に記載した文献の 2 ページ右欄 5 行ないし 8 行、及び、7 ページ右欄 2 行ないし 8 ページ左欄 3 行を参照のこと。
正解訳文	See, if necessary, the following documents, page 2, right column, lines 5 to 8, and page 7, right column, line 2 to page 8, left column, line 3.
審決③ 最終学習後	If necessary, refer to page 2, right column, lines 5 to 8, and page 7, right column, line 2 to page 8, left column, line 3 of the document described

統合システム	below.
	BLEU : 0.7628 (+0.6744) RIBES : 0.9105 (+0.1777)
	If necessary, see lines 5-8 in the right column on page 2 and line 2 in the right column on page 7 to line 3 in the left column on page 8 of the document described below.
	BLEU : 0.0884 (-0.6744) RIBES : 0.7329 (-0.1777)

上掲の文番号#1552 は、全書類統合システムで BLEU スコアが最も下落していた文である。BLEU スコアは最終学習後の 0.7628 から 0.0884 へ、**-0.6744** と激しく下落した。だが、実際の機械翻訳文を比較してみると、語順は大きく変化しているものの、採用された訳語自体はほぼ同じである。唯一、「参照のこと」の訳語が refer to から see に変化しているが、これも完全に同義の範囲内である。どちらの翻訳文も原文の内容を正確に伝達しており、実際の品質は完全に同等と見なせる。つまり、本例は、単なる表層的な語順の違いにより自動評価スコアが不当に変動していたものであった。

続いて、BLEU スコアが上例に次いで下落していた文を示す。

【文番号：審決-#906】 BLEU 下落第 2 位 (**-0.6451**)

原文	本願補正発明と引用発明を対比する。
正解訳文	The Amended Invention and the Cited Invention are compared.
審決③	The Amended Invention and the Cited Invention are compared.
最終学習後	BLEU : 1.0000 (+0.6451) RIBES : 1.0000 (+0.0543)
統合システム	The Amended Invention and Cited Invention will be compared.
	BLEU : 0.3549 (-0.6451) RIBES : 0.9457 (-0.0543)

こちらも、BLEU スコアは最終学習後の 1.0000 から 0.3549 (**-0.6451**) へと激減している。ただし、機械翻訳文の差異は、Cited Invention に定冠詞 the が付いているか否かと、「対比する」に相当する訳語の時制が現在形 are compared から未来形 will be compared へと変化したのみである。審決と他の審査書類との間で冠詞の有無や時制に特段の違いがあるとは考えられず、この変化は、学習データを統合したことに起因する変化というよりは、個々のシステムでもよく見られた「みんなの自動翻訳」の不安定な訳語変化にすぎないと推察される。双方の機械翻訳文から理解される内容に違いはなく、実質的な品質差はない。本例も、正解訳文と完全一致でなくなったという表層的な変化によってスコアが極端に変動したにすぎないと見なせる。

次に、RIBES スコアが最も大きく下落した文#1605 を見てみる。

【文番号：審決-#1605】RIBES 下落第 1 位 (-0.4834)

原文	TOC 濃度の測定には、SHIMADZU 社製、TOC-VCPH を使用した。
正解訳文	For measuring the TOC concentration, Shimadzu-made TOC-V. CPH was used.
審決③ 最終学習後	For the measurement of TOC concentration, TOC-VCPH manufactured by SHIMADZU Corporation was used.
	BLEU : 0.1768 (+0.0430) RIBES : 0.8546 (+0.4834)
統合システム	TOC-VCPH manufactured by SHIMADZU Corporation was used for the measurement of the TOC concentration.
	BLEU : 0.1338 (-0.0430) RIBES : 0.3711 (-0.4834)

本例では RIBES スコアが最終学習後との比較で-0.4834 と最も大きく下落した。ただし機械翻訳文を見ると、使われている訳語は完全に同じで、単に for the measurement of the TOC concentration の配置が文頭から文末に移動し、TOC concentration に定冠詞 the が付加されたのみの変化である。前例では統合システムで定冠詞が消失していたが、本例では逆方向の変化が生じており、これらがやはり「みんなの自動翻訳」特有の不安定な訳語変化によるものであることがわかる。双方の翻訳文とも文意は完全に同一であり、どちらも原文の内容を完璧に伝達している。つまり、本例も、翻訳文の前半と後半が入れ替わったことで正解訳文と語順が大幅に異なってしまう、極端なスコア下落を招いたにすぎない。よって、本例の RIBES スコアも、表層的な変化に起因する不当な下落である。

最後に、RIBES スコアが上例に次いで下落した文を示す。

【文番号：審決-#1275】RIBES 下落第 2 位 (-0.4496)

原文	請求項 1 に係る発明は、甲第 2 号証～甲第 4 号証のいずれかに記載された発明であるから、特許法第 29 条第 1 項第 3 号に該当し、特許を受けることができない。
正解訳文	The invention according to Claim 1 is an invention described in any of Evidence A No. 2 to Evidence A No. 4, falls under Article 29(1)(iii) of the Patent Act, and, therefore, should not be granted a patent.
審決③ 最終学習後	The invention according to Claim 1 is an invention described in any of Evidence A No. 2 to Evidence A No. 4, and thus falls under Article 29 (1) (iii) of the Patent Act, and the appellant should not be granted a patent for it.

	BLEU : 0.8114 (+0.2838) RIBES : 0.9685 (+0.4496)
統合システム	The invention according to Claim 1 shall not be granted a patent under the provision of Article 29 (1) (iii) of the Patent Act for the reason that the claimed invention has been deemed to be identical with the invention described in any one of Evidence A No. 2 to Evidence A No. 4.
	BLEU : 0.5276 (-0.2838) RIBES : 0.5189 (-0.4496)

本例でも RIBES スコアが**-0.4496** と大きく下落している。本例は前 3 例と異なり、いくつかの訳語に有意な変化が生じている。まず「発明であるから」の「である」という語が、**is** という直訳から **has been deemed to be identical with** (同一と見なされる) という意識的な表現に変化した。正解訳文が前者の表現を採用しているためスコアの下落要因となっているが、原文の主旨からはむしろ後者のほうが正確であり、少なくとも同等である。

また、「特許を受けることができない」という箇所が、最終学習後の翻訳文では the appellant should not be granted a patent for it (請求人は、それについて特許を受けることができない) と、主語 the appellant を補足した形で訳しているのに対し、全書類統合システムでは (The invention according to Claim 1) shall not be granted a patent と、原文どおりの直訳調で訳している。この箇所に関しても伝わる意味は双方同じであり、むしろ原文に存在しない主語を補った最終学習後の翻訳文よりも、統合システムの訳し方のほうが正解訳文の should not be granted a patent に近い。ただし、全体的に見ると、個別システムの翻訳文は前半が正解訳文と完全一致しており、また総語数も 45 語と正解訳文 (40 語) に近いのに対し、統合システムの翻訳文は正解訳文との語順の違いが大きく、かつ、総語数も 54 語と極端に多い。こうした正解訳文との表層的なかい離によって、自動評価スコアが大きく下落したものと考えられる。したがって、本例に関しても、全書類統合システムの実質的な翻訳品質は全く劣っていない。

このように、全書類統合システムで翻訳したことで最もスコアが低下した 4 例を見たが、各例とも採用された訳語自体は審決③システムの最終学習後の翻訳文とほぼ同一であり、ごく少数の訳語変化も、全て同義の範囲内であった。各例とも、スコアの極端な下落は、正解訳文と語順が異なるという表層的な要因によるものであり、実質的な翻訳品質は個別システムと完全に同等であった。

つまり、これら 4 例を見る限り、全書類統合システムを用いたことによる自動評価スコアの低下は、語順や訳語の表層的な変化が正解訳文から離れる方向に偏ったためと推察される。少なくとも、これら 4 例には、学習データを統合したことによる実質的な品質低下要因は存在していない。

とはいえ、一定規模の文集合の平均値による自動評価スコアの相対比較結果には高い信頼性があることも事実である。このため、特段の品質低下要因は存在しないにせよ、学習データを統合したことに起因する何らかの品質低下が実際に生じている可能性も否定はできない。そこで、実際に品質低下が生じているものと仮定して、全書類統合システムによる審決の実質的な翻訳品質がどの程度の水準であるかを推定する。

全書類統合システムによる審決③の評価用コーパスの自動評価スコアは BLEU が 0.4207、RIBES が 0.8280 であった。このスコアは、審決④システムの第 2 回学習後のスコア（⇒5.2.1.1.）に最も近く、その差異は BLEU が-0.0010、RIBES が+0.0011 と誤差レベルである。そして、審決④の第 2 回学習後の内容伝達レベルが 4.24 である（⇒5.3.4.1.）ことから、全書類統合システムの実質的な翻訳品質もこれと同程度、すなわち 4.24 相当と推定される。個別システム（審決③）の最終学習後の内容伝達レベルは 4.42 であり、これとの比較では -0.18 程度の品質低下が生じていることになる。

以上、本事業で作成した全ての対訳コーパスを一括で学習させた全書類統合システムで得られる各種書類の機械翻訳品質をシミュレートした。その結果、審査書類①及び審査書類②に属する各書類に関しては個別システムと同水準の翻訳品質が得られるが、審決のみは若干の品質低下が生じる可能性があることが判明した。具体的には、個別システム（審決③）との比較で内容伝達レベル-0.18 相当の品質低下が生じる懸念がある。

ただし、全書類統合システムによる審決の翻訳品質は、内容伝達レベル 4.24 相当と推定されるため、依然としてレベル 4 を大きく上回っており、機械翻訳としては十分に高品質といえる。また、本項で取り上げた各事例を見る限り、実質的な品質低下が実際に生じているかどうかは明らかではない。

全書類統合システムの可否は、こうした審決における若干の品質低下の可能性と、複数のシステムを統合することによるコスト削減とのトレードオフとして判断すべきである。

7. 本調査のまとめ

本調査分析では、拒絶査定、拒絶理由通知書、補正却下の決定、前置報告書の4種の審査書類に由来する対訳コーパスを一括で学習させた審査書類①システム、意見書由来のコーパスのみを学習させた審査書類②システム、そして審決由来のコーパスのみを学習させた審決③システムと、これに審決逆翻訳コーパスを追加学習させた審決④システムの全4種の機械翻訳システムについて、本事業で作成した対訳コーパスの学習による機械翻訳品質の改善状況を調査してきた。本章で、上記4種のシステムが最終的に到達した翻訳品質、各システムに学習データとして適用した各種対訳コーパスの作成意義、そして残存する課題について、横断的に考察する。

7.1. 各システムで達成された翻訳水準について

本事業は、日本の審査書類の機械翻訳品質が現状では不十分であること、そしてニューラル機械翻訳に審査書類／審決由来の対訳コーパスを大量に学習させることで高品質な機械翻訳が得られることを前提に実施した。本報告書に記した各システムの調査分析結果により、これら本事業の前提が的確であり、かつ、本事業で作成した審査書類対訳コーパス及び審決対訳コーパスが企図したとおりの品質改善効果を奏することが実証された。

下表に、本事業で調査対象とした4種のシステムそれぞれの、コーパス学習前後の内容伝達レベル及び自動評価スコアを一覧で示す。

表 7.1.1. システム別 コーパス全件学習前後の自動評価及び内容伝達レベル一覧

システム	学習データ量	指標	無学習時	最終学習後	向上度
審査書類①	1,810,685 文	内容伝達 Lv	3.14	4.69	+1.55
		BLEU	0.3160	0.5082	+0.1922
		RIBES	0.7558	0.8456	+0.0898
審査書類②	179,830 文	内容伝達 Lv	2.99	4.70	+1.71
		BLEU	0.3162	0.4777	+0.1615
		RIBES	0.7630	0.8416	+0.0786
審決③	156,152 文	内容伝達 Lv	2.52	4.42	+1.90
		BLEU	0.2725	0.4327	+0.1602
		RIBES	0.7316	0.8336	+0.1020
審決④	2,677,978 文 ※審決逆翻訳コーパスを含む	内容伝達 Lv	2.52	4.31	+1.79
		BLEU	0.2725	0.4234	+0.1509
		RIBES	0.7316	0.8285	+0.0969

上表のとおり、4種類のシステムの全てにおいて、表に記載した文数の対訳コーパスの学習の結果、内容伝達レベルがそれぞれ無学習時のレベルから1.5ポイント以上の大幅な向上を示し、最終的には「重要情報の80%以上が正確に伝達される」品質である内容伝達レベル4を大きく上回る機械翻訳品質に到達した。機械翻訳として極めて高品質であり、本事業で作成した各種コーパスが高い翻訳品質改善効果を奏し、本事業の目指す日本審査書類及び審決の高品質な機械翻訳文の提供を実現するものであることを示す結果といえる。

また、本調査分析で構築した4種のシステムとも学習後期には品質評価結果が現状維持状態に移行しており、品質向上の最大の要因であった特許・審査書類用語への対応がほぼ網羅的に完了していることがうかがえた。このことから、各システムの翻訳品質は審査書類／審決対訳コーパスの学習によって達成可能な上限に達しつつあり、後はマイナーな用語・表現に改善の余地を残すのみの状態であると結論される。

7.2. 各種対訳コーパスの潜在的な翻訳改善効果について

本事業で構築し、段階的な学習を施した4種のシステム、すなわち審査書類①、審査書類②、審決③／審決④は、いずれも学習初期に大幅にスコアが向上したが、学習中期以降は、大規模アップデート直後を除けば、ほぼ現状維持状態で推移した。各システムは学習データの量が大きく異なるが、現状維持状態に移行したタイミングはほぼ同時期であった。

この結果から、各システムとも学習初期の時点で、それぞれが対象とする書類に頻出の特許・審査書類用語や表現の大部分がカバーされたものと見なせる。具体的には、人手翻訳評価で内容伝達レベルが4を超えた時点、すなわち、審査書類①／②であれば第5回学習後、審決③／④であれば第2回学習後の学習データ量であれば、各種の審査書類／審決をレベル4程度の高品質で機械翻訳するために必要な学習データ量を確実に満たしているといえる。具体的には、審査書類①が492,000件、審査書類②が48,000件、審査書類③／④が124,654件（審決逆翻訳コーパスは除く）となる。

ただし、このことは、上記データ量に達した後に作成された各種対訳コーパスの意義が減じたことを意味するものではない。機械翻訳エンジンの学習において、学習初期にまず頻出の用語が一斉に改善されて全体的な翻訳精度が大きく向上し、その後は、そこから漏れたマイナーな用語の改善に移行していくことは必然である。このため、同量の学習データを追加しても、全体品質に表れる改善効果が急激に小さくなることも、本調査分析に限らず必至である。しかし、学習早期に達成される頻出の用語・表現のカバーは、機械翻訳の利用性においては最低限の要件にすぎず、さらなる改善のためには、こうしたマイナーな用語も地道に改善していかなければならないことは、言うまでもない。

本調査分析においては、こうしたマイナーな用語の改善が全体品質、すなわち評価スコアに反映されるためには、まずは限られた文数の評価用コーパス中で使用されているマイナーな用語が、学習データ中でも（偶然に）カバーされており、両者が合致する必要がある。本事業の場合、本事業の場合、審査書類対訳コーパスを作成する際に重複文をあらかじめ除外しているため⁷³、合致の確率はさらに低くなる。そして、仮に合致して翻訳品質が改善したとしても、マイナーな用語であり、かつ評価用コーパス自体が非重複文で構成されているため、改善は局所的であり、全体品質に与える改善効果は非常に見えにくくなる。

本事業において、審査書類対訳コーパスを非重複文のみとしたのは、限られた文数で、可能な限り広範囲の用語・表現を効率的にカバーするためである。このため、これを評価用コーパスと学習データとに分割して使用した本調査分析では、両者の合致率が低くなることは避けられない。実用時においては、審査書類対訳コーパスに含まれる文は全て非重複文であることから、基本的には全ての文が、それぞれ何らかのマイナーな用語・表現の訳質改善に寄与する可能性を等しく有している。そして、こうした文は、可能な限り広範に作成しておくほど、それを学習させた機械翻訳エンジンの潜在的な翻訳性能が向上することは確実である。

その意味において、本事業は、非重複文のみで対訳コーパスを大量に作成するという方針により、限られた文数で最も広範囲の潜在的な翻訳改善効果を効率的に取得することができたと結論される。

7.3. 各システムで未解決の課題について

本調査分析で構築した4種のシステムは、審査書類／審決対訳コーパスの学習によって到達可能なほぼ上限の品質水準に到達したと見られる。ただし、各システムの調査分析では、対訳コーパスの学習では改善困難であったいくつかの課題が検出された。以下、これらの課題についてまとめる。

7.3.1. 最終学習後も残存した特徴的なエラー

各システムの最終学習後の機械翻訳文で検出されたエラーの中には、複数の文で発生しており、発生する原文の条件や、発生したエラーの状況から特定のパターンに括られるものがいくつか見られた。以下に列挙する。

⁷³ 審決対訳コーパスは審決全文をコーパス化したため重複文も含んでいる。

7.3.1.1. カギカッコで括られた原文において翻訳文が途絶するエラー

原文全体がカギカッコで括られている場合に、機械翻訳文が文頭から 254～255 文字に達した時点で途絶してしまうエラーである。本項ではこれ以降、「カギカッコ途絶エラー」と称する。実例を以下に示す。

A913-#1083

原文	・「炭酸を含有する電子材料用洗浄水の余剰分を戻すポイントは、脱炭酸装置の上流であれば特に制限はないが、特に前処理工程と一次純水製造部との中間に位置する水槽が適当である。」(段落【0 0 0 7】)
機械翻訳文 (最終学習後)	"The point for returning the surplus of the washing water for electronic materials containing carbonic acid is not particularly limited as long as it is upstream of the decarboxylation device, but in particular, a water tank located between the pretreatm

上例では、原文の「前処理工程」に相当する pretreatment が pretreatm と単語の途中で途絶し、それ以降の内容が失われている。出力された文字数は 254 文字である。

こうした「カギカッコ途絶エラー」は、審査書類①／②では第 3 回学習後以降、審決③／④でも時期にあたる第 2 学習後以降に発生されるようになり、以降、最終学習後も全てのシステムで複数件検出されている。学習データの内容や量が大きく異なる各システムで共通して発生していることから、「みんなの自動翻訳」自体のロジックに起因するものと考えられる。

本エラーは、原文が下記のいずれかの条件に該当し、かつ、機械翻訳文が 254 文字以上となる場合に発生している。

- ・タイプ A：冒頭に中点（・）が打たれ、直後の本文全体がカギカッコで括られ、その後にカッコ補記が存在する

(※上例が該当)

- ・タイプ B：上記タイプ A と同様だが、カッコ補記を伴わない

・「すなわち、本願新請求項 1 の発明は、走行体にカーブを走行させることなく、…給電部を配置することが技術的な特徴です。」

- ・タイプ C：本文全体がカギカッコで括られ、その後にカッコ補記が存在する

「交換機または回線に呼が集中すると過負荷となり、通信網が異常となっ

て、疎通が不可能な輻輳状態になる。…このような通信網の品質低下を抑えるために輻輳時の対策が必要である。」(段落 0 0 0 2)

- ・タイプ D：上記タイプ C と同様だが、カッコ補記を伴わない

「よって、本願発明では、モータの電氣的時定数を転流間隔時間の 1 / 6 にするための…請求項 1 の従属項として定義したものである。」

「みんなの自動翻訳」で見られた通常の「欠落エラー」には、学習データの不足に起因するものと、それまで問題なく訳されていた語が突発的に欠落するものとが混在していた。本調査分析では、後者をエンジンのロジックに起因する「欠落エラー」として扱った。ただし、これらのエラーにおいても、上例のように単語が途中で途絶したり、特定の文字数以降の翻訳文が全て消失したりするような現象は見られない。したがって、「カギカッコ途絶エラー」は、カテゴリとしては「エンジンのロジックに起因する欠落エラー」ではあるが、他の「欠落エラー」とは根本的に性質が異なるものといえる。

審査書類には出願書類や先行技術文献などからの引用文も多く、引用であることを示すために上記の各タイプの様式を用いることも少なくない。このため、比較的影響範囲の広いエラーである。

7.3.1.2. 同一文でランダム的に発生する途絶エラー

上記「カギカッコ途絶エラー」と同様、翻訳文が 254 文字に達した時点で途絶してしまうエラーである。以下、一例を示す。

【審決-#910】

原文	引用発明における「前記マルチフィラメントと前記被覆層との間に配置されたホットメルト接着剤 A-1：三井化学（株）製の「アドマー QE-060」で形成された接着層」は、本願補正発明における「前記補強コードと前記被覆用樹脂層との間に配置されて前記補強コードと前記被覆用樹脂層とを接合する接合用樹脂層」に相当する。
正解訳文	"An adhesive layer made from a hot melt adhesive A-1 of 'ADMER QE-060' manufactured by Mitsui Chemicals, Inc. disposed between said multifilament and said covering layer" of the Cited Invention corresponds to "a resin layer for adhesion disposed between said reinforcing cord and said resin layer for covering to adhere said reinforcing cord and said resin layer for covering" of the Amended Invention.

審決③ 第3回学習後 欠落：有	"A hot melt adhesive a- 1 arranged between the multifilament and the covering layer : An adhesive layer formed of" ADMER QE 060 "manufactured by Mitsui Chemicals, Inc." in the Cited Invention corresponds to "a bonding resin layer arranged between the rei
審決④ 第3回学習後 欠落：無	The "hot melt adhesive a- 1 disposed between the multifilament and the coating layer : An adhesive layer formed of' ADMER QE - 060' manufactured by Mitsui Chemicals, Inc." in the Cited Invention corresponds to the "bonding resin layer disposed between the reinforcing cord and the coating resin layer to bond the reinforcing cord and the coating resin layer" in the Amended Invention.

本例も、審決③の最終学習後の機械翻訳文において、翻訳文が 254 文字で途絶し、原文の「補強コード」に相当する reinforcing cord が rei と、単語の途中までしか出力されていない。なお、審決④では、同じタイミングで同じ原文を機械翻訳しているにもかかわらず、文末まで翻訳文が出力されている。

上例をはじめ、本エラーが発生する各文ともカギカッコは使用されているが、前項に示した「カギカッコ途絶エラー」の発生条件（タイプ A～D）には該当しておらず、発生条件が特定できない。また、「カギカッコ途絶エラー」が条件に合致する文で毎回確実に発生していたのに対し、本エラーは、上例のとおり、同一の入力文でもシステムによってエラーが発生する場合としない場合があり、かつ、同じシステムでも学習前後でエラーが発生したり解消したりすることもあるなど、明らかにランダム性が強い。実際、上例も、前回（第2回）学習後は審決③では翻訳文は途絶せず、審決④で途絶するという、第3回後とは正反対の結果となっている。

こうした状況から、本エラーは、翻訳文が 254 文で途絶するという状況自体は「カギカッコ途絶エラー」と同じだが、本質的には別種のエラーであると考えられる。翻訳文が 254 文字で途絶する現象は、単に「みんなの自動翻訳」が、機械翻訳処理が何らかの理由により完了できなかった場合に、容量 254 文字相当の作業用バッファの内容を出力する構造であるためと推察される。

つまり、本エラーは、必ずしも発生条件が一意に定まるものではなく、ニューラル機械翻訳のランダム性に起因する種々の要因で機械翻訳処理が完了できなかった場合に発生するものであって、その際の出力文字数が共通しているにすぎない可能性が高い。

7.3.1.3. カンマで区切られた複数の数字が合計 3 つ以上のときに連結されるエラー

原文中で複数の数値がカンマで区切られて並び、数字が合計で 3 つ以上である場合に、3 桁以上の 1 つの数値として誤訳されてしまうエラーである。例えば「請求項 1 4, 1 6」であれば、Claims 14, 16 ではなく、カンマで区切られた 4 つの数字 1、4、1、6 が連結されて Claim 1416 と 1 つの数値として連結される。

各システムで大規模アップデート直後から突如発生しており、条件に合致する文では確実に発生していることから、こちらも「みんなの自動翻訳」自体のロジックに起因するエラーと見なされる。

7.3.1.4. 和暦→西暦の誤変換

「みんなの自動翻訳」では原文中の和暦を西暦に変換して翻訳文に反映しているが、まれに誤った西暦年が採用されることが判明した。これらの和暦→西暦の誤変換は、同一文中において数値（すなわち西暦年）の配置の入れ替わりと、日本語文で多用される「同年」という表現の翻訳難易度に起因するものと考えられる。

文中の数値が入れ替わるエラー自体は、ニューラル機械翻訳では比較的良好に見られる。文中に年が複数含まれている場合、それらが入れ替わって結果的にそれぞれが誤った西暦となることがある。一方、「同年」という表現は、英文では具体的な西暦年で表すのが一般的であるが、その際に誤った西暦年が採用される場合がある。和暦→西暦の誤変換は、これら二つの要因でそれぞれ発生するため、発生数が多くなっているものと考えられる。以下、実例を示す。

【文番号：#626】

原文	その後、平成 2 7 年 5 月 1 1 日付けで審理事項通知がされ、平成 2 7 年 5 月 2 4 日付け（同年 5 月 2 5 日差出し）で被請求人より上申書が提出され、平成 2 7 年 6 月 1 5 日付けで被請求人より口頭審理陳述要領書が提出され、平成 2 7 年 6 月 2 2 日差出しで請求人より口頭審理陳述要領書が提出された。
正解訳文	Notification of matters to be examined were notified as of May 11, 2015. A written statement was submitted by the demandee as of May 24, 2015 (sent on May 25, 2015). An oral proceedings statement brief was submitted by the demandee as of June 15, 2015. An oral proceeding statement brief was sent and submitted by the demandant on June 22, 2015.
審決③	Thereafter, a notice of matters to be examined was issued on May 11,

第3回学習後 数値エラー：有	2015, a written statement was submitted by the demandee on May 24, 2015 (submitted on May 25, 2015), an oral proceedings statement brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2019.
審決④ 第3回学習後 数値エラー：有	After that, a notification of matters to be examined was issued on May 11, 2015, a written statement was submitted by the demandee on May 24, 2015 (delivered on May 25, 2016), an oral proceedings statement brief was submitted by the demandee on June 15, 2015, and an oral proceedings statement brief was submitted by the demandant on June 22, 2015.

上例では、原文に「(同年5月25日差出し)」という文言が存在する。ここでの「同年」は、直前の「平成27年」を指し、翻訳文では西暦2015年に変換される。審決③はこの箇所を(submitted on May 25, 2015)と正しく訳しているが、審決④では、(delivered on May 25, 2016)と、西暦年が誤変換されている。

原文には「平成27年」が4回、「同年」が1回出現するが、全て西暦2015年に相当し、審決④で採用された2016年にあたる要素は原文中に存在しない。このことから、この誤変換は、「同年」を「平成27年」を指すものと解釈できず、何らかの理由で2016年と解釈した結果であると考えられる。

「同年」という語は、英語に直訳すると of the same year 等となるが、この訳語を上例の「同年5月25日」のような日本語文での省略用法に適用すると、May 25 of the same year となり、具体的な西暦年で訳すよりも大幅に冗長なものになってしまう。このため、英文では、既出の年を再度述べる文脈でも具体的な西暦年を用いるのが一般的である。よって、こうしたデータを学習した機械翻訳エンジンも、「同年」を具体的な西暦で翻訳する傾向が強くなる。

しかしながら、この場合、機械翻訳エンジンは、「同年」が具体的に指す西暦年を、文ごとに正確に選定しなければならない。現状においては、この選定を誤る場合があり、このため、西暦の誤変換が散見されるものと推察される。

なお、上例では、審決③にも和暦→西暦の誤変換が見られる。具体的には、文末の「平成27年6月22日差出し」が June 22, 2019 と誤訳されている。これは一見すると「平成27年」が西暦2019年と誤訳されたように見える。だが、これは実際には、審決④と同じく

「同年」が誤解釈されて 2019 となり、かつ、ニューラル機械翻訳特有の数値の入れ替わりによって、これと文末の「平成 27 年」に対応する 2015 とが入れ替わって出力されたものと考えられる。

文中における数値の入れ替わり、そして「同年」に対応する西暦年の正確な判定とも、学習データでの対応は困難である。ニューラル機械翻訳技術のさらなる発展が待たれる。

7.3.2. 同一案件内の訳語ゆれの発生

ニューラル機械翻訳では、各用語の訳語は、大量の学習データ中で、その用語が入力文と最も近い使われ方をしている文例の対訳文で採用されている訳語が採用される傾向が強い。これは、同じ用語でも文脈に応じて最適の訳語に訳し分けることができるという、ニューラル機械翻訳の利点に通じる。ただし、その反面、同じ案件に繰り返し出現し、同じ訳語を採用すべき語が、文によって異なる訳語で訳されるという状況、すなわち用語の訳ゆれが発生する可能性もはらんでいる。

本調査分析でも、審決③／④の最終学習後の機械翻訳文を確認した結果、審判番号 2017-800128 号(評価用コーパス文番号#402～#617)の審決における「突出部」(56 文中、projecting portion が 27 文、protrusion portion が 17 文、protrusion が 11 文、protrusion part が 1 文)や 2017-006367 号(同#746～#938)における「タイヤ骨格体」(tire frame が 14 文、tire skeleton が 4 文)、2018-700001 号(同#1230～#1313)における「被覆層」(covering layer が 15 文、coating layer が 13 文)、2017-012323 号(同#1563～#1738)における「泡立ち」(foaming が 20 文、bubbles が 5 文、bubbling が 4 文)など、同一学習回・同一案件内で技術用語の訳ゆれが複数件検出された。

上記 4 案件で採用された各訳語を見ると、おおむね同義の範囲内かつ頻用の訳語ではある。それでも、特に技術用語は、同じ用語が案件内で複数の訳語に訳ゆれしていると、読み手を無用に混乱させる場合が多く、好ましくはない。

こうした訳語変化は、冒頭で述べたニューラル機械翻訳エンジンの訳語選定メカニズムに起因するものであり、本調査分析で多数見られた学習前後の技術／一般用語の不安定な訳語変化と同根である。このため、上述した審決③／④システムに限らず、全システムで発生する可能性が高い。

7.3.3. 固有名詞の翻訳難易度の高さ

審決③／④の評価用コーパスには商標審決由来の文が含まれており、これらの文の翻訳精度が、特許審決由来の文に比して有意に低かった。その一因として、商標審決における固

有名詞の多用が挙げられる。

固有名詞は無限に存在し、学習データでその全てを網羅することは事実上不可能である。このため、いかなる機械翻訳方式であっても、固有名詞の機械翻訳品質がその他の語句に比べて低くなることは、少なくとも現状においては、不可避といえる。

そして、商標審決は、その性質上、固有名詞が多用される傾向が他の書類よりも強い。商標は、それ自体が固有名詞であることが多く、つまり審理の主題となることが多いため、審決内で繰り返し言及される傾向にある。その分、商標審決は、特許を対象とした審決や審査書類に比べて、固有名詞の翻訳難易度の高さの影響を受ける度合いが強くなる。商標審決をはじめとする商標関連の書類に機械翻訳を導入する際には、この点に留意が必要である。