公表版

令和６年度
日インドネシア語の対訳コーパス及び
辞書の整備に関する調査事業

調 査 報 告 書

令和７年３月21日

一般財団法人 日本特許情報機構

内容

１．はじめに

1.1. 調査の目的

特許庁では、対訳コーパス[1]を学習した機械翻訳エンジンを用い、日本の審査情報等の海外発信において日英機械翻訳、外国公報の検索照会において中日・韓日機械翻訳を提供してきているが[2]、一方で、世界における日本企業の現地法人数について、ASEAN の割合が 2020 年まで 10 年連続で拡大[3]、日本から ASEAN 各国への特許出願件数が増加傾向[4]等のデータが示すとおり、日本企業の海外ビジネス展開にあたって、米欧中韓以外の国への関心も高まりを見せ、これが維持されている。

そのような状況の中、日米欧中韓の知財庁（以下、「五庁」という。）で主に使用される言語（日英中韓）以外の言語を対象とした機械翻訳の精度を高め、五庁以外の知財庁の特許情報の日本語への機械翻訳、日本の審査情報等の現地語への機械翻訳を提供することが望まれるが、日英中韓以外の言語について、機械翻訳の精度に資する対訳コーパス作成は国内外で進んでいない。

また、「令和５年度 特許情報の機械翻訳における多言語対応に向けた課題検討に関する調査事業」[5]では、経済的・技術的情報を総合的に考慮した結果、特許情報に特化した対訳コーパスを整備するにあたって、特にインドネシア語が有力な候補として挙げられた。

本調査は、こうした状況を踏まえ、機械翻訳を介した ASEAN 諸国の特許情報へのアクセス性向上及び ASEAN 知財庁における日本の審査結果・公報のより一層の活用促進のため、インドネシア語に着目し、日本語及びインドネシア語（以下、「日インドネシア語」とする。）の機械翻訳の精度向上のための対訳コーパス及び辞書の整備に資することを目的に実施した。

---

[1] 異なる言語の文と文を対訳にしてまとめた文書のあつまり。

[2] https://www.jpo.go.jp/support/j_platpat/kaizen20200518.html

[3] 第 51 回 海外事業活動基本調査（2021 年調査）
https://www.meti.go.jp/press/2022/05/20220530001/20220530001-1.pdf

[4] ASEAN の知財概況（2022 年） ジェトロシンガポール事務所・バンコク事務所 知的財産部
https://www.wipo.int/edocs/mdocs/mdocs/ja/wipo_webinar_wjo_2022_5/wipo_webinar_wjo_2022_5_inf.pdf

[5] https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/kikai_honyaku/2023_01.pdf

1.2. 調査の概要

本調査では、日インドネシア語の対訳コーパス及び対訳辞書について、特許・実用新案公報から高い精度で効率的に作成する手法を国内外の公開情報から広く収集のうえ比較し、最善の手法を選定してこれらを実際に作成した。

具体的には、対訳コーパスの取得目標を「一定以上の精度を満たす 100 万文対以上（1 文と複数文の対、複数文と複数文の対についても、自動作成により対となった場合には 1 文対とカウントする。）」と定め、これを達成するための最善の作成手法を採用した。

そのうえで、作成過程の分析や、得られた対訳コーパスに対する人手による修正と解析、さらに当該対訳コーパスを機械翻訳エンジンに試験的に学習させ、学習前後の機械翻訳結果の精度を評価してその学習効果を測ること等により、日インドネシア語の対訳コーパスの整備に関する現状の問題点や課題を総合的に分析するとともに、今後の解決策を提言した。

なお、対訳辞書に関しては、対訳コーパス作成の過程で特許・実用新案公報に頻出する用語（一般・技術用語及び特許用語）を収集し、人手による修正と採否判定を行う形で、約 2,000 語の日インドネシア対訳辞書を作成した。

1.3 調査の構成

本調査は、以下の各手順で実施した。

① 日インドネシア語の対訳コーパス作成のためのデータの取得と加工［第 2 章］
② 日インドネシア語の文分割に関する調査と最適な手法の決定［第 3 章］
③ 日インドネシア語の文アライメントに関する調査と最適な手法の決定［第 4 章］
④ 日インドネシア語の対訳コーパスの作成と人手確認・修正、分析［第 5〜6 章］
⑤ 日インドネシア語公報に特化した対訳辞書の作成［第 7 章］
⑥ 日インドネシア語対訳コーパスによる機械翻訳エンジンの学習効果の評価［第 8 章］
⑦ 日インドネシア語対訳コーパスの整備に関する課題の分析と解決策の提言［第 9 章］

次章より、上記手順ごとの詳細をまとめる。

２．日インドネシア語の対訳コーパス作成のためのデータの取得と加工

本調査は、ファミリー関係にある日本とインドネシアの特許・実用新案公報データをソースに、一定以上の精度を満たす 100 万文対以上の日インドネシア対訳コーパスを取得することを目標とした。

これを達成するための第一の工程として、データソースに用いる大量の日インドネシア公報のファミリー紐付けを行い、続く第二の工程として、ファミリーと特定した日インドネシア公報それぞれから対訳コーパスに加工するためのテキストデータを抽出した。

本章にて、上記各工程の実施結果についてまとめる。

2.1. 日インドネシア公報のファミリー紐付け（ファミリーリストの作成）

はじめに、本調査で実施した日インドネシア公報のファミリー紐付け（ファミリーリストの作成）の概要及び結果を示す。

2.1.1. ファミリー紐付けの対象範囲

ファミリー紐付けは、特許庁から貸与されたインドネシア特許・実用新案公報フロントページの全件（97,500 件）と、インドネシア知的財産総局の公報データ提供サイト[6]に収録されている 2000 年以降の特許・実用新案公報の全件（174,729 件）から取得した書誌情報及び優先権情報（優先国及び優先権番号）を対象とした。

インドネシア公報の書誌情報及び優先権情報は、特許庁から貸与された公報フロントページ PDF（図 2-1）ならびにインドネシア知的財産総局公報データ提供サイト（図 2-2）から該当箇所を機械的に抽出する形で取得した。書誌情報としては出願番号、出願日、文献番号（公開番号）、公開日、特許分類情報を、優先権情報としては優先国、優先番号、優先日を取得した。

[6] https://pdki-indonesia.dgip.go.id/

図 2-1　インドネシア公報から書誌情報、優先権情報取得

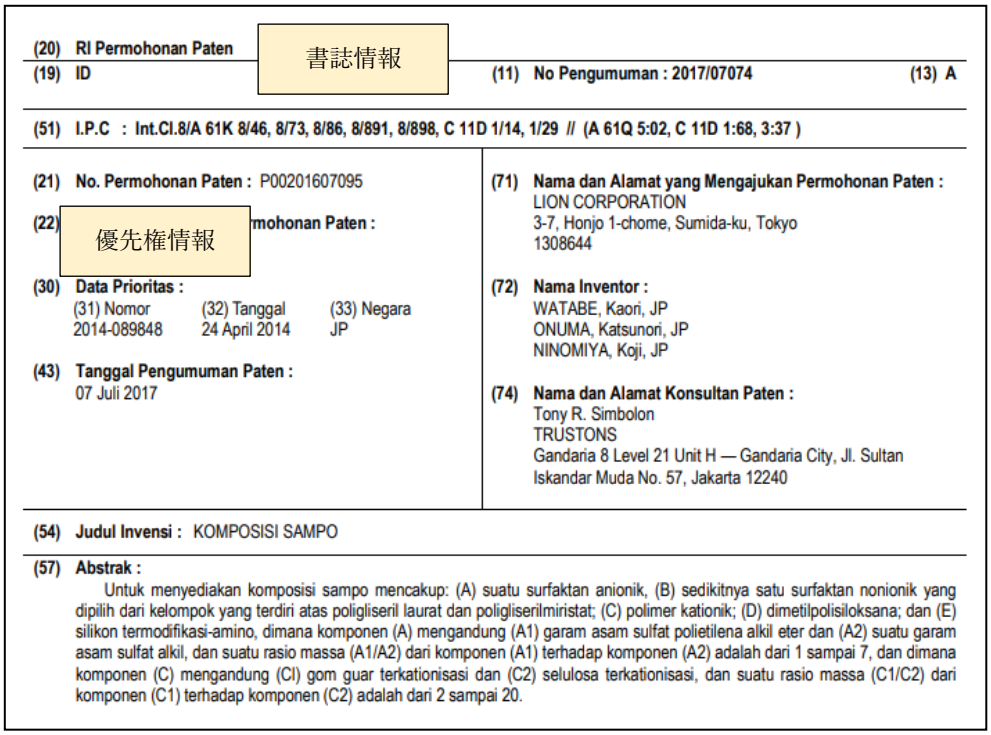
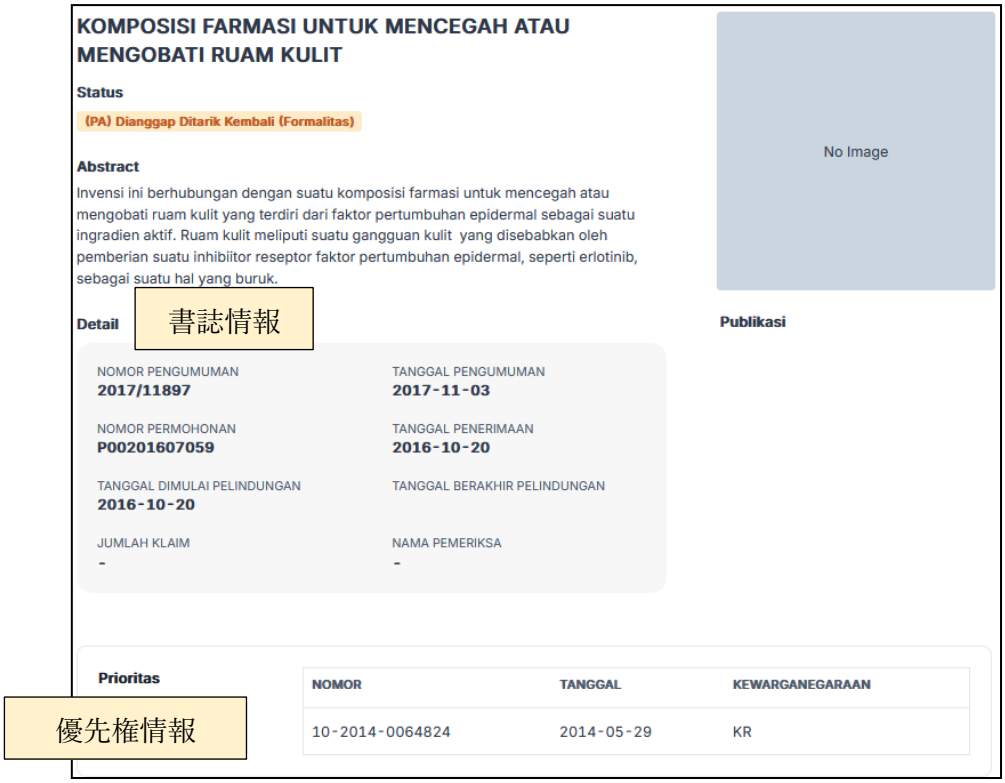


図 2-2　インドネシア文献の書誌情報（インドネシア知的財産総局公報データ提供サイト）



2.1.2. ファミリー紐付けの手順

ファミリー紐付けは、インドネシア公報フロントページ PDF 及びインドネシア知的財産総局公報データ提供サイトから取得した全ての優先権情報（すなわちインドネシア文献由来の優先権情報）を DOCDB のファミリーID と照合し、ファミリーに日本の優先権情報が含まれているものを対象とした。こうして各インドネシア公報とファミリー関係にある日本公報を特定し、そのうえで特許庁から貸与された日本公報データから書誌情報として出願番号、出願日、文献番号（公開番号)、公開日及び特許分類情報を取得した。

DOCDB のファミリーID を用いた上記手法では、インドネシア公報に日本の優先権情報が記載されている場合に加え、インドネシア公報が第三国への出願を介して間接的に日本公報と優先権関係にある場合（つまり、インドネシア公報自体には日本の優先権情報が記載されていない場合）にも双方を対応づけることができ、より網羅的なファミリー紐付けが行える。本調査では、この手法によって 63,406 件の日インドネシア文献対をファミリーとして紐付けることができた。

本調査では、これに加えて特許庁から提供された 24,810 文献対のファミリー情報を用い、全 88,216 文献対を日インドネシアのファミリー文献対として特定した。

2.1.3. ファミリー紐付けの結果（ファミリーリストの作成）

上記手順で紐付けたインドネシアと日本のファミリー公報 88,216 件について、両国の公報の出願番号、出願日、文献番号（公開番号)、公開日、優先権情報、特許分類情報を記載したファミリーリストにまとめた。

なお、上記手順で作成したファミリーリストには、1 文献に対して複数の文献が紐付けられるものも存在した（22,180 件)。この場合、対訳コーパスの作成対象は、複数の文献のうち公開日が最も古いものを使用した。結果、対訳コーパスの作成対象として 66,036 件の日インドネシア公報の文献対が特定された。

2.2. 日インドネシア公報からのテキストデータの抽出と加工

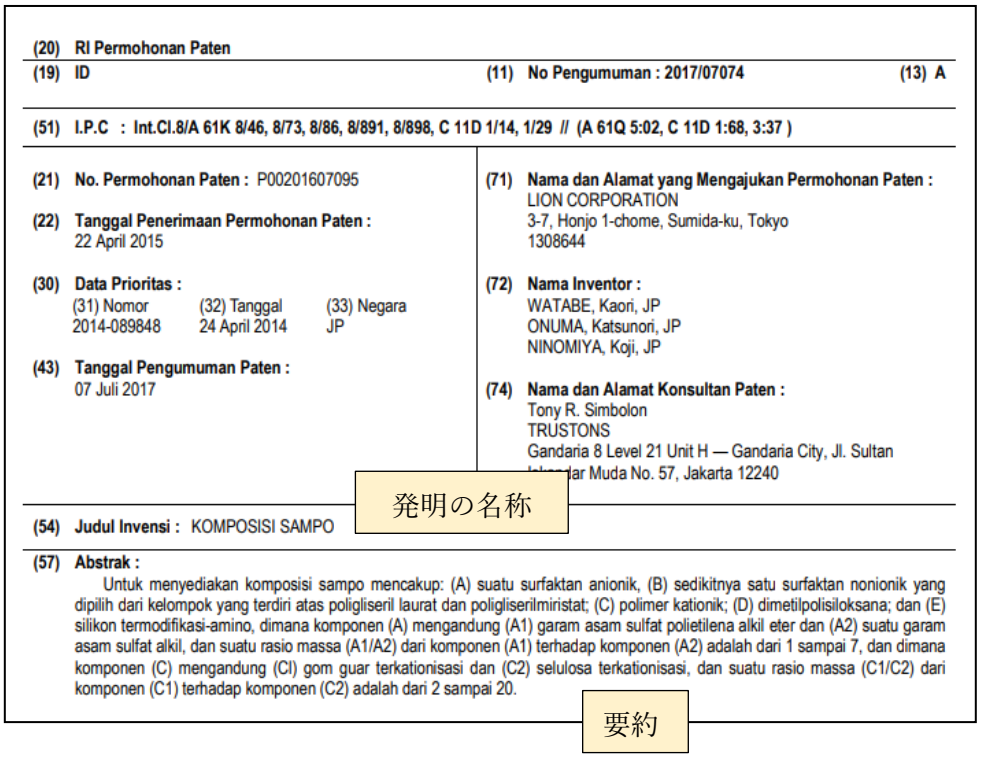
ファミリーとして紐付けた 66,036 件の日インドネシア文献対の各件から、対訳コーパスの作成対象として「要約」「発明の名称」「明細書」「特許請求の範囲」の全 4 項目のテキストデータを項目別に抽出した。本項で、その詳細を示す。

2.2.1. インドネシア公報からの「発明の名称」及び「要約」のテキスト抽出

インドネシア公報の「発明の名称」及び「要約」のテキストデータは、特許庁から貸与されたインドネシア公報フロントページの PDF ファイル、もしくはファミリー紐付け時にインドネシア知的財産総局の公報データ提供サイトからダウンロードした書誌情報から取得した。

インドネシア公報フロントページ（PDF 形式）においては、「発明の名称」「要約」とも文献ごとに出現位置が異なるため、それぞれに対応する見出し語（「(54) Judul Invensi」、「(57) Abstrak」）の出現位置を手掛かりとして、目的のテキストを特定した（図 2-3）。

図 2-3　インドネシア公報から「要約」「発明の名称」のテキストデータを取得



なお、PDF からテキストデータを抽出すると、上図の「要約」のように文の途中で改行されている場合、テキストが一行ずつ区切られて（つまり一文が複数に分割されて）しまう。このため、抽出したテキストデータから、項目（「発明の名称」及び「要約」）ごとに全ての改行を除去した。

一方、インドネシア知的財産総局公報提供サイト由来の文献は、ファミリー紐付けに使用するために書誌情報や優先権情報をダウンロードした際、あわせて「発明の名称」及び「要約」を取得しており、これをそのまま使用した。公報提供サイト由来のテキストデー

タは改行除去が不要であるため、公報フロントページ PDF と文献が重複する場合は、こちらを優先的に採用した。

なお、インドネシア公報の「要約」の文末には、「(Gambar X)」のような代表図を示す文言が記載される場合がある。同様の記載は日本の要約でも「【選択図】X」等として存在する場合もあるが、実内容がなく見出しのみであること（図自体はイメージデータでテキスト抽出されない)、文言の対応度が低く（Gambar X は「図 X」を意味するのみ）学習データとして利用性が低いこと、そしてファミリー文献同士でも日本、インドネシアで記載の有無がまちまちで対応度が悪いことから、ノイズデータと判断し、抽出後に全て除去した。

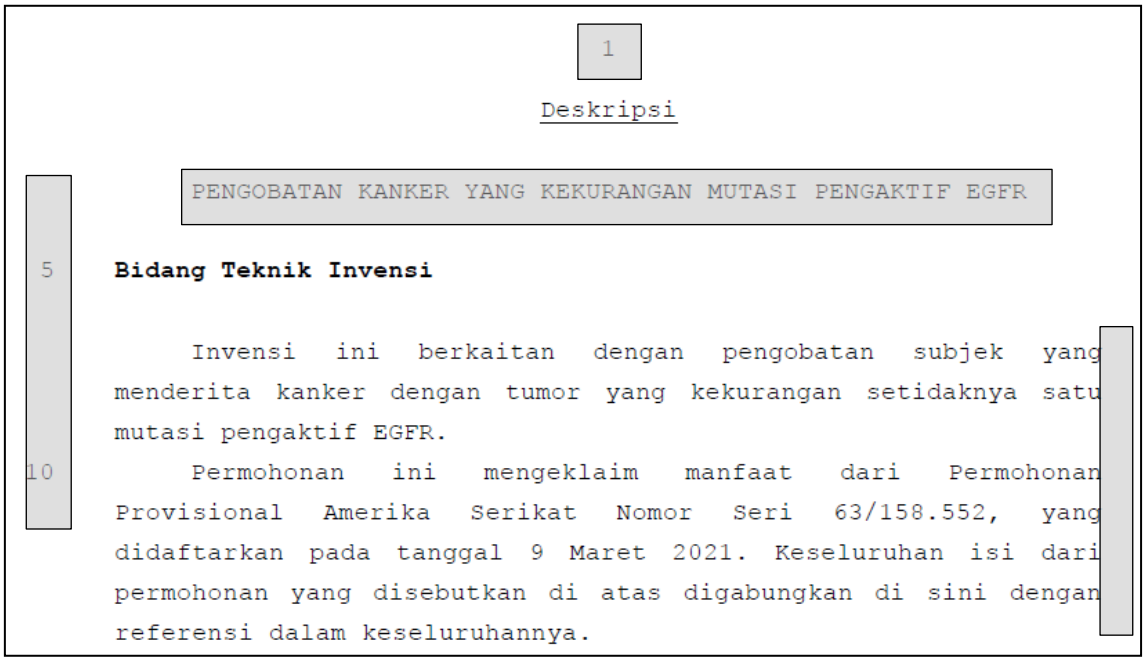
2.2.2. インドネシア公報からの「明細書」及び「特許請求の範囲」のテキスト抽出

インドネシア公報の「明細書」及び「特許請求の範囲」は、特許庁から貸与されたフロントページ PDF ファイル、インドネシア知的財産総局公報データ提供サイトから抽出したデータのいずれからも取得できない。このため、ファミリーリスト中のインドネシア公報のうち、公開日が最も新しい 25,000 件を対象に、公報全文の PDF ファイルが入手可能であるかをインドネシア知的財産総局公報データ提供サイトで逐一確認のうえダウンロードした。これにより、20,971 件のインドネシア公報全文 PDF ファイルを取得した。

こうして取得した各 PDF ファイルから、見出し語「Deskripsi」および「Klaim」を手がかりに「明細書」部分と「特許請求の範囲」部分を特定し、それぞれのテキストデータを抽出した。

これらのテキストデータの抽出に際しては、PDF フロントページと同様の改行の除去に加えて、各ページ中のページ番号ならびにページの欄外に 5 行おきに表示される行数表示の除去も必要であった。さらに、インドネシア公報の「明細書」冒頭には「発明の名称」が再掲されており、日本公報と対応しないためこれもノイズとして除去した（次図 2-4)。

図 2-4　インドネシア PDF 公報の「明細書」部分（※網掛け部分をノイズとして除外）



1

Deskripsi

PENGOBATAN KANKER YANG KEKURANGAN MUTASI PENGAKTIF EGFR

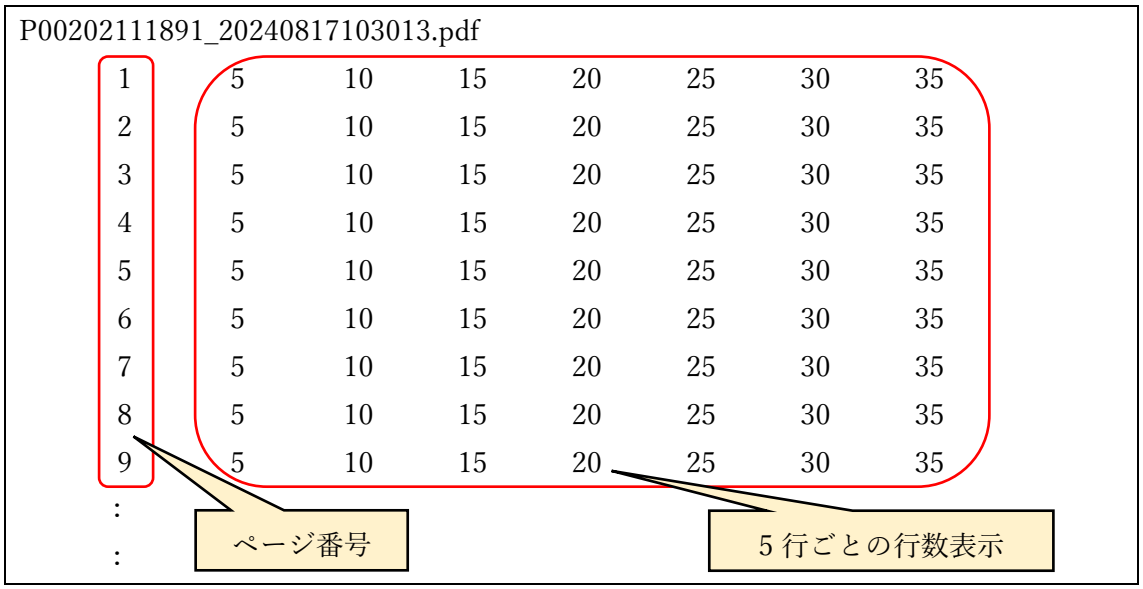
Bidang Teknik Invensi

Invensi ini berkaitan dengan pengobatan subjek yang menderita kanker dengan tumor yang kekurangan setidaknya satu mutasi pengaktif EGFR.

Permohonan ini mengeklaim manfaat dari Permohonan Provisional Amerika Serikat Nomor Seri 63/158.552, yang didaftarkan pada tanggal 9 Maret 2021. Keseluruhan isi dari permohonan yang disebutkan di atas digabungkan di sini dengan referensi dalam keseluruhannya.

これらのノイズの除去にあたっては、案件単位で除去したノイズ（ページ番号および行数表示）の内容を出力し（下図 2-5)、想定通りの除去が行われているかを確認した。

図 2-5　ノイズ除去結果リスト



P00202111891_20240817103013.pdf

1	5	10	15	20	25	30	35
2	5	10	15	20	25	30	35
3	5	10	15	20	25	30	35
4	5	10	15	20	25	30	35
5	5	10	15	20	25	30	35
6	5	10	15	20	25	30	35
7	5	10	15	20	25	30	35
8	5	10	15	20	25	30	35
9	5	10	15	20	25	30	35
:							
:							

ページ番号

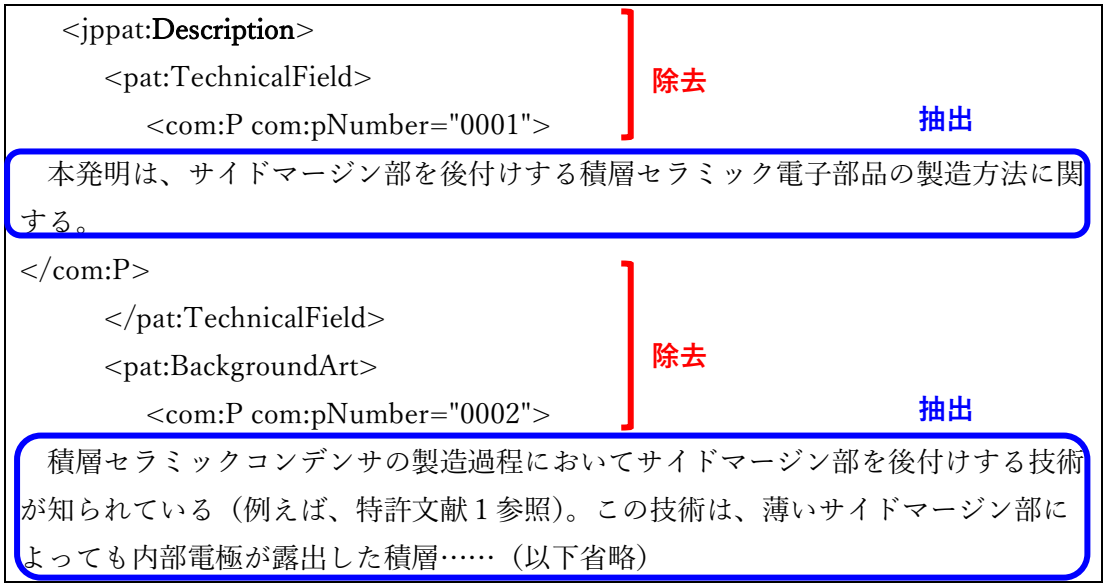
5 行ごとの行数表示

除去処理が正常に行われている場合、前図のように、各行の冒頭にページ番号、続いて5から35まで5刻みの行数表示が並ぶ。このパターンに合致しない場合はエラーメッセージを表示するようにした。本リストでエラーメッセージが発生したものをサンプリングで目視確認した結果、エラーは全て行数表示が出力されないケースであったが、これらはページ全体に表や図などのイメージが貼り付けられているため行数表示自体が存在しないものであり、正常な処理の結果であった。この結果により、不要箇所の除去処理が想定どおりに動作したことが確認された。

2.2.3. 日本公報からの「発明の名称」「要約」「明細書」「特許請求の範囲」のテキスト抽出

特許庁から貸与された日本公報（SGML又はXMLファイル）からの各項目のテキスト抽出は、公開されているデータ仕様[7]を参考に「発明の名称」、「要約」、「明細書」、「特許請求の範囲」の各項目を表わすタグ(<InventionTitle>、<Abstract>、<Description>、<Claims>)を手がかりにテキストデータを取得し、ノイズ（改行、各種タグ）を除去する形で行った（図2-6）。

図2-6　日本公報の例（XMLファイル）



「要約」においては、インドネシア公報と同じ理由から、末尾の「【選択図】X」をノイズとして一律で除去した。また、SGML／XMLファイルでは、記号「<」、「>」、「&」がそれぞれ「<」、「>」、「&」と実体参照文字で表現されているため、それぞれ半角記号表記への置換を行った。

[7] https://www.jpo.go.jp/system/laws/koho/shiyo/index.html

2.2.4. テキストデータ抽出結果

2.2.1.～2.2.3.項に示した各ソースからのテキスト抽出の結果、インドネシア、日本それぞれ項目ごとに以下の件数のテキストデータが抽出された。

表 2-1　インドネシア及び日本文献 記載項目別テキスト抽出結果

項目	インドネシア	日本
発明の名称	66,036 件	64,555 件
要　　約	62,996 件	64,447 件
明　細　書	20,667 件	19,244 件
特許請求の範囲	20,860 件	19,234 件

なお、インドネシア公報、日本公報とも、この時点では各項目とも全文が（改行の除去により）連結された一個のデータとなる（このため各項目とも 1 文献につき 1 件となっている)。

本来、インドネシア、日本ともに「発明の名称」と「要約」の取得件数はファミリーリスト掲載の 66,036 件（⇒2.1.3.)、「明細書」と「特許請求の範囲」はインドネシア公報の全文 PDF ファイルが入手できた 20,971 件（⇒2.2.2.）となるべきであるが、ウェブサイト由来の項目においては一部情報の欠落、PDF ファイル由来の項目においては項目の特定失敗（手掛かりとなる見出し語の不在や誤記が原因）により、取得できなかったものも若干数存在した。

こうして得られたテキストデータは、次章以降に記す後続工程で文単位に分割し（文分割)、さらにインドネシアと日本の文単位データ同士を対応づけて（文アライメント)、日インドネシア対訳コーパスに加工していくことになる。

３．日インドネシア語の文分割に関する調査と最適な手法の決定

前工程で取得した日インドネシア特許・実用新案文献由来のテキストデータから対訳コーパスを作成するためには、まず各項目の全文が連結されたテキストデータを一文単位に適切に分割する必要がある。本調査では、そのための最適な文分割手法について調査し、最適な手法を決定した。本章でその詳細を記す。

3.1. 文分割手法の調査の概要

本調査では、公開情報に基づいて日インドネシア語の文分割を自動的に実施する手法を広く収集し、各手法について、概要、サポート言語、想定される精度、課題、コスト、文分割スコア（文分割の確からしさを示す情報）の定義、（既存のツールの場合には）プログラミング言語、ライセンス条件、他言語（例えば、ベトナム語、タイ語）への適用可能性等をとりまとめた。

公開情報の調査は海外の情報も調査範囲とし、日本語での調査に加えて英語及びインドネシア語での調査も実施した。具体的な調査対象としては、arXiv.org7 を含む論文誌（プレプリントを含む）及び学会講演要旨集、Github、Hugging Face、ニュースリリース、及び、Google 等の検索エンジンの検索結果を調査した。

そして、調査により収集した各文分割手法について、想定される精度や課題、さらには予備実験の結果等を比較のうえ、本調査で用いる文分割の手法候補として日本語とインドネシア語それぞれ 3 手法を選定した。その後、各手法候補を用いて実際の特許文献の文分割処理を実施し、その結果を精密に比較して、本調査に最適なインドネシア語用及び日本語用の文分割手法をそれぞれ選定した。

3.2. 文分割手法の方式

公開情報調査により、文分割手法には、大別して「①：人間が決めたルールに基づき分割を実現するもの（ルール方式)」と「②：機械学習の結果（モデル）に基づき分割を実現するもの」の二種類が存在することが判明した。後者はさらに「②-1：文分割の位置などの情報を付加したデータをモデルに使用する方式（機械学習［教師あり］方式)」と「②-2：特別な情報のないプレーンな文書データをモデルに使用する方式（機械学習［教師なし］方式)」とに細分化される。

3.3 日インドネシア語の文分割手法一覧

本調査で検出したインドネシア語及び／又は日本語の文分割手法、全 16 種の概要を下記一覧表に示す。

表 3-1　日インドネシア語に使用可能な文分割手法一覧

No.	手法名	方式	サポート言語		
			イン	日	英
1	spaCy	複数方式[8]	○	○	○
2	id-sentence-segmenter	ルール	○	×	×
3	wtpsplit	機械学習(教師なし)	○	○	○
4	ja_sentence_segmenter	ルール	×	○	×
5	Stanza	機械学習(教師あり)	○	○	○
6	Splitta	機械学習(教師あり)	×	×	○
7	Ersatz	機械学習(教師あり)	×	○	○
8	Punkt	機械学習(教師なし)	×	×	○
9	CoreNLP	ルール	×	×	○
10	sentence-splitter	ルール	×	×	○
11	GiNZA	機械学習(教師あり)	×	○	×
12	pySBD	ルール	×	○	○
13	syntok	ルール	×	×	○
14	kuzukiri	ルール	×	○	×
15	SKBI 規則による文分割	ルール	○	×	○
16	EYD 規則による文分割	ルール	○	×	×

なお、インドネシア語はアルファベットで記載され、ピリオドが文区切り記号に用いられる点で英語と性質が類似している。このため、手法が豊富な英語の文分割手法が利用できる可能性が高い。このため、英語をサポートする主要な文分割手法も調査結果に加えた。

[8] spaCy は言語により用いる方式が異なり、依存関係解析を使用した文分割、統計手法を利用した文分割、ルールベースによる文分割を言語に応じて使い分けている。

3.4 予備実験の実施

収集した各手法とも、公開情報で得られる精度評価は英語を対象としたものが大半であり、インドネシア語や日本語の文分割精度を示す情報は十分に得られなかった。かつ、評価されているデータのジャンルもニュース記事のテキストを用いたものが多く、本調査で用いる特許文献由来のデータでの精度評価は存在しなかった。こうした状況の中、本調査の用途に沿った手法候補を的確に選定するため、特許文献を用いた予備的な実験を実施し、日インドネシア特許文献に対する処理精度を比較する必要があった。本項でこの予備実験の詳細を示す。

3.4.1. サンプル文献データ

予備実験に用いたサンプル文献データは、2.1.項に記したファミリー紐付け工程で取得した日インドネシアのファミリー文献対から無作為に選定した１文対（インドネシア公開公報：P00201907657、日本公開公報：JP2020037938A）から抽出した「要約」「明細書」「特許請求の範囲」のテキストデータを使用した。なお、「発明の名称」は常に１文で文分割が不要な項目につき省略した。

3.4.2. インドネシア文分割手法

インドネシア文分割手法の予備実験は、公開情報調査で収集した手法のうち、まずはインドネシア語のサポートを明示している４手法（下表 1〜4）を対象とした。さらに、英語をサポートする手法のうち、比較的新しい２手法（同 5〜6）も実験対象に加えた。インドネシア語はアルファベット文字である点やピリオドを文末記号に用いる点で英語と類似しており、英語をサポートする手法でも適切な文分割が行われる可能性が高いためである。

表 3-2　予備実験の対象としたインドネシア文分割手法

	手法名	言語	方式	備考
1	spaCy[9]	インドネシア	ルール	R5 年度調査で使用
2	Id-sentence-segmenter[10]	インドネシア	ルール	
3	wtpsplit[11]	インドネシア	機械学習［教師なし］	
4	Stanza[12]	インドネシア	機械学習［教師あり］	
5	Ersatz[13]	英語	機械学習［教師あり］	
6	pySBD	英語	ルール	

[9] https://spaCy.io/

[10] https://github.com/yudanta/id-sentence-segmenter

[11] https://github.com/bminixhofer/wtpsplit?tab=readme-ov-file

[12] https://stanfordnlp.github.io/Stanza/

[13] https://github.com/rewicks/Ersatz

これら 6 手法を用いて、サンプルのインドネシア文献（P00201907657）の全文テキストデータの文分割処理をそれぞれ実施した。

3.4.3. インドネシア文分割手法の実験結果

6 種類のインドネシア文分割手法によるサンプル文献の文分割結果に基づき、各手法の文分割の傾向及びその精度、処理に要した時間を分析した。本項にその結果を示す。

3.4.3.1. 項目（要約、明細書、特許請求の範囲）別の傾向

予備実験用インドネシア文献の「要約」「明細書」「特許請求の範囲」それぞれに対する各手法の文分割処理結果の特徴を本項にまとめた。

＜要約＞

「要約」の文分割では、wtpsplit を除く全てのツールが同一の適切な分割結果となった。wtpsplit のみ、下例のとおりコロンの直後での過分割（不要な分割）が見られた。

［コロン直後での過分割の実例］

インドネシア文	参考訳（機械翻訳）
Gambar yang dipilih:	代表図：
Gambar 1	画像 1

＜明細書＞

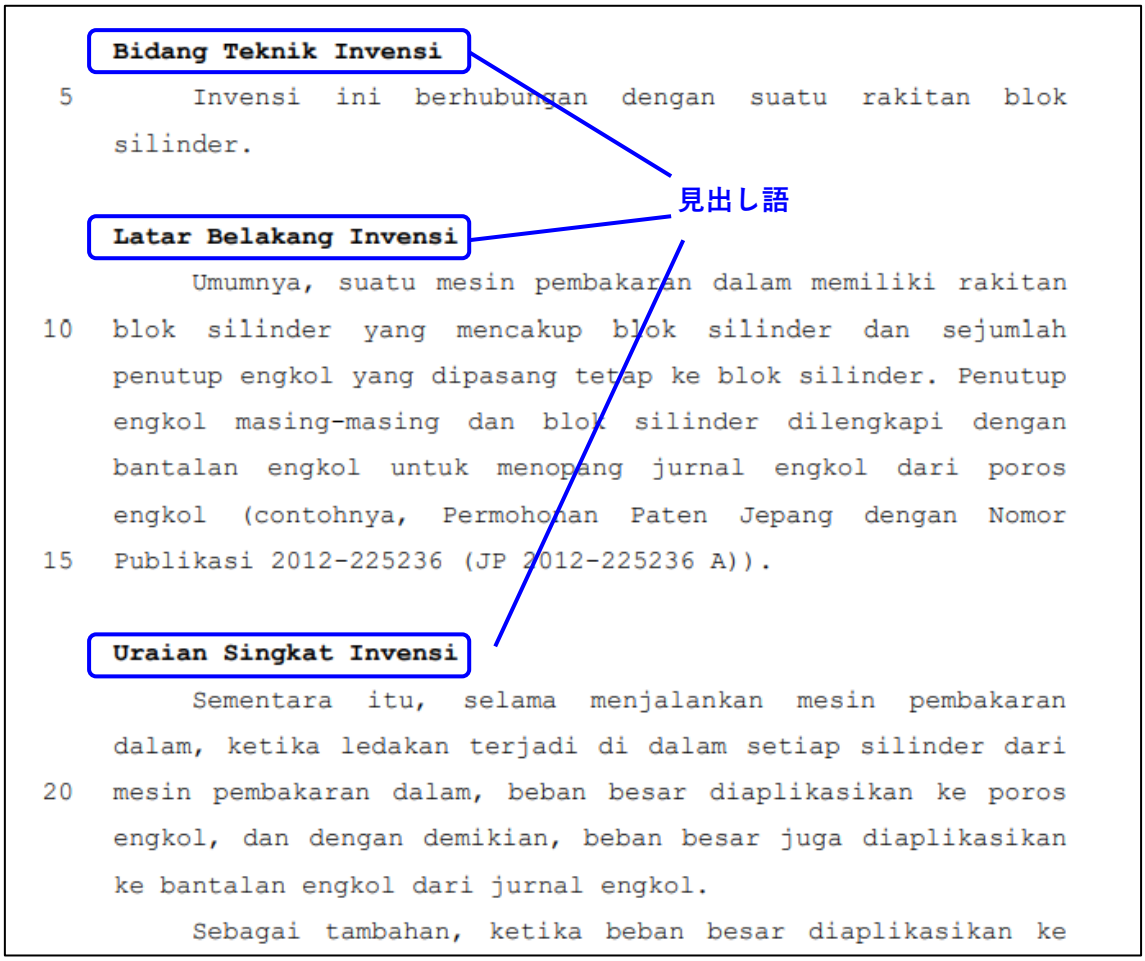
「明細書」の文分割結果も、「要約」と同様、wtpsplit を除く 5 手法がおおむね同一の内容となった。wtpsplit は「明細書」においてもコロン直後での過分割が見られた。

「明細書」の文分割結果における各手法共通の不備として、「見出し語」と本文との連結が挙げられる。

「明細書」は、他の項目に比べて「見出し語」が頻用される傾向にある。見出し語とは、例えば次ページ図 3-1 に示すようなものを指す。

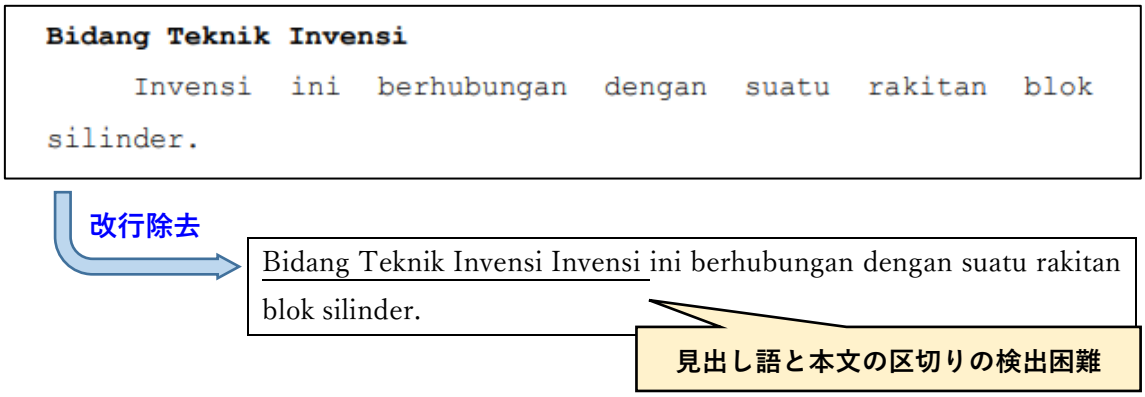
第 2 章で述べたとおり、PDF ファイルからテキストデータを抽出すると、各行末尾の改行により一文が分断されてしまう。このため、項目中の全ての改行を除去して全文を 1 つのデータファイルに連結している。各手法に見られた「見出し語」と本文との誤連結は、この処理が原因である。

図 3-1　インドネシア明細書における見出し語の例



通常の文であれば、項目中の改行を除去して全データを連結しても、文を区切る記号（インドネシア語の場合はピリオド）を的確に認識できれば、一文単位への再分割が可能である。しかしながら、見出し語の場合、文末にピリオドを有さないものが大半である。このため、文分割処理では、見出し語と直後の本文とが連結されると、次ページ図 3-2 のように区切りがわからなくなり、文分割処理でのリカバリは不可能に近い。

図 3-2　インドネシア見出し語と本文の連結結果の例



実際、予備実験結果においても、大半の手法において見出し語（特に上例のように末尾にピリオドや閉じカッコ区切り記号を伴わないもの）は全て本文と連結された。特許明細書からテキストデータを取得する際の大きな課題といえる。

なお、wtpsplit のみは、サンプル文献中の見出し語の多くで、本文との適切な再分割がなされていた。wtpsplit はコロン直後での過分割など他の手法より文を細かく区切る傾向が見られたが、見出し語の処理に限れば、この傾向が奏功している。

＜特許請求の範囲＞

特許請求の範囲においては、インドネシアの請求項冒頭に付される請求項番号（「1.」「2.」など）の直後で文を分割するか否かで、各手法の結果に相違が生じた。

spaCy は請求項番号の直後で必ず文分割を行い、Id-sentence-segmenter と pySBD は常に請求項番号と直後の本文とを連結した一文として出力した。これらに対し、その他の手法は請求項番号の直後で文分割を行うケースと行わないケースとが混在しており、一貫した方針が見出せなかった。なお、Ersatz は、請求項番号を本文と分割したうえで、直前の請求項の末尾に連結する傾向が顕著であった。

［請求項番号直後での過分割(spaCy)］

2.
Rakitan blok silinder (2) sesuai dengan klaim 1, yang dicirikan dengan masing-masing dari penutup engkol pusat dan penutup engkol samping memiliki lubang, dan lubang tersebut menembus penutup engkol (20).

［請求項末尾に次の請求項番号を連結(Ersatz)］

Rakitan blok silinder (2) sesuai dengan klaim 7, yang dicirikan dengan alur secara berurutan disediakan pada kedua permukaan samping dari masing-masing penutup engkol pusat dan penutup engkol samping untuk memiliki bentuk simetris, permukaan samping tersebut ditempatkan di dalam arah penjajaran dari silinder (11). 9.
Rakitan blok silinder (2) menurut salah satu dari klaim 1 sampai 8, yang dicirikan dengan sedikitnya sebagian dari lubang atau alur dipasang untuk menumpang tindih salah satu yang bersesuaian dari bantalan engkol jika dilihat di dalam arah pemasangan di mana penutup engkol (20) dipasang ke blok silinder (10). 10.

3.4.3.2. 手法ごとの特徴的な傾向

各項目を通じて各手法に見られた特徴的な傾向を本項にまとめる。

＜spaCy＞

spaCy は昨年度特許庁殿調査事業で使用されたツールである。spaCy の文分割結果においては、ピリオドの後で分割されない「未分割」が散見された（10 カ所)。「未分割」とは、本来分割されるべき箇所で分割されていない文分割の不備を指す。spaCy で見られた上記未分割はいずれも、図番号直後のピリオドが未分割となっていた。一例を示す。

［文末の図番号＋ピリオドでの未分割（spaCy)］

Gambar 14A adalah tampak samping dari penutup pusat dan penutup samping, dan Gambar 14B adalah penampang mendatar yang diambil sepanjang garis XIV-XIV dari Gambar 14A. Sebagaimana ditunjukkan pada Gambar 14A, Gambar 14B, dalam modifikasi pertama, masing-masing dari penutup pusat (20#3) dan penutup samping (20#1, 20#5) dilengkapi dengan tiga lubang (60') yakni lubang pertama (61'), lubang kedua (62'), dan lubang ketiga (63'), sebagaimana dengan perwujudan di atas.

また、これらとは別に、ピリオド直後で分割すべきところ、その後に続くダブルクォーテーションの直後で誤分割しているケースが 1 件存在した。

［ダブルクォーテーション直後での過分割（spaCy)］

Sebagai tambahan, arah yang tegak lurus terhadap "arah depan-belakang" dan "arah tinggi" diacu sebagai "arah lateral". "
Arah depan-belakang", "arah tinggi", dan "arah lateral" tidak harus menetapkan arah pemasangan dari rakitan blok silinder.

<Id-sentence-segmenter>

spaCy と同様、図番号直後のピリオドにおける未分割が見られた（10 カ所）。

［文末の図番号＋ピリオドでの未分割（Id-sentence-segmenter)］

Konfigurasi dari mesin pembakaran dalam yang dilengkapi dengan rakitan blok silinder sesuai dengan perwujudan ini akan diMT(JtoE)laskan dengan mengacu pada Gambar 1 dan Gambar 2. Gambar 1 adalah tampak perspektif terurai dari rakitan blok silinder sesuai dengan perwujudan ini.

また、重大な問題として、文献の途中までで分割結果が出力されなくなる事象が見られた。具体的には、下に示したように明細書末尾の 10 文（赤字部分）が出力されなかった。

［分割結果における文の不出力（Id-sentence-segmenter)］

Gambar 20 adalah bagan alir yang menunjukkan prosedur pembuatan dari rakitan blok silinder yang dilengkapi dengan penutup (20) yang ditunjukkan di dalam Gambar 3A, Gambar 3B, Gambar 4A, Gambar 4B. Pertama, dalam langkah S11, penutup yang tidak memiliki lubang (60) (atau tanpa alur (70)) diproduksi.
Pertama, dalam langkah S11, penutup yang tidak memiliki lubang (60) (atau tanpa alur (70)) diproduksi.
: (8 文省略) :
Walaupun perwujudan yang disukai sesuai dengan invensi ini telah diMT(JtoE)laskan di atas, invensi ini tidak terbatas pada perwujudan-perwujudan tersebut, dan berbagai modifikasi dan perubahan dapat dibuat dalam ruang lingkup klaim-klaim.

＜wtpsplit＞

wtpsplit は他のツールに比べて文を細かく分割する傾向が顕著であった。代表例はコロン、セミコロンでの分割であり、請求項等では文区切りとして妥当な範囲と見なせるものもあったが、明細書などで用いられる一般的な文においては過分割となるものが大半であった。さらに、下例のようにカンマの直後で分割したことで過分割となるケースも見られた。

［カンマ直後での過分割（wtpsplit)］

Penutup engkol masing-masing dan blok silinder dilengkapi dengan bantalan engkol untuk menopang jurnal engkol dari poros engkol (contohnya,
Permohonan Paten MT(JtoE)pang dengan Nomor Publikasi 2012-225236 (JP 2012-225236 A)).

＜Stanza＞

文分割すべきピリオド位置で分割しない「未分割」が下例を含め 7 カ所で見られた。

［文末ピリオドでの未分割（Stanza)］

Gambar 3A adalah tampak samping dari penutup pusat dan penutup samping , dan Gambar 3B adalah penampang mendatar yang diambil sepanjang garis III-III dari Gambar 3A. Dalam perwujudan ini , penutup pusat (20#3) dan sejumlah penutup samping (20#1 , 20#5) memiliki bentuk yang sama .

＜Ersatz（英語用分割）＞

Ersatz は英語用の文分割ツールであるが、インドネシア語と言語的特徴が類似することから有用である可能性があるため実験対象に加えた。結果、ピリオドの後で分割されない未分割が見られたが、2 カ所と少量であり、インドネシア語をサポートしている各手法と比べて特段の遜色はなかった。

［文末ピリオドでの未分割（Ersatz)］

Sebagai tambahan, alur (70) disediakan untuk simetris ke satu dengan yang lain terhadap bidang Z. Oleh karena itu, kemampuan deformasi dari penutup (20) dikonfigurasikan untuk simetris di dalam arah depan-belakang dan arah lateral.

また、特許請求の範囲において、請求項番号と本文とが分割される傾向が見られた。

<pySBD（英語用分割）>

pySBD も英語用の文分割ツールであり、「要約」、「明細書」では Ersatz と完全に同一の結果であったが、「特許請求の範囲」においては、Ersatz で見られた請求項番号直後での分割が発生しておらず、より良好であった。

3.4.3.3. 各手法の処理時間比較

文分割手法の選択においては処理効率も重要となる。このため、実験用文献の「明細書」部分の分割処理に各手法が要した時間を比較した。下表にその結果を示す。

表 3-3　インドネシア文分割手法の処理時間比較（実験用「明細書」全文）

ツール名	方式	処理時間（分:秒）
id-sentence-segmenter	ルール	0:00.03
pySBD	ルール	0:02.27
Ersatz	機械学習［教師あり］	0:02.65
spaCy	ルール	0:03.50
Stanza	機械学習［教師あり］	0:12.01
wtpsplit	機械学習［教師なし］	0:18.90

上表のとおり、最も処理時間が短かったのはルール方式の id-sentence-segmenter の 0.03 秒であり、最も長かったのが機械学習［教師なし］方式の wtpsplit の 18.9 秒であった。単純計算では、本調査で処理することとなる約 2 万件のインドネシア公報明細書全文の文分割を wtpsplit で行うには、約 100 時間が必要となる。

3.4.3.4. 評価候補 3 手法の選定

予備実験の目的は、本調査に最適の文分割の手法の選定のため、インドネシア語、日本語それぞれについて、評価対象とする手法候補を 3 手法ずつ決定することである。

予備実験の対象とした 6 種のインドネシア文分割のうち、見出し語の再分割という長所はあるものの、不要な過分割が顕著であり全体的な処理精度が最も劣る wtpsplit が最初に除外される（wtpsplit は処理に要する時間も他に比して長い）。

残る 5 手法の文分割の傾向はおおむね共通しており、出力結果も差異はわずかであったが、こうした少量の差異のうち、id-sentence-segmenter のみで見られた「明細書の中途で出力が中断するエラー」は、頻発した場合、得られる対訳コーパスの量を著しく減少させる可能性がある。他に比して重大度が高い不備と判断し、選外とした。また、Ersatz に見られた「請求項番号を分割し、それを直前の請求項の文末に連結する」という不備も、文アライ

メント処理でのリカバリが困難な性質の不備である[14]ため、選外とした。

これにより、残る 3 手法すなわち pySBD、spaCy、Stanza を、本調査の文分割手法候補とする。各手法とも少量の過分割、未分割は見られたが、選外の 3 手法に比べると発生頻度や重大性が低く、かつ過分割の多くは文アライメント処理でリカバリされる可能性がある。なお、Stanza は候補手法に含めたが、他の 2 手法に比べ、リカバリ困難な未分割が多く見られ、かつ処理時間を要するという弱点がある。

3.4.4. 日本文分割手法

日本文の文分割手法に関しては、公開情報調査で取得した 8 種のうち kuzukiri と ja_sentence_segmenter を除いた 6 種を予備実験の対象とした[15]。下表に一覧を示す。

表 3-4　予備実験の対象とした日本文分割手法

	手法名	言語	方式	備考
1	spaCy	日本	機械学習［教師あり］[16]	
2	wtpsplit	日本	機械学習［教師なし］	
3	Stanza	日本	機械学習［教師あり］	
4	Ersatz	日本	機械学習［教師あり］	
5	pySBD	日本	ルール	
6	GiNZA[17]	日本	機械学習［教師あり］	R5 年度調査で使用

これら 6 手法を用いて、サンプルの日本文献（JP2020037938A）の全文テキストデータの文分割処理をそれぞれ実施した。

---

[14] 請求項番号と本文を分割するのみであれば、後続の文アライメント処理で再連結され適切な一文（請求項番号＋本文）にリカバリされる可能性があるが、前請求項の末尾に連結された場合、文アライメント処理でのリカバリは望めない。

[15] kuzukiri は「令和 5 年度特許情報の機械翻訳における多言語対応に向けた課題検討に関する調査事業」において小数点で文分割が行わる課題が報告されており、本調査には不適と判断されたため選外とした。ja_sentence_segmenter は、予備実験以降も継続された公開情報調査において検出され、手法一覧に追加したものである。

[16] spaCy は言語により用いる方式が異なっている。インドネシア語はルール方式であったが、日本語の場合は機械学習が用いられる。

[17] https://megagonlabs.github.io/GiNZA/

3.4.5. 日本文分割手法の実験結果

日本文分割手法の予備実験も、インドネシア文の手法と同様、サンプル文献（日本公開公報 JP2020037938A）を上記 6 手法それぞれで文分割処理し、それぞれの傾向と精度、処理時間を比較する形で実施した。本項にその結果をまとめる。

3.4.5.1. 項目（要約、明細書、特許請求の範囲）別の傾向

＜要約＞

「要約」部分の文分割結果は 6 手法で完全に一致し、処理内容も適切であった。

＜明細書＞

「明細書」部分においても、wtpsplit 以外の 5 手法の文分割結果はほぼ一致し、その内容も適切であった。wptsplit のみ、インドネシア文の実験時と同様、不要な過分割が他に比して頻発していた。

なお、インドネシア文分割において各手法で見られた「見出し語が直後の本文と連結される」事象について、日本文献は XML 形式であり、タグで設定されている汎用的な見出し（例：【発明の概要】、【技術分野】など）はテキスト抽出時に本文との区切りを保持するようにした。このため、こうした見出しではインドネシア語のような本文との誤連結は発生しない。ただし、各文献で任意に用いられる見出し語（＝タグで設定されていない見出し語）に関しては、インドネシア文献と同様、テキスト抽出時に本文と連結されることが避けられない。

＜特許請求の範囲＞

特許請求の範囲においても、wtpsplit を除く各手法の文分割結果はほぼ同一かつ適切であり、wtpsplit のみで過分割が多発した。

3.4.5.2. 手法ごとの特徴的な傾向

本項では、各手法において他との相違が見られた文分割結果について主なものを挙げる。ただし、前述のとおり、wtpsplit 以外の 5 手法に関しては相違自体がごく少量であった。

＜spaCy＞

spaCy では、明細書部分において文末のカッコ補記表現において、開きカッコ直後で過分割される不備が 2 カ所見られた。

［文末カッコ補記における過分割（spaCy)］

加えて、本実施形態では、これら三つの孔部６０は、上下方向に見たときに、いずれも凹部２１の横方向における両端（すなわち、クランク軸受２２の横方向における両端）の内側に配置される（
すなわち、図３（Ａ）における領域Ｙ内に配置される)。

また、句点「。」の直前での過分割が 3 カ所見られた。

［句点の直前での過分割（spaCy)］

次いで、ステップＳ１３において、未加工キャップ及び加工済みキャップがシリンダブロックに組み付けられる
。

＜wtpsplit＞

wtpsplit では、下例のように他の手法では見られない過剰な分割が多数見られた。

［文の区切りとして適さない位置での過分割（wtpsplit)］

（５）
一つの前記
クランクキャップに設けられた複数
個の前記
孔部は互いに同一形状を有する、
上記（３）
又は（４）に記載のシリンダブロック組立体。

［引用文献番号の途中での過分割（wtpsplit)］

特開
２０１２－２２５２３６号公報

過分割は、後続の文アライメント処理でリカバリ（再連結）される余地があるため、リカバリ不可能な未分割より不備の重大性は小さいともいえる。ただし、文アライメント処理で「何文までを連結対象とするか」はパラメータによる制御が必要であり、その範囲を広げるほど処理負荷が指数関数的に増大していく（4.5.5.1.項で後述）。一文を細切れで出力する傾向がある wtpsplit は、文アライメント処理においてパラメータ値を極大に設定する必要が生じ、処理効率の観点できわめて不利である。

<Stanza>

Stanza の文分割結果に他の手法に見られない特徴的な誤りは見られなかった。ただし、下例のように文分割結果を半角スペースで細かく区切った形で出力するため、後処理で除去する必要がある。

［文分割結果をスペースで区切って出力（Stanza)］

以下␣、␣図面␣を␣参照␣し␣て␣実施␣形態␣に␣つい␣て␣詳細␣に␣説明する␣。

※「␣」は半角スペースを表す。

＜Ersatz＞

Ersatz の文分割結果もおおむね正確であったが、下例のように未分割（分割すべき句点で分割されない事象）が 2 カ所見られた。

［分割すべき句点（下線部）での未分割（Ersatz)］

図 11 からわかるように、3 番ジャーナル 31#3 では、2 番シリンダ 11#2 及び 3 番シリンダ 11#3 で爆発が生じたときに最も大きな摩擦損失が生じる。3 番ジャーナル 31#3 では、1 番ジャーナル 31#1 と同様に、摩擦損失は、全てのキャップ 20 の孔部 60 が設けられていない場合に比べて、1 番、3 番及び 5 番のキャップ 20 のみに孔部 60 が設けられている場合及び全てのキャップ 20 に孔部 60 が設けられている場合の方が小さいことがわかる。

なお、Ersatz は、入力文の全角英数字記号を全て半角に変換して出力するという、他手法では見られない特徴を有することが判明したが、後続の文アライメント処理への特段の悪影響はないと考えられる。

＜pySBD＞

pySBD 文分割結果は前出の Stanza と完全に一致しており、顕著な問題は見られなかった。かつ、Stanza に見られた出力結果への半角スペースの挿入もなく、より扱いやすい。

＜GiNZA＞

GiNZA は令和 5 年度の特許庁調査[18]で使用された手法である。同調査の報告書に「明らかに文区切りでないところで区切ってしまう問題」が指摘されていたが、現行バージョン（5.2）においてもこれに該当する過分割が少数ながら見られた。

［文の中途での過分割①（GiNZA)］

（１１）前記
中央クランクキャップと二つの前記側方クランクキャップとは互いに同一形状を有する、上記（1）～（１０）のいずれか一つに記載のシリンダブロック組立体。

［文の中途での過分割②（GiNZA)］

また、図中の＃
１、＃２、＃３、＃４は、１番シリンダ１１＃１での爆発時期、２番シリンダ１１＃２での爆発時期、３番シリンダ１１＃３での爆発時期、４番シリンダ１１＃４での爆発時期をそれぞれ示している。

一方、他のツールでは対応できなかった「テキスト抽出時に本文と連結されてしまった任意の見出し語」を再分割できた箇所が存在した。

［見出し語と本文の再分割①（GiNZA)］

＜クランクキャップの構成＞
次に、図３及び図４を参照して、クランクキャップ２０の構成について具体的に説明する。

ただし、リカバリは一部の見出し（＜＞で括られたもの）に限られ、かつこれに該当する見出しでも分割位置が不正確となる場合もあった。

［見出し語と本文の再分割②（GiNZA)］

＜内燃機関の構成
＞図１及び図２を参照して、本実施形態に係るシリンダブロック組立体を備える内燃機関の構成について説明する。

[18] 特許情報の機械翻訳における多言語対応に向けた課題検討に関する調査事業。1.1.項の脚注 5 参照。

3.4.5.3. 各手法の処理時間比較

日本文分割手法についても、実験対象の文献（明細書部分）の分割処理に要した時間を比較した。下表にその結果を示す。

表 3-5　日本文分割手法の処理時間比較（実験用「明細書」全文）

ツール名	方式	処理時間（分:秒）
pySBD	ルール	0:00.11
Ersatz	機械学習［教師あり］	0:02.79
GiNZA	機械学習［教師あり］	0:06.15
wtpsplit	機械学習［教師なし］	0:06.18
Stanza	機械学習［教師あり］	0:16.84
spaCy	機械学習［教師あり］	1:38.86

最も処理時間が短かったのはルール方式の pySBD の 0.11 秒、最も長かったのが機械学習［教師あり］方式の spaCy の 1 分 38 秒 86 であった。インドネシア文に比べて全体的に処理時間を要しており、spaCy の 2 万文献対での想定所要時間は 544 時間（約 22.7 日）となる。

3.4.5.4. 評価候補 3 手法の選定

日本文分割手法に関しても、本調査で使用する手法候補として評価対象とする 3 手法を本実験結果に基づき選定する。

日本文の分割結果においても、wtpsplit は他の手法に比べて極端な過分割が頻発するため最初に選外となる。残る 5 手法は処理結果に特徴的な差異は見られず、いずれも良好な精度であった。その中で、spaCy は他に比べて極端に処理時間を要しており、処理効率の観点で除外される。

残る Stanza、Ersatz、pySBD、GiNZA の 4 手法の比較では、相対的に誤分割が少なかった pySBD が第一候補となり、出力結果に挿入される半角スペースを除去する必要があるものの分割精度自体は pySBD と同様に最上位であった Stanza が次点となる。Ersatz と GiNZA の比較では、本文と連結された見出し語の一部を再分割するという他に見られない特徴が見られた後者（GiNZA）を優位と見て、pySBD、Stanza、GiNZA の三手法を候補とした。

3.4.6. 文分割処理実施における課題

予備実験の実施結果から、各手法を通じて文分割処理の実施の際に課題となる事象が検出された。以下に示す。

＜課題１：テキスト抽出時に見出し語と直後の本文が連結される＞

明細書部分では各所に見出し語が使用される。見出し語は末尾にピリオドや句点を伴わないため、テキスト抽出時に直後の本文と連結された後、文分割処理で再分割することが難しい。

ただし、多くの文献で共通して使用される見出し語に関しては、あらかじめリストアップしておき、該当する文字列が文中に存在する場合、その前後で再分割する後処理を施すことが可能である。この処理により、頻出の見出し語に関しては、本文と切り離すことができる。

インドネシア公報約 1,000 件を対象に検出した見出し語を以下に示す。

表 3-6　インドネシア公報における頻出の見出し語（全 42 種）

No.	インドネシア見出し語	日本見出し語（参考訳）
1	ACUAN SILANG KE PERMOHONAN TERKAIT	関連出願の相互参照
2	BIDANG TEKNIK	技術分野
3	Bidang Teknik Invensi	発明の技術分野
4	BIDANG TEKNIK INVENSI	発明の技術分野
5	BIDANG TEKNIS	技術分野
6	Bidang Teknis Invensi	発明の技術分野
7	BIDANG TEKNIS INVENSI	発明の技術分野
8	Daftar Kutipan	引用文献リスト
9	Daftar Tanda Acuan	参考文献リスト
10	Daftar Tanda-tanda Referensi	符号の説明
11	Efek Invensi	発明の効果
12	Efek Menguntungkan dari Invensi	本発明の効果
13	Efek Menguntungkan Invensi	発明の効果
14	KLAIM PRIORITAS	優先権主張
15	KLAIM PRIORITAS DI BAWAH	優先権主張
16	Latar Belakang Invensi	発明の背景
17	LATAR BELAKANG INVENSI	発明の背景
18	LINTAS REFERENSI KE PERMOHONAN TERKAIT	関連出願の相互参照

19	Masalah Teknik	技術的課題
20	Masalah Teknis	技術的課題
21	Masalah Yang Akan Diselesaikan Melalui Invensi	発明が解決しようとする課題
22	Pemecahan Masalah	課題を解決するための手段
23	Pengaruh Invensi yang Menguntungkan	発明の効果
24	Penyelesaian Masalah	発明が解決しようとする課題
25	Penyelesaian Untuk Memecahkan Masalah	課題を解決するための手段
26	Permasalahan Teknis	技術的課題
27	REFERENSI SILANG	関連出願の相互参照
28	REFERENSI SILANG / KLAIM PRIORITAS UNTUK APLIKASI TERKAIT	関連出願の相互参照/優先権主張
29	REFERENSI SILANG BERIKAITAN DENGAN REFERENSI	関連出願の相互参照
30	REFERENSI SILANG DENGAN APLIKASI TERKAIT	関連出願の相互参照
31	REFERENSI SILANG DENGAN PERMOHONAN TERKAIT	関連出願の相互参照
32	REFERENSI SILANG KE PERMOHONAN TERKAIT	関連出願の相互参照
33	REFERENSI SILANG PADA APLIKASI TERKAIT	関連出願の相互参照
34	REFERENSI SILANG UNTUK APLIKASI TERKAIT	関連出願の相互参照
35	REFERENSI SILANG UNTUK PERMOHONAN TERKAIT	関連出願の相互参照
36	REFERENSI-SILANG DENGAN PERMOHONAN TERKAIT	関連出願の相互参照
37	RUJUKAN SILANG KE PERMOHONAN TERKAIT	関連出願の相互参照
38	RUJUKAN SILANG UNTUK APLIKASI TERKAIT	関連出願の相互参照
39	Solusi untuk masalah	課題を解決するための手段
40	Uraian Lengkap Invensi	発明の完全な説明
41	Uraian Singkat Gambar	図の簡単な説明
42	Uraian Singkat Invensi	発明の簡単な説明

本調査で行う文分割処理では、この後処理を施すこととする。ついては、後続で実施するインドネシア文分割処理手法候補の評価も、この後処理を施す前提で行う。

＜課題２：請求項番号の取り扱い＞

インドネシア文献データは、各請求項の冒頭に「数字＋ピリオド」の形式で請求項番号が付されており、そのままテキスト抽出される。この箇所が、手法により請求項本文と連結される場合と、請求項番号のみで分割される場合がある。

一方、日本文献データでは、請求項番号は XML タグ化されており、このため請求項本文と分割した状態で抽出される。このため「インドネシア文の冒頭が「数字＋ピリオド」のものについて再分割する」という、課題１と同種の後処理を施せば、日本文とインドネシア文の双方を、請求項番号と請求項本文とが分割された状態とすることができる。

ただし、この場合、請求項番号部分から生成される対訳コーパスは、インドネシア語が「数字＋ピリオド」、日本語が「【請求項 N】(N は数字)」という、機械翻訳の学習データとして無用なものとなる。つまり、後処理を実施しても、対訳コーパスから除外すべきノイズデータが各文献から多数生成されるのみである。

このため、請求項番号は、インドネシア文分割処理の実施後に、請求項部分から取得した各文冒頭から「番号＋ピリオド」を削除するのが最善と考える。そのうえで、日本文ではタグ化された請求項番号を取得対象外とすることで、双方とも請求項本文のみを文分割結果として後続の文アライメント処理に渡すことができる。

後続で実施するインドネシア文分割処理手法候補の評価は、この後処理を施す前提で行う。

＜課題３：図番号の後のピリオドで分割されない（未分割）＞

インドネシア文分割手法の出力の１文中に、パターン「Gambar**X.Y**」(X は英数からなる 1 単語、Y は大文字の英字)が出現する場合、XY 間のピリオドの直後で分割する後処理を追加することで対処可能と考えられる。

［GambarX.Y に該当する図番号直後のピリオドでの未分割の事例］

Gambar 3A adalah tampak samping dari penutup pusat dan penutup samping , dan Gambar 3B adalah penampang mendatar yang diambil sepanjang garis III-III dari Gambar 3A. Dalam perwujudan ini , penutup pusat (20#3) dan sejumlah penutup samping (20#1 , 20#5) memiliki bentuk yang sama .

後続で実施するインドネシア文分割処理手法候補の評価は、この後処理を施す前提で行う。

3.5. 文分割手法候補の人手評価

本調査に最適の文分割手法の選定のため、インドネシア文、日本文それぞれ3種の文分割手法候補に対し、人手による評価を実施した。本項にその結果をまとめる。

3.5.1. インドネシア文分割手法候補3種の評価結果

評価対象とするインドネシア文分割手法の候補は、予備実験の結果を踏まえ、下記の3手法とした。

① pySBD
② spaCy
③ Stanza

3.5.1.1 評価方法

評価は、本調査で取得した日インドネシアのファミリー特許文献から任意に選定した下記インドネシア文献 5 件を評価用文献に用い、各文献から抽出した全文テキストを各手法候補で文分割処理した結果を評価対象とした。具体的には、各文献からそれぞれ60文ずつを評価対象文に選定し、それらが過不足なく一文として分割されているかを評価した。

表 3-7　評価用インドネシア文献（5 文献）

	インドネシア文献	IPC	（参考）対応日本文献
①	P00202203385	B23K	2022-002855
②	P00202301594	A01H	2022-024515
③	P00202304729	H04W	2022-073428
④	P00202300918	C08F	2022-012848
⑤	P00202301067	B60R	2022-034129

評価対象文には、文中にピリオドやコロン、セミコロン、カンマを含む文や長大な文、特徴的な表記の化学物質名が列挙された文など、文分割の難易度が高いと予測される文を多数含めた。これらに、特殊な事例としてピリオドを伴う微生物名を含む1文を加えた全301文を評価対象とした。

各文の評価は人手で行い、過不足なく文分割ができている「正常」、一文が過剰に分割されている「過分割」、複数文が分割されず連結されている「未分割」のいずれかの判定とした。

ただし、見出し語と直後の本文とが連結されているタイプの「未分割」は、テキストデータ取得時に不可避的に生じるものであり、文分割処理で機械的に再分割することが事実上

不可能であるため（3.5.1.4.項で詳述）、別項目「見出し連結」としてカウントし、文分割手法の評価においては「正常」に準ずる扱いとした。

なお、過分割、未分割、見出し連結は 1 文において同時に発生することもある。この場合は、最上位から未分割＞過分割＞見出し連結の順で 1 種のみをカウントした。

3.5.1.2 評価結果

以下、手法候補 3 種の人手評価結果（5 文献×60 文＋ 1 ）を示す。

［① pySBD］…正解率[19]：78.4%（準正解率[20]：89.4%）

文献	正常	見出し連結	過分割	未分割	合計
①	48	9	3	0	60
②	45	7	6	2	60
③	54	1	4	1	60
④	41	7	9	3	60
⑤	47	9	4	0	60
微生物	1	0	0	0	1
合計	236	33	26	6	301

［② spaCy］…正解率：74.1%（準正解率：80.1%）

文献	正常	見出し連結	過分割	未分割	合計
①	44	3	13	0	60
②	40	3	14	3	60
③	55	1	2	2	60
④	38	3	14	5	60
⑤	46	8	4	2	60
微生物	0	0	1	0	1
合計	223	18	48	12	301

[19] 「正解率」は、評価対象 301 文中の「正常」の比率とした。

[20] 「準正解率」は、評価対象 301 文中の「正常」及び「見出し連結」の比率とした。

［③ Stanza］…正解率：76.7%（準正解率：86.3%）

文献	正常	見出し連結	過分割	未分割	合計
①	39	8	1	12	60
②	47	7	3	3	60
③	42	1	1	16	60
④	49	8	2	1	60
⑤	53	5	0	2	60
微生物	1	0	0	0	1
合計	231	29	7	34	301

各手法候補の正解率の比較では、pySBD が最も高く 78.4%（301 文中 236 文）、次いで Stanza が 76.7%（同 231 文）、spaCy が 74.1%（223 文）の順であった。いずれも 75%周辺で顕著な差は見られなかった。ただし、文分割処理においては不可抗力[21]といえる「見出し連結」を正解に含めた準正解率は、pySBD が 89.4%、Stanza が 86.4%なのに対し spaCy は 80.1%にとどまり、やや差が開いた。他の 2 手法では（不可避的な）見出し連結が生じた文において、spaCy ではより重大な過分割もしくは未分割が生じるケースが多かったことを示している。この点において spaCy は他の 2 手法にやや劣る。

[21] 見出し語は末尾に区切り記号を有さないことが多いため、公報データからのテキスト抽出時に直後の本文と連結されてしまうと文分割処理でのリカバリは困難である。

3.5.1.3. 各ツールの特徴

各手法候補とも、文分割に失敗する特徴的な記載パターンが存在した。以下、その内容を示す。

<pySBD>

pySBD における特徴的な文分割不正のパターンを以下に示す。

【カッコ補記での過分割】

pySBD では、文中のカッコ補記部分で過分割されるケースが散見された。以下、一例を示す。

［文献① 明 52］

対象文	文分割結果
Kasus dimana nilai rata-rata dari rasio area senyawa- senyawa berbasis CuSn terkelupas dari antarmuka yang bergabung (nilai rata-rata untuk lima sampel) adalah kurang daripada 1% dianggap sebagai “〇” (baik), sedangkan kasus dimana nilai rata- rata dari rasio area senyawa-senyawa berbasis CuSn (nilai rata-rata untuk lima sampel) adalah 1% atau lebih dianggap sebagai “×” (bersifat cacat).	Kasus dimana nilai rata-rata dari rasio area senyawa- senyawa berbasis CuSn terkelupas dari antarmuka yang bergabung (nilai rata-rata untuk lima sampel) adalah kurang daripada 1% dianggap sebagai “〇”
	(baik), sedangkan kasus dimana nilai rata-rata dari rasio area senyawa-senyawa berbasis CuSn
	(nilai rata-rata untuk lima sampel)
	adalah 1% atau lebih dianggap sebagai “×”
	(bersifat cacat).

本例は文中にカッコ補記が 4 箇所存在するが、うち 3 箇所で不要な文分割（過分割）が発生している。具体的にはカッコ補記の直前で不要な文分割同様の事象は、文献①の明 54、55、文献②の請 05、明 22、明 39、文献⑤の明 41、42、47 でも見られた。ただし、文中にカッコ補記が必ず過分割されるわけではなく、上例でも、最初のカッコ補記（青字）では過分割は生じていない。

カッコ補記過分割は、文献②、文献⑤を見ると、(a)、(b)の場合に必ず分割されるようである（下例参照）。

［文献② 請 05］

対象文	文分割結果
Biji bunga matahari menurut salah satu dari klaim 1 sampai 3, dimana sekuensnya terdiri dari (a) atau (b) di bawah ini: (a) sekuens basa yang diwakili oleh SEQ ID NO: 1, atau (b) sekuens basa yang memiliki setidaknya 80% identitas dengan sekuens yang diwakili oleh SEQ ID NO: 1. 5.	Biji bunga matahari menurut salah satu dari klaim 1 sampai 3, dimana sekuensnya terdiri dari
	(a) atau
	(b) di bawah ini:
	(a) sekuens basa yang diwakili oleh SEQ ID NO: 1, atau
	(b) sekuens basa yang memiliki setidaknya 80% identitas dengan sekuens yang diwakili oleh SEQ ID NO: 1. 5.

このことから、pySBD は、(a)、(b)のような箇条書き番号に多用される文字列はその直前で文分割を行う設定になっている可能性がある。ただし、冒頭に挙げた事例はこの条件には合致せず、どのような基準で文分割が行われたのかは判然としない。

【文末の数値の過分割と次文の冒頭への連結】

pySBD では、文末が単独の数字＋ピリオドの場合に、その直前で過分割され、次文の冒頭に連結されるケースが多数見られた。以下、一例を示す。

［文献③ 明 47］

対象文[22]	文分割結果
Pemrosesan bagan alir ini dimulai sebagai respons terhadap peralatan komunikasi 102 yang memulai pemrosesan di S505 pada Gambar 5.	Pemrosesan bagan alir ini dimulai sebagai respons terhadap peralatan komunikasi 102 yang memulai pemrosesan di S505 pada Gambar
Pertama, peralatan komunikasi 102 menentukan apakah interval transmisi dari frame penemuan FILS telah berlalu sejak transmisi terakhir dari frame suar	5. Pertama, peralatan komunikasi 102 menentukan apakah interval transmisi dari frame penemuan FILS telah berlalu sejak transmisi terakhir dari frame suar atau frame

[22] 「対象文」に示した（文分割処理前の）テキストデータは、実際には全ての文が連結された状態であるが、「文分割結果」との対比を容易にするため、あるべき文分割位置で区切って表示している。このため枠線は点線とした。一方、文分割結果は実際に文分割処理された位置で区切っており、枠線は実線としている。後続の事例も同様。

atau frame penemuan FILS S601.	penemuan FILS S601.

これもカッコ補記の過分割と同様、単独の数字＋ピリオドを箇条書き番号と見なしている可能性が考えられる。同様のケースは、文献③の明 53、54、文献④の明 12、28、49、51、52、58 でも見られた。

なお、本例（2 文目末尾）もそうだが、文末が数字＋ピリオドであっても、単独の数字でない（つまり箇条書き番号として頻用されない）文字列である場合は、過分割は起こらないようである。

＜spaCy＞

spaCy における特徴的な文分割不正のパターンを以下に示す。

【文頭のカッコ直後で過分割され前文末尾に連結】

spaCy では、文頭がカッコから始まる文において、開きカッコの直後での過分割が確実に発生していた[23]。以下、一例を示す。

［文献① 明 23］

対象文	文分割結果
Selanjutnya, unsur-unsur yang disertakan dalam paduan patri menurut perwujudan ini akan diMT(JtoE)laskan.	Selanjutnya, unsur-unsur yang disertakan dalam paduan patri menurut perwujudan ini akan diMT(JtoE)laskan. (
(1) Ag: 3,1 sampai 4,0% massa	1) Ag: 3,1 sampai 4,0% massa Ag memiliki efek untuk meningkatkan sifat …[24]
Ag memiliki efek untuk meningkatkan sifat …	

上例「対象文」の 2 文目は冒頭にカッコ付番「(1)」が付された見出し語であるが、「文分割結果」では、先頭の「(」の直後で過分割され、これが前文の末尾に連結されてしまっている。文頭がカッコではじまる全ての文献の全ての文で発生しており、「[　]」（文献①明 58、④明 44、46、48、49、⑤明 33 など）や「“　”」（文献②明 38、④明 23 など）など多種のカッコ表記で発生するなど発生頻度が高い。これが spaCy の評価が伸び悩んだ大きな理由

[23] 予備実験では文末のカッコ補記の開きカッコ直後での過分割を指摘したが、これも当該カッコ補記が次文冒頭部と見なされた可能性が高い。

[24] 本行の文分割結果は、見出し語と、次文（Ag memiliki efek untuk meningkatkan sifat …）とが連結されてしまっている。これが「２．人手評価」で述べた「見出し連結」である。この後の事例でも同様の状況のものがある。

であるが、過分割の中では対訳コーパスの精度への悪影響は比較的低いタイプともいえる。

なお、このエラーでは、対象文２行目は冒頭カッコの過分割であるが、１行目は文末に次文の冒頭カッコが連結されているため「未分割」と判定される。つまり、１、２文目がいずれも評価対象文であった場合、１文目が未分割、２文目が過分割となる。

【冒頭の数字＋ピリオドでの過分割】

文中ピリオドでの過分割は各ツールで散見された。その中には、どのような条件で発生したのか判然としないものも多いが、spaCy の場合、見出し語など冒頭に数字＋ピリオドを伴う文において、ピリオドの直後で確実に過分割される。以下に例示する。

［文献① 明 22］

対象文	文分割結果
1. Paduan patri	1.
Paduan patri menurut perwujudan ini biasanya adalah …	Paduan patri Paduan patri menurut perwujudan ini biasanya adalah …

上例の対象文「1. Paduan patri」は項番付きの見出し語であり、途中で分割すべきではないが、spaCy は「1.」の直後で過分割している。本例を含め、少なくとも「文頭が数字１文字＋ピリオド」に該当する文においては、全ての文で同種の過分割が発生していた（①明 46、②明 30）。

さらに類似のケースとして、文献④の明 13、14、15、27 のように「<N. …>」や「<N.N …>」などというタイプや、文献②の明 20、21 のように、「Item N. …」というタイプの見出し語もことごとく過分割されていた。以下、一例を示す。

［文献④ 明 13～14］

対象文	文分割結果
<1. Sistem produksi pulp dan sistem perlakuan lindi hijau menurut perwujudan ini>	<1.
<1-1. Ringkasan sistem produksi pulp>	Sistem produksi pulp dan sistem perlakuan lindi hijau menurut perwujudan ini> <1-1.
Gambar 1 adalah diagram konfigurasi skematis sistem produksi pulp (1) yang memiliki sistem perlakuan lindi hijau	Ringkasan sistem produksi pulp> Gambar 1 adalah diagram konfigurasi skematis sistem produksi pulp (1) yang memiliki sistem

menurut satu perwujudan dari invensi ini, meskipun invensi ini tidak terbatas padanya.	perlakuan lindi hijau menurut satu perwujudan dari invensi ini, meskipun invensi ini tidak terbatas padanya.

【文末の dll.での未分割】

インドネシア語の「dll.」は、英語の「etc.」に相当する略語であり、ピリオドを伴う。文中で使用されている場合はそこで分割すべきではないが、spaCy の場合、dll.が文末であっても文分割の対象外とされてしまうと見られる。実例を示す。

［文献③ 明 08］

対象文	文分割結果
IEEE adalah singkatan dari Institute of Electrical and Electronics Engineers, dan seri standar IEEE 802.11 mencakup standar seperti IEEE 802.11a/b/g/n/ac/ax dll.	IEEE adalah singkatan dari Institute of Electrical and Electronics Engineers, dan seri standar IEEE 802.11 mencakup standar seperti IEEE 802.11a/b/g/n/ac/ax dll.
Dalam IEEE 802.11ax yang diMT(JtoE)laskan dalam PTL 1 mencakup standar yang memungkinkan peningkatan kecepatan komunikasi dalam kondisi kemacetan selain untuk mencapai throughput puncak tinggi yang distandarisasi.	Dalam IEEE 802.11ax yang diMT(JtoE)laskan dalam PTL 1 mencakup standar yang memungkinkan peningkatan kecepatan komunikasi dalam kondisi kemacetan selain untuk mencapai throughput puncak tinggi yang distandarisasi.

評価対象文で dll.が末尾に来る文は上例と同③の明 26 の 2 文のみであったが、双方とも未分割が発生していた（後者は Stanza でも発生）。このことから、dll.は文分割の対象外となっていると推定される。

【文中の Co., Ltd.での過分割】

spaCy では、「Co., Ltd.」およびその類似語での過分割も目立った。以下、一例を示す。

［文献⑤ 明 46］

対象文	文分割結果
Kemudian, getaran benturan diterapkan ke sisi dalam dari porsi dinding-sisi (7b) dari bagian luar (7) dari spesimen uji (21) dengan palu benturan (yang dibuat oleh Ono Sokki Co., Ltd.: GK-3100), suatu gaya benturan dan akselerasi yang dihasilkan pada spesimen uji (21) dibawa ke alat analisis FFT (yang dibuat oleh Ono Sokki Co., Ltd.: CF-7200A), dan suatu fungsi respons frekuensi dihitung.	Kemudian, getaran benturan diterapkan ke sisi dalam dari porsi dinding-sisi (7b) dari bagian luar (7) dari spesimen uji (21) dengan palu benturan (yang dibuat oleh Ono Sokki Co.,
	Ltd.:
	GK-3100), suatu gaya benturan dan akselerasi yang dihasilkan pada spesimen uji (21) dibawa ke alat analisis FFT (yang dibuat oleh Ono Sokki Co.,
	Ltd.:
	CF-7200A), dan suatu fungsi respons frekuensi dihitung.

上例の対象文は文中に「Co., Ltd.:」が 2 回出現するが、どちらも「Co.,」の直後と「Ltd.:」の直後で過分割が発生している。どちらも、ピリオド直後ではなく、前者はカンマ、後者はコロンの直後で区切られており、単なる文中ピリオドでの過分割ではない。

spaCy では文献①明 49 の「CO.,／LTD.)／…」、文献④明 41 の「…Ltd.,／…」などで類似の事象が見られる（／は過分割箇所を示す）。だが、Co.や Ltd.の直後で文を分割すべきケースは皆無と思われ、どのような基準でこうした過分割が生じているのかは不明である。

＜Stanza＞

Stanza における特徴的な文分割不正のパターンを以下に示す。

【文末の略語＋ピリオドの未分割】

Stanza の文分割不正で特に多発していたのが、「文末が略語＋ピリオド」である場合にピリオド直後で文分割されず、次の文と連結されて「未分割」となるパターンである。一例を示す。

［文献③ 明 23］

対象文	文分割結果
Sebagai contoh, peralatan komunikasi 102 dan peralatan komunikasi 103 dapat membentuk link ketiga di pita 6 GHz selain link pertama 104 di pita 2,4 GHz dan link kedua 105 di pita 5 GHz.	Sebagai contoh, peralatan komunikasi 102 dan peralatan komunikasi 103 dapat membentuk link ketiga di pita 6 GHz selain link pertama 104 di pita 2,4 GHz dan link kedua 105 di pita 5 GHz. Alternatifnya, link dapat dibuat melalui sejumlah kanal berbeda yang termasuk dalam pita frekuensi yang sama.
Alternatifnya, link dapat dibuat melalui sejumlah kanal berbeda yang termasuk dalam pita frekuensi yang sama.	

文献③はこの未分割条件に合致する（つまり文末が略語である）文が多数含まれていたため、当文献の未分割のカウントは、pySBD が 1 文、spaCy が 2 文なのに対し、Stanza は 16 文と突出した。文献①も同様の状況であり、他の 2 ツールがともに 0 文であったのに対し、Stanza の未分割カウントは 12 文となった。

評価対象文のうち、Stanza で未分割となった文末の略語を列挙すると、文献①では「Sn.」「As.」「ECU.」「SnSb.」「Ni.」、文献②では「Es2.」「”Hs.”」、文献③では「GHz.」「TU.」「DVD.」「AP2.」「FILS.」「STA.」「S702.」「S703.」、文献⑤でも「Hz.」「MPa.」と非常に多岐にわたる。

【冒頭の数字＋ピリオドでの過分割】

spaCy において顕著であった「冒頭に数字＋ピリオドを伴う文がピリオドの直後で過分割される」不備（⇒P.36）は、Stanza においても散見された。ただし、spaCy のように条件に該当する文全てで発生するわけではなく、同一と思われる条件下でもランダム的な発生状況であった。例えば、案件②の明 24 では「Item10.」とその直後の本文が過分割されているが、直後の明 25 では「Item11.」と直後の本文は過分割されていない。文献①においても、明 22 では「1. Paduan patri」が spaCy と同様に「1.」の直後で過分割されているが、同じ

文献の明 46「4. Bentuk-awal Patri」では（spaCy と異なり）過分割は発生していない。

3.5.1.4. 見出し連結

見出し連結とは、文末にピリオド等の区切り記号を有さない「見出し語」と、その次の文とが連結してしまっているケースである。以下に一例を示す。

［文献① 明 12］

対象文	文分割結果
Masalah Yang Akan Diselesaikan Melalui Invensi Di sisi lain, di bagian sambungan patri, suatu fenomena yang disebut sebagai “pengelupasan” telah diamati dimana lapisan paduan terkelupas dari antarmuka dengan perlakuan penuaan.	Masalah Yang Akan Diselesaikan Melalui Invensi Di sisi lain, di bagian sambungan patri, suatu fenomena yang disebut sebagai “pengelupasan” telah diamati dimana lapisan paduan terkelupas dari antarmuka dengan perlakuan penuaan.

特許文献から取得したテキストデータは、インドネシア語、日本語とも、項目（「発明の名称」「要約」「明細書」「特許請求の範囲」）単位で全文を連結している。特にインドネシア公報は、PDF データから OCR でテキスト抽出を行うため、抽出した時点では文が各行末で寸断された状態となっている。文分割処理に供するにあたっては、まずはこれらを全て連結する必要がある。

通常の文であれば、文末にピリオド等の区切り記号があるため、全テキストを連結しても支障なく文分割処理が行える。ただし、文末に区切り記号を持たない見出し語の場合、一度連結されてしまうと、次文との区切りがわからず、機械的に再分割することはほぼ不可能となる。このため、文分割結果においては、上例のように見出し語と直後の本文が連結された「見出し連結」が多発することとなる。

これら「見出し連結」は、状況としては「未分割」に該当する。だが、本調査のテキスト取得方法においては不可避の事象であり、有効な対処策が存在しない。つまり文分割処理にとっては入力データの問題といえ、再分割できる可能性がほぼないため、文分割ツールの性能判定において通常の未分割と同等に扱うのは不当である。このため別カウントとし、ツールの正解率算出において「正解」としても「不備」としても取り扱えるようにした。

3.5.1.5. 結論（本調査に最適のインドネシア文分割手法）

3.5.1.3.項に示したとおり、評価対象とした 3 種の文分割手法（pySBD、spaCy、Stanza）には、それぞれ他では見られない特徴的な不備が見られた。その重大度を比較すると、まず Stanza で多発した「文末の略語＋ピリオドの未分割」は、対象となる略語が広範囲であるため発生頻度が高く、かつ、対訳コーパスが複数文：複数文になってしまう[25]デメリットも大きいため、最も重大な不備と見なすべきである。このため、正解率、準正解率は第 2 位ながら、本調査のツール候補としては最初に除外される。

pySBD と spaCy の比較では、それぞれ「カッコ補記での過分割（pySBD)」、「冒頭の数字＋ピリオドでの過分割（spaCy)」という、文アライメント処理でリカバリ（再連結）され得るタイプの不備と、「文末の数値の過分割と次文の冒頭への連結（pySBD)」、「文頭のカッコ直後で過分割され前文末尾に連結（spaCy)」という、文アライメントでのリカバリがほぼ不可能な不備とが発生していた。

このうち後者、すなわち文アライメントでの正常化が不可能な不備の比較では、文頭のカッコが全文末尾に連結されるのみで文意理解にほぼ影響のない spaCy のカッコ過分割はごく軽微な不備といえる。これに比べると、pySBD の文末数字の過分割は文意理解への悪影響をもたらす。だが、spaCy には dll.での未分割や Co.,Ltd.の過分割など上記以外にも特定の条件下で確実に生じる不備が多く見られ、結果、正解率及び準正解率も pySBD の 78.4%／89.4%に比べて 74.1%／80.1%とかなりの差が開いた。これらの状況から総合的に判断して、本調査で用いるインドネシア文分割手法には pySBD を採用するのが最善であると結論される。

[25] 未分割によりインドネシア文が複数文となると、それと対応させるべく文アライメント処理で複数の日本文が対応付けられる可能性が高くなる。

3.5.2. 日本文分割手法候補 3 種の評価結果

インドネシア文分割手法と同様、日本文についても 3 種の文分割手法候補に対して評価を実施した。手法候補は、予備実験の結果を踏まえ下記の 3 手法とした。

① GiNZA
② pySBD
③ Stanza

3.5.2.1. 評価方法

各手法候補の評価は、インドネシア文分割手法の評価で評価対象とした 5 文献とそれぞれファミリー関係にある日本文献 5 文献（下表）を対象とした。

表 3-8　評価用日本文献（5 文献）

	日本文献	IPC	（参考）対応インドネシア文献
①	2022-002855	B23K	P00202203385
②	2022-024515	A01H	P00202301594
③	2022-073428	H04W	P00202304729
④	2022-012848	C08F	P00202300918
⑤	2022-034129	B60R	P00202301067

インドネシア文献と同様、各文献の全文テキストに対し、3 種の手法候補を用いて文分割処理を行ったうえで、1 文献あたり 60 文ずつを評価対象文として選定し、それらが過不足なく文分割されているかを評価した。

評価対象文は必ずしもインドネシア文分割手法の評価で選定した文と対訳関係にある文とはせず、日本文として文分割の難易度が高そうなものを意図的に選定した。ただし、インドネシア文で文分割の難易度が高そうな文は、対応する日本文も同様の特徴を有することが多く、結果的には双方が選定されるケースが多かった。また、インドネシアで特殊な事例として追加した「ピリオドを伴う微生物名を含む 1 文」に対応する日本文も評価対象文とし、評価対象文数を 301 文に揃えた。

評価方法はインドネシア文分割手法と同様とし、各手法による各評価対象文の文分割結果に対して、「正常」、「過分割」、「未分割」、「見出し連結」のいずれに該当するかを人手で評価した。

3.5.2.2. 評価結果

以下、3 種の文分割手法候補それぞれの人手評価結果を示す。

［① GiNZA］…正解率：76.4%（準正解率：80.4%）

文献	正常	見出し連結	過分割	未分割	合計
①	47	5	8	0	60
②	39	1	20	0	60
③	57	0	2	1	60
④	36	4	20	0	60
⑤	50	2	8	0	60
微生物	1	0	0	0	1
合計	230	12	58	1	301

［②pySBD］…正解率：71.8%（準正解率：81.1%）

文献	正常	見出し連結	過分割	未分割	合計
①	34	8	18	0	60
②	46	2	10	2	60
③	51	0	9	0	60
④	35	13	12	0	60
⑤	49	5	6	0	60
微生物	1	0	0	0	1
合計	216	28	55	2	301

［③ Stanza］…正解率：83.1%（準正解率：95.7%）

文献	正常	見出し連結	過分割	未分割	合計
①	45	12	3	0	60
②	50	3	4	3	60
③	60	0	0	0	60
④	41	17	2	0	60
⑤	53	6	1	0	60
微生物	1	0	0	0	1
合計	250	38	10	3	301

各手法の正解率の比較では、Stanza が正解率 83.1%（301 文中 250 文）で最上位となった。以下、GiNZA が正解率 76.4%（230 文）、pySBD が 71.8%（216 文）と続く。インド

ネシア文分割手法に比べ、各手法の差が大きく開いた。

「見出し連結」をカウントに含めた準正解率の比較でも Stanza が 95.7%で最上位となり、他の 2 手法（GiNZA：80.4%、pySBD：81.1%）とはさらに大差となった。

3.5.2.3. 各ツールの特徴

日本文分割手法においても、不備を生じる特徴的な条件がそれぞれの手法ごとに存在した。以下、その内容を示す。

＜GiNZA＞

前項で示したとおり、日本文分割手法候補 3 種の評価結果は、準正解率において上位 1 種と下位 2 種とに明確に分かれた。GiNZA は下位グループに属したが、これは不可解な基準の過分割が多発したことが大きな理由である。GiNZA は正解率では第 2 位であるものの、過分割の発生頻度は 19.3%（301 文中 58 文）と最多であった。本項で状況を説明する。

【化学物質名の過分割】

評価対象文献①は化学物質名を列挙する文を多く含んでおり、評価対象文にも 4 文が選ばれた（明 39、40、41、43）。GiNZA では、このうち明 41 を除く 3 文において、多重的な過分割が発生していた。一例を示す。

［文献① 明 43］

対象文	文分割結果
アルコール系溶剤としてはイソプロピルアルコール、1，2－ブタンジオール、イソボルニルシクロヘキサノール、2，4－ジエチル－1，5－ペンタンジオール、2，2－ジメチル－1，3－プロパンジオール、2，5－ジメチル－2，5－ヘキサンジオール、2，5－ジメチル－3－ヘキシン－2，5－ジオール、2，3－ジメチル－2，3－ブタンジオール、1，1，1－トリス（ヒドロキシメチル）エタン、2－エチル－2－ヒドロキシメチル－1，3－プロパンジオール、2，2′－オキシビス（メチレン）ビス（2－エチル－1，3－プロパンジ	アルコール系溶剤としてはイソプロピルアルコール、1，2－ブタンジオール、イソボルニルシクロヘキサノール、2，4－ジエチル－1，5－ペンタンジオール、2，2－ジメチル－1，3－プロパンジオール、2，5－ジメチル－2，5－ヘキサンジオール、2，5－ジメチル－3－ヘキシン
	－2，5－ジオール、2，3－ジメチル－2，3－ブタンジオール、1，1，1－トリス（ヒドロキシメチル）エタン、2－
	エチル－2－
	ヒドロキシメチル
	－1，3－プロパンジオール、2，2′－
	オキシビス（メチレン）ビス（2－エチル－

オール)、2，2－ビス（ヒドロキシメチル）－1，3－プロパンジオール、1，2，6－トリヒドロキシヘキサン、ビス［2，2，2－トリス（ヒドロキシメチル）エチル］エーテル、1－エチニル－1－シクロヘキサノール、1，4－シクロヘキサンジオール、1，4－シクロヘキサンジメタノール、エリトリトール、トレイトール、グアヤコールグリセロールエーテル、3，6－ジメチル－4－オクチン－3，6－ジオール、2，4，7，9－テトラメチル－5－デシン－4，7－ジオール等が挙げられる。	1，3－プロパンジオール)、2，2－ビス（ヒドロキシメチル）－1，3－プロパンジオール、1，2，6－トリヒドロキシヘキサン、ビス［2，2，2－トリス（ヒドロキシメチル）エチル］
	エーテル、1－エチニル－
	1－
	シクロヘキサノール、1，4－シクロヘキサンジオール、1，4－シクロヘキサンジメタノール、エリトリトール、トレイトール、グアヤコールグリセロールエーテル、3，6－ジメチル－4－
	オクチン－3，6－ジオール、2，4，7，9－テトラメチル－5－
	デシン－4，7－ジオール等が挙げられる。

上例では 1 文が 11 文に過分割されている。その内容を見ると、ハイフォンの直前・直後での過分割が大半である。だが、文中の全てのハイフォンが対象となっているわけではなく、かつ分割点がハイフォンの直前の場合と直後の場合が混在するなど、分割基準が判然としない。この文献では評価対象文 4 文のほかにも同じように化学物質名を列挙する文が多数存在するが、ほぼ全ての文で同様の（ハイフォン周辺での）過分割が発生していた。

【特定の文言での過分割】

GiNZA の文分割結果を見ると、理由は不明ながら、過分割を起こしやすい特定の文言が存在するようである。例えば文献②では、「配列」の直前で過分割される文が明 39、40、41、58 と 4 文存在した。また同じ文献では「ヒマワリ」の直後での過分割も要 02、明 32、43、47 とやはり 4 文存在する。ただし、「配列」や「ヒマワリ」でも過分割されない場合もあり（数としては過分割されないほうが多い）、分割の発生条件が不明である。類似の事象は、文献⑤の「ハット断面部材」などでも見られており、GiNZA には無用な過分割が発生する傾向が他のツールよりも強いと結論される。基準が不明瞭であるため対策も立てにくい。

＜pySBD＞

pySBD の正解率は 3 手法の最下位であった。GiNZA と同じく過分割が多発したためであるが、分割基準が不明確であった GiNZA と異なり、pySBD の過分割には「ピリオド直後での過分割」という明確な傾向が見られた。

【ピリオド直後での過分割】

pySBD はピリオドを無条件で文分割記号として扱っていると見られる。このため、全ての小数点で過分割が発生することとなった。以下、一例を示す。

［文献① 明 18］

対象文	文分割結果
本発明のようにＡｇ：３．１～４．０質量%、Ｃｕ：０．６～０．８質量%、Ｂｉ：１．５～５．５質量%、Ｓｂ：１．０～６．０質量%、Ｃｏ：０．００１～０．０３０質量%、Ｆｅ：０．０２～０．０５質量%、残部Ｓｎからなる、はんだ合金を採用することで、高いヒートサイクル特性を維持しつつ、継手再溶融時や高温負荷時等に化合物の合金中への遊離を抑制等することによって音響品質に悪影響が出ることを防止したはんだ合金等を得ることができる。	本発明のようにＡｇ：３． １～４． ０質量%、Ｃｕ：０． ６～０． ８質量%、Ｂｉ：１． ５～５． ５質量%、Ｓｂ：１． ０～６． ０質量%、Ｃｏ：０． ００１～０． ０３０質量%、Ｆｅ：０． ０２～０． ０５質量%、残部Ｓｎからなる、はんだ合金を採用することで、高いヒートサイクル特性を維持しつつ、継手再溶融時や高温負荷時等に化合物の合金中への遊離を抑制等することによって音響品質に悪影響が出ることを防止したはんだ合金等を得ることができる。

ピリオドは小数点以外にも見出し語の項番や英数字記号など出現頻度が高い。これらがことごとく過分割されることで、pySBD における過分割の発生率は 18.3%（301 文中 55 文）と、GiNZA とほぼ同等の高頻度となった。

＜Stanza＞

Stanza の正解率は 83.1%で 3 手法の首位であった。過分割の発生率も 3.3%と他の 2 手法より顕著に低い。なお、過分割の過半数は、GiNZA や pySBD では見られなかった「カッコ補記内の句点での過分割」であった。

【カッコ補記内の句点での過分割】

カッコ補記内の句点は分割すべきでないが、Stanza ではこれに該当する全ての評価対象文（文献①明 09、11、文献②明 33、46、51、文献④明 20）で過分割が発生していた。以下一例を示す。

［文献② 明 09］

対象文	文分割結果
自動車には、プリント基板に電子部品をはんだ付けした電子回路（以下、車載電子回路という。）が搭載されている。	自動車には、プリント基板に電子部品をはんだ付けした電子回路（以下、車載電子回路という。
	）が搭載されている。

【その他の過分割】

Stanza では上記以外にも、文末周辺のカッコ補記で冒頭の「(」の直後で過分割される不備が 3 件（文献①明 60、④明 53、⑤明 20）と文中コロン直後で過分割される不備が 1 件（文献②明 59）と、少量の過分割が発生している。ただし、いずれも同条件の他の文では不備は発生しておらず、突発的な発生と見られる。

3.5.2.4. 結論（本調査に最適の日本文分割手法）

人手評価の結果、日本文分割手法においては、各手法とも不備の大半は「過分割」であり、「未分割」はごくまれにしか発生しなかった。インドネシア文の場合、文分割記号であるピリオドは文末以外にも多用されるが、日本文の文分割記号である句点は文末以外で使用されることがないため、文分割の難易度は相対的に低いと考えられる。その結果、日本語文分割手法は、最上位 Stanza の正解率が 83.1%、準正解率が 95.7%と、インドネシア文分割手法の最上位 pySBD の 78.4%／89.4%よりも顕著に高くなった。

こうした状況につき、日本語文分割ツールの優劣は、不要な過分割の多寡が事実上の焦点となった。下位グループとなった GiNZA と pySBD は、ハイフォンやピリオド、特定の文言などによって文分割を行う設定がなされていると見受けられるが、それが奏功した（つまり他の手法で不備となった文で正解した）ケースは見当たらず、全て過分割となってしまっていた。その結果、正解率、準正解率とも首位の Stanza とは大きく差が開いた。

Stanza は、他の 2 手法では見られなかった「カッコ補記内の句点での過分割」が該当文すべてで発生するという課題は見られたものの、他の手法で正解率を低下させた「句点以外での過分割」はごく少量にとどまっており、ほぼ「句点での分割」のみを行う設定と見られる。そして Stanza の正解率を見る限り、日本語文に対しては特殊な文分割設定をせず、シンプルに句点での分割のみを行うことが良好な結果をもたらすと結論される。本調査においても、日本語文分割ツールは Stanza を使用するのが最善である。

４．日インドネシア語の文アライメントに関する調査と最適な手法の決定

4.1. 調査の概要

対訳コーパスを作成するためには、文分割処理の次工程として、文単位に分割された日本文とインドネシア文を内容に応じて対応づける「文アライメント処理」が必要である。本調査では、日インドネシア間の文アライメントを自動的に実施する手法についても比較検討を実施した。

まずは公開情報に基づき、各手法の概要、想定される精度、課題、コスト、文アライメントスコア（文アライメントの確からしさを示す情報）の定義、プログラミング言語、ライセンス条件、他言語（例えば、ベトナム語、タイ語）への適用可能性、などについて調査した。

調査は、文分割手法の調査と同様、日本語での情報に加えて、海外（英語、インドネシア語）の情報も調査範囲とし、具体的な調査対象として、arXiv.org7 を含む論文誌（プレプリントを含む）及び学会講演要旨集、Github、Hugging Face、ニュースリリース、及び、Google 等の検索エンジンの検索結果を調査した。

そして、調査で取得した各手法の想定精度や課題、さらには予備実験の結果等に基づき、本調査で行う文アライメントの候補手法として 3 手法を選定し、各手法候補を用いて評価用文献に対する文アライメント処理を実施し、その結果に基づき、本調査で採用する文アライメント手法を選定した。

4.2. 文アライメント手法の方式

調査の結果、文アライメントを実現する方式として以下の 4 種が存在することがわかった。

表 4-1　文アライメントを実現する 4 種の方式

方式	概要
ベクトル化方式	インドネシア語の文、日本語の文をそれぞれベクトル表現に変換し、双方のベクトルデータの類似度を計算して、その類似度に基づきアライメントデータを作成する方法。
英語への機械翻訳方式	インドネシア語の文又は日本語の文を英語に機械翻訳し、日本語－英語の文アライメントツールまたはインドネシア語－英語の文アライメントツールを使ってアライメントデータを作成する方法。

汎用方式	入力する文の言語が限定されていないアライメントツールを使ってアライメントデータを作成する方法。
辞書方式	辞書を用いたアライメントツールを使ってアライメントデータを作成する方法

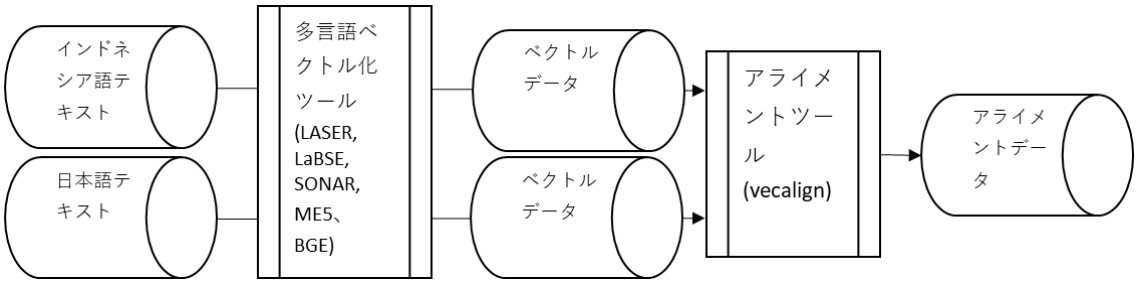
以下、各方式の調査結果について詳述する。

4.2.1. ベクトル化方式

4.2.1.1. ベクトル化方式とは

ベクトル化方式は、下図に示すように、入力文をベクトル化する多言語ベクトル化ツールと、ベクトル化されたデータを入力しアライメントデータを作成するツールとで構成される。

図 4-1　ベクトル化方式（概要図）



多言語ベクトル化ツールのうち日本語、インドネシア語の双方をサポートするものとして LASER、LaBSE、SONAR、ME5、BGE の 5 種類が存在する。一方、アライメントデータを作成する手法としては、vecalign と Bertalign が存在するが、Bertalign は日本語、インドネシア語に対応しておらず、本調査の用途には不適であった。

4.2.1.2. 多言語ベクトル化ツール

多言語ベクトル化ツールは、多言語で記載されたデータを学習してモデルを作成し、作成したモデルを使って対象とする文のベクトル化を行う。以下に、各ツールの学習で使用されたインドネシア語と日本語のデータを示す。また、各ツールの精度を参考情報として示す。

表 4-2　多言語ベクトル化ツールの学習データ量及び学習精度

ツール	学習データ量	精度
LASER	Europarl, United NationsA、OpenSubtitles2018, Global Voices、Tanzil、および Tatoeba のコーパスを組み合わせたデータを使用。 インドネシア語: 4.3M 文 日本語: 3.2M 文	文類似エラー率 インドネシア語→英語: 5.80 日本語→英語: 5.40[26]
LaBSE	CommonCrawl data、Wikipediadata を使用。 インドネシア語: 250M 文 日本語: 1,400M 文 (LaBSE の論文図 7)	日本語、インドネシア語に関する精度の記載は見られなかったが、LASER との比較で LaBSE の精度が高いとの記載あり[27]。
SONAR	NLLB data を使用。 インドネシア語 417M 文 日本語 269M 文	文類似エラー率(98 言語)にて SONAR は LASER、LaBSE よりエラー率が低いと報告されている[28]。 SONAR エラー率：0.1 LASER3 エラー率：1.1 LaBSE エラー率：1.5

[26] Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zeroshot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguistics, 7:597–610.

[27] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

[28] Paul-Ambroise Duquenne and Holger Schwenk and Benoit Sagot, 2023, Sentence-Level Multimodal and Language-Agnostic Representations, arXiv

ME5	学習データのうち日本語とインドネシア語のデータが含まれているデータは以下のとおり。 (1) xP3 data インドネシア語 4.6G バイト。 (2) NLLB data 日本語 269M 文 インドネシア語 417M 文	英語の埋め込みモデルにおいて ME5 は、LaBSE や BGE より精度が高いと報告されている[29]。
BGE	学習データのうち日本語とインドネシア語が含まれているデータは以下のとおり。 (1) xP3 data インドネシア語 4.6G バイト。 (2) NLLB data 日本語 269M 文 インドネシア語 417M 文 (3) CCMatrix data 日本語とインドネシアとの対訳データ 770 万文	埋め込みモデルを使用した多言語検索の評価でインドネシア語、日本語ともに ME5 より精度が高いことが報告されている[30]。

4.2.1.3. アライメントツール vecalign

ベクトル化方式におけるアライメントツールは、多言語ベクトル化ツールが出力したインドネシア文のベクトル化データ及び日本文のベクトル化データに対し、ベクトルの類似度計算を行い、類似度に基づいて両者を対応づけてアライメントデータを作成する役割を果たす。

前項に示した多言語アライメントツールと組み合わせて使用できるアライメントツールとして vecalign が存在した。vecalign は、インドネシア文のベクトルと日本文のベクトルのコサイン類似度にインドネシア文、日本文の文数を加味したスコア関数に基づいて類似度を計算している。文のアライメントには Fast Dynamic Time Warping と呼ばれるアルゴリズムを使用している。

[29] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report, arXiv

[30] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian and Zheng Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, arXiv

4.2.1.4. ベクトル化方式の課題

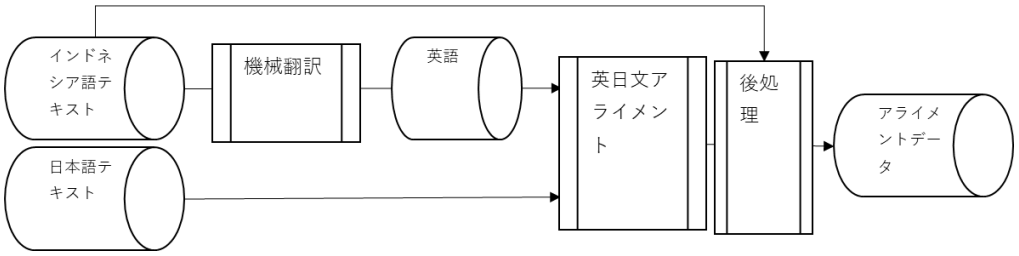
ベクトル化方式の課題として、ベクトル化の処理の負荷が大きく、他の方式に比べて大きなリソースを要することが挙げられる。特に、文のオーバーラップ（複数の文を連結してアライメント候補に加える）の許容数を増加させるほど負荷は大幅に増大する。

4.2.2. 英語への機械翻訳方式

4.2.2.1. 英語への機械翻訳方式とは

英語への機械翻訳方式は、インドネシア語または日本語を英語に機械翻訳し、日本語－英語アライメントツールまたはインドネシア語－英語アライメントツールと組み合わせて日インドネシア語のアライメントデータを作成する方式である。下図に、インドネシア語を英語に機械翻訳する場合の処理フローを示す。

図 4-2　インドネシア語から英語への機械翻訳方式（概要図）



上図に示した方式では、まずインドネシア語テキストを英語に機械翻訳し、機械翻訳した英語テキストと日本語テキストで英日アライメント処理を行う。その後、英日のアライメント結果に対し、英語テキストを翻訳前のインドネシア語テキストに置き換える後処理を行うことで、日本語とインドネシア語のアライメントデータが作成されることになる。

4.2.2.2. 文アライメントツール

本方式で使用するアライメント処理は、日本語－英語のアライメントツールまたはインドネシア語－英語のアライメントツールとなる。いずれも、ベクトル化方式と同様、vecalignの使用が考えられる。

なお、特許庁「平成２８年度ベトナム・タイ語の対訳コーパス・辞書の自動作成に向けたツール等の検証調査」では、日本語－英語を対象としたアライメントツールとして国立研究開発法人情報通信研究機構（NICT）が作成したアライメントツールの利用が報告されている。同ツールはその内部で辞書を使用しアライメントを行う辞書方式である。インドネシア語を英語に機械翻訳して日本語－英語のアライメントを行う場合、このツールを利用することも可能であるが、有償につき別途 NICT との契約が必要となる。

4.2.2.3. 英語への機械翻訳方式の課題

大量の日本語又はインドネシア語データの機械翻訳処理を行うため、一定の処理時間を要し、かつ機械翻訳の利用コストが発生する。

4.2.3. 汎用方式

4.2.3.1. 汎用方式とは

汎用方式は、入力する文の言語が限定されていないアライメントツールを使ってアライメントデータを作成する方法である。汎用方式は、①文の単語数などの統計情報のみに基づいてアライメント処理を行うもの(hunalign, Gargantua, Microsoft Bilingual Sentence Aligner(Robert C Moore. 2002))と、②インドネシア語または日本語を他方の言語に機械翻訳し、同一言語の単語の一致率に基づいてアライメントを行うもの(GSK2017-A 文アライメントツール（ealign)、Bleualign)とがある。

図 4-3　①汎用方式（概要図）

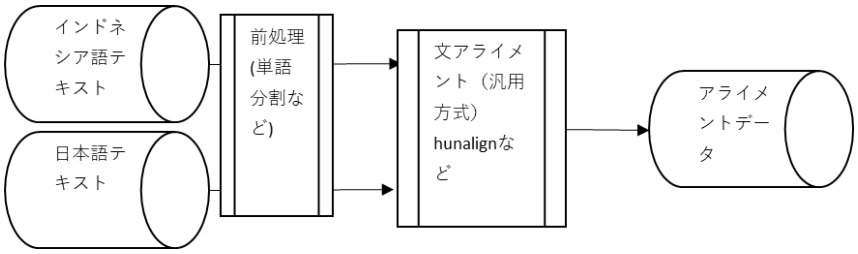
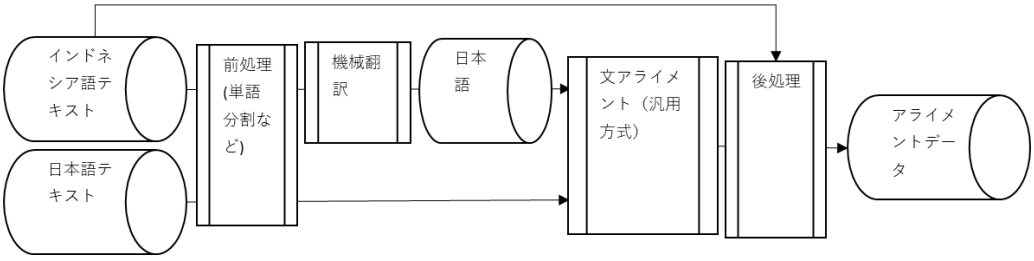


図 4-4　②機械翻訳＋汎用方式（概要図）



4.2.3.2. 汎用方式の課題

機械翻訳を用いる場合（上記②）は、英語への機械翻訳方式と同様、機械翻訳処理のための一定の処理時間と利用コストが発生する。

4.2.4. 辞書方式

4.2.4.1. 辞書方式とは

辞書方式は、インドネシア語と日本語との対訳辞書を用いてアライメントを行う方法である。この方式のアライメントツールとしては Champollion と hunalign が存在した。また、インドネシア語と日本語との対訳辞書として、GSK2006-A-5 CICC 専門語辞書[31]（有償）が存在した。

4.2.4.2. 辞書方式の課題

辞書方式にはインドネシア語と日本語の対訳辞書が必須となるが、上記「GSK2006-A-5 CICC 専門語辞書」はコンピュータ、電気関係の専門用語のみを収録した 2.7 万語の小規模な辞書であり、全技術分野をカバーする必要のある特許文献の対訳コーパス作成においては不足である。これ以外の日インドネシア対訳辞書は現状では入手困難と見られ、大規模な対訳辞書の調達が大きな課題となる。

[31] https://www.gsk.or.jp/catalog/gsk2006-a-5/

4.3. 日インドネシア語の文アライメント手法一覧

本調査で検出したインドネシア語及び／又は日本語の文アライメント手法、全 16 種の概要を一覧表にまとめる。

表 4-3　日インドネシア語に使用可能な文アライメント手法一覧

No.	手法名	方式	機能	サポート言語		
				イン	日	英
1	LASER	ベクトル化	ベクトルデータ作成	○	○	○
2	LaBSE	ベクトル化	ベクトルデータ作成	○	○	○
3	SONAR	ベクトル化	ベクトルデータ作成	○	○	○
4	Multilingual E5 Text Embeddings	ベクトル化	ベクトルデータ作成	○	○	○
5	BGE M3-Embedding	ベクトル化	ベクトルデータ作成	○	○	○
6	vecalign	ベクトル化	アライメント	○	○	○
7	BlueAlign	汎用	アライメント	○	○	○
8	Bertalign	ベクトル化	アライメント	×	×	×
9	Champollion	対訳辞書	アライメント	○	○	○
10	hunalign	汎用／対訳辞書	アライメント	○	○	○
11	Gargantua	汎用	アライメント	○	○	○
12	GSK2017-A 文アライメントツール（ealign）	汎用	アライメント	○	○	○
13	NICT 文アライメント	対訳辞書	アライメント	×	×	○
14	GSK2006-A-5 CICC 専門語辞書	対訳辞書	ベクトルデータ作成	○	○	○
15	Microsoft Bilingual Sentence Aligner（Robert C Moore. 2002）	汎用	アライメント	○	○	○
16	lrec2016sentalign	対訳辞書	アライメント	○	○	○

4.4. 予備実験の実施

文アライメント手法に関しても、公開情報に基づく調査で得た情報のみでは、本調査で実施する日インドネシア特許文献の対訳コーパス作成における相対的な優劣を判断することは困難であった。このため、候補手法を絞り込むための文アライメント精度等の情報を得るべく、文分割手法と同様にサンプル文献を用いた予備実験を実施した。本項でその詳細を示す。

4.4.1. サンプル文献データ

文分割手法の予備実験で使用したものと同一のファミリー特許文献（インドネシア特許公開公報 P00201907657 と日本の特許公開公報 JP2020037938A）の全文を用いた。それぞれから抽出した全文テキストデータに対し、第 3 章で述べたとおり本調査用として選定されたインドネシア文分割ツール pySBD と日本文分割ツール Stanza を用いて文分割済みデータを作成し、これに対して実験対象の各アライメントツールで文アライメントを実施した結果を評価対象とした。

4.4.2. 実験対象とした日インドネシア文アライメント手法

予備実験の対象には、4.2.項に示した 4 種類の文アライメント方式それぞれに属する代表的な手法（ツール）を選定した。次ページの表 4-4 に一覧を示す。

ベクトル化方式としては、公開情報調査における学習データ量や精度情報に基づき、3 種のベクトル化ツール（LASER、LaBSE、BGE）とアライメント作成ツール vecalign とを組み合わせた各手法（表中①、②、③）を実験対象とした。

英語への機械翻訳方式としては、インドネシア語を Google 翻訳で英語に機械翻訳したデータと英日アライメントツールとを組み合わせた方式（同④）と、日本語を Japio-AI 翻訳[32]で英語に機械翻訳したデータと英インドネシアアライメントツールとを組み合わせた方式（同⑤）を調査対象とした。

汎用方式としては、候補中で比較的新しい Microsoft Bilingual Sentence Aligner(Robert C Moore. 2002)を調査対象とした（同⑥）。機械翻訳と汎用方式を組み合わせて使用する方式としては、Bluealign（同⑦）と GSK2017-A 文アライメントツール(e-align)を調査対象とした（同⑧）。

なお、辞書方式に関しては、同方式の課題として述べたとおり（⇒4.2.4.2.）、事実上の必

[32] https://japio.or.jp/service/service06.html

須要件となる大規模な日インドネシア対訳辞書が入手困難であり、実験対象に加えることができなかった。

表 4-4　予備実験の対象とした文アライメント手法（全 8 種）

ツール略称	方式	概要
① LASER	ベクトル化方式	ベクトル化データ作成: LASER（日本語、インドネシア語） ベクトルデータのアライメント: vecalign
② LaBSE	ベクトル化方式	ベクトル化データ作成: LaBSE（日本語、インドネシア語） ベクトルデータのアライメント: vecalign
③ BGE	ベクトル化方式	ベクトル化データ作成: BGE（日本語、インドネシア語） ベクトルデータのアライメント: vecalign
④ MT（ItoE）	英語への 機械翻訳方式	機械翻訳：Google 翻訳（インドネシア語→英語） ベクトル化データ作成:LASER（英語、インドネシア語） ベクトルデータのアライメント: vecalign
⑤ MT（JtoE）	英語への 機械翻訳方式	機械翻訳：Japio-AI 翻訳（日本語→英語） ベクトル化データ作成:LASER（英語、インドネシア語） ベクトルデータのアライメント: vecalign
⑥ Moore	汎用方式	アライメント：Microsoft Bilingual Sentence Aligner （Robert C Moore. 2002）
⑦ Bleualign	機械翻訳 ＋汎用方式	インドネシア日機械翻訳(Google 翻訳) アライメント：Bleualign
⑧ealign	機械翻訳 ＋汎用方式	インドネシア日機械翻訳(Google 翻訳) アライメント：GSK2017-A 文アライメントツール （e-align）

上記ツールのうち⑥Moore は、文アライメント結果として出力される日本文が半角スペースが細かく区切られる。また⑧ealign は、文対が複数文である場合、文の間に「///」が自動的にセットされる。本実験は、これらを機械的に除去したうえで評価を実施した。

4.4.3. 日インドネシア文アライメント手法の実験結果

予備実験では、対象に選定した全８種類の日インドネシア文アライメント手法について、サンプル文献に対する文アライメントの精度を比較した。あわせて、処理に要した時間を測定し、その処理効率を調査した。本項にその結果を示す。

4.4.3.1. 文アライメント精度の評価方法

サンプル文献の全文を対象に、あるべき文アライメント結果（正解データ）を人手で作成し、文対ごとに、文アライメント結果がこれと過不足なく合致していれば「正解」、何らかの誤りがあれば「誤り」と判定して、各手法の正解率を比較した。

なお、文分割手法の課題として挙げた「見出し連結」（⇒3.4.6.＜課題１＞）については、テキストデータ抽出時に不可避的に発生する事象であり、文分割処理でのリカバリも事実上不可能である。このため、文アライメントツールの入力データである文分割済みデータの時点で既に発生している。つまり、文アライメント処理の問題で生じたものではないため、文アライメントの評価では「誤り」とせず、「見出し連結」として別カウントとした。ただし、頻出の見出し語（例えば「Uraian Singkat Invensi／【発明の概要】」）は、文分割後の後処理で本文と分割を行っており、これらがアライメント時に再度本文と連結された場合は「誤り」にカウントした。

［注記］

予備実験用の正解データは、「あるべき文アライメント結果」、つまり、当該ファミリー文献の状況において最善のアライメント結果とした。このため、双方の文の記載内容に相違点があっても、前後関係などから両者をアライメントするのが妥当であれば、それを正解としている。以下、一例を示す。

［請 12］

Rakitan blok silinder (2) menurut salah satu dari klaim 1 sampai 6, yang dicirikan dengan masing-masing dari penutup engkol pusat dan penutup engkol samping memiliki alur, dan alur tersebut disediakan pada permukaan samping dari masing-masing penutup engkol pusat dan penutup engkol samping, permukaan samping tersebut ditempatkan di dalam arah penjajaran dari silinder (11).	前記除去部分は、前記クランクキャップの前記シリンダの整列方向に位置する側面に形成された溝を含む、請求項１～６のいずれか１項に記載のシリンダブロック組立体。

上例はサンプル文献の請求項 12 に該当する。この文の記載内容には少なからぬ相違があ

り、同一とはいえないが、前後の請求項の対応関係から、文アライメント処理上はこの対応づけが最善であるため、アライメントツールの評価が目的である予備実験では、これも正解とした。

4.4.3.2. 文アライメント難易度が高い事例

実験対象としたファミリー文献には、文アライメント処理の難易度が高い文対がいくつか存在した。以下、それらの状況を説明する。

＜不対応文＞

ファミリー文献においても、一方のみに新規に付け加えられた文など、他方の文献に対になるべき文が存在しない場合がある。実験対象文献においても、日本文献の要約１文目［要2］と３文目［要４］など、インドネシア文献に対応する文がない「不対応文」が存在した。

［要2～4］

要2	（不対応）	製造コストを抑制しつつクランクジャーナルクランクジャーナルとクランク軸受との間の摩擦損失を低減させる。
要3	Suatu rakitan blok silinder (2) mencakup blok silinder (10) yang memiliki silinder (11) dan sejumlah penutup engkol (20) yang dipasang tetap ke blok silinder (10).	シリンダブロック組立体２は、一列に並んだ偶数個のシリンダ１１を有するシリンダブロック１０と、シリンダの整列方向に一列に並んでシリンダブロックに固定される複数のクランクキャップ２０と、を備える。
要4	（不対応）	各クランクキャップ及びシリンダブロックにはクランクシャフト３を回転可能に支持するクランク軸受が設けられる。

文アライメント処理としては、要 2、要 4 はインドネシア文と対応づけず、要 3 のみを１：１で対応させるのが正解であるが、要２や要４が要３（や要５以降）と連結されて１：N の文対とさせてしまう誤りが想定された。

上例の場合、１文対をはさんで前後に不対応文が存在するという非常に難易度の高いケースであったため、2 つの不対応文とも適切に処理できたツールは存在しなかったが、②LaBSE、③BGE、⑦Bleualign、⑧ealign の 4 ツールは、要 2 に対しては不対応文として正

しく処理できていた。

＜N：１対応＞

ファミリー文献では、一方が１文で記述した内容を、他方では複数文で記述することがある。実験対象文献にも、例えば［明 99～明 100］など、日本文が１文で述べた内容をインドネシア文は 2～3 文で述べているケースが数件見られた。

［要 2～4］

明 99	Permukaan bawah dari blok silinder (10) dilengkapi dengan sejumlah cerukan setengah lingkaran (12), dan bantalan engkol (13) disediakan pada setiap cerukan (12).	シリンダブロック１０の下面には、複数の半円状の凹部１２が形成され、この凹部１２にクランクシャフト３を回転可能に支持するためのクランク軸受１３が設けられる。 ※青字部分が２文目に相当。
明 100	Bantalan engkol (13) menopang poros engkol (3), sedemikian sehingga poros engkol dapat diputar.	

このような文は１：１で対応している文よりもアライメントの難易度は高くなるが、大半のツールは良好な成功率で対応づけに成功していた。具体的には、2:1～3:1 対応の 7 文[33]中、②LaBSE が全 7 文、③BGE、④MT(ItoE)、⑧ealign が 6 文を正解しており、以降も①LASER が 4 文、⑤MT（JtoE）が 3 文、⑦Bleualign が 2 文成功した。唯一、⑥Moore のみは文の消失が発生しており（後述）、正解は 0 文であった。

一方、日本文が複数となる１：N 文は１文［明 58～77］のみ存在した。インドネシア文がセミコロンで接続された長文で、これが日本文 20 文と対応する形であったが、文が長大すぎるためか、各ツールとも文対応が大幅に乱れた。各ツールにおけるオーバーラップ文数（⇒4.2.1.4.）の上限値に抵触したものと考えられる。

＜見出し語の不一致＞

特許文献では、インドネシア公報、日本公報とも、見出し語が多用される。そして、インドネシア文献の見出し語は文分割されず本文と連結されることが避けられない。これを改善するため、頻出の見出し語については、文分割後に強制的に本文と分割する後処理を施した（⇒3.4.6.＜課題１＞）。

[33] 明 38～9、52～3、108～9、119～20、123～4、180～1、206～7 の 7 文対。明 38～9 のみ 3:1 対応で、他は 2:1 対応。

この措置により、頻出の見出し語が本文と分割され、適切に対応づけられるケースが多く確認された。具体的には［明 28］の「Uraian Singkat Invensi／【発明の概要】」や［明 78］の「Uraian Lengkap Invensi／【発明を実施するための形態】」などが挙げられる。

その反面、実験対象文献においては、日本の見出し語に対応するインドネシア見出しが存在しないケースが目立った。具体的には、［明 29］「【発明が解決しようとする課題】」や［明 36］「【課題を解決するための手段】」、［明 55］「【発明の効果】」などが該当する。この場合、日本語見出しは 4.1 項で述べた「不対応文」の一種となる。

実際のアライメント結果も他の不対応文と同様、前後の文と誤って連結されてしまうことが多かった。例えば［明 29］［明 55］とも、正解したのは⑦Bleualign のみであった。［明 36］も、⑦Bleualign と⑧ealign のみが正解であり、正解率は全体的に低かった。

4.4.3.3. 各手法の正解率

評価対象データ 347 文における各手法の正解・誤りのカウント結果を下表に示す。

表 4-5　予備実験における文アライメント手法 8 種の正解率

	正解	見出し連結	誤り	消失	正解率	備考
① LASER	287	10	50	0	82.7%	
② LaBSE	294	10	43	0	84.7%	
③ BGE	295	10	42	0	85.0%	4 位
④ MT(ItoE)	296	10	41	0	85.3%	2 位
⑤ MT(JtoE)	282	10	55	0	81.3%	
⑥ Moore	250	8	7	82	72.0%	スコアなし
⑦ Bleualign	296	10	39	2	85.3%	2 位、スコアなし
⑧ ealign	299	10	38	0	86.2%	1 位

正解率は⑧ealign が最も優秀で 86.2%、次いで④MT（ItoE）と⑦Bleualign が 85.3%で同率 2 位、③BGE が 85.0%で第 4 位となった。以降も②LaBSE が 84.7%、①LASER が 82.7%、⑤MT（JtoE）が 81.3%と、7 ツールが 80%台となった。

唯一、⑥Moore のみは、対訳文が消失してしまう事象が 82 文と頻発したため、正解率は 72.0%にとどまった。

4.4.3.4. 正解スコア／誤りスコア分布

8 手法のうち、⑥Moore と⑦Bleualign を除く 6 手法は、アライメント結果に対して自らスコアを付している（以下、「文アライメントスコア」という）。下表に、各ツールの「正解」文および「誤り」文に付された文アライメントスコアの範囲と平均値を示す。

なお、6 手法中、⑧ealign のみはスコアの数値が大きいほど精度が良いことを示すが、他の 5 手法は数値が小さいほど精度が良いことを示す。

表 4-6　予備実験における文アライメント手法 6 種の文アライメントスコア比較

	正解文			誤り文			スコア設定
	最低値	最高値	平均値	最低値	最高値	平均値	
① LASER	1.2672	0.2043	0.4460	0.8139	0.3675	0.5785	大＜小
② LaBSE	0.9190	0.0980	0.3035	1.0561	0.2949	0.6322	
③ BGE	0.8983	0.2141	0.4996	0.9309	0.4980	0.7463	
④ MT(ItoE)	0.8809	0.2785	0.5413	0.9497	0.2060	0.5734	
⑤ MT(JtoE)	1.2414	0.1621	0.3835	0.8073	0.2987	0.5693	
⑥ Moore	－	－	－	－	－	－	－
⑦ Bleualign	－	－	－	－	－	－	－
⑧ ealign	0.2500	1.0000	0.6501	0.0513	0.5376	0.2711	小＜大

各ツールの正解文と誤り文の平均値を比較すると、スコア設定が「大＜小」である①～⑤はいずれも正解文のほうが誤り文よりもスコアが低く、「小＜大」である⑧は正解文のほうが誤り文より文アライメントスコアが高くなっている。このことから、各ツールの付すスコアには一定の信頼性があると見なせる。

本調査では、アライメントを行った各文対に対してその対応精度の優劣を機械的に特定する必要がある。このため、文アライメントスコアの付与とその精度はツール選定の重要な要素となる。上表の結果では、特に⑧ealign と②LaBSE は正解文と誤り文の平均値に倍以上の開きがあり、特に信頼性が高いとみられる。③BGE、⑤MT(JtoE)もこれに準ずる。

ただし、スコアの最高値、最低値の範囲を見ると、各ツールとも正解文と誤り文のスコア範囲にかなりのオーバーラップが見られる。例えば⑧ealign は正解文のスコア範囲が 0.2500～1.0000 なのに対し、誤り文の範囲は 0.0513～0.5316 であり、0.2500～0.5316 の範囲にオーバーラップが見られる。他のツールのオーバーラップ状況もこれと同等かそれ以上である。この結果からは、スコア範囲のみでアライメントの正否を区別することは難しいと判断される。

4.4.3.5.文長比の比較

予備実験で検出された重大な誤り文は、一方が無文であるものを除けば、その大部分は、一方の文献のみに存在する（つまり対応する文をもたない）新規文（例えば［要 2］や[要 4]）がその前後の文と連結されてしまい内容の相違をもたらすケースや、1：N や N：1 対応の文（例えば［明 38～39]）に対して 1 ： 1 の対応づけをしてしまい、一方に内容の不足をもたらすケースである。

こうした場合、内容に大きな過不足が生じることから、日インドネシア文対の通常の文長比を大きく逸脱すると考えられる。したがって、正解文と誤り文の文長比の傾向から、アライメントの正否を判定する基準を策定できる可能性がある。下表左側に、各ツールの正解文の文長比（インドネシア文の文字数÷日本文の文字数）を算出した。

各ツールとも正解文はおおむね同一であるため、正解文長比の最小値・最大値、平均値ともほぼ同じ数値となった。そこで、平均値を中央値として前後に約 1.25 ずつ広げた文長比 1.5 以上 4.0 未満の範囲で、全正解文数のどの程度をカバーできるかと、その範囲に混入してしまう誤り文数の数を調べた（下表右側)。

表 4-7　予備実験での正解文長比と文長比 1.5～4.0 での正解文カバー率／誤り文混入率

	正解文長比			文長比範囲 1.5～4.0 における			
	最小値	最大値	平均値	正解文数	カバー率	誤り文数	混入率
① LASER	1.109	4.870	2.777	276	96.2%	2	4.0%
② LaBSE	1.109	4.870	2.772	275	93.5%	2	4.7%
③ BGE	1.109	4.870	2.774	275	93.5%	2	4.8%
④ MT(ItoE)	1.109	4.870	2.768	279	94.3%	3	7.3%
⑤ MT(JtoE)	1.109	4.870	2.770	268	95.0%	4	7.3%
⑥ Moore	1.538	4.870	2.769	247	98.8%	5	71.4%
⑦ Bleualign	1.109	4.870	2.768	278	93.9%	8	20.5%
⑧ ealign	1.109	4.870	2.772	278	93.0%	1	2.6%

その結果、文長比範囲 1.5～4.0 の範囲設定により、各ツールとも正解文の 95%前後がカバーされ、かつ誤り文（見出し連結文と N:N 対応文は除く）の混入は⑥Moore と⑦Bleualign を除き 1～4 文とごく少量に抑えられることがわかった。

大半のツールで最大値（つまりインドネシア文が日本文に対して過度に長大）となった文長比 4.870 の文対は、冒頭で「内容は同一ではないがアライメントとしては正解である」とした［請 12］である。一方、最小値（日本文がインドネシア文に対して過度に長大）は［明

319］の 1.109 であるが、こちらも内容に下示のとおり顕著な相違を含む文対であった（赤字部分がインドネシア文に存在しない)。

［明 319］

Gambar 19 adalah tampak samping dari penutup pusat dan penutup samping.	図１９（A）は、中央キャップ及び側方キャップの側面図であり、図１９（B）は、図１９（A）の線B－Bに沿って見た断面平面図である。

これらの状況は、文長比が平均値から離れるほど、その文対の内容には相違があらわれる確率が高くなる傾向を示している。

予備実験後に行う文アライメント手法候補の人手評価（4.5.項で後述）では、まず評価対象文として「自動評価等で対応が取れていると判定された文対」を定める。上記予備実験の結果から、本調査では文長比範囲による判定(1.5 以上 4.0 未満を対応が取れているとする）を行うこととした。

4.4.3.6. スコア分布及び文長比の分析における補記

4.4.3.4.項に示した誤りスコア分布、ならびに 4.4.3.5.項に示した誤り文数のカウントでは、以下の文は対象外とした（正解としてもカウントしていない)。

① 一方が無文（対応文が存在しない）の文対
② N：N の文対
③ アライメント処理の失敗による見出し連結文対

本調査では、作成した対訳コーパス中の各文対のアライメントの正否（つまり、対応精度の程度）を機械的に判定する必要がある。本実験で行ったスコア分布や文長比の分析も、そのための基準の策定に資するべく実施した。

この観点において、上記①は一方が無文であるため判別は容易であり、あえてスコアや文長比で判定する必要がない。かつ、文アライメントスコアを正否判定の基準を作成するために正解文、誤り文の平均スコアを取得する際も、①に該当する文はスコアがゼロまたは極小値となるため、「スコアが小さいほど対応精度が良い（＝スコアが大きいほど対応精度が悪い)」というスコア付けをするツールが大多数ななか、誤り文の平均スコア算出にこれらを加えると、本来正解文より高くあるべき平均スコアが（スコア 0 の文対によって）不当に低くなり、基準としての利用性が損なわれる懸念がある。

②に挙げた「N:N の文対」とは、本来 1：1 でアライメント可能であるのに処理の失敗で N:N 対応となった文対を指す。アライメント処理としては改善の余地があるため「誤り」としたが、内容的には過不足なく対応しており、文長比や文アライメントスコアは正常値となる。誤り文の平均スコアや文長比範囲を取得するのは、内容に過不足や不一致があるという基準で正解文と区別するためであり、②はこれらの点では誤り文の特徴を有していないため、誤り文の検出のための文長比やスコアの範囲・平均値算出ではむしろノイズとなる。

同様に③「アライメント処理の失敗による見出し連結」も、（双方とも連結されている場合は）内容の不一致や文長比の異常とはならないため、②と同様の理由で対象外とした。

4.4.3.7. 評価候補 3 手法の選定

予備実験の目的は、本調査で使用する文アライメント手法候補として評価対象とする 3 手法を決定するための第一次選考である。

処理精度の観点からは、予備実験での正解率が 8 手法中唯一 80%を下回った⑥Moore がまず除外される。Moore は、後続工程での利用が想定される文アライメントスコアを出力できない点でも他の手法に劣る。同じく⑦Bleualign も、正解率は同率 3 位と優秀であったが、文アライメントスコアを出力できないため、やはり本調査には不向きである。

これらを除外した 6 手法での正解率の上位は⑧ealign の 86.2%、④MT(ItoE)の 85.3%、③BGE の 85.0%となる。一方、文アライメントスコアの信頼性は⑧ealign と②LaBSE が特に優れており、③BGE、⑤MT(JtoE)がこれに準じた。

なお、本調査では、手法候補中に「多言語の埋め込みモデルを活用した手法」と「英語への機械翻訳を介した手法」を必ず 1 種類ずつ選定する必要がある。前者は①〜③、後者は④〜⑤が該当する。

この要件に則り、多言語の埋め込みモデルを活用した手法（①〜③）からは、正解率が最上位で、文アライメントスコアの信頼度も比較的高い③BGE を選定した。一方、英語への機械翻訳を介した手法（④〜⑤）の比較では、正解率は④MT(ItoE)が 85.3%と⑤MT(JtoE)の 81.3%を上回るが、文アライメントスコアの利用性に関しては、④MT(ItoE)は正解文の平均スコア 0.5413 に対して誤り文は 0.5734 と数値の差がごくわずか（±0.0321）であり、本調査で後に行う文対応の正誤判定の基準に用いるには精度が不十分となる懸念がある。その点、⑤MT(JtoE)は正解文の平均スコアが 0.3835、誤り文が 0.5693 と顕著な差（±0.1858）がある。本調査での用途から総合的に判断すると、英語への機械翻訳を介した手法としては

⑤MT(JtoE)を選定するのが適当と判断した。

以上から、本調査の文アライメント手法候補には、多言語埋め込みモデル（ベクトル化方式）からは③BGE、英語への機械翻訳方式からは⑤MT(JtoE)を選定した。残る 1 手法は、全手法で最も正解率が高かった⑧ealign（機械翻訳＋汎用方式）を選定した。

4.5. 文アライメント手法候補の人手評価

本調査に最適の文アライメント手法を決定すべく、予備実験で選出した 3 種の文アライメント手法候補に対し、人手による評価を実施した。本項にその結果をまとめる。

4.5.1. 人手評価の概要

まずは人手評価の概要について示す。

4.5.1.1. 評価対象とした手法候補

評価対象とする文アライメント手法候補は、予備実験の結果を踏まえ、下記の 3 手法とした。

表 4-8　人手評価の対象とした文アライメント手法 3 種

手法	方式及び使用ツール
① BGE	ベクトル化方式
② MT(JtoE)	英語への機械翻訳方式（JapioAI 翻訳：日本語⇒英語） BGE×vecalign
③ ealign	汎用方式×機械翻訳（Google 翻訳：インドネシア語⇒日本語） MT×ealign

このうち①BGE が「多言語の埋め込みモデルを活用した手法」、②MT(JtoE)が「英語への機械翻訳を介した手法」に相当する。なお、MT(JtoE)は予備実験時は LASER×vecalign を用いたが、実験の結果、LASER よりも BGE のほうが高精度なベクトル化処理となる可能性が高いと判断されたため、本評価では BGE×vecalign としている。

4.5.1.2. 評価用文献

人手評価用文献には、文分割手法の評価時に使用した日インドネシアのファミリー文献 5 文献対（下表①～⑤）にさらに 2 文献対（⑥～⑦）を追加した 7 文献を用いた。各文献から抽出したテキスト全文に対し、インドネシア文献は pySBD、日本文献は Stanza で文分割した結果を各アライメントツールの入力データとした[34]。

[34] なお、インドネシア文献の見出し語はテキスト抽出時に本文と連結されてしまうため、インドネシア文に対しては、文分割処理後に定番の見出し語（全 42 種。3.4.6.項の表 3-6 参照）を本文と切り離す後処理を施したデータを用いた。

表 4-9　評価用インドネシア／日本ファミリー文献（7 文献）

	インドネシア文献	日本文献	IPC
①	P00202203385	2022-002855	B23K
②	P00202301594	2022-024515	A01H
③	P00202304729	2022-073428	H04W
④	P00202300918	2022-012848	C08F
⑤	P00202301067	2022-034129	B60R
⑥	P00202210342	2023-514859	G10Q
⑦	P00202213742	2023-530932	F16C

なお、文アライメント処理は、各文献から取得した全文テキストデータを「要約」「発明の名称」「明細書」「特許請求の範囲」に切り分けたうえで、それぞれ個別に実施した。

4.5.1.3. 評価の手順

人手評価は、まず「第 1 の評価」として、評価対象文献 7 文献対の全文における各手法候補の文アライメントの正解率を調査した。これにより、各手法候補のアライメント精度を比較した。

続く「第 2 の評価」では、各文献対から選定した 420 文対（60 文対×7 文献）を対象に、各文対の対応度（内容の一致度）を人手で精密に評価した。そのうえで、対応度ごとの文長比や文アライメントスコアの分布及び平均値を調査した。

そして、第 1、第 2 の評価結果から総合的に判断し、本調査に最適の文アライメント手法を 3 種の候補から選定した。そのうえで、その文アライメント手法の使用を前提に、以下の二点についてさらに検討した。

① 本調査で作成した対訳コーパスの各文対に対し、その対応度を機械的に判別するための適切な指標を検討・策定した。

② 選定した文アライメント手法においてユーザが調整可能なパラメータを調査し、最適のパラメータ設定を検討・決定した。

これら一連の評価により、本調査での対訳コーパス作成に用いる最適の文アライメント手法ならびにパラメータ値を決定するとともに、作成した対訳コーパスを文の対応度によってふるい分けするための具体的な指標を得た。

4.5.2. 第 1 の評価

第 1 の評価では、まずは評価用 7 文献の全文数に対する「自動評価等で対応が取れていると判定された文対の数の割合」を手法ごとに算出した[35]。この割合、つまりアライメント処理で正しい対応づけがなされたと見なされる文対の割合が高いほど、本調査の文アライメント手法として有用であると判断される。

4.5.2.1. 文長比範囲の指標による正解率

第 1 の評価における、正しい対応づけか否かの判定基準は、予備実験での検討結果に基づき、「インドネシア文の日本文に対する文長比が 1.5 以上 4.0 未満」という指標を採用した。予備実験の結果からは、この指標により、適切な文アライメント[36]がなされている文対の 90%以上がカバーされ、かつ重大な不対応の混入を少量に抑えられると予測される（⇒ 4.4.3.5.）。

下表に、各手法による 7 文献全文のアライメント結果に対し、上記指標に該当する文対を「正解（＝最善な対応が取れている）」と見なした場合の正解率を示す。表中の「総文数」は、人手で作成した各文献の「あるべきアライメント結果（＝正解）」に基づく。

表 4-10　文アライメント手法 3 種の文献別正解率

文献	総文数	BGE		MT(JtoE)		ealign	
		正解数	正解率	正解数	正解率	正解数	正解率
①	272	233	85.66%	233	85.66%	229	84.19%
②	193	163	84.46%	166	86.01%	160	82.90%
③	287	259	90.24%	259	90.24%	259	90.24%
④	413	317	76.76%	316	76.51%	314	76.03%
⑤	216	156	72.22%	155	71.76%	154	71.30%
⑥	112	101	90.18%	101	90.18%	94	83.93%
⑦	142	134	94.37%	134	94.37%	130	91.55%
合計	1,635	1,363	83.36%	1,364	83.43%	1,340	81.96%

本指標による正解率は MT(JtoE)が 83.43%と最良であり、他の 2 手法も BGE が 83.36%、ealign が 81.96%と僅差で続いた。ただし、ここで用いた「文長比 1.5 以上 4.0 未満」という指標は、極端な文長比から内容に大規模な不対応が生じている可能性が高い文対を排除

[35] 本調査の調達仕様書に記載された要件である。

[36] ここで言う「適切な文アライメント」とは、文献対の状況から最も妥当な文同士が対応付けられていることを指し、文対の内容が過不足なく完全に同一であるかは問わない。

することが主眼であり、文長比に表れないタイプの誤り文の混入は避けられない。このため、相応の誤差を想定しておくべきであり、ごく僅差であった上記結果のみで 3 手法の優劣を判断することは困難である。

4.5.2.2. 目視判定による正解率

この結果を受け、より正確な比較を行うため、各手法のアライメント結果全文について、目視でアライメントの正否を判定した。判定は、人手による前記「あるべきアライメント結果」と合致したものを「正解」とし、それ以外は「誤り」とした。ただし、インドネシア文においてマイナーな見出し語が本文と連結する事象はテキスト抽出の手法上不可避であるため、これに該当し、かつ見出し連結以外の不一致がない文対は「見出し連結」として別カウントとした。以下、結果を示す[37]。

表 4-11　文アライメント手法別「正解」「見出し連結」「誤り」文数比較

文献	BGE			MT(JtoE)			ealign		
	正解	見出し連結	誤り	正解	見出し連結	誤り	正解	見出し連結	誤り
①	196	24	25	193	23	29	199	23	36
②	151	6	25	157	6	19	141	5	51
③	263	0	18	263	0	18	262	0	22
④	260	15	83	254	14	89	264	5	120
⑤	182	10	11	181	10	11	168	8	33
⑥	102	0	7	99	0	9	88	0	26
⑦	136	0	4	136	0	4	129	0	15
合計	1,290	55	173	1,283	53	179	1,251	41	303
正解率	78.90%			78.47%			76.51%		
準正解率	82.26%			81.71%			79.02%		

上表の正解率は、前表に示した「あるべきアライメント結果」の総文数 1,635 文を分母とした正解率である。各手法とも「文長比 1.5 以上 4.0 未満」という指標での正解率から 5%ほど低下している。この差異は主に、文長比への影響が小さい誤り文が目視判定で排除されたことによる。

なお、「見出し連結」は目視判定では正解としなかったが、見出しと本文が連結しているのみで文内容への悪影響は小さく、また文分割の時点で連結されたものはアライメント処

[37] 各ツールの出力した文対レコードのカウント結果である。複数文を 1 レコードで出力したり、1 文が分割されて複数のレコードで出力されたりするため、同一文献でも各ツールで合計文数が異なっている。

理でのリカバリ（再分割）は不可能であるため、アライメント手法の評価においては不問とすべきとも考えられる。このため、見出し連結を正解に加算した「準正解率」を別途算出した。

目視判定による手法間の比較では、正解率、準正解率とも BGE が MT(JtoE)を逆転して首位となったが、依然としてごく僅差である。一方、ealign は今回も最下位となり、上位 2 手法に比べて誤り文の多さも目立つ。これは過分割された文を正しく再連結できなかったケースが多いことを示しており、この観点においても上位 2 手法にやや劣る。

目視判定による評価は、前項に示した「文長比 1.5 以上 4.0 未満」の指標による判定よりも精度が明らかに高くなるはずであるが、どちらの評価結果においても、3 手法の相対的な優劣、そして各手法の正解率とも、おおむね合致した。このことから、今回用いた「文長比 1.5 以上 4.0 未満」という正否判定指標は、簡易的ながら十分に有用であるといえる。

4.5.2.3. 各手法における典型的な誤り

前項に示した目視判定では、各手法のアライメント結果の全文対について、文対応の正誤を判定した。その過程では、各手法における典型的な誤り文のパターンが検出された。本項にまとめる。

<全手法共通の誤りパターン>

アライメント処理に供する文分割データは全手法候補、同じものを使用している。この文分割データ自体が誤っていると、文アライメント処理でのリカバリは不可能なことも多い。その結果、全手法で同一の誤り文が発生しているケースが散見された。以下、その典型パターンを示す。

①文末の数値＋ピリオド直前の誤分割

本調査に使用するインドネシア文分割ツール pySBD は、文末が数値＋ピリオドである場合、その直前で文分割を行い、数値＋ピリオドを次文の文頭に連結してしまう課題がある。このような状態のデータをアライメント処理で修正することは不可能であるため、全ての手法で同一の誤り文が発生することとなる[38]。一例を示す。

[38] 唯一の対処策は 2 文を連結して 2 ： 2 の対応とすることだが、3 手法でそのような処理をしたものは存在しなかった。

［文献③明 265～6］

Pemrosesan pada S704 hingga S708 serupa dengan pemrosesan pada S601 hingga S605 pada Gambar	Ｓ７０４～Ｓ７０８の処理は、図６のＳ６０１～Ｓ６０５の処理と同様である。
6. Setelah frame suar ditransmisikan pada S708, proses dilanjutkan ke S506 pada Gambar	Ｓ７０８において、Ｂｅａｃｏｎフレームを送信したら、図５のＳ５０６の処理に進む。

上例の１文目は本来「Gambar 6.（図６）」の直後で文分割するのが正しいが、pySBD には数値＋ピリオドの直前で誤分割し、次文の冒頭に（あたかも箇条書き番号のように）連結させる傾向があり、これが原因で「6.」が２文目の冒頭にセットされてしまっている。文アライメント処理では入力された文を分割することはできないため、この「6.」を文頭から切り離して前行末尾に移設するリカバリは不可能である。事実、３手法候補いずれも上掲のとおりの処理結果となり、１文目、２文目ともに「誤り」と判定された。

②不対応文の前後文への連結

ファミリー文献であっても、どちらか一方のみに記載され、対応づけるべき相手が存在しない「不対応文」が含まれることがある。目視判定では、この場合、対応文不在とし、１：０（もしくは０：１）としたアライメント結果を正解と判定した。

BGE と MT(JtoE)は比較的こうした不対応文にも適切に対処していたが（例：文献①要2)、比較的難易度の高い処理であり、状況によっては全ツールが不対応文を前後いずれかの文に連結してしまっているケースも存在した。以下、一例を示す。

［文献③明 285～6］

Penemuan ini juga dapat diimplementasikan oleh suatu rangkaian (misalnya, ASIC) yang merealisasikan satu atau lebih fungsi. Penemuan ini tidak terbatas pada perwujudan yang diuraikan di atas, tetapi berbagai perubahan dan modifikasi dapat dilakukan tanpa menyimpang dari inti dan ruang lingkup penemuan ini.	また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

本例はインドネシア文の１文目（前半の黒字部分）と日本文とが１：１対応であり、インドネシア文の２文目（後半の赤字部分）は同内容の文が日本文献に存在しない不対応文であ

るが、3 手法とも上記のように両文を連結し、2：1 の誤ったアライメント処理をしている。

＜ealign に特有の誤り＞

4.5.2.2.項の目視判定結果に示されたように、ealign は他の 2 手法に比べて正解率・準正解率がやや劣り、かつ誤り文のカウントが突出して多い。これは、ealign に特有の誤りパターンが存在することに起因する。以下にその内容を記す。

①過分割された文のリカバリの失敗

文分割ツール pySBD は、前項①で述べた「文末の数字＋ピリオド」以外にも、「文頭にあると箇条書き番号に見える文字列」の直前で過分割を行う傾向がある。その中には、アライメント処理で再連結することでリカバリできるものもある。

こうしたものについて、BGE と MT(JtoE)は比較的良好にリカバリしていたのに対し、ealign は過分割された文をそのまま使用する傾向が見られた。この違いが、正解文や誤り文（過分割された文は誤りカウントも複数となる）のカウント差に表れている。以下に実例を示す。

［文献⑦明 85］

Permukaan-permukaan (22,	
23) terdiri dari alur- alur berbentuk cincin (24, 25) pada bagian tengah, yang dihubungkan melalui bukaan (26) yang lewat melalui selongsong (21).	表面（２２、２３）は、ブッシュ（２１）を貫通する開口部（２６）を介して接続された、中心部分に環状の溝（２４、２５）を備える。

上例は本来は 1：1 対応であるが、ealign のアライメント結果のみ、インドネシア文が 2 文に過分割されている。原因は pySBD の文分割処理での過分割[39]にあるが、他の 2 手法（BGE と MT(JtoE)）はこれを再連結して正しい文対を生成できている。その結果、この文は BGE と MT(JtoE)では正解文が＋1 カウントされたのに対し、ealign では誤り文が＋2 カウントされることとなった。同様の例は、文献②の明 87 や④明 21、⑤明 87 など各文献で散見され、ealign が過分割のリカバリ能力で上位 2 手法にやや劣ることが示された。

[39] 2 文目の冒頭「23)」は箇条書き番号に見えるため分割されたと推測される。ただし、同一文中に存在する「25)」では過分割されておらず、基準が不明瞭である。

②見出し等における対応誤り

ealign 特有の対応誤りでは、難易度が極めて低いと思われる条件でのアライメント失敗も挙げられる。その多くは、見出し語や発明の名称など、短文かつ末尾にピリオドを有さないものであった。事例を示す。

［文献④明 27］

	【発明が解決しようとする課題】
Masalah Teknik	

［文献②名 1］

	ヒマワリ種子
BIJI BUNGA MATAHARI	

前者は見出し語、後者は発明の名称であるが、いずれも前後の文に連結されているわけではなく、対応づけの難易度はごく低いと考えられるが（後者は文献の 1 文目でもある）、ealign では特に見出し語において、このような不可解なアライメント失敗が見られた。ごく限定的な発生であり、ealign も大半の見出し語や発明の名称は正しくアライメントできているが、少量とはいえ他の 2 手法にない誤りであり、正解率の低下と、誤り文数の増加（上記 2 例とも誤りカウントは＋ 2 となる）の一因となっている。

<BGE と MT(JtoE)の誤り >

BGE と MT(JtoE)は正解率、準正解率ともほぼ同等であった。BGE、MT(JtoE)ともにアライメントツールは vecalign を使用しており、両者の違いは、インドネシア文と対応づけするデータに日本文を直接用いるか（BGE）、英語に機械翻訳した文を用いるか（MT(JtoE)）である。全文の目視判定の結果からは、各文献におけるアライメント結果自体がほぼ同内容であることが確認できた。かつ、両手法とも、他の手法で見られない特有の誤りパターンは見出せなかった。

4.5.2.4. 文アライメント処理における課題

全手法共通の課題として、4.5.2.3.項に示した典型的な誤りパターンが挙げられる。このうち「①文末の数値＋ピリオド直前の誤分割」については、文字数上は軽微な差異であるため、文長比による排除が困難である。

ただし、このような文は、「文末がピリオドで終わらない」という特徴がある。各手法のアライメント結果において、この特徴を有する文は、定番見出し語を除けば、本エラーが発生した文におおむね限定できる。このため、こうした文を特定する際には、文長比ではなく、

「インドネシア文末がピリオド以外」という条件での機械的な判別がより効率的かつ網羅的である。

4.5.2.5. 処理時間の比較

続いて、三手法候補それぞれが評価用 7 文献全文の文アライメント処理に要した実時間に基づき、本調査で実施することとなる約 2 万文献対の推定処理時間を試算した。下表に結果を示す。

表 4-12　文アライメント手法の処理時間比較（評価用 7 文献対）

方式	評価（7 文献対）での処理時間	試算（2 万文献ペア）
BGE	・ベクトルデータ作成(BGE): 38 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:15 日(5GPU) ・アライメント処理:4 日 ●合計：19 日
MT(JtoE)	・日英機械翻訳：13 分 ・ベクトルデータ作成(BGE):38 分 ・アライメント処理(vecalign):2 分	・日英機械翻訳：7 日(4GPU) ・ベクトルデータ作成:15 日(5GPU) ・アライメント処理:4 日 ●合計：26 日
ealign	・イン日機械翻訳：4 分 ・アライメント処理（ealign）：2 分	・イン日機械翻訳：8 日 ・アライメント処理：5 日 ●合計：13 日

単純計算による処理時間の比較では、ベクトルデータ作成を行わない ealign が最も短くなる。ただし、ealign が使用する Google 翻訳（インドネシア⇒日）は、過去の使用時には、大量のデータを投入すると処理が頻繁に中断され、人手による対応が頻発した。このため、実際には試算結果より所要時間を要する可能性が高く、かつその程度も不確定である。

ealign が最も処理時間が短いのは事実だが、前述のとおり Google 翻訳処理での処理中断という不確定要素が存在し、一定のリスクを考慮しなければならない。かつ、elaign は文アライメント精度が他の 2 手法より明らかに劣る。

BGE と MT(JtoE)は、ベクトルデータ作成処理（所要時間：約 15 日）およびアライメント処理（約 4 日）は共通であり、両者の違いは MT(JtoE)のみ日英機械翻訳処理（約 7 日）を要する点である。双方のアライメント精度がほぼ同等（わずかに BGE が優勢）であることに鑑みて、両者の比較では BGE を採用するのが妥当である。

4.5.2.6. 第 1 の評価の結論

第 1 の評価では、BGE、MT(JtoE)、ealign の三候補について、文長比範囲の指標による正解率と、全文の目視判定による正解率を算出・比較した。それぞれの結果からは、BGE もしくは MT(JtoE)が本調査の選択肢となると判断される。ealign も極度の性能差はないが、正解率でやや劣り、かつ 4.5.2.3.項の＜ealign に特有の誤り＞（p.73）で述べたとおり、他の 2 手法では見られない特有の不備も検出されているため、積極的に選択する理由を欠く。

一方、処理時間の比較では、ealign が最も所要時間が短いと試算された。ただし、ealign には Google 翻訳における処理中断の懸念があり、実際の所要時間がどの程度になるかは不確定である。所要時間が予測できる BGE と MT(JtoE)の比較では、機械翻訳処理を要さない分、BGE のほうが所要時間が 7 日ほど短く有利である。BGE は文アライメント精度でも MT(JtoE)と同等以上であり、本調査で用いる手法として最も適当と考える。

4.5.3. 第 2 の評価

第 2 の評価では、第 1 の評価で用いた 7 文献のアライメント結果から「対応が取れていると判定された文対」を 1 文献あたり 60 文対ずつ選定し、各手法の処理結果に対して人手による精密な評価を実施した。以下、その概要と結果を示す。

4.5.3.1. 評価方法

人手評価は、日本語とインドネシア語の双方を解する評価者を起用し、人手評価対象文 420 文対（60 文対×7 文献）の対応精度に応じて、以下の基準でスコアを付す形で実施した。

【対応精度スコア】
- スコア S：文対の文の内容が、100%一致している
- スコア A：文対の文の内容が、90%以上一致している
- スコア B：文対の文の内容が、80%～89%の範囲で一致している
- スコア C：文対の文の内容が、50%～79%の範囲で一致している
- スコア D：文対の文の内容が、49%以下しか一致しない

さらに、何らかの不一致が検出された文対に対しては、その原因を特定し、不備の類型化の分析（⇒第 6 章）に備えた。

4.5.3.2. 人手評価対象文の選定方法

人手評価対象文は、各ツールのアライメント結果から「対応が取れていると判定された文」を 1 文献あたり 60 文ずつ、全 420 文選定した。

選定は原則ランダムに行ったが、文の長短や項目（発明の名称、要約、明細書、特許請求の範囲）、技術分野などに偏りが生じないように留意した。なお、選定する文は、手法候補間の比較のため、原則同一の文とした[40]。

各手法候補の文アライメント精度の比較は第 1 の評価でより大規模に実施しており、すでに結論はほぼ出ている。このため第 2 の評価では、本調査で作成する対訳コーパスの各文対に対し、機械的に前掲の対応精度スコア（A～D[41]）を付すための「対応精度を示す指標」を定めることの比重が大きくなる。具体的には、文長比および各アライメントツールが出力する文アライメントスコア等を組み合わせて、対訳コーパスの各文対を A～D それぞ

[40] ただし、各手法のアライメント状況によっては、同じ文でも異なる対応づけがされていることはありえる。

[41] 対応精度を示す指標の策定においては、スコア S はスコア A に包含させた。

れの対応精度スコア群に的確に分別する指標を作らなければならない。

そのためには、各対応精度スコアのサンプルを十分な数、取得する必要がある。つまり、人手評価対象文に選定する 420 文対には、スコア A〜D それぞれに該当する文対がなるべく均等に含まれることが望ましい。

ただし、本調査の仕様により、選定対象となる母集合は「インドネシア文の日本文に対する文長比が 1.5 以上 4.0 未満」に該当する文対となる。かつ、候補手法間の比較のため、原則として同一の文を選定することが求められている[42]。

第 1 の評価で明らかになったとおり、「文長比 1.5 以上 4.0 未満」の文対集合における文アライメントの正解率は 80%近い。したがって、ここから無作為にサンプルを採取すると、文対応が良好（つまり対応精度スコア A）な文対ばかりが選定される可能性が高い。

そこで、対応度の低い文対を選定する確率を高めるべく、人手評価対象文の選定にあたっては、各文献の「文長比 1.5 以上 4.0 未満」の文対集合から、以下の各条件に該当する文対を優先的に選定することとした[43]。

① 文長比が最も小さい 5 文対
② 文長比が最も大きい 5 文対
③ 文アライメントスコアが最も高い 5 文対
④ 文アライメントスコアが最も低い 5 文対

これにより、1 手法、1 文献あたり最大 20 文対が選定される。数値からは極端なデータの比重が大きいとも映るが、実際には各条件とも複数の手法で同一文が該当することが多く、①〜④に属する人手評価対象文は各文献とも全体の半数前後にとどまった。例えば文献①では、BGE で上記①〜④に該当する全 19 文（1 文は①と④の双方に該当）のうち、14 文が MT(JtoE)と、8 文が ealign とそれぞれ重複している。結果、①〜④の条件で選ばれた文対は 3 手法合計で 35 文にとどまり、残り 25 文はランダム的に選定した。

このような基準で人手評価対象文を選定した結果、420 文対の全てが、全手法共通の評価対象文となった。ただし、手法ごとのアライメント結果の違いにより、同じ日本文に対して

---

[42] 手法間でアライメント結果が異なるため、ある手法では文長比が 1.5〜4.0 の範囲内であっても、他の手法では範囲外となるケースも存在する。こうした文が優先選定条件①〜④に該当する場合は、例外的に 1.5〜4.0 の文長比の範囲外の文対も選ばれる。

[43] ただし、見出し語のみの文対は評価対象として不適なため、5 文対に該当する場合には除外して次点の文を採用した。

異なるインドネシア文が対応付けられていたり（またはその逆）、過不足があったり等の差異はありえる。以下に、同じ日本文に対して異なるインドネシア文が対応付けられている事例を示す。

[文献①明 42]

日本文	特許文献 1 のようにはんだ合金にＮｉが含有されていると「遊離」が生じやすくなる。
BGE	Ketika paduan patri mengandung Ni di dalamnya seperti dalam Dokumen Paten 1, “pengelupasan” lebih mungkin terjadi.
MT(JtoE)	Ketika paduan patri mengandung Ni di dalamnya seperti dalam
ealign	Dokumen Paten 1, “pengelupasan” lebih mungkin terjadi.

上例は、BGE は日本文とインドネシア文が過不足なく対応しているのに対し、MT(JtoE)と ealign はそれぞれインドネシア文が過分割され、その前半（MT(JtoE)）、または後半(ealign)のみが対応付けられている。

人手評価対象文の選定結果は別添資料⑥「文アライメント人手評価結果」のとおりである。

4.5.3.3. 第 2 の評価（人手評価）の結果

下表に、BGE、MT(JtoE)、ealign それぞれの、評価対象文 420 文対に対して付された人手評価結果（対応精度スコア S～D。⇒4.5.3.1.）の集計結果を示す。評価対象文ごとの評価結果は別添資料⑥「文アライメント人手評価結果」を参照されたい。

表 4-13　文アライメント手法別 対応精度スコア（S～D）集計結果

スコア	BGE		MT(JtoE)		ealign	
S	307	73.1%	300	71.4%	299	71.2%
A	57	13.6%	62	14.8%	51	12.1%
B	39	9.3%	39	9.3%	35	8.3%
C	8	1.9%	7	1.7%	15	3.6%
D	9	2.1%	12	2.9%	20	4.8%

対応精度スコア A 以上の構成率の比較では、BGE が 86.7%（S：73.1%＋A：13.6%）、MT(JtoE)が 86.2%とほぼ同等であり、ealign のみ 83.3%とやや劣る。３手法候補の相対的順位は第 1 の評価と合致しており、そこでの「(正解率の観点からは）BGE もしくは MT(JtoE)を選定すべき」との結論の妥当性を補強する結果が得られた。

また、上記結果は、日インドネシア特許文献から対訳コーパスを取得した場合、「文長比1.5〜4.0」という条件で絞り込むことによって、BGE であれば 86.7%(S:73.1%＋A:13.6%)、MT(JtoE)であれば 86.2%、ealign であれば 83.3%程度の、スコア A 以上の文対の構成比をもつ対訳コーパスを作成できることを示している[44]。

本調査では、第 2 章冒頭で述べたとおり、「一定以上の精度を満たす 100 万文対以上の日インドネシア対訳コーパスを取得する」ことを目標としている。具体的には、日インドネシアのファミリー特許文献に由来した、スコア A 以上の文対の構成比が 93%以上となる『対訳コーパス A』の作成が求められる。

このため、今回の人手評価で得られた文対ごとのスコア分布を精緻に分析し、スコア A 以上の構成比が 93%以上となるような最良の「対応精度を示す指標」を定めることが、第 2 の評価の主な目的となる[45]。具体的には、「①文長比の範囲を 1.5〜4.0 からさらに絞り込む」とともに、「②各手法が付す文アライメントスコアに対しても範囲を設ける」ことで、一定の規模を維持しつつ、スコア A 以上の文対の構成比を高めていくことが考えられる。以下、それぞれについて検討する。

[44] 厳密には、人手評価対象文は、文長比 1.5〜4.0 という基準による母集合から、半数弱を前記優先条件①〜④に該当した文、残り半数強をランダムに選定している。このうち前者は文長比が極端な数値であるもの（①②）や、文アライメントスコアが特に低いもの（④）を選んでおり、他の文に比べて何らかの不備が生じている確率が高い。つまり、人手評価対象文 420 文は、文長比 1.5〜4.0 に属する全文対の集合に比べて不備の含有率が高いと考えられる。換言すれば、文長比 1.5〜4.0 の全文対集合における対応精度スコア A 以上の構成率は、各手法とも前掲のパーセンテージよりもさらに高くなる可能性が高い。

[45] 三手法の文アライメント精度の優劣は、各文献 60 文のみの第 2 の評価よりも、第 1 の評価で実施した全文献の全文に対する目視判定による正解率の比較のほうが信頼性が高い。

4.5.3.4. 文長比範囲の考察

各手法の 420 文対の対応精度スコアについて、1.5～4.0 の文長比範囲を 0.5 刻みに細分化してスコア A 以上／B 以下の件数をカウントした結果を下表に示す。

表 4-14　文長比範囲別 対応精度スコア（A 以上／B 以下）分布

	BGE		MT(JtoE)		ealign	
以上～未満	A 以上	B 以下	A 以上	B 以下	A 以上	B 以下
～1.5*	0	2	2	0	0	11
1.5～2.0	43	7	46	11	46	10
2.0～2.5	74	14	75	12	75	13
2.5～3.0	118	8	117	9	110	11
3.0～3.5	74	12	68	13	68	9
3.5～4.0	50	11	50	11	47	13
4.0～	5	2	4	2	4	3

*計測不能（対応文なし）を含む。

上表を見ると、各手法とも、A 以上の文対は文長比 2.5～3.0 を頂点として前後に逓減していくベルカーブ状の分布を示しているのに対し、B 以下の文対は 1.5～4.0 の各範囲でおおむね均一に分布していることがわかる。したがって、文長比範囲を 2.5～3.0 を中心に狭めていくことで、A 以上の文対の比率をさらに高めることができる。

例えば、BGE で文長比範囲を 2.0～3.5 とすることで、件数が 300 文対（全 420 文対の 71.4%）に絞られ、そこでの A 以上の比率は 88.7%に高まる。さらに 2.5～3.0 まで絞れば A 以上の比率は 93.7%となり目標値 93%に到達するが、反面、件数は 126 文対（全文対の 30%）と大幅に減少する。30%の採用率で 100 万文対を達成できるかは本調査で作成する対訳コーパス全体の規模次第であるが、本項では、A 以上の比率 93%を堅持しつつ、可能な限り多数の対訳コーパスを採用できる最適の指標を追求する。

4.5.3.5. 文アライメントスコア範囲の考察

評価対象のアライメント手法候補はいずれも、自身が生成した各文対に対し、その対応の確からしさを示す「文アライメントスコア」を付す機能を有する。この文アライメントスコアを、前項で調査した文長比範囲と掛け合わせてさらに絞り込むことで、対応精度スコア A 以上の構成比を 93%以上に高められるかを検討した。

下表に、各手法がアライメント処理の際に各文対に対して自ら付した「文アライメントスコア」の最高値、最低値および平均値を示す。

表 4-15　文アライメント手法別 文アライメントスコア最高／最低／平均値比較

文アライメントスコア	BGE		MT(JtoE)		Ealign	
	A 以上	B 以下	A 以上	B 以下	A 以上	B 以下
最高値	0.1118	0.2852	0.0725	0.2555	1.0000	0.9545
最低値	0.8697	1.0380	0.8756	1.1484	0.1923	0.0037
平均値	0.4392	0.5675	0.3631	0.4949	0.6742	0.5484

※BGE と MT(JtoE)は文アライメントスコアが小さいほど良好、ealign のみ大きいほど良好。

各手法とも、文アライメントスコアの平均値を見ると対応精度 A 以上の文対のほうが B 以下の文対よりも良いスコアとなっており、一定の精度を有しているといえる。ただし、両者のスコア範囲は大きくオーバーラップしており、特定の文アライメントスコア値をもって A 以上と B 以下との境界とすることは困難である。

続いて、文長比と同様、文アライメントスコアの範囲を 0.1 単位に細分化して、それぞれに属する対応精度 A 以上／B 以下の文対数を比較した。結果を下表に示す。

表 4-16　文アライメントスコア範囲別 対応精度 A 以上／B 以下の文対分布

文ｱﾗｲﾒﾝﾄｽｺｱ	BGE		MT(JtoE)		ealign	
以上～未満	A 以上	B 以下	A 以上	B 以下	A 以上	B 以下
～0.1000	0	0	4	0	0	7
0.1000～0.2000	13	0	44	0	2	0
0.2000～0.3000	46	1	66	7	6	2
0.3000～0.4000	76	5	110	9	11	6
0.4000～0.5000	114	12	77	19	42	6
0.5000～0.6000	72	19	48	14	34	15
0.6000～0.7000	32	11	7	3	85	15
0.7000～0.8000	9	1	5	2	84	13
0.8000～0.9000	2	5	1	2	60	4
0.9000～1.0000	0	1	0	1	24	2
1.0000～	0	1	0	1	2	0
合計	364	56	362	58	350	70

上表のとおり、各手法とも A 以上の分布ピーク（青字）と B 以下の分布ピーク（赤字）とは 1 段ずれているものの、いずれも広い範囲になだらかに分布しており、A 以上と B 以

下との間に明確な境界は存在していない。ただし、各手法とも極端に悪いスコア範囲では総じて B 以下の構成比が高くなっており、例えば BGE であれば、スコア 0.8000 以上の範囲に属する文対を除外することで、A 以上も 2 文対（全体の 0.5%）失われるものの、B 以下は 7 文対（全体の 12.5%）と、より多く排除することができる。同様に MT(JtoE)であれば 0.7000 以上を対象外とすることで A 以上が 6 文対(1.7%)に対して B 以下を 6 文対(10.3%)、ealign であれば 0.1000 以下を対象外とすることで B 以下のみを 7 文対 (10.0%)、それぞれ排除できる。

こうした文アライメントスコア範囲を特定した除外は、スコアの精度が高いほど効率が良くなる。上記の結果から見て、各手法のスコア精度に大きな差はなく、『対訳コーパス A』の作成における利用性は同等といえる。

4.5.3.6. 第 2 の評価の結論

4.5.2.項に示した第 1 の評価では、本調査で用いるべき文アライメント手法の候補として BGE、MT(JtoE)、ealign の三手法の文アライメント結果の正解率と処理時間を比較した。その結果、正解率では BGE と MT(JtoE)がほぼ互角でありいずれかを選択すべきこと、そして処理時間の比較では、MT(JtoE)が日本文原文の英語への機械翻訳処理を必要とするのに対し、BGE はこうした機械翻訳を介さず原文を直接アライメント処理できるため、効率面でより有利であると結論した。

この結果を受けて実施した第 2 の評価では、精緻な人手評価によって、第 1 の評価で行った手法候補間の精度の優劣判定が的確であることが裏付けられた。また、『対訳コーパス A』作成のための「対応精度を示す指標」に用いる文アライメントスコアの有用度の比較においては、三手法の文アライメントスコア精度に特段の優劣がないと判明した。

これら第 1、第 2 の評価結果を総合的に判断して、本調査で用いる文アライメント手法には、正解率が最上位かつ処理効率にも優れ、文アライメントスコアの利用性も問題のない BGE を選択するのが最善であると結論される。

4.5.4. 対応精度スコア（A～D）を判定する「対応精度を示す指標」の策定

第 1～第 2 の評価により、本調査で用いる文アライメント手法に BGE を選定した。この選定結果を受け、4.5.3.4.項における文長比範囲の考察と、4.5.3.5.項における BGE の文アライメントスコア範囲の考察結果をかけ合わせて、対応精度スコア A 以上が 93%以上となる『対訳コーパス A』を機械的に抽出するための最適の「対応精度を示す指標」を策定した。以下にその結果を示す。

4.5.4.1. 対応精度スコア A 以上

まずは対応精度スコア A 以上の指標を策定する。まずは人手評価結果 420 文における、スコア A 以上の分布状況を示す。

＜文アライメントスコア×文長比範囲の分布状況＞

対応精度スコア A 以上を示す指標として、まずは 4.5.3.5.項の考察結果に基づき、BGE の付す文アライメントスコアが 0.8000 以上の文対を除外した。そのうえで、4.5.3.4.項の分析結果に基づいて、文長比の範囲をスコア A 以上の文対のボリュームゾーンである 2.0～3.5 に限定し、詳しい分布を調べた。下表に示す。

表 4-17　文長比範囲（詳細）別 対応精度スコア A 以上／B 以下の文対の分布

文長比 (以上～未満)	スコア A	スコア B 以下	(内訳)				
			S	A	B	C	D
2.0～2.1	15	2	15			2	
2.1～2.2	12	4	7	5	2	1	1
2.2～2.3	16	4	14	2	2	1	1
2.3～2.4	13	1	10	3		1	
2.4～2.5	18	3	17	1	3		
2.5～2.6	20	3	16	4	3		
2.6～2.7	22	1	17	5			
2.7～2.8	17	1	15	2	1		
2.8～2.9	34		29	5	1		
2.9～3.0	25	3	25		3		
3.0～3.1	17		12	5			
3.1～3.2	19	3	14	5	3		
3.2～3.3	9	4	7	2	4		
3.3～3.4	17	2	14	3	2		
3.4～3.5	12	3	9	3	3		

上表のとおり、文アライメントスコア 0.8000 未満の集合においては、文長比 2.3 未満と 3.1 以上の各エリアでスコア B 以下の構成率が高いことがわかる。これらのエリアを除外して文長比を 2.3 以上 3.1 未満に絞ることで、文対数は 178 対、このうち A 以上が 166 対となり、構成比は 93.3%と目標値 93%に到達する。この「BGE の文アライメントスコア 0.8000 未満 × 文長比 2.3 以上 3.1 未満」という範囲設定が、対応精度 A 以上の構成比を 93%以上としつつ、文対数を最も多く確保できる、最善の指標である。

この指標を適用すると、人手評価文 420 文対のうち 178 対、つまり約 42.4%が『対訳コーパス A』として選定されることとなる。評価対象文献 7 文献の総文対数は 1,635 文対であり（⇒4.5.2.1.）、1 文献平均 234 文対とすると、評価対象文の母集合である「文長比 1.5～4.0」に該当する文対数はこのうち 83.36%にあたる約 195 文対、そしてこのうちの 42.4%にあたる 83 文対が、1 文献から対訳コーパス A として選定される文数となる。

本調査では、対訳コーパスのソースとして約 2 万対の日インドネシア公報全文テキストを取得している（⇒2.2.4.）。このため、対訳コーパス A 候補として 83 文対×2 万文献＝166 万文対程度が取得できる計算となる[46]。このうち A 以上の文対は構成比 93.3%であるため 154 万文強となる。これを BGE の文アライメントスコアでソートし、対応精度上位と見られる 100 万文を『対訳コーパス A』に採用することで、対応精度スコア A 以上の文対の構成比は確実に 93%を上回ると予測される。

＜対応精度スコア A の指標の結論＞

上記考察により、本調査の『対訳コーパス A』は、以下の方式で作成することとする。

・文アライメント手法には BGE を用いる
×
・文アライメントスコアは 0.8000 未満の範囲とする（※スコア 0 .0000 は除外）
×
・文長比（インドネシア文長÷日本文長）は 2.3 以上 3.1 未満の範囲とする
×
・文アライメントスコアの良い順（数値の小さい順）に上位 100 万文対を選定する

[46] これに加え、約 6 万件の公報フロントページ由来のデータも対訳コーパスのソースとするため、取得できる文対数はさらに多くなる。

4.5.4.2. 対応精度スコア B、C、D

前項にて、「対応精度スコア A 以上を示す指標」について具体的に定めた。これと矛盾のない形で、対応精度スコア B～D それぞれの指標についても検討する。

まずは、BGE の人手評価結果 420 文における、対応精度スコア B、C、D それぞれの状況を示す。

<文長比範囲ごとの分布状況>

BGE の人手評価結果 420 文のうち、対応精度スコア B、C、D を付された文対の文長比範囲ごとの分布状況を下表に示す。

表 4-18　文長比範囲別 対応精度スコア B～D の文対の分布

文長比	スコア B		スコア C		スコア D	
総数	39 文対		8 文対		9 文対	
～1.5	0		0		2	22.2%
1.5～2.0	3	7.7%	1	12.5%	3	33.3%
2.0～2.3	4	10.3%	4	50.0%	2	22.2%
2.3～2.4	0		1	12.5%	0	
2.4～2.5	3	7.7%	0		0	
2.5～2.6	3	7.7%	0		0	
2.6～2.7	0		0		1	11.1%
2.7～2.8	1	2.6%	0		0	
2.8～2.9	0		0		0	
2.9～3.0	3	7.7%	0		0	
3.0～3.1	0		0		0	
3.1～3.5	12	30.8%	0		0	
3.5～4.0	10	25.6%	1	12.5%	0	
4.0～	0		1	12.5%	1	11.1%

スコア A 以上の文長比範囲として定めた「2.3 以上 3.1 未満」のエリアでのスコア B 以下の分布率は、B が 25.6%（10／39 件）、C が 12.5%（1/8 件）、D が 11.1%（1/9 件）といずれも低値であり、スコア B 以下の大部分はこのエリア外に分布していることが示された。この状況を踏まえて、スコア B、C、D の文長比範囲は、スコア A のエリアの外縁から徐々に広げていくのが妥当と考える。

まず、文長比 1.5 未満と 4.0 以上の極値は、後者にスコア C が 1 件含まれるものの、スコ

アＤとするのが妥当である。したがって、スコアＢとＣとの境界線は、文長比範囲 1.5～2.3 の区間と 3.1～4.0 の区間にそれぞれ引くことになる。

まず文長比 3.1～4.0 の区間には、スコアＢは 22 件、スコアＣは 1 件のみ存在する。区間を等分に分割すると、3.1～3.55 がスコア B、3.55～4.0 がスコアＣのエリアとなる。この設定の場合、スコアＢの 22 件中 14 件（63.6%）がエリア内に入る。スコアＢのうち 8 件（36.4%）がスコアＣとなってしまうため、やや厳しい設定となるが、スコアＣはデータが 1 件のみ（文長比 3.76）と少ないため状況把握が難しく、スコアＣの混入を避けるためには厳しい設定としておくほうが安全である。このため、スコアＢとＣの境界は 3.55 とするのが妥当と考える。

一方、文長比 1.5～2.3 の区間については、同様に等分に分割した場合、スコアＢが 1.9～2.3、スコアＣが 1.5～1.9 のエリアとなるが、このエリア設定では、スコアＣの 5 件全件がスコアＢエリアに属してしまう。今回のデータ分布からは、この区間でスコアＢとＣとを妥当に振り分ける境界は存在しない。したがって、本調査においては、この区間は全てスコアＣとせざるを得ない。

上記より、対応精度スコア B～D の文長比範囲はそれぞれ以下とするのが妥当と結論される。

・スコアＢ：文長比範囲 3.1 以上 3.55 未満
・スコアＣ：文長比範囲 1.5 以上 2.3 未満　および　3.55 以上 4.0 未満
・スコアＤ：文長比範囲 1.5 未満　および　4.0 以上

＜文アライメントスコアの状況＞

続いて、BGE の文アライメント結果に対する人手評価でスコア B、C、D を付された文対の文アライメントスコアの状況を下表にまとめる。

表 4-19　文アライメントスコア範囲別 対応精度スコア B～D の文対の分布

ｱﾗｲﾒﾝﾄｽｺｱ	スコア B		スコア C		スコア D	
総数	39 文対		8 文対		9 文対	
～0.1000						
0.1000～0.2000						
0.2000～0.3000	1	2.6%				
0.3000～0.4000	4	10.3%	1	12.5%		
0.4000～0.5000	9	23.1%	3	37.5%		
0.5000～0.6000	17	43.6%	1	12.5%	1	11.1%
0.6000～0.7000	7	17.9%	2	25.0%	2	22.2%
0.7000～0.8000	1	2.6%				
0.8000～0.9000			1	12.5%	4	44.4%
0.9000～1.0000					1	11.1%
1.0000～					1	11.1%
最高値／最低値	0.2852	0.7053	0.3622	0.8463	0.5401	1.0380
平均値	0.5193		0.5480		0.7941	

スコア A 以上の範囲設定では、文アライメントスコアは 0.8000 未満を条件とした。上表のとおり、スコア B は全件、スコア C もその大半（87.5%）が 0.8000 未満に属している。このため、文アライメントスコア 0.8000 以上は文長比によらず全てスコア D と判定するのが妥当である。

なお、前項では文長比範囲 1.5～2.3 中でスコア B と C とを分離できなかったが、文アライメントスコアと掛け合わせることで可能となる可能性がある。このため、上表の対象としたスコア B 及び C の文対のうち、文長比 1.5～2.3 の範囲に絞って文アライメントスコア分布を調査した。下表に示す。

表 4-20　文アライメントスコア別 対応精度スコア B／C の文対分布

ｱﾗｲﾒﾝﾄｽｺｱ	スコア B		スコア C	
総数	7 文対		5 文対	
～0.1000				
0.1000～0.2000				
0.2000～0.3000	1	14.3%		
0.3000～0.4000	2	28.6%		
0.4000～0.5000	2	28.6%	3	60.0%
0.5000～0.6000	2	28.6%	1	20.0%
0.6000～0.7000			1	20.0%
0.7000～0.8000				
最高値／最低値	0.2852	0.5857	0.4148	0.6336
平均値	0.4291		0.5095	

上表のとおり、今回の評価対象文 420 文においては、文アライメントスコア 0.4000 未満のエリアにスコア B は 3 文対（42.9%）分布しているが、スコア C は 1 文対も分布していない。サンプル数は極めて少ないものの、今回の結果からは、文長比 1.5～2.3 の範囲では、文アライメントスコア 0.4000 未満をスコア B、0.4000 以上（0.8000 未満）をスコア C と判定するのが妥当といえる。

＜対応精度スコア B～D の指標の結論＞

このように定めた文長比範囲および文アライメントスコアの範囲を掛け合わせた結論は以下となる。

・スコア B：文長比範囲 1.5 以上 2.3 未満 × 文ｱﾗｲﾒﾝﾄｽｺｱ 0.4000 未満
　　又は　文長比範囲 3.1 以上 3.55 未満 × 文ｱﾗｲﾒﾝﾄｽｺｱ 0.8000 未満

・スコア C：文長比範囲 1.5 以上 2.3 未満 × 文ｱﾗｲﾒﾝﾄｽｺｱ 0.4000 以上 0.8000 未満
　　又は　文長比範囲 3.55 以上 4.0 未満 ×文ｱﾗｲﾒﾝﾄｽｺｱ 0.8000 未満

・スコア D：スコア A～C に該当しない全ての文対

上記指標を用いた場合、人手評価対象文 420 文のうちスコア B〜D に区分される文対数はそれぞれ以下となる。

・スコア B：141 文対（33.6%）
・スコア C：88 文対（21.0%）
・スコア D：13 文対（3.1%）

このうちスコア B は、本調査で『対訳コーパス A』と別途に作成する『対訳コーパス A-B』の対象となる（5.5.2.項で後述）。ソースとなるファミリー文献対は 2 万件、1 文献あたりの平均文対数は 195 文対と仮定されるため（⇒4.5.4.1.）、スコア B として取得できる対訳コーパス数は 2 万文献×195 文対×33.6%＝131 万文対程度と予測される。

同様の計算方法により、スコア C の取得見込件数は 2 万文献×195 文対×21.0%＝81.9 万文対程度となる。スコア D については、文長比 1.5〜4.0 の範囲における取得見込件数は 2 万文献×195 文対×3.1%＝約 12.1 万件だが、文長比 1.5〜4.0 の範囲外の文対が 1 文献あたり 39 件発生する見込み[47]であり（⇒4.5.4.1）、これらは全てスコア D であるため別途 2 万文献×39 文対＝78 万文対が発生する。このためスコア D の取得件数は 12.1 万件＋78 万件＝90.1 万件程度と見込まれる。

[47] 1 件あたり平均文数 234 文から文長比 1.5〜4.0 に属する 195 文を除いた文数。

4.5.5. BGE のパラメータ調整

本調査では文アライメントとして BGE を使用することと定めた。BGE には、ユーザによる調節が可能で、かつ文アライメント結果に影響を与える可能性のあるパラメータが存在する。BGE の使用に先立ち、適切なパラメータ設定について検討した。

4.5.5.1. BGE で調節可能なパラメータ

BGE でユーザが調節可能なパラメータには、下表に示す 3 種が存在する。

表 4-21　ユーザが調節可能な BGE のパラメータ 3 種

	パラメータ名称	パラメータの概要
①	--num_overlaps NUM_OVERLAPS	ベクトル化データの生成にあたり、本パラメータを 1 に設定すると 1 文のみをベクトル化し、2 以上に設定すると、1 文でのベクトル化に加え、その文と前後の文を合わせた（オーバーラップさせた）別パターンのベクトル化データが作成される。多数のベクトル化データ候補を生成することで、その中から最もベクトルが近似するものと対応付けることができるためアライメント精度が向上するが、反面、パラメータ値を上げるほど処理に要する時間が長くなる。
②	--alignment_max_size ALIGNMENT_MAX_SIZE	サイズ N 文対 M 文 の範囲でアライメントを行う。指定する値は N と M を加算した値を指定する。上記①で設定したパラメータ値が上限となる。
③	--one_to_many [ONE_TO_MANY]	1 対複数（例: 1:1、1:2、... 1:M）のアライメントを実行する。（デフォルト: なし）

①は、概要欄に記載したとおり、ベクトル化データ候補の生成数を指定するパラメータである。たとえば 4 を指定すると、インドネシア文献と日本文献それぞれの各文について、文 A のみ、文 A＋文 B、文 A＋B＋C、文 A＋B＋C＋D といった最大 4 重のオーバーラップまで許容した多数のパターンのベクトル化データが生成される。これにより、文アライメント処理では、多数パターンのインドネシア文×日本文の組み合わせの中から最良の文アライメントを選ぶことができる。１：N や N：N 関係の文対をはじめ、文分割処理で一文が寸断された場合も、このパラメータ値の範囲内であれば、アライメント処理でリカバリされる可能性がある。

つまり、本パラメータは基本的には数値が大きいほどアライメント精度は高まると考えられる。このため、第 1 ～第 2 の評価時の BGE は、本パラメータに 10 を指定していた。

だが、この数値を大きくしたことで、ベクトル化処理に要する時間も大幅に増加し、ealignとの処理時間の差（⇒4.5.2.5.）となったと考えられる。つまりパラメータ①は、トレードオフの関係にあるアライメント精度と効率のバランスを調整するパラメータであるといえる。

パラメータ②は、アライメントする文数の最大値、つまり 1 文対として対応づけられるインドネシア文と日本文の合計数を指定するパラメータである。その数値は、パラメータ①で指定した数値以下しか許容されていない。パラメータ①の設定により作成されるベクトル化データ候補を最大限に活用するにはパラメータ②も①と同値とする必要があり、それ以外の選択肢は考えにくい。第 1 〜第 2 の評価時の設定値も、①と同値の 10 としている。

パラメータ③は、一方の文を 1 文に固定したうえで、その相手として何文まで許容するかを指定するパラメータである。作成するコーパスが、「1 ：N は許容されるが N：N は許容されない」という特殊な条件である場合に有用である。ただし、本調査で用いる文分割手法はインドネシア文用（pySBD)、日本文用（Stanza）とも特定の条件で一文が過分割される事象が見られ、文アライメント処理でのリカバリが必要となる。文対の一方を 1 文に固定するパラメータ③は、どのように設定してもこうした過分割のリカバリが制限されることとなるため、本調査ではデフォルト（設定なし）のままとするのがよい。

4.5.5.2. パラメータ①および②の調整結果

前項に示した 3 種のパラメータの用途に鑑み、BGE のパラメータ調整は、パラメータ①および②の値を現状の 10 から変動させることで（パラメータ②は①と同値とする)、文アライメントの精度と処理効率のバランスを最適化する設定を見極める目的で実施することとした。

BGE におけるパラメータ①のデフォルト値は 4 であり、現状の 10 という設定は、効率性を考慮せず精度を極大化したものといえる。この数値を下げていくことで、文アライメントの実質的精度を損わずに処理効率を改善できる最も実用的な設定値を調査した。

具体的には、BGE のパラメータ値①をデフォルト値の 4 、及び現状値 10 とデフォルト値 4 の中間にあたる 7 とし、パラメータ②はそれぞれ①と同値とした 2 種類の設定（パラメータ 4、パラメータ 7）を用いて、先の評価で用いた人手評価対象文献 7 文献および人手評価対象文 420 文の文アライメント結果の変化および処理効率を調査し、現状値との優劣を判定した。以下の各表にその結果を示す。

表 4-22　目視判定による文アライメント正解率[48]比較（パラメータ 10/7/4）

文献	パラメータ 10			パラメータ 7			パラメータ 4		
	正解	見出し連結	誤り	正解	見出し連結	誤り	正解	見出し連結	誤り
①	196	24	25	196	24	25	191	22	33
②	151	6	25	149	6	27	148	6	28
③	263	0	18	263	0	18	261	0	20
④	260	15	83	260	15	83	261	18	79
⑤	182	10	11	182	10	11	182	10	11
⑥	102	0	7	104	0	5	98	0	11
⑦	136	0	4	136	0	4	136	0	4
合計	1,290	55	173	1,290	55	173	1,277	56	186
正解率	78.90%			78.90%			78.10%		
準正解率	82.26%			82.26%			81.53%		

表 4-23　人手評価対象文 420 文におけるスコア別カウント結果（パラメータ 10/7/4）

文献	パラメータ 10		パラメータ 7		パラメータ 4	
S	307	73.1%	307	73.1%	301	71.7%
A	57	13.6%	57	13.6%	56	13.3%
B	39	9.3%	39	9.3%	40	9.5%
C	8	1.9%	8	1.9%	12	2.9%
D	9	2.1%	9	2.1%	11	2.6%

表 4-24　処理時間比較（パラメータ 10/7/4）

ﾊﾟﾗﾒｰﾀ	評価（7 文献対）での処理時間	試算（2 万文献対）
10	・ベクトルデータ作成(BGE): 38 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:15 日(5GPU) ・アライメント処理:4 日 ●合計：19 日
7	・ベクトルデータ作成(BGE): 24 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:10 日(5GPU) ・アライメント処理:4 日 ●合計：14 日
4	・ベクトルデータ作成(BGE): 16 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:6 日(5GPU) ・アライメント処理:4 日 ●合計：10 日

[48] 正解率、準正解率はそれぞれ正解文、正解文＋見出し連結文の、総文数 1,635 に対する比率を示す。

【パラメータ値７】

パラメータ①及び②の値を１０から７に下げた結果、評価対象文献７文献、1,635 文中、アライメント結果に変化が生じたのは４文（文献②の明 99〜100、⑥の請 09〜10）のみであった。このうち２文（文献②明 99〜100）は評価が低下（正解⇒誤り）したが、残る２文（⑥請 09〜10）は改善（誤り⇒正解）しているため、トータルの正解文数や正解率は変化しなかった。この結果から、パラメータ値を１０から７に下げることで、アライメント精度を損なわずに処理時間を短縮させることが可能であると結論される。具体的には、前掲の［処理時間比較］表に示したとおり、2 万文献を対象とした総処理時間で約 5 日間（19 日→14 日）の短縮、すなわち約 26.3%の効率化となると試算される。

【パラメータ値４】

パラメータ①及び②をデフォルト値の４まで下げた結果、1,635 文中 36 文でアライメント結果に変化が生じた。これにより正解文が 13 文減少し、誤りが 13 文増加して、正解率は 78.90%から 78.10%に 0.80 ポイント低下した。人手評価文 420 文における A 以上の構成率も 86.7%から 85.0%に低下しており、パラメータ値 10 と比較すると若干の精度低下が見られた。

一方、パラメータ値を大きく下げたことで処理効率は大幅に改善し、パラメータ値 10 と比較して 2 万文献の総処理時間は 19 日から 10 日とほぼ半減（約 47.4%の短縮）する。本調査では実質的なアライメント精度の維持を優先し、パラメータ値４は不採用とするが、処理効率の観点からは、正解率のわずかな低下（-0.80 ポイント）とのトレードオフで総処理時間を半減させることができ、きわめて実用的な設定値といえる。

4.5.5.3. パラメータ①及び②の追加調整結果

前項のパラメータ調整結果から、パラメータ値を 7 に設定することでパラメータ値 10 と同等の処理結果をより効率よく得られること、そしてパラメータ値 4 ではさらに処理効率が良くなるものの、文アライメント精度には若干の低下が生じることが確認された。このため、パラメータ値 10 と同等の精度を維持しつつ、最良の処理効率を得られるパラメータ設定は 7 ～ 5 の範囲となる。

この結果を受け、パラメータ 6 および 5 についても同様の評価を実施した。以下、結果を示す。

表 4-25　目視判定による文アライメント正解率比較（パラメータ 7/6/5）

文献	パラメータ 7			パラメータ 6			パラメータ 5		
	正解	見出し連結	誤り	正解	見出し連結	誤り	正解	見出し連結	誤り
①	196	24	25	192	24	29	190	24	31
②	149	6	27	149	6	27	147	6	29
③	263	0	18	263	0	18	261	0	20
④	260	15	83	260	15	83	260	15	83
⑤	182	10	11	182	10	11	182	10	11
⑥	104	0	5	104	0	5	100	0	9
⑦	136	0	4	136	0	4	136	0	4
合計	1,290	55	173	1,286	55	177	1,276	55	187
正解率	78.90%			78.65%			78.04%		
準正解率	82.26%			82.02%			81.41%		

表 4-26　人手評価対象文 420 文におけるスコア別カウント結果（パラメータ 7/6/5）

文献	パラメータ 7		パラメータ 6		パラメータ 5	
S	307	73.1%	306	72.9%	302	71.9%
A	57	13.6%	57	13.6%	56	13.3%
B	39	9.3%	38	9.1%	42	10.0%
C	8	1.9%	8	1.9%	9	2.1%
D	9	2.1%	11	2.6%	11	2.6%

表 4-27　処理時間比較（パラメータ 7/6/5）

ﾊﾟﾗﾒｰﾀ	評価（7 文献対）での処理時間	試算（2 万文献対）
7	・ベクトルデータ作成(BGE): 24 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:10 日(5GPU) ・アライメント処理:4 日 ●合計：14 日
6	・ベクトルデータ作成(BGE): 22 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:9 日(5GPU) ・アライメント処理:4 日 ●合計：13 日
5	・ベクトルデータ作成(BGE): 19 分 ・アライメント処理(vecalign):2 分	・ベクトルデータ作成:8 日(5GPU) ・アライメント処理:4 日 ●合計：12 日

【パラメータ値 6】

パラメータ値を 7 から 6 に下げた結果、アライメント結果に変化が生じたのは評価対象文 1,635 文中 4 文（文献①の明 224～5、232～3）のみであった。いずれも正解から誤りに劣化している。これにより全体の正解率は 78.90%から 78.65%へ、0.25 ポイントの低下となった。一方、2 万文献の推定処理時間はパラメータ 7 の 14 日からさらに 1 日（7.1%）短縮される。

【パラメータ値 5】

パラメータ値 5 では、パラメータ値 7 との比較で 14 文のアライメント結果に変化が生じ、変化量が多くなりつつある。うち 12 文が正解から誤りに劣化しており[49]、正解率は 78.04%と 0.86 ポイント低下した。この正解率はパラメータ値 4 より若干（-0.006 ポイント）ながら悪化している。ただし、パラメータ値 7 からの変化量自体は大幅に少なく（パラメータ値 4 の変化文数は 36 文)、今回の結果は偶然の偏りと見るべきであろう。ここまでの調査で、パラメータ値を下げたことによる変化の大半は劣化となることが示されており、より多くのサンプルで比較すれば､パラメータ値 4 よりは高い正解率となる可能性が高い。なお、推定処理時間はパラメータ 7 から 2 日間（14.3%）の短縮となる。

[49] 残り 2 文（文献②明 99～100）は、パラメータ値 10 の時点から誤りであり、パラメータ値 5 でも（アライメント結果は変化したが）誤りであったため、カウント上は不変であった。

4.5.5.4. パラメータ調整の結論

BGE のパラメータ①及び②のパラメータ値を 10 から 7、6、5、4 に下げた際の精度と処理効率を調査した結果、パラメータ値 7 であれば精度を低下させずに処理効率を大幅に向上させられることが確認された。またパラメータ 6 以下でも、正解率の低下は 1 ポイント未満であるのに対し、処理効率は 5 日（26.3%）～9 日（47.4%）の大幅な短縮が果たされ、いずれも十分な実用性を有する設定であることが確認された。

この結果を踏まえ、改めてパラメータ①及び②の意味について考察する。前述したとおり、パラメータ①はベクトル化データでオーバーラップさせる文の最大数を設定するもの、パラメータ②はアライメント可能な対訳文（本調査では日本文とインドネシア文）の上限を定めるものである。パラメータ②をパラメータ①より大きな値に設定することは許容されていない。

たとえばパラメータ①、②とも 5 に設定した場合、ベクトル化データは（パラメータ①により）日本文、インドネシア文とも最大 5 文をオーバーラップさせた候補が生成される。そして、これらベクトル化データ同士の文アライメントは、パラメータ②の設定により、両者の合計で 5 文、すなわち 1 ： 4 から 4 ： 1 まで許容される[50]。

ただし、通常の特許文献で、一方が 1 文で記述している内容について、他方で 4 文以上を費やすような極端なバランスの文対応はほぼ皆無と予測される。仮に存在しても、そのような文対は学習データとして利用性が極めて低い。したがって、本パラメータ設定は、事実上、文分割処理の失敗によって過分割された文のリカバリをどの程度までカバーするかを設定するものとなる。たとえば、インドネシア文が 5 つに過分割されていた場合、パラメータ①の値が 5 以上であればその 5 つをまとめたベクトル化データ候補が生成でき、それが選ばれれば過分割は完全にリカバリされる。反面、パラメータ値が 4 以下であると、ベクトル化データとして 5 つをまとめた候補が生成されないため、リカバリは不可能となる。

さらに、この（5 分割をまとめた）ベクトル化データ候補が日本文 1 文とアライメントされるには、パラメータ②の数値は 6 （ 5 文＋ 1 文）以上でなければならない。

これまでの調査で、過分割の大半はインドネシア文で発生することがわかっている。したがって、パラメータ①及び②の設定は、実質的には「過分割されたインドネシア文を文アライメントでリカバリ（再連結）できる範囲」を定めるものと見なせる。具体的には、設定値を、リカバリさせる過分割の上限プラス 1 とする（5 分割までリカバリさせるには、6 に設

[50] 0 ： 5 や 5 ： 0 という特殊な文対応を除く。

定する）形となる。

今回の結果を見ると、パラメータ値を 6 以上に設定すると、文分割結果はほとんど変化しなくなる。つまり、インドネシア文が 5 つ以上に過分割される確率は約 0.2〜0.4%とごく小さく、これらが失われても、対訳コーパスの取得量への影響は無視できる範囲と考えられる。

これに対し、2 万文献の推定処理時間ではパラメータ値 6 は 7 に比べて約 1 日（7.1%）の効率化が図れる。これらから総合的に判断すると、日インドネシア特許文献の対訳コーパス作成においては、パラメータ値 6 が精度と処理効率のバランスが最も優れた設定であると結論される。

BGE のパラメータ②設定値（10、7、6、5、4）ごとの文アライメント結果の詳細は別添資料⑥「文アライメント人手評価結果」の「BGE パラメータ変更」シートを参照されたい。

５．日インドネシア語の対訳コーパスの作成と人手確認・修正、分析

日インドネシアのファミリー文献から取得したテキストデータ（⇒第 2 章）に対し、本調査に最適と判断した文分割手法（⇒第 3 章）と文アライメント手法（⇒第 4 章）を用いて、日インドネシア語の対訳コーパスを作成した。本章にその詳細をまとめる。

5.1. 対訳コーパス作成の概要

本調査では、以下の手順に沿って 2 種類の対訳コーパスを作成した。

① ファミリー文献から取得したインドネシア語及び日本語のテキストデータを、第 3 章で最適と判断した文分割手法である pySBD（インドネシア語）と Stanza（日本語）を用いて分割し、インドネシア語および日本語の文単位データを取得する。

② インドネシア語及び日本語の文単位データを、第 4 章で最適と判断した文アライメント手法（BGE、パラメータ値 6）を用いて日インドネシア対訳文データを取得する。

③ 日インドネシア対訳文データの各文対に対し、第 4 章 4.5.4.項で設定した「対応精度を示す指標」、具体的には日インドネシア文長比と BGE が付した文アライメントスコアの範囲に基づき、対応精度が良好な順にスコア A～D に分別する。

④ スコア A が付された対訳文データを、文アライメントスコアの良好な順、すなわち文対応精度が最も優れていると見なされる順にソートし、上位 100 万文対を『対訳コーパス A』に選定する。

⑤ スコア A 及び B を付された対訳文データ、すなわち対応精度が十分に良好と見なされる文対の全件を『対訳コーパス A-B』に選定する。

⑥ 『対訳コーパス A』100 万文対中、相対的に対応精度が低いと推定される 7 万文対に対して、人手による確認・修正作業を行う。

⑦ 上記⑥で修正対象となった文対の傾向を分析し、『対訳コーパス A』の他の文対に適用可能な修正方法があれば、これを適用する。

⑧ 『対訳コーパス A』及び『対訳コーパス A-B』に対し、上記⑥～⑦で修正された全文対を修正後の内容に置換して、『対訳コーパス A＋』『対訳コーパス A-B＋』とする。

次項より、上記手順①～⑧の実施結果を順に示す。

5.2. インドネシア語及び日本語テキストデータの文分割処理結果

本工程では、ファミリー関係にある日インドネシア特許文献（公報全文またはフロントページ）由来のテキストデータに対して文分割処理を施し、インドネシア語、日本語それぞれの文単位データを取得した。

5.2.1. 対象データの件数

文分割の対象とした特許文献テキストデータ件数は下表のとおりである。原則、公報 1 文献あたり、全文データであれば「発明の名称」「要約」「明細書」「特許請求の範囲」各 1 件、フロントページであれば「発明の名称」と「要約」それぞれ 1 件のテキストデータが取得されることになる。

表 5-1　項目別 文分割対象テキストデータ件数（インドネシア／日本）

項目	インドネシア	日本
発明の名称	66,036 件	64,555 件
要約	62,996 件	64,447 件
明細書	20,667 件	19,244 件
特許請求の範囲	20,860 件	19,234 件

なお、文献によっては項目が欠落しているものや、インドネシア文献において各項目の検出の手がかりに用いた見出し語が不使用又はイレギュラーな記載であったために項目を検出できなかったもの、そして日本文献において電子出願以前の公報でありテキストが取得できないものや再公表公報が存在しないものが存在したため、各項目とも日本とインドネシアのデータ件数に相違が生じている。

5.2.2. 文分割の結果

前項に示した言語・項目別のテキストデータについて、個別に文分割処理を施し、それぞれ文単位データに加工した。文分割用ツールは、インドネシア語には pySBD、日本語には Stanza を使用した。処理結果を次ページの表に示す。

表 5-2　項目別 文分割処理結果（インドネシア／日本）

項目	インドネシア	日本
発明の名称	66,461 文	64,558 文
要約	183,764 文	187,687 文
明細書	8,449,992 文	7,526,556 文
特許請求の範囲	487,176 文	450,733 文

5.2.3. 極端な長文データの除外

前項に示した文単位データ件数は、一部、極端な長文を除外したものである。これは、主にインドネシア文献由来のテキストデータにおいて、日本のテキストデータに存在しない長大な文が存在することに対処として実施した措置である。

日インドネシアのファミリー文献からそれぞれテキストを取得したところ、明細書部分において、インドネシア語のテキスト量が日本語に対して極端に多く、両者にアンバランスが生じていることが判明した。原因を調査したところ、明細書末尾に添付されることの多い「配列表」や「表」のデータがインドネシア語のみでテキスト化されることが原因とわかった。

日本文献の場合、配列表や表はイメージデータ化されており、テキストデータとして採取されない。これに対し、インドネシア文献は、配列表や表データも全てテキスト化されており、このため PDF からのテキスト採取時に本文と区別なく対象となってしまっていた。

こうした配列表や表データは、日本文献からは（イメージデータであるため）テキスト抽出されず、文対として成立しないため、対訳コーパスとならない。このため、コーパス作成においては排除すべきノイズである。

加えて、極端な長文は、たとえ対訳コーパス化されたとしても、機械翻訳の学習データとして使用する際には排除するのが一般的である。極端な長文を学習させると、一般的な文長の文の翻訳精度がむしろ悪化するためである。

さらに、テキストデータのボリュームは後続の文アライメント（特にベクトルデータ作成）処理の所要時間に直接影響する。このため、極端な長文は文アライメント処理前にあらかじめ除外しておくのが最も効率的である。

本調査では、上記判断に基づき、取得した文単位データから「文長が 300 文字を超える」ものを、インドネシア文、日本文を問わず一律に除外した。前項に示したカウント値は、これらを除外した後の数値である。

除外された長文は、インドネシア 15,875 文、日本文 3,022 文であった。それぞれの内訳はインドネシア文では明細書 13,502 文、特許請求の範囲 2,372 文、要約 1 文であり、日本文は明細書 2,061 文、特許請求の範囲 961 文であった。インドネシアの明細書由来の文に大きく偏っているのは、前述した配列表や表データのテキストデータの混入によるものである。

5.3. 日インドネシア文の文アライメント処理結果

5.3.1. 文アライメント処理結果

文アライメント処理では、同一文献から取得したインドネシア語と日本語それぞれの項目（「発明の名称」「要約」「明細書」「特許請求の範囲」）別の文単位データを、文アライメントツールである BGE を用いて対応付け、日インドネシア文対に加工した。下表に、項目別に取得した文対数を示す。

表 5-3　項目別 文アライメント処理結果（取得文対数）

項目	取得文対数
発明の名称	64,333 件
要約	170,069 件
明細書	440,776 件
特許請求の範囲	7,122,173 件
合計	7,797,351 件

5.4. 「対応精度を示す指標」による対応精度スコア A～D の付与結果

前項で取得した全文対 7,797,351 件に対し、本調査で定めた下記「対応精度を示す指標」に則り、対応精度の良否に応じてスコア A～D を付与した。

5.4.1. 対応精度スコア A～D の定義

日インドネシア各文対の対応精度のスコア A～D は、以下のとおりとした。

・文対の文の内容が、90%以上一致している（スコア A）
・文対の文の内容が、80%～89%の範囲で一致している（スコア B）
・文対の文の内容が、50%～79%の範囲で一致している（スコア C）
・文対の文の内容が、49%以下しか一致しない（スコア D）

5.4.2. 対応精度を示す指標

各文対の対応精度（スコア A～D）を示す指標は、第 4 章での検討結果に基づき、以下のとおり定めた。

・スコア A：文長比（インドネシア文長÷日本文長）は 2.3 以上 3.1 未満
　　×　文アライメントスコア 0.8000 未満*

・スコア B：文長比範囲 1.5 以上 2.3 未満 × 文アライメントスコア 0.4000 未満*
　　又は　文長比範囲 3.1 以上 3.55 未満 × 文アライメントスコア 0.8000 未満*

・スコア C：文長比範囲 1.5 以上 2.3 未満 × 文アライメントスコア 0.4000 以上 0.8000 未満*
　　又は　文長比範囲 3.55 以上 4.0 未満 ×文アライメントスコア 0.8000 未満*

・スコア D：スコア A～C に該当しない全ての文対

＊スコア 0.0000 は対象外（スコア D）とした。

5.4.3. 対応精度スコア（A～D）別の文対数

上記指標に則り各文対に該当する対応精度スコアを付した。結果を下表に示す。

表 5-4　文対応スコア別の該当文対数及び項目別内訳

スコア	該当文対数	発明の名称	要約	特許請求の範囲	明細書
A	3,118,196	19,732	67,469	160,790	2,870,205
B	1,861,863	10,805	32,884	40,808	1,777,366
C	1,211,696	12,875	28,075	113,761	1,056,985
D	1,605,596	20,921	41,641	125,417	1,417,617
総計	7,797,351	64,333	170,069	440,776	7,122,173

なお、日本語・インドネシア語とも完全一致する重複文対については、2 件目以降は無条件でスコア D とした。これは、上記スコアに基づいて作成する『対訳コーパス A』及び『対訳コーパス A-B』において、同内容の文対の重複を排除するためである。これにより、567,244 文対が重複排除によりスコア D となった。

5.5. 『対訳コーパス A』及び『対訳コーパス A-B』の作成結果

5.5.1. 『対訳コーパス A』

『対訳コーパス A』は、対応精度スコア A を付された文対中で最も精度のよい 100 万文対を選抜して作成した。文対間の精度比較は、文アライメントツール BGE が各文対に付した文アライメントスコアを用い、スコア値が良い順（数値が小さい順）にソートして上位 100 万文を選定した。

『対訳コーパス A』100 万文対の詳細を以下に示す。

表 5-5　『対訳コーパス A』100 万文対の項目別内訳

項目	文対数	構成比
発明の名称	2,425 文対	0.2%
要約	17,449 文対	1.7%
特許請求の範囲	6,593 文対	0.7%
明細書	973,533 文対	97.4%
合計	1,000,000 文対	100.0%

続いて、本調査で作成した対訳コーパス全件の文対数とその内訳を示す。この構成比が、日インドネシア特許文献から取得される各項目の実際の構成比である。

表 5-6　本調査で作成した対訳コーパス全文対の項目別内訳

項目	文対数	構成比
発明の名称	64,333 文対	0.8%
要約	170,069 文対	2.2%
特許請求の範囲	440,776 文対	5.7%
明細書	7,122,173 文対	91.3%
合計	7,797,351 文対	100.0%

双方を比較すると、全文対の構成比に比べて『対訳コーパス A』では明細書の構成率が 6.1 ポイント高く、その他の三項目（発明の名称、要約、特許請求の範囲）の構成率はそれぞれ-0.6、-0.5、-5.0 ポイント低い。

『対訳コーパス A』の構成比が全文対での構成比よりも低い項目は、精度の良い対訳コーパスが得にくい項目と見なせる。本調査においては特許請求の範囲（5.7%→0.7%）の構成率の低下が特に目立つ。この結果は、特許請求の範囲から取得された文対は、他の項目に比べて文対応精度が低いものが多いことを示している。これは、特許請求の範囲の性質上、ファミリー文献間であっても他の項目より各請求項の内容や順序などが変化し易いこと、及び記載様式や文構造が国ごとに異なる（例えば日本の請求項は体言止めで書かれるが他国はそうではない等）ことが主な理由と考えられる。要約や発明の名称に関しても、後者の理由、すなわち記載様式等が国により異なるという性質はある程度有する。これに対し、明細書は自由記載で特段の様式を有さないため、ファミリー文献間での対応が総じて他の項目より良好であると推測される。

5.5.2. 『対訳コーパス A-B』

『対訳コーパス A-B』は、対応精度スコア A 及び B が付された全文対とした。内訳を以下に示す。

表 5-7 『対訳コーパス A-B』全文対の項目別内訳

項目	文対数	構成比
発明の名称	30,537 文対	0.6%
要約	100,353 文対	2.0%
特許請求の範囲	201,598 文対	4.0%
明細書	4,647,571 文対	93.3%
合計	4,980,059 文対	100.0%

『対訳コーパス A-B』の各項目の構成比はちょうど『対訳コーパス A』とコーパス全文対との中間といえる。『対訳コーパス A』に比して明細書の構成率が 4.1 ポイント下がり、他の三項目は順に 0.4、0.3、3.3 ポイント上がって、それぞれ全文対の構成比に接近している。

各コーパスの選定方法により、全文対の平均的な対応精度は『対訳コーパス A』、『対訳コーパス A-B』、全文対の順に低くなっている。その前提でこれら 3 種のコーパスにおける各項目の構成比の推移は、前項で述べた項目間の相対的な対応精度の良否と整合しており、考察内容が妥当であることを示している。

5.6. 人手確認・修正作業結果

『対訳コーパス A』100 万文対中、相対的に対応精度が低いと見なされる 7 万文対に対し、人手による内容の同一性確認と、内容に不一致がある場合は同一内容となるよう人手による修正を実施した。本項にその詳細を示す。

5.6.1. 人手確認対象 7 万文対の選定

本調査では、100 万文対の『対訳コーパス A』のうち、対応精度が相対的に低いと見られる 7 万文対をピックアップして、文対の内容が過不足なく対応しているかの確認と、内容が同一でない場合には同一となるように修正を行った。

人手確認・修正作業の対象 7 万文対は、基本的には『対訳コーパス A』100 万文対の中で文アライメントスコアが最も下位のものから順に選定する予定であった。しかしながら、人手評価において指摘された内容不一致の状況から、一部、物理的な条件で優先的に確認・修正対象とすべき不備が検出された。

人手評価で対応精度スコア A：「文対の文の内容が、90%以上一致している」と判定された文対、すなわち、「ほぼ正しく対応しているが、スコア S のように 100%一致してはおらず、若干（10%未満）の相違がある」と判定された文には、一つの顕著な不一致のパターンが存在した。それは、「インドネシア文末の数値＋ピリオドが欠落して次文の冒頭に付される」ことによる不一致である。以下、一例を示す。

［文献③明 234：文末の数値＋ピリオドが欠落し次文冒頭に付されるケース］

ｲﾝﾄﾞﾈｼｱ文	Pemrosesan bagan alir ini dimulai sebagai respons terhadap peralatan komunikasi 102 yang memulai pemrosesan di S505 pada Gambar	
日本文	本フローチャートの処理は、通信装置１０２が図５のＳ５０５の処理を開始したことに応じて開始される。	
スコアとその理由	A	日本文の「図５」に対応すべきインドネシア文末の「Gambar 5.」の「5.」が欠落（次文の冒頭と連結されている）。

本例では、インドネシアの文末は本来、日本文の「図５」に対応した「Gambar 5.」であるべきだが、実際には末尾の「5.」が消失して「Gambar」となっている。消失した「5.」は、次のインドネシア文の冒頭に（あたかも箇条書き番号のように）連結されている。このため、本文、次文とも日本文との小規模な不一致があり、人手による修正が必要である。

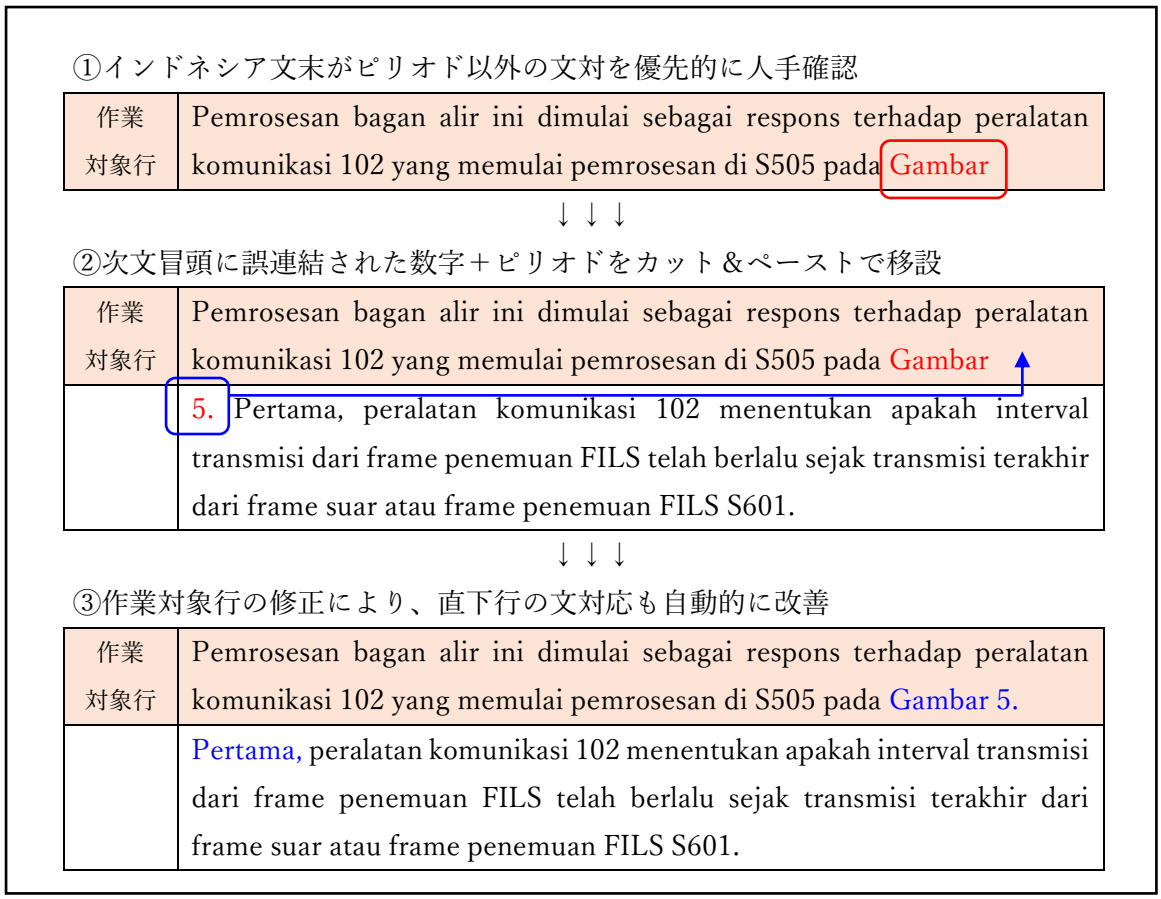
この不備は、本調査用として選定したインドネシア文分割手法 pySBD の課題であり（⇒3.5.1.3.）、本番運用でも一定の頻度で発生することが予測される。事実、人手評価対象文のうち 5.4.1.項に定めた指標で『対訳コーパス A』と判定される 176 文対のうち 6 文対、約 3.4%がこの不備に該当しており、単純計算では、100 万文対の『対訳コーパス A』のうち 3.4 万文程度がこの不備に該当する可能性がある。

人手確認・修正作業の対象を『対訳コーパス A』のうち文アライメントスコアの下位 7 万文としたとき、この不備が発生した全ての文が確実に含まれる保証はない。文字数としてはごく小規模な相違であり、文長比、文アライメントスコアとも極端な数値とならない可能性が高いためである。実際、上掲した事例の文アライメントスコアは 0.4201 と『対訳コーパス A』のボリュームゾーンに属し、ソート順で下位 7 万文対には含まれない見込みである。

スコア A と判定された文対にも一定量含まれることとなるこの不備を確実に人手修正の対象とするには、「インドネシア文の文末がピリオドで終わらないもの」という条件を設け、これに該当するものを優先的に選定するのが効率的である。これにより、不備が発生した全ての文対を確実に修正対象とすることができる。かつ、修正作業を次図のように「次文の冒

頭部の数字＋ピリオドをカット＆ペーストで本文末尾に移動させる」形で行わせることで、次文対の不一致も（修正対象 7 万文対に選ばれているか否かによらず）自動的に解消される。

図 5-1　インドネシア文末の数字＋ピリオド過分割の人手修正イメージ



ただし、「インドネシア文の末尾がピリオドで終わらない」という条件には、発明の名称のほぼ全件も合致する。したがって本調査では、発明の名称はこの条件での選定の対象外とした。

このように、人手確認修正作業の対象文は、まずは『対訳コーパス A』100 万文対から「インドネシア文の末尾がピリオドで終わらない（ただし発明の名称を除く）」という条件で 8,679 文対を選定した。その後、これらを除いた 991,536 文対を文アライメントスコア順にソートし、下位の 61,321 文対（スコア範囲 0.0864～0.3795）を選定して全 7 万文対とした。

なお、仮に 7 万文対すべてを文アライメントスコアの下位順に選定した場合、「インドネシア文末がピリオド以外」の 8,679 文対のうち選ばれるのは 957 文対のみであり、残る

7,722 文対は選外となる。この結果からも、「インドネシア文末の数値＋ピリオド直前の誤分割」は文アライメントスコアには表れにくい不備であり、「インドネシア文の末尾がピリオド以外である」という物理的条件で特定する必要があったとわかる。

下二表に、人手確認対象に選定した文対の種別と、項目ごとの内訳を示す。

表 5-8 『対訳コーパス A』人手確認対象 7 万文対の選定条件別件数

選定条件	文対数
① インドネシア文末がピリオド以外	8,679 文対（12.4%）
② 文アライメントスコア下位順	61,321 文対（87.6%）
合　　計	70,000 文対

表 5-9 『対訳コーパス A』人手確認対象 7 万文対の項目別内訳

項目	文対数
発明の名称	215 文対（0.3%）
要約	1,820 文対（2.6%）
明細書	902 文対（1.3%）
特許請求の範囲	67,063 文対（95.8%）
合計	70,000 文対

5.6.2. 人手確認・修正作業の方針

人手確認・修正作業は、原則「文対の実質的内容が一致しているか否か」を基準とし、不一致である場合は、双方の内容が同一となるよう修正を施した。修正に際しては、『対訳コーパス A』の機械翻訳の学習データとしての利用性を最良とすることを基本方針とした。

本調査で作成する対訳コーパスは、日本やインドネシアの特許文献を機械翻訳するための学習データとして用いられることが想定される。機械翻訳の学習データは「翻訳対象の文献で使われる用語や表現、文体で書かれている」必要がある。その点、本調査の対訳コーパスの各文対は実際の日インドネシア特許文献から採取されたものであり、（文対応に不備がない限り）この要件を完璧に満たす。

その一方で、特許対訳コーパスは、文対が必ずしも逐語訳になっているとは限らない。実質的な記載内容は同じでも、使われる言い回しや語の出現順、文体、文数（日本文は 1 文だがインドネシア文は 2 文で書かれている等）など、言語ごとの性質の違いや国ごとの

記載様式の違いなどにより、翻訳の観点から見て逐語訳（直訳文）とはなっていないものも多数存在する。

しかしながら、これを完璧な逐語訳、直訳文に修正することは、実際の特許文献から採取した文に大幅な改変を加え、「特許文献には存在しない文」に書き替えることにつながる。特許文献に存在しない文は、程度にもよるが、学習データとしての価値が損なわれる危険性がある。

このことから、人手確認・修正作業における「内容の一致／不一致」の判定、すなわち「修正の要否」の判断は、文対が「逐語訳であるか」ではなく、「双方から読み取れる実質的な内容が同一であるか否か」を基準とした。そのうえで、修正を要すると判断した場合は、なるべく原文の特許文としての特徴を損なわない最小限の修正を行うよう努めた。

5.6.3. 修正された文対数

人手確認対象の 7 万文対のうち、不備が検出された文対数は以下のとおりであった。

表 5-10 『対訳コーパス A』人手確認対象 7 万文対の不備の有無別カウント結果

項目	不備あり	不備なし	合計
7 万文対全件	13,748 文対	56,252 文対	70,000 文対
	19.6%	80.4%	－
①インドネシア文末がピリオド以外	6,507 文対	2,172 文対	8,679 文対
	75.0%	25.0%	－
②文アライメントスコアが下位	7,241 文対	54,080 文対	61,321 文対
	11.8%	88.2%	－

7 万文対全件のうち、何らかの不備が検出された（つまり何らかの修正がなされた）文対は 13,748 文対、不備発生率は 19.6%であった。選定条件別の内訳では「①インドネシア文末がピリオド以外」の 8,679 文対での不備発生率が 75.0%（6,507 文対）と極度に高く、「②文アライメントスコア下位」61,321 文対での発生率 11.8%（7,241 文対）と大きな差が生じている。

この結果から、本調査で用いた対訳コーパス作成手法においては「インドネシア文末がピリオド以外」である文対に不備が多いことが確認された。つまり、この条件で文対を特定し修正を施す（または除外する）ことで、対訳コーパス全体の対応精度を効率的に改善することが可能である。

上記集計結果に基づいて『対訳コーパス A』100 万文対全件における不備発生率を試算すると、まず「①インドネシア文末がピリオド以外」に属する文は人手確認対象とした 8,679 文対が『対訳コーパス A』における全件であり、ここでの不備発生数 6,507 文対が不備の総数となる。一方、それ以外の全文対、すなわちインドネシア文末がピリオドである 991,321 文対における不備発生率は、「②文アライメントスコア下位」61,321 文対での不備発生率 11.8%以下であると考えられる[51]。したがって、991,321 文×11.8%＝116,975 文対を、末尾ピリオド以外の 6,507 文対と合算した 123,482 文対、つまり全体の 12.3%程度が、『対訳コーパス A』の推定不備率の最大値と見なせる。

試算された『対訳コーパス A』の不備発生率 12.3%は低い数値とは言い難い。ただし、これらの不備には、誤字脱字やピリオドの打ち忘れといった原文自体の問題に起因する（つまり文分割や文アライメント処理の問題ではない）ものも多数含まれる[52]。かつ、文長比や文アライメントスコアに表れない小規模・軽微な不備が多いと考えられる。人手確認・修正作業で検出された不備の類型別分析は、次章（第 6 章）で詳述する。

5.6.4. 派生的に修正された文対数

対訳コーパスにおける不一致は、インドネシア文末の数字＋ピリオドの誤分割に代表されるような文分割の不備や、対応させるべき文がずれてしまう文アライメントの不備など、前後の文から文字列を移動させたり対応のずれを正したりするだけで（つまり、文言の追加や変更を要さずに区切り方の修正のみで）修正できるものが多い。このため人手確認・修正作業では、人手確認の対象文対だけでなく、文献全体を作業者に提供し、必要に応じて前後の文からカット＆ペーストで修正を行わせる原則とした。

その結果、5.6.1.項の図 5-1 に示したように、人手確認対象行を修正することにより、その直前または直後の文も自動的に修正されるケースが多数発生した。

このように派生的に修正された文対の総数は 6,767 件であり、このうち『対訳コーパス A』に属する文対が 689 件、『対訳コーパス A-B』に属する文対が 4,451 件(『対訳コーパス A』の 689 件を含む)存在した。

---

[51] 11.6%は 991,321 文対のうち文アライメントが最も低かった 61,321 文対での不備発生率であり、残る 931,321 文対の文対応はより良好と推定されるため、不備発生率も低下する可能性が高い。

[52] 後述の 6.3.5.項の試算では、原文データが原因の主要な不備の発生率は 9.0%程度と見られ、全不備の発生率 12.5%の約 72%を占めている。

なお、修正の結果、人手確認対象行の前後行において文対が双方とも空欄となる場合がある[53]。このような実体のない文対は各コーパスから除外する必要がある。本調査では、上記 6,767 件とは別に 15 件発生した。このうち『対訳コーパス A』に属する文対が 1 件存在したため、当該文対を除外した。

5.6.5. 修正結果に基づき追加で対応した文対数

人手確認修正作業で修正された各文対の修正内容を確認した結果、いくつか、発生の有無を機械的に推定可能な不備パターンが検出された。これを受け、下表に示す①～③の 3 種の不備パターンについて、『対訳コーパス A』及び『対訳コーパス A-B』から該当する文対を特定し、それぞれ対処（修正または除去）を行った。

表 5-11 『対訳コーパス A』及び『対訳コーパス A-B』の不備パターン別追加対応文対数

不備パターン	対処内容	コーパス A	コーパス A-B
① 文冒頭に段落番号が存在	修正	9,939 文対	84,002 文対
② 英文の除去	除去	229 文対	1,084 文対
③ インドネシア文に行番号が混入	修正	19 文対	194 文対

※コーパス A-B の文対数にはコーパス A の文対数も含めている。

各不備パターンの詳細については、第 6 章の 6.4 項「検出された不備パターンへの機械的対応」で詳述する。

[53] 例えばインドネシア文が過分割されて 2 行に分かれており、日本文は過分割されず 1 行目のみにセットされている（つまり 2 行目の日本文欄は空行）ような場合、2 行目のインドネシア文を 1 行目の末尾にカット＆ペーストする修正が施される。その結果、2 行目は日本文、インドネシア文とも空欄となる。

5.7. 『対訳コーパス A＋』及び『対訳コーパス A-B＋』の作成結果

5.5.項に記載した内容で作成した『対訳コーパス A』及び『対訳コーパス A-B』に対し、5.6.項に示した人手確認・修正作業の各種結果を反映して『対訳コーパス A＋』及び『対訳コーパス A-B＋』を完成させた。

具体的には、各コーパスに対して、下表のとおりの修正・除外を行った。

表 5-12 『対訳コーパス A』及び『対訳コーパス A-B』の修正内容別件数

修正内容		コーパス A （1,000,000 文対）	コーパス A-B （4,980,059 文対）
人手修正	⇒5.6.3.	13,748 文対	13,748 文対
派生的に修正	⇒5.6.4.	689 文対	4,451 文対
修正による空行の除外		1 文対 削除	1 文対 削除
減少分補充		1 文対 補充	－
追加修正① 段落番号カット	⇒5.6.5 ※詳細は 6.4 項参照	9,939 文対	84,002 文対
追加修正② 英文除去		229 文対 削除	1,084 文対 削除
減少分補充		229 文対 補充	－
追加修正③ 行番号カット		19 文対	194 文対
修正後件数		1,000,000 文対 （コーパス A＋）	4,978,974 文対 （コーパス A-B＋）

派生的に修正された文対のうち、文対の双方が空欄となったため『対訳コーパス A＋』から除外されたものが 1 件存在した（⇒5.6.4.）。本調査では、『対訳コーパス A+』の総数を 100 万文対に保つため、『対訳コーパス A-B＋』中の『対訳コーパス A＋』に属していない文対から文アライメントスコア最上位の 1 文対を『対訳コーパス A＋』に補充した。

また、追加修正②で行った英文の除去においても、『対訳コーパス A＋』から 229 文対、『対訳コーパス A-B＋』から 1,084 文対が除去された。これに対しても、『対訳コーパス A＋』の総数を 100 万文対に維持するため、『対訳コーパス A-B＋』の中から『対訳コーパス A＋』に属していない文対を文アライメントスコア上位順に 229 文対補充した。

こうして『対訳コーパス A＋』の総数を『対訳コーパス A』と同じ 100 万文対に維持した。一方、『対訳コーパス A-B＋』の総件数は、双方空欄の文対 1 文対と、英文として除去された 1,084 文対、計 1,085 文対が『対訳コーパス A-B』（4,980,059 文対）から減少し、全 4,978,974 文対となった。

6．人手確認対象 7 万文対における不備の傾向

人手確認・修正作業（⇒5.6.）では、対象 7 万文対について、内容の不一致が検出された場合（すなわち何らかの修正を要した場合)、その不一致の原因となった不備を判定する作業を実施した。本章にその結果をまとめる。

6.1 不備の類型化

人手確認・修正作業における不備判定に備え、対訳コーパスにおいて発生し得る代表的な不備の類型化を行った。具体的には、文アライメント手法候補の人手評価（第 2 の評価。⇒4.5.3.）で評価対象とした 420 文について、不備の状況を精緻に分析し、カテゴライズした結果、不備の主要な類型として以下があることが確認された。

① 文分割の不備
　①-1 過分割
　①-2 見出し語の不適な連結
　①-3 文末の数字＋ピリオド過分割
② 文アライメントの不備
③ 文内容の相違
　③-1 内容の意図的な変更
　③-2 文章や語句の誤り
　③-3 様式の相違
　③-4 ヘッダ・フッタ等の混入（テキスト抽出エラー）
　③-5 データ形式に起因する相違
④ その他の相違

まず本項で、各類型の説明と具体例を示す。

6.1.1.「① 文分割の不備」

文対応の不正の原因には、文献単位のテキストデータを一文単位に区切る「文分割処理（⇒第 3 章)」の失敗に起因するものがある。これを「① 文分割の不備」とした。

具体的には、インドネシア文、日本文のいずれかで（または双方で)、文が一文単位に適切に区切られておらず、途中で切れていたり、余計な部分が付随している（その結果として、文対の実質的内容が相違している）ケースを指す。

例えば下例の「対象行」（人手確認対象となった行。以下同じ）は、インドネシア文に対応する日本文が存在しない。インドネシア文は文法的に中途半端であり、前行と連結することでようやく一文として成立し、かつ前行の日本文とも同内容となる。

	Terutama, dalam "arah tinggi", sisi di mana blok silinder ditempatkan relatif terhadap penutup engkol diacu sebagai sisi	特に、「上下方向」において、相対的にクランクキャップに対してシリンダブロックが位置する側を上方、相対的にシリンダブロックに対してクランクキャップが位置する側を下方と称する。
対象行	ditempatkan atas, dan relatif sisi terhadap di blok mana penutup engkol diacu sebagai silinder sisi bawah.	

こうした状況は、インドネシア文献の文分割の失敗によるもの（本来 1 文として扱われるべきものが不当に区切られた）と考えられ、「文分割の不備」に相当する。「文分割の不備」に該当する文対は、このように文法的に異常な区切られ方になっているものが大半となる。

なお、「文分割の不備」の典型パターンとして、以下に示す 3 種のサブカテゴリを設け、不備判定においては、このうち最も合致するものを選択することとした。

6.1.1.1.「①-1 過分割」

「過分割」とは、一文が複数に分断され、その結果として文対が不対応となっているケースを指す。具体的には、（ア）当該文が過分割されて内容の一部が前後の行に誤連結されている場合と、（イ）前後の文が過分割されて内容の一部が当該行に誤連結されている場合とに大別される。以下、実例を示す。

文１	Biji bunga matahari menurut klaim 1, dimana sekuens tersebut panjangnya 400 sampai 800 basa. Biji bunga matahari menurut salah satu dari klaim 1 sampai 3, dimana sekuensnya terdiri dari (a) atau (b) di bawah ini:	前記配列が 400～800 塩基長である、請求項１に記載のヒマワリ種子。
文２	(a) sekuens basa yang diwakili oleh SEQ ID NO: 1, atau	前記配列が、下記（a）又は（b）：（a）配列番号１で示される塩基配列、又は（b）配列番号１で示される塩基配列に対して 80%以上の同一性を有する塩基配列を含む配列である、請求項１～３のいずれかに記載のヒマワリ種子。
文３	(b) sekuens basa yang memiliki setidaknya 80% identitas dengan sekuens yang diwakili oleh SEQ ID NO: 1. Biji bunga matahari menurut klaim 4, dimana identitasnya paling sedikit 90%.	前記同一性が 90%以上である、請求項４に記載のヒマワリ種子。

上例では、日本の「文２」と対応付けられるべきインドネシア文（赤字部分）が過分割され、その一部が「文１」の末尾と「文３」の冒頭に誤連結されている。したがって本例は、「文２」が前記（ア）「当該文が過分割されて内容の一部が前後の行に誤連結されている」ケース、「文１」と「文３」が前記（イ）「前後の文が過分割されて内容の一部が当該行に誤連結されている」ケースにそれぞれ該当する。よって、文１～文３のいずれが人手確認対象行であっても、不備判定は「①-1 過分割」となる。

6.1.1.2.「①-2 見出し語の不適な連結」

本調査で作成する対訳コーパスは、テキストデータ抽出の必要上、見出し語と本文とが連結される事象が頻発することが避けられない（⇒3.5.1.4.）。

こうした「見出し連結」は、厳密には文分割の失敗といえるが、現状の対訳コーパス作成技法においては不可避の事象であり、かつ見出し語が双方の文頭にある限り内容の不一致とはならないため、文中において見出し語と本文との区切りが判別できれば、不備とは見なさない方針とした。

ただし、見出し語と本文との区切りが不明瞭（見出し語がカッコ等で括られておらず、本文と直接連結されている）であるなど、見出し語と本文との未分割が原因で文対が不自然な状態となっており、何らかの修正を要する場合は、不備と見なして「①-2 見出し語の不適な連結」と判定した。典型例を示す。

文 1	Alat Deteksi 207 Semua peralatan yang dapat menjalankan program, dapat melakukan konfirmasi dan perbaikan informasi server dengan perangkat serupa, termasuk PC, smartphone, dan tablet yang dapat terhubung ke Internet sebagai perangkat yang digunakan di toko, dan saat digunakan tidak ada batasan jumlah perangkat	207 端末機店舗で使用する機器でインターネット接続ができる PC、スマートフォン及びパッドを含む類似機器。サーバーの情報を照会し修正できるプログラムが作動できるすべての装備が可能であり、使用台数も原則的に制限がない。

インドネシア文の冒頭の「Alat Deteksi 207」は見出し語である（下図参照）。日本文冒頭の「207 端末機」も同じく見出し語である。

図 6-1　インドネシア PDF 公報 P00202303466 の見出し語の状況

Alat Deteksi 207

Semua peralatan yang dapat menjalankan program, dapat melakukan konfirmasi dan perbaikan informasi server dengan perangkat serupa, termasuk PC, smartphone, dan tablet yang dapat terhubung ke Internet sebagai perangkat yang digunakan di toko, dan saat digunakan tidak ada batasan jumlah perangkat

上図のとおり、PDF 公報では見出し語（黄マーカ）と本文とは改行され、分離されている。しかしながら、日インドネシア対訳コーパスでは、テキストデータ抽出処理の必要上、見出し語が直後の本文と連結されてしまう。その結果、本例のように（カッコや末尾記号を伴わず）文字のみで記述された見出し語であると、上例のように、見出し語と本文との区切りが不明瞭な、不自然な文となる。これは日本語見出し「207 端末機」も同様である。

こうした場合、実質的な内容の不一致はないものの、文として不自然であり修正を要することから「①-2 見出し語の不適な連結」と不備判定した。なお、修正は原則、双方から見出し語を除去する方針とした。

6.1.1.3.「①-3 文末の数字＋ピリオド過分割」

本調査に用いたインドネシア文分割手法 pySBD には、インドネシア文末尾が「数字＋ピリオド」である場合、その直前で過分割され、次文の冒頭に誤連結される傾向が見られる（⇒5.6.1.）。これに該当するケースは「①-3 文末の数字＋ピリオド過分割」とした。

文 1	Kasus dimana nilai rata-rata dari rasio rongga adalah 5% atau kurang ditentukan sebagai rasio rongga yang rendah, dan ditunjukkan oleh “◎” pada Tabel	ボイド率の平均値が 5 %以下の場合には、ボイド率が低いと判断し、表 1 では「◎」で示している。
文 2	1. Kasus dimana nilai rata-rata dari rasio rongga melebihi 5% ditentukan sebagai tidak rendah dalam rasio rongga, dan ditunjukkan oleh “○” pada Tabel 1.	ボイド率の平均値が 5 %超過の場合には、ボイド率が低くはないと判断し、表 1 では「○」で示している。

上例では、「文 1 」のインドネシア文末尾が「Tabel 1.（日本文の「表 1 」に対応)」で区切られるべきところ、文末の数字＋ピリオド（つまり「1.」）が過分割され、次文の冒頭に（あたかも箇条書き番号のように）誤連結されている。結果的に、文 1， 2 とも日本文との対応に過不足が生じており、修正（「1.」の文 1 末尾への移設）が必要である。

このような場合は、文 1 、文 2 のいずれが確認対象行であっても「文末の数字＋ピリオド過分割」と判定される。なお、pySBD による文分割結果には、「数字＋ピリオド」以外にも、英数字付番など、次文の冒頭に配置すると箇条書き番号のように見える文字列が文

末にある場合、これが過分割されて次文冒頭に誤連結される傾向が見られた。根本的には同種の不備であるが、「数字＋ピリオド」以外は「①-1 過分割」に分類した。

6.1.1.4.「未分割」の扱い

文分割処理における典型的な不備としては、区切られるべき箇所で分割がなされず、結果として複数文が連結されたままとなる「未分割」も挙げられる。

しかしながら、対訳コーパスに加工された後のデータから、未分割の有無を判定することは困難である。仮にある文対においてインドネシア文が２文であったとしても、それが未分割によるものか、それとも分割されていた２文が文アライメント処理で再連結された結果かを判別することはできない。

かつ、本調査の対訳コーパスではＮ：Ｎの対応づけが許容されており、たとえ一方が複数文であったとしても、他方も複数文が対応づけられ、双方の内容が合致していれば、正常な文対として取り扱われる。つまり「未分割」は、仮に発生したとしても、後続の文アライメント処理でリカバリ可能な性質を有する。

つまり、こうした状況で文対の不一致が発生した場合、その原因は未分割であるとも、文アライメント処理でのリカバリ失敗であるともいえる。本調査においては、混乱を避けるため、こうした場合は全て「② 文アライメントの不備」として取り扱うこととし、「未分割」は不備判定の選択肢から外した。

6.1.2.「② 文アライメントの不備」

インドネシア文、日本文とも文分割は正しく行われている（適切に一文単位に区切られている）にもかかわらず、両者の対応づけが前後にずれているような場合は、「文アライメント（文の対応づけ）の不備」に該当する。

文アライメントの不備の多くは、インドネシア文と日本文とが１：１の対応関係でなく、同じ内容を双方が異なる文数で記述していることによって生じる。以下、典型的な事例を示す。

文１	Oleh karena itu, dalam perwujudan ini, rakitan blok silinder (2) mencakup lima penutup (20) yang disusun sebaris di dalam arah penjajaran dari silinder (11).	したがって、本実施形態では、シリンダブロック組立体２は、シリンダ１１の整列方向に一列に並んだ五つのキャップ２０を備え、これらキャップ２０は上下方向に見たときに、クランクシャフト３の軸線方向（シリンダ１１の整列方向）において各シリンダ１１の両側に一つずつ配置される。
文２	Penutup (20) disusun sedemikian sehingga salah satu dari penutup (20) dipasang pada masing-masing dari kedua sisi dari masing-masing silinder (11) di dalam arah aksial dari poros engkol (3) (arah penjajaran dari silinder (11)) ketika dilihat dalam arah tinggi.	

インドネシア文も日本文も一文単位に適切に文分割されているが、両者の対応づけが誤っているため「文アライメントの不備」に相当。

文アライメントが正しいか否かは、「ファミリー文献の各文が、最も適切なかたちで対応づけられているか否か」で判断される。よって、上例のように同じ内容が異なる文数で書かれている場合には、異なる数の文が対応づけられているのが正しい状況である。上例であれば以下が本来あるべき形である。

文１	Oleh karena itu, dalam perwujudan ini, rakitan blok silinder (2) mencakup lima penutup (20) yang	したがって、本実施形態では、シリンダブロック組立体２は、シリンダ１１の整列方向に一列に並んだ五つ

	disusun sebaris di dalam arah penjajaran dari silinder (11). Penutup (20) disusun sedemikian sehingga salah satu dari penutup (20) dipasang pada masing-masing dari kedua sisi dari masing-masing silinder (11) di dalam arah aksial dari poros engkol (3) (arah penjajaran dari silinder (11)) ketika dilihat dalam arah tinggi.	のキャップ２０を備え、これらキャップ２０は上下方向に見たときに、クランクシャフト３の軸線方向（シリンダ１１の整列方向）において各シリンダ１１の両側に一つずつ配置される。

なお、「ファミリー文献の各文が、最も適切なかたちで対応づけられているか否か」という基準からは、文対の内容が完全に同一でなくとも、記載内容がおおむね類似していたり、文献中の出現順序（たとえば前後行は正しく対応している）などから見て最も妥当な文と対応づけられている限り、内容不一致の原因は文アライメント処理ではないため、「文アライメントの不備」とは判定されない。一例を示す。

文１	Gambar 16A adalah tampak yang menunjukkan konfigurasi dari penutup engkol pusat sesuai dengan modifikasi ketiga; Gambar 16B adalah tampak yang menunjukkan konfigurasi dari penutup engkol samping sesuai dengan modifikasi ketiga;	図１６は、第３変形例に係る中央クランクキャップ及び側方クランクキャップの構成を示す図である。
文２	Gambar 17A adalah tampak yang menunjukkan konfigurasi dari penutup engkol pusat, penutup engkol samping, dan penutup engkol antara sesuai dengan modifikasi keempat; Gambar 17B adalah tampak yang menunjukkan konfigurasi dari penutup engkol pusat, penutup engkol samping, dan penutup engkol antara sesuai dengan modifikasi keempat;	図１７は、第４変形例に係る中央クランクキャップ及び側方クランクキャップと中間クランクキャップとの構成を示す図である。

文３	Gambar 18A adalah tampak yang menunjukkan konfigurasi dari penutup engkol pusat, penutup engkol samping, dan penutup engkol antara sesuai dengan modifikasi kelima; Gambar 18B adalah tampak yang menunjukkan konfigurasi dari penutup engkol pusat, penutup engkol samping, dan penutup engkol antara sesuai dengan modifikasi kelima;	図１８は、第５変形例に係る中央クランクキャップ及び側方クランクキャップと中間クランクキャップとの構成を示す図である。

上例は、「文１」の日本文が「図１６」の説明であるのに対し、対応するインドネシア文は「図 16A（Gambar 16A)」と「図 16B（Gambar 16B)」をそれぞれ説明した２文が連結されている。文２～３についても同様の状況である。

各文対とも実質的内容に相違があるため、修正を要する。このため何らかの不備判定が必須となる。だが、文アライメント処理に関しては、前後の文の内容や文献中での位置関係から見て、現状の対応づけが最も妥当である。このため、本例は文アライメントの不備と見なすべきではない。(後述の③「文内容の相違」の「③-1 内容の意図的な変更」と判定するのが適切である。)

6.1.3.「③ 文内容の相違」

「文内容の相違」は、「文対の実質的内容が同一ではない（したがって修正を要する)」ことを指す。ただし、前述した「文分割の不備」や「文アライメントの不備」がその根本原因であれば、そちらに判定するため、不備判定における「文内容の相違」は、「文分割処理やアライメント処理の失敗によるものではない、修正を要する内容の相違全般」を指すものとなる。

典型的なケースとしては、インドネシアと日本の文献の一方のみに、全く新しい文が書かれているようなケースが挙げられる。この場合、双方の文献が正しく文分割されていても、アライメント処理で対応づけるべき文が他方に存在しない。結果として、一方が空欄となった文対（例１)、もしくは「新しい文」が直前や直後の文に連結された文対（例２）が発生することとなる。

［インドネシア文献に新規文（青字）が存在した場合の文対の現れ方］

例 1	Contoh "kopolimer" meliputi …… meskipun kopolimer tidak terbatas padanya.	当該「共重合体」として、例えば……などが挙げられるが、当該共重合体は、これらに限定されない。
	Monomer dapat tertaut silang tambahan atau sejenisnya disukai monomer akrilamida.	（空欄）

例 2	Contoh "kopolimer" meliputi …… meskipun kopolimer tidak terbatas padanya. Monomer dapat tertaut silang tambahan atau sejenisnya disukai monomer akrilamida.	当該「共重合体」として、例えば……などが挙げられるが、当該共重合体は、これらに限定されない。

ただし、例 1 のように一方が空欄であったり、例 2 のように一方の文の冒頭や末尾に他方と対応しない文が存在する文対は、「文アライメントの不備」や「文分割の不備」が生じた場合にも発生し得る。つまり、不備の種類は当該行のみを見ても特定できず、常に前後行の状態を見て判断する必要がある。

例えば、前記②で説明した「文アライメントの不備」によっても、作業対象行の外見が例 1 の 2 文目や例 2 と同じになる場合はあり得る。だが、文アライメントの失敗であれば、こうした場合も新規文（Monomer dapat……monomer akrilamida.）に対応する日本文は前後行に存在する。これに対し、「文内容の相違」、すなわち新規文（Monomer dapat……monomer akrilamida.）がインドネシア文献のみに存在するような場合は、これに対応する日本文は前後行にも存在しない。対応させるべき文が存在しない以上、この対応不正はアライメント処理が原因ではなく、「文アライメントの不備」には該当しない。

換言すれば、「文内容の相違」とは、前述の 2 種の不備（「① 文分割の不備」「② 文アライメントの不備」）と異なり、一方の文献に存在する文や語句が何らかの理由により他方の文献には存在せず、その結果、文対に修正を要するレベルの内容の相違が生じているケース全般を指す。

なお、「5.6.2. 人手確認・修正作業の方針」の項で述べたとおり、文の内容の一致、不一致は、「双方が逐語訳（直訳文）であるか否か」ではなく、「双方から実質的に同一の内容が読み取れるか否か」で判断される。たとえば、一方が 1 文で表現している内容を他方では複数文で表現しているような場合、それは直訳文の関係ではないが、双方から実質的に

同じ内容が読み取れる限り、「文内容の相違」とは判定されない。文数の違いが許容される以上、当然ながら文構造や文法上の相違や、言語の性質に基づく相違、内容理解にほぼ影響しない語句の有無（たとえば文脈上明らかな語句の省略など）も許容される。

文内容の相違が発生する主な原因は複数想定される。このため、不備判定の選択肢として、「③ 文内容の相違」の下階層にも５つのサブカテゴリ（③-1～③-5）を設けた。判定作業では、文内容の相違に該当する場合、これら５つのサブカテゴリから該当するものが適用される。

6.1.3.1.「③-1 内容の意図的な変更」

一方の文献のみに文もしくはその一部が追加（または一方のみで削除）されている場合など、双方の文で内容が意図的に変更されている場合は、「内容の意図的な変更」に相当する。

文１	dan Gambar 20 adalah bagan alir yang menunjukkan prosedur pembuatan dari rakitan blok silinder yang dilengkapi dengan penutup yang ditunjukkan di dalam Gambar 3A, Gambar 3B, Gambar 4A, Gambar 4B.	図２０は、図３及び図４に示したキャップを備えるシリンダブロック組立体の製造手順を示すフローチャートである。

上例の文対の内容はほぼ同じだが、日本文が「図３及び図４」の説明なのに対し、インドネシア文は「図 3A、3B、4A、4B（Gambar 3A, Gambar 3B, Gambar 4A, Gambar 4B)」の説明となっている。このように双方の文の内容が意図的に（つまり誤記などではなく）変更されており、合致させるために修正を要するものは、「③-1 内容の意図的な変更」と判定される。

6.1.3.2.「③-2 文章や語句の誤り」

一方、明らかな誤記や誤訳などによって双方の内容が（意に反して）異なってしまっている場合は、「文章や語句の誤り」と判定される。

文 1	Di sisi lain, dan jurnal keempat (31#4), pemiringan jurnal (31) itu kecil, sehingga jurnal tersebut kecil kemungkinan berkontak dengan bantalan engkol (22). Dengan demikian, rugi-rugi gesekan karena kontak sebagian menjadi lebih kecil.	一方、４番ジャーナル３１＃４では、ジャーナル３１は傾きが小さいことからクランク軸受２２には接触しにいため、部分的な接触に伴う摩擦損失は小さい。

上記文対の内容は実質的に同一であるが、日本文に「接触しにい」という記載ミスがあり、「接触しにくい」と修正する必要がある。このような意図せぬ誤記や誤訳などは、実質的内容の相違とは言い難いものが大半だが、修正が必須であるため、「内容の相違」の一種と見なし「③-2 文章や語句の誤り」と不備判定する。

本例と前項③-1 の事例は、ともに「文内容の相違」ではあるが、本例の相違は意図せぬ誤記であり、前例③-1 のような「内容の意図的な変更」とは性質が異なる。このため、別のサブカテゴリとして取り扱うこととした。これは後述する③-3～③-5 も同様である。

6.1.3.3.「③-3 様式の相違」

たとえば一方のみに見出しや行番号などが付されているなど、国ごとの公報様式の違いや、作成者の記述の好みなどにより実質的な内容に差異が生じ、修正を要する場合は、不備判定は「③-3 様式の相違」とした。一例を示す。

文 1	Konfigurasi dari mesin pembakaran dalam yang dilengkapi dengan rakitan blok silinder sesuai dengan perwujudan ini akan dijelaskan dengan mengacu pada Gambar 1 dan Gambar 2.	＜内燃機関の構成＞図１及び図２を参照して、本実施形態に係るシリンダブロック組立体を備える内燃機関の構成について説明する。

上例の本文はインドネシア文、日本文とも同じ内容だが、日本文のみ「＜内燃機関の構成＞」という見出しが付随しており、両文の内容に大きな差異が生じている。このため、見出しを除去する修正が必要である。修正を要する相違であるため不備判定が必須となるが、本例は「③-2 文章や語句の誤り」ではなく、かつ「③-1 内容の意図的な変更」とも

性質が異なる。国ごと、作成者ごとの明細書の記載様式の違い等に起因する自然発生的な差異の一種と見なすのが妥当である。こうしたケースを分類するため、「③-3 様式の相違」というサブカテゴリを設けた。

上例は「①-2 見出し語の不適な連結」と混同しやすいが、「見出し語の不適な連結」はあくまで「文分割の不備」のサブカテゴリであり、見出し語がうまく分割できていないことが問題視される。これに対し上例は、見出し語が一方の文献のみでしか使われていないという根本的な内容不一致に起因しており、文分割以前の問題である。

6.1.3.4.「③-4 ヘッダ、フッタ等の混入（テキスト抽出エラー）」

本調査では、特許文献からテキストを抽出する際、ヘッダやフッタ、ページ番号などを機械的に除去している（⇒2.2.2.）。ただし、場合によっては、こうした不要な文字列が本文に混入してしまうケースも考えられる。「③-4 ヘッダ、フッタ等の混入（テキスト抽出エラー）」は、この状況を想定して設定した。以下、この不備の発生イメージを示す。

文 1	Cerukan (12) dan bantalan engkol (13) disusun sebaris di dalam arah aksial dari poros engkol (3). Lebih lanjut, ketika dilihat dalam arah tinggi, satu cerukan (12) dan satu bantalan　－ 55 －　engkol (13) disediakan pada setiap sisi dari masing-masing silinder (11) di dalam arah aksial dari poros engkol (3).	凹部１２及びクランク軸受１３は、クランクシャフト３の軸線方向において一列に並んで配置されると共に、上下方向に見たときに、クランクシャフト３の軸線方向において各シリンダ１１の両側に一つずつ配置される。

対訳コーパスを作成するためのテキストデータは、特許文献から抽出される。その際、ページ中の本文以外の文字記号、たとえばヘッダやフッタ、ページ番号や段落番号、脚注などが誤って抽出され、本文中に混入してしまう可能性がある。上例は、インドネシア文の「bantalan　engkol　（クランク軸受け）」の途中に「－ 55 －」というページ番号が混入しているケースを模したものである。

この場合、「－ 55 －」を除去して正しい文に修正することとなるが、不備のカテゴリとしては「③-1 内容の意図的な変更」、「③-2 文章や語句の誤り」、「③-3 様式の相違」のいずれにも該当しない。このため、別途「③-4 ヘッダ・フッタ等の混入(テキスト抽出エラー)」という選択肢を設けた。

6.1.3.5.「③-5 データ形式に起因する相違」

本調査では、インドネシア文は PDF 公報からテキスト抽出して取得している。このため、日本公報では（イメージデータとなっているため）テキスト化されない箇所が、インドネシア文のみでテキスト化される場合がある。主なものとして、表・テーブルや、配列表などが挙げられる。以下、一例を示す。

文 1	[Tabel 1] Komposisi Patri (% massa) Rumus Struktur (akustik) TCT Rongga Sn Ag Cu Bi Sb Co Fe As Ni Contoh 1 sisa 3,4 0,7 3,2 3 0,008 0,025 - - 0,049 ◎ Contoh 2 sisa 3,1 0,7 3,9 3,2 0,006 0,028 - - 0,046 ◎ - Contoh 3 sisa 4,0 0,8 2,9 3 0,01 0,03 - - 0,060 ◎ - Contoh 4 sisa 3,5 0,6 4,2 5 0,008 0,025 - - 0,049 ◎ - Contoh 5 sisa 3,4 0,8 3 3,6 0,01 0,024 - - 0,054 ◎ - Contoh 6 sisa 3,3 0,7 1,5 4,2 0,01 0,026 - - 0,056 ◎ - Contoh 7 sisa 3,3 0,7 5,5 2,8 0,006 0,026 - - 0,044 ◎ - Contoh 8 sisa 3,5 0,7 4,8 1,0 0,008 0,025 - - 0,049 ◎ - Contoh 9 sisa 3,5 0,7 2,8 6,0 0,01 0,028 - - 0,058 ◎ - Contoh 10 sisa 3,4 0,6 2,8 3 0,001 0,025 - - 0,028 - Contoh 11 sisa 3,3 0,7 3,4 2,8 0,03 0,027 - - 0,117 - Contoh 12 sisa 3,4 0,7 4 3,2 0,01 0,02 - - 0,050 ◎ - Contoh 13 sisa 3,3 0,8 3,6 3,2 0,01 0,05 - - 0,080 ◎ - Contoh 14 sisa 3,5 0,7 3,8 2,8 0,008 0,026 0.004 - 0,050 ◎ ◎ Contoh 15 sisa 3,5 0,7 3,4 2,8 0,002 0,02 - - 0,026 - Selain itu, telah berhasil dikonfirmasi bahwa hasil uji siklus panas (TCT) dianggap sebagai “◎”, yang bahkan lebih baik, dalam aspek dimana jumlah (Fe + 3 × Co) dari nilai % massa Fe dan nilai tiga kali % massa Co (ditunjukkan sebagai “Rumus” pada Tabel-tabel 1 dan 2) adalah 0,03 sampai 0,1.	また、Ｆｅの質量%の値とＣｏの質量%の３倍の値との和（Ｆｅ＋３×Ｃｏ）（表１及び表２では「式」として示している。

上例は、日本文と対応するインドネシア文は末尾の「Selain itu,…sampai 0,1.」のみであり[54]、冒頭から延々と続く赤字部分は、対応する日本文が存在しない。これは、日本公報ではイメージデータであるためテキストとして抽出されなかった「表１」が、インドネシア公報ではテキスト化されていたために抽出され、結果として不一致となっているものである。

このような公報のデータ形式（主にイメージデータとテキストデータの相違）に基づく内容の不一致と見なせる場合は、「③-5 データ形式に起因する相違」とし、一方のみに存在する内容を除去する修正を施した。イメージデータ部分がテキスト化された場合、上例のように解釈困難な語の羅列となることが多く、対訳コーパスとして無益有害なものとなるためである。上例の場合、インドネシア文から「表１」に該当する部分を削除する修正が施される。

6.1.4.「その他の相違」

修正を要するレベルの内容の相違であるが、ここまでに定義したいずれの不備カテゴリ（①-1〜①-3、②、③-1〜③-5）にも該当しないものが発生する可能性も考えられる。こうした場合に備え、「④ その他の相違」を別途設けた。

6.1.5. 複数の不備が発生している場合

たとえば文分割の不備（①）が発生した場合、連鎖的に文アライメントも不正（②）となり、内容も不一致（③）となるというケースが考えられる。このように複数の不備が連鎖的に発生している場合は、より根源的な不備（すなわち①＞②＞③）を優先して判定した。

なお、このような連鎖的なものではなく、１つの文対で別種の不備が同時発生しているケースも想定される。例えば、同じ文対で意図的な内容の変更（③-1）と誤記（③-2）が併発するようなことも考えられる。このような場合は、判定者の判断により、内容の相違をもたらす重大度が最も大きいと感じる不備を選択した。

[54] この部分に関しても双方の内容に少なからぬ不一致が存在するが、前後の関係からは両者を対応づけるのが文アライメント処理としては最善である。

6.2 類型別の不備の発生状況

人手確認・修正の対象とした 7 万文対における各類型の不備の発生状況を本項にまとめる。

6.2.1. 人手確認対象 7 万文対における類型別の不備判定結果

人手確認対象 7 万文対において不備判定された文対数を類型別に下表に示す。

表 6-1 『対訳コーパス A』人手確認対象 7 万文対における類型別の不備発生数

不備の類型	発生数と発生頻度	言語別の修正数と比率	
		ｲﾝﾄﾞﾈｼｱ	日本
①-1 過分割	1,449	1,310	338
	2.1%	90.4%	23.3%
①-2 見出し語の不適な連結	1,528	1,364	1,066
	2.2%	89.3%	69.8%
①-3 文末の数字＋ピリオド過分割	3,313	3,313	331
	4.7%	100.0%	10.0%
② 文アライメントの不備	498	450	102
	0.7%	90.4%	20.5%
③-1 内容の意図的な変更	1,367	915	534
	2.0%	66.9%	39.1%
③-2 文章や語句の誤り	3,490	3,271	292
	5.0%	93.7%	8.4%
③-3 様式の相違	1,797	582	1,353
	2.6%	32.4%	75.3%
③-4 ヘッダ、フッタ等の混入	5	5	0
	0.01%	100.0%	0.0%
③-5 データ形式に起因する相違	240	224	59
	0.3%	93.3%	24.6%
④その他の相違	61	58	13
	0.1%	95.1%	21.3%
合　計	13,748	11,492	4,088
	19.6%	83.6%	29.7%

表中、「発生数と発生頻度」下段のパーセンテージは 7 万文対における当該不備の発生率である。一方、「修正数と修正率」下段のパーセンテージは、当該不備が発生した全文対におけるインドネシア文と日本文それぞれの修正率を示している[55]。

集計の結果、10 種の不備類型のうち特に発生数が多かったのは「③-2 文章や語句の誤り」3,490 件と「①-3 文末の数字＋ピリオド過分割」3,313 件であった。双方で全不備の 49.5%とほぼ半数を占めており、かなり突出した発生数となっている。

このうち後者「①-3 文末の数字＋ピリオド過分割」については、事前に多発が予測されたため、『対訳コーパス A』100 万文対から、この不備の特徴である「インドネシア文末がピリオドでない」8,679 文対の全てを優先的に人手確認対象とした（⇒5.6.1.）。つまり、あらかじめ標的として対象文対が選定されており、他の不備より発生数が多くなることは当然といえる。

これに対し、発生数第 1 位の「③-2 文章や語句の誤り」は、必ずしもインドネシア文末がピリオドでなくなる性質の不備ではない。したがって、こちらの不備は『対訳コーパス A』の全件（100 万文対）においても、7 万文対とおおむね同程度の発生率となると考えられる。つまり、コーパス全体としては「①-3 文末の数字＋ピリオド過分割」よりもはるかに多発している可能性が高い。なお、「③-2 文章や語句の誤り」と判定された 3,490 文対で修正されたインドネシア文と日本文のカウント数を見ると、前者が 3,271 文対（修正率 93.7%）、後者が 292 文対（同 8.4%）と、インドネシア文に対する修正が圧倒的に多かった。

全類型を合算した 7 万文対全件での不備発生率は、表の最下段に示したとおり 19.6%であった。ただし、この 7 万文対は「インドネシア文末がピリオド以外」の全件など、不備が発生している可能性が高い文対を優先的に選んだ結果であり、この不備発生率を『対訳コーパス A』全 100 万文対にそのまま適用すべきではない。事実、より精密な試算では、『対訳コーパス A』の不備発生率は最大 12.3%程度と推定されることは、5.6.3.項で述べたとおりである。

さらに言えば、10 種に類型化した不備のうち、本調査で用いた文アライメント作成手法が原因のものは、文分割処理の失敗による「①-1 過分割」と「①-3 文末の数字＋ピリオド過分割」、そして文アライメント処理の失敗による「② 文アライメントの不備」の 3 種

[55] 一文対で日本文とインドネシア文の双方が修正される場合もあるため、双方の修正率の合計は 100%を超える場合が多い。

のみといえる。テキストデータ抽出処理を含めても「①-2 見出し語の不適な連結」が加わる程度であり、その他の不備（③-1〜④）はいずれもソースデータ自体の問題である。

6.2.2. 文対選定理由別の不備判定結果

人手確認対象 7 万文対には、「①インドネシア文末尾がピリオド以外であるもの」8,679 文対と「②文アライメントスコアの下位順」61,321 文対という 2 種類の選定理由によるものが混在している。そして 5.6.3.項に示した集計の結果、前者での不備発生率が 75.0%なのに対し、後者は 11.8%と極端な差異が生じていることが判明した。

もともと、選定条件①は、インドネシア文分割手法 pySBD において「文末の数字＋ピリオドの直前での過分割（つまり類型①-3)」の多発が予測されたことから設定されたものであり、この不備に関しては選定条件①に偏って発生するのは当然である。だが、5.6.3.項で示したとおり、選定条件①での不備発生数は全種合計で 6,507 件であり、このうち「①-3 文末の数字＋ピリオド過分割」の発生数は 3,313 件と半数強にとどまる。

この状況は、10 種の不備類型の中に、発生すると「①インドネシア文末がピリオド以外」という状況をもたらす性質のものが他にも存在することを示している。つまり、類型①-3 以外にも「インドネシア文末がピリオド以外」という条件で検出しやすいタイプの不備が存在する可能性が高い。

そこで、前項に示した類型別の不備発生件数を、さらに選定条件別に「①インドネシア文末がピリオド以外」と「②文アライメントスコア下位」とを分けて集計し、各類型における状況を比較した。次表に集計結果を示す。各類型とも、上段が「①インドネシア文末ピリオド以外」、下段が「②文アライメントスコア下位」の集計値である。また、「発生数／発生率」の右欄には、①は全 8,679 文対、②は全 61,321 文対における各不備の発生率を示している。

表 6-2 『対訳コーパス A』類型別不備発生数（イン末尾ピリオド以外／文スコア下位）

不備の類型	条件	発生数／発生率		修正	
				ｲﾝﾄﾞﾈｼｱ	日本
①-1 過分割	①	1,133	13.1%	1,055	206
	②	316	0.5%	255	132
①-2 見出し語の不適な連結	①	636	7.3%	630	281
	②	892	1.5%	734	785
①-3 文末の数字＋ピリオド過分割	①	3,080	35.5%	3,080	311
	②	233	0.4%	233	20
② 文アライメントの不備	①	373	4.3%	359	51
	②	125	0.2%	91	51
③-1 内容の意図的な変更	①	196	2.3%	172	52
	②	1,171	1.9%	743	482
③-2 文章や語句の誤り	①	631	7.3%	622	27
	②	2,859	4.7%	2,649	265
③-3 様式の相違	①	327	3.8%	307	86
	②	1,470	2.4%	275	1,267
③-4 ヘッダ、フッタ等の混入	①	1	0.01%	1	0
	②	4	0.01%	4	0
③-5 データ形式に起因する相違	①	116	1.3%	115	12
	②	124	0.2%	109	47
④その他の相違	①	14	0.2%	14	4
	②	47	0.1%	44	9
合　計	①	6,507	75.0%	6,355	1,030
	②	7,241	11.8%	5,137	3,058

上表のとおり、選定条件①「インドネシア文末がピリオド以外」と②「文アライメントスコア下位」の不備発生率を比較すると、全類型において前者が後者より高くなった。このことから、「インドネシア文末がピリオド以外」という条件設定は、対訳コーパスの不備の絞り込みに極めて有効であることが伺える。

ただしその一方で、選定条件①と②の不備発生率の差異は、類型によって条件①に極端に偏っているもの（①-1～①-3、②、③-5）と、条件①、②ともおおむね近似の発生率であり、このため母集合の大きい条件②のほうが発生件数が多くなっているもの（③-1～③-4、④）とに二分された。

前者、すなわち選定条件①「インドネシア文末がピリオド以外」に極度に偏って発生している不備としては、「①-3 文末の数字＋ピリオド過分割」、「①-1 過分割」、「②文アライメントの不備」の 3 種が挙げられる。それぞれの選定条件①の発生率は条件②の 86.3 倍、25.4 倍、21.5 倍と極端な差異が生じている。もともと条件①のターゲットであった「①-3」が高率になることは当然の結果であるが、①-1 と②については、それぞれ「文の途中（ピリオド以外の箇所）での過分割」、「文末への見出し語の誤連結」といった典型パターンが条件①に合致したことで、高い発生率となったと考えられる。詳しくは 6.3.項「類型別の不備の実例と考察」で分析する。

これら 3 種の不備は、前項で「本調査で用いた文アライメント作成手法が原因」の不備として挙げた 3 種と同一である。こうした（ソースデータの問題ではない）不備が発生した文対を効率的に絞り込む手段として、「インドネシア文末がピリオド以外」という条件（選定条件①）が非常に有効であることがわかる。

6.2.3. 言語別の不備判定結果

前々項及び前項に示した表 6-1、6-2 の右欄には、不備判定された各文対において、インドネシア文と日本文のどちらが修正されたかも示している。

表 6-1 に示したとおり、不備判定された 13,748 文対のうち、インドネシア文は 11,492 文が修正され、これに対して日本文の修正は 4,088 文にとどまり、双方には約 3 倍の開きがある。

7 万文対の選定条件の一方である「①インドネシア文末がピリオド以外」は、もともとインドネシア文の不備を想定したものであった。このため、こちらの条件においては、インドネシア文の修正が大半となるのは妥当である。事実、選定条件①で不備判定された 6,507 文対における修正数はインドネシア文が 6,355 文（構成比 97.7%）とほぼ全件であり、1,030 文（15.8%）であった日本文とは極端な開きがある。

一方、選定条件②「文アライメントスコア下位」で不備判定された 7,241 文対に関しては、インドネシア文の修正が 5,137 文（構成比 70.9%）、日本文の修正が 3,058 文（同 42.2%）であった。選定条件①に比べ双方の差は大幅に縮まったが、依然としてインドネシア文の修正が顕著に多い。

インドネシア文の修正（すなわち不備）が多い理由として、6.2.1.項では「③-2 文章や語句の誤り」に対する修正がインドネシア文に大きく偏っていたことを指摘した。6.4.2.項で後述する英文の混入（「④ その他の相違」に該当）などもこの理由に該当する。

ただし、ここまでの調査分析から、選定条件や類型によらず総じてインドネシア文に不備が多かった理由として、さらに二点考えられる。

一つは、インドネシア文の文区切り記号であるピリオドが、日本の句点と異なり文末以外でもさまざまな用途で使用されるため、テキストデータ上での文末の特定（すなわち文分割）の難易度が高いことである。「①-1 過分割」や「② 文アライメントの不備」が日本文に比べて多発したのは、このことが一因であろう。

もう一つは、インドネシア文が主に PDF 公報由来であったため、テキスト抽出の際、行番号やイメージデータ部分など不要な文字列が日本文よりも混入しやすい状況であった点が挙げられる。「③-4 ヘッダ、フッタ等の混入」や「③-5 データ形式に起因する相違」はこのことが直接の原因である。

このように総じてインドネシア文の修正率が高い中、「①-2 見出し語の不適な連結」と「③-3 様式の相違」のみ、選定条件②での日本文の修正率がインドネシア文を上回っている。それぞれ明らかな理由があり、前者については日本文献のみで使われる見出し語が多かったため（この場合、通常は日本文から見出し語を除去して内容を一致させた)、後者については、本類型の典型パターン「文対の一方のみに段落番号が存在する」ケースの多くが日本文であったため（日本文から段落番号を除去して内容を一致させた）である。

6.3. 類型別の不備の実例と考察

本項にて、7 万文対における各種不備の実例と発生傾向について類型別にまとめる。

6.3.1. 「① 文分割の不備」

文分割の不備に属する 3 種のサブカテゴリ「①-1 過分割」、「①-2 見出し語の不適な連結」及び「①-3 文末の数字＋ピリオド過分割」のそれぞれについて、実例と全体的な傾向を考察する。

6.3.1.1. 「①-1 過分割」

「過分割」は、文分割処理において区切る必要のない箇所で文が分割され、文アライメント処理でもリカバリされなかったケースが該当する。

＜不備の典型例＞

「過分割」に該当する実例には、例えば以下のものがあった。

［P00202003642_JP2021504484A SEQ:31］

修正前	ｲﾝﾄﾞﾈｼｱ	Namun, deregulasi jalur EGFR/EGF dengan ekspresi berlebih atau aktivasi konstitutif mendorong proliferasi sel tumor, invasi, dan berhubungan dengan prognosis yang buruk pada banyak keganasan (Yarden, Y.
	日　本	しかしながら、過剰発現または構成的活性化によるＥＧＦＲ／ＥＧＦ経路の調節解除は、腫瘍細胞の増殖、侵攻を促進し、さらには多くの悪性腫瘍における予後不良と関連付けられている（
修正後	ｲﾝﾄﾞﾈｼｱ	Namun, deregulasi jalur EGFR/EGF dengan ekspresi berlebih atau aktivasi konstitutif mendorong proliferasi sel tumor, invasi, dan berhubungan dengan prognosis yang buruk pada banyak keganasan (Yarden, Y. dan Sliwkowski, M., Ulasan Alam. 2: 127- 137, 2001).
	日　本	しかしながら、過剰発現または構成的活性化によるＥＧＦＲ／ＥＧＦ経路の調節解除は、腫瘍細胞の増殖、侵攻を促進し、さらには多くの悪性腫瘍における予後不良と関連付けられている（Yarden, Y. and Sliwkowski, M., Nature Reviews. 2: 127-137, 2001）。

また、前後行で発生した過分割により、当該行の文頭や文末に前後の文の一部が誤連結されているケースも、「過分割」の一種として多数検出された。一例を示す。

［P00201600911_JP2016525164A SEQ:507］

修正前	ｲﾝﾄﾞﾈｼｱ	Namun, biasanya, suhu permukaan sekitar 20° C sampai 60° C lebih tinggi dari suhu komposisi untuk membentuk suatu film elastomerik. Tahap
	日　本	しかし、通常は、表面温度はエラストマーフィルムを形成するための組成物の温度よりも約 20℃～60℃高い。
修正後	ｲﾝﾄﾞﾈｼｱ	Namun, biasanya, suhu permukaan sekitar 20° C sampai 60° C lebih tinggi dari suhu komposisi untuk membentuk suatu film elastomerik. ~~Tahap~~
	日　本	しかし、通常は、表面温度はエラストマーフィルムを形成するための組成物の温度よりも約 20℃～60℃高い。

さらに、インドネシア文の典型的な過分割パターンにつき、別項目とした後述の「①-3 文末の数字＋ピリオド過分割」と同系統の事例も数多く検出された。一例を示す。

［P00202306616_JP2024503614A SEQ:171］

修正前	ｲﾝﾄﾞﾈｼｱ	Misalnya, suatu antena dari unit antena kedua 210-2 dapat menempatkan suatu interval unit (misalnya, jarak yang diketahui, jarak yang setara dengan panjang gelombang atau beberapa panjang gelombang dari sinyal komunikasi) jauh dari antena dari unit antena pertama 210-1 dalam arah
	日　本	例えば、第２のアンテナユニット２１０－２のアンテナは、第１のアンテナユニット２１０－１のアンテナから x 方向にユニット間隔（例えば、既知の距離、通信信号の波長に相当する距離、又は通信信号の波長の倍数）離れて位置付けられ得る。
修正後	ｲﾝﾄﾞﾈｼｱ	Misalnya, suatu antena dari unit antena kedua 210-2 dapat menempatkan suatu interval unit (misalnya, jarak yang diketahui, jarak yang setara dengan panjang gelombang atau beberapa panjang gelombang dari sinyal komunikasi) jauh dari antena dari unit antena pertama 210-1 dalam arah x.
	日　本	例えば、第２のアンテナユニット２１０－２のアンテナは、第１のアンテナユニット２１０－１のアンテナから x 方向にユニット間隔

		（例えば、既知の距離、通信信号の波長に相当する距離、又は通信信号の波長の倍数）離れて位置付けられ得る。

上例のインドネシア文（修正前）は、文末の「x.」、つまり「英字＋ピリオド」の直前で過分割されている。本調査で用いたインドネシア文分割ツール pySBD には、文末の文字＋ピリオドが、次の文の冒頭に配置するとあたかも箇条書き番号のように見えるものであると、過分割して次文冒頭に誤連結する傾向が見られる。本例の「x.」もこれに相当する。「数字＋ピリオド」ではないため「①-3 文末の数字＋ピリオド過分割」ではなく「①過分割」にカウントしたが、おそらく同じロジックで分割されていると見られる。

日本文の過分割も若干数確認された。その大半は、文末のカッコ補記の途中で過分割されるパターンであった。一例を示す。

［P00202203142_JP2022545457A SEQ:652］

修正前	ｲﾝﾄﾞﾈｼｱ	Secara mengejutkan, mencit menerima EGFR/LGR5 bispesifik, yang terdiri atas MF3755 dan MF5816 + irinotekan pengobatan kombinasi memiliki volume tumor yang lebih rendah dibandingkan semua kelompok mencit lainnya (gambar 3b, 3c).
	日　本	驚くべきことに、ＭＦ３７５５及びＭＦ５８１６＋イリノテカン併用治療を含む、二重特異性ＥＧＦＲ／ＬＧＲ５を受けるマウスは、他の群の全てのマウスと比較すると低い腫瘍体積を有した（
修正後	ｲﾝﾄﾞﾈｼｱ	Secara mengejutkan, mencit menerima EGFR/LGR5 bispesifik, yang terdiri atas MF3755 dan MF5816 + irinotekan pengobatan kombinasi memiliki volume tumor yang lebih rendah dibandingkan semua kelompok mencit lainnya (gambar 3b, 3c).
	日　本	驚くべきことに、ＭＦ３７５５及びＭＦ５８１６＋イリノテカン併用治療を含む、二重特異性ＥＧＦＲ／ＬＧＲ５を受けるマウスは、他の群の全てのマウスと比較すると低い腫瘍体積を有した（図３Ｂ、図３Ｃ）。

上例の日本文（修正前）は、文末にあるべきカッコ補記「（図３Ｂ、図３Ｃ）。」が開きカッコ直前で過分割されている。同様のケース（つまり日本文の末尾が開きカッコで終わっている文対）は、人手確認対象 7 万文中 91 文で見られた。全件が不備と判定され、うち 85 件が「過分割」と判定された。

＜発生頻度＞

人手確認対象 7 万文対における「過分割」の発生頻度を下表に示す。あわせて、7 万文対のうち「インドネシア文の末尾がピリオドで終わらないもの」8,679 文対と、「『対訳コーパス A』における文アライメントスコア下位のもの」61,321 文対のそれぞれにおける発生頻度も示す。

表 6-3　7 万文対における「①-1 過分割」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	1,449	1,310	338
	2.1%	90.4%	23.3%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	1,133	1,055	206
	13.1%	93.1%	18.2%
文アライメントスコア下位のもの（61,321 文対）	316	255	132
	0.5%	80.7%	41.8%

7 万文対における「過分割」の発生数は 1,449 件、発生率は 2.1%であった。この発生率を 100 万文対の『対訳コーパス A』に単純に適用すると、約 2.1 万文対において過分割が発生する試算となる。

ただし、上表に示した内訳のとおり、過分割の発生数 1,449 件のうち約 78%に相当する 1,133 件は「インドネシア文がピリオドで終わらないもの」という条件の文対から検出されている。『対訳コーパス A』中でこの条件に該当する文対は全件が 7 万文対に含まれるため、上記 1,133 件が、『対訳コーパス A』でこの条件に該当する全文対の不備発生数となる。

一方、「文アライメントスコア下位のもの」61,321 文対での過分割の発生件数は 316 件、発生率は 0.5%である。『対訳コーパス A』で人手対象外となった 930,000 文対の中に「インドネシア文がピリオドで終わらないもの」は存在しないため、これらの文対での不備発生率は、「文アライメントスコア下位」61,321 文対と同じ 0.5%程度と考えるべきである。

したがって、『対訳コーパス A』100 万文対のうち、「インドネシア文がピリオドで終わらないもの」8,679 件以外の 991,321 件における「過分割」の発生頻度は 0.5%、4,957 件程度と推定される。これに、「インドネシア文がピリオドで終わるもの」8,679 件での発生

数 1,133 件を足した 6,090 件、発生頻度では 0.6%程度が、『対訳コーパス A』100 万文対における「①-1 過分割」の推定発生件数となる。

なお、過分割はインドネシア文、日本文の双方で発生したが、上表のとおり、全体的にはインドネシア文に偏って発生している（1,310 件、全件の 90.4%）。インドネシア文の過分割の典型パターンである「①-3 文末の数字＋ピリオドの直前での過分割」が別カウントであるにもかかわらず、その他の過分割でもこのような比率となったことから、事実上、文分割処理における過分割の大半はインドネシア文で発生しているといえる。

その大きな理由は、6.2.3.項で指摘した、インドネシア文と日本文の文分割の難易度の差によるものと考えられる。日本語の場合、句点（。）は文区切り専用の記号であるため、句点直後で文分割すればおおむね正解となる。これに対しインドネシア文のピリオド（.）は、文区切り記号以外にも多種多様な使われ方がされるため、文区切り記号か否かの都度の判断が必要となる。現状のインドネシア文分割手法は、本調査で使用した pySBD を含め、この難易度の高さに十分に対応できているとは言い難く、改善の余地を有する。

6.3.1.2.「①-2 見出し語の不適な連結」

本調査の文分割処理の大きな課題として、見出し語と直後の本文が不可避的に連結されてしまう点が挙げられる（⇒3.5.1.4.）。人手確認対象 7 万文対においても、これに該当する文対が数多く見られた。

＜不備の典型例＞

「見出し語の不適な連結」の典型例を以下に示す。

［P00202002649_JPWO2019078208A1 SEQ:14］

修正前	ｲﾝﾄﾞﾈｼｱ	Latar Belakang Invensi Berbagai faktor nutrisi diperlukan untuk pertumbuhan tanaman. Kekurangan beberapa dari faktor tersebut dikenal menghalangi pertumbuhan tanaman.
	日　本	背景技術植物が生長するには種々の栄養要素が必要であるが、そのいくつかの要素が不足すると植物の生育に支障を来すことが知られている。
修正後	ｲﾝﾄﾞﾈｼｱ	~~Latar Belakang Invensi~~Berbagai faktor nutrisi diperlukan untuk pertumbuhan tanaman. Kekurangan beberapa dari faktor tersebut dikenal menghalangi pertumbuhan tanaman.

	日　本	~~背景技術~~植物が生長するには種々の栄養要素が必要であるが、そのいくつかの要素が不足すると植物の生育に支障を来すことが知られている。

上例の文対（修正前）はインドネシア文、日本文とも見出し語（Latar Belakang Invensi／背景技術）と本文が連結されている典型的なケースである。見出し語がカッコで括られていれば本文との区切りが明瞭であるため修正を要さず、不備とならないが、本例はどちらも本文との区切りが不明瞭で、文として不自然である。

このように見出し語と本文とが連結されており、かつ区切りが不明瞭である場合は、文分割の不備と見なして「①-2 見出し語の不適な連結」にカウントした。この不備に該当した文対は、上例のように双方から見出し語を除去し、自然文同士の文対に修正する原則とした。

＜発生頻度＞

人手確認対象 7 万文対における「見出し語の不適な連結」の発生頻度を下表に示す。

表 6-4　7 万文対における「①-2 見出し語の不適な連結」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	1,528	1,364	1,066
	2.2%	89.3%	69.8%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	636	630	281
	7.3%	99.1%	44.2%
文アライメントスコア下位のもの（61,321 文対）	892	734	785
	1.5%	82.3%	88.0%

7 万文対における「見出し語の不適な連結」の発生数は 1,528 件、発生率は 2.2%であった。その内訳を見ると、「インドネシア文がピリオドで終わらないもの」での発生件数は 636 件（7.3%）、「文アライメントスコア下位のもの」での発生件数は 892 件（1.5%）であり、「①-1 過分割」ほど前者への極端な偏りは見られない。見出し語の不適な連結は文頭で発生することが多く、文末のピリオドの有無とは直接関係しないためと考えられる。

全 1,528 文対でのインドネシア文と日本文の修正数はそれぞれ 1,364 件と 1,066 件であり、言語的にも大きな偏りはない。また、見出し語の不適な連結に対しては上例のように双方から見出し語を除去するケースが大半であるため、インドネシア文、日本文ともに修正率は 89.3%、69.8%と高くなっている。

最後に、「過分割」と同様の方式で『対訳コーパス A』における「見出し語の不適な連結」の発生件数を精密に試算する。「文アライメントスコア下位のもの」61,321 件での見出し語の不適な連結の発生件数が 892 件、発生率が 1.5%であるため、『対訳コーパス A』100 万文対のうち「インドネシア文がピリオドで終わらないもの」8,679 件を除く 991,321 件での発生頻度も 1.5%、14,870 件程度と考えるべきである。これに、「インドネシア文がピリオドで終わるもの」全 8,679 件での発生数 636 件を足した 15,506 件、発生頻度では約 1.5%が、『対訳コーパス』100 万文対における「見出し語の不適な連結」の発生予測数となる。

6.3.1.3.「①-3 文末の数字＋ピリオド過分割」

インドネシア文分割手法 pySBD の最も特徴的な過分割のパターンである。インドネシア文末が数字＋ピリオドで終わる場合、その直前で過分割され、文末の数字＋ピリオドが次文冒頭に誤連結される。

＜不備の典型例＞

まずは「文末の数字＋ピリオド過分割」の典型例を示す。なお、着色した行（1 文目）は人手確認対象行であることを示している。

［P00202001912_JP2021042527A SEQ:88］

修正前	ｲﾝﾄﾞﾈｼｱ	Suatu lubang laluan 7H dibentuk sehingga dapat melewati dinding pendukung
		7. Lubang laluan 7H adalah suatu lubang melingkar yang berpusat pada pusat aksial X10.
	日　本	支持壁７には、挿通孔７Ｈが貫通するように形成されている。
		挿通孔７Ｈは、軸心Ｘ１０を中心とする丸穴である。
修正後	ｲﾝﾄﾞﾈｼｱ	Suatu lubang laluan 7H dibentuk sehingga dapat melewati dinding pendukung 7.
		~~7.~~ Lubang laluan 7H adalah suatu lubang melingkar yang berpusat pada pusat aksial X10.
	日　本	支持壁７には、挿通孔７Ｈが貫通するように形成されている。
		挿通孔７Ｈは、軸心Ｘ１０を中心とする丸穴である。

修正前のインドネシア文は、人手確認対象文の文末となるべき数字＋ピリオド「7.」の直前で過分割が発生し、「インドネシア文末がピリオドで終わらない」文対となっている。一方、文末から分離された「7.」は次文の冒頭に誤連結されている。

このような場合は、対象行を「①-3 文末の数字＋ピリオド過分割」と不備判定したうえで、修正後のインドネシア文として示したように、次文冒頭にある数字＋ピリオドを対象行の文末にカット＆ペーストする形の修正を行う。こうすることで、次行も同時に改善させることができる。

なお、下例のように、前文の末尾の数字＋ピリオドが文頭に誤連結された文のほうが人手確認の対象行となる場合もある。こうした場合も、内容の不一致の原因は前文で発生した「文末の数字＋ピリオド過分割」であるため、不備判定は本カテゴリとした。

［P00201809818_JP2019525897A SEQ:1091］

修正前	ｲﾝﾄﾞﾈｼｱ	Puncak protein di-dekonvolusi menggunakan fungsi MassLynx MaxEnt
		1. Kromatografi cairan fase terbalik (LC)-MS Sampel dianalisis dengan menggunakan kromatografi cairan fase terbalik.
	日　本	タンパク質ピークを、ＭａｓｓＬｙｎｘＭａｘＥｎｔ１関数を使用してデコンボリューションした。
		逆相液体クロマトグラフィー（ＬＣ）－ＭＳサンプルを、逆相液体クロマトグラフィーを使用して分析した。
修正後	ｲﾝﾄﾞﾈｼｱ	Puncak protein di-dekonvolusi menggunakan fungsi MassLynx MaxEnt 1.
		~~1.~~ Kromatografi cairan fase terbalik (LC)-MS Sampel dianalisis dengan menggunakan kromatografi cairan fase terbalik.
	日　本	タンパク質ピークを、ＭａｓｓＬｙｎｘＭａｘＥｎｔ１関数を使用してデコンボリューションした。
		逆相液体クロマトグラフィー（ＬＣ）－ＭＳサンプルを、逆相液体クロマトグラフィーを使用して分析した。

＜発生頻度＞

人手確認対象 7 万文対における「文末の数字＋ピリオド過分割」の発生頻度を下表に示す。

表 6-5　7 万文対における「①-3 文末の数字＋ピリオド過分割」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	3,313	3,313	331
	4.7%	100.0%	10.0%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	3,080	3,080	311
	35.5%	100.0%	10.1%
文アライメントスコア下位のもの（61,321 文対）	233	233	20
	0.4%	100.0%	8.6%

上表のとおり、7 万文対における「文末の数字＋ピリオド過分割」の発生数は 3,313 件、発生率は 4.7%と、各種不備のなかでも特に多い（全カテゴリ中第 2 位）。そして、その大多数 3,080 件は「インドネシア文がピリオドで終わらない文対」から検出されている。

これは、そもそも人手確認対象 7 万文対において「インドネシア文がピリオドで終わらない文対」を優先的に選定した理由がこの「文末の数字+ピリオド過分割」への重点対応のためであり、順当な結果である。

一方、「文アライメントスコア下位のもの」、すなわちインドネシア文末がピリオドで終わる 61,321 文対においても、この不備が発生しているものが 233 件検出された。これらは、２つめの典型例として示した「前文末尾から過分割された数字＋ピリオドが当該文の冒頭に誤連結された」ケースにあたる。本来、文末の数字+ピリオドが過分割された文と、過分割された数字＋ピリオドが文頭に誤連結された文は同数発生するが、後者は前者のように機械的に特定できないため、ごく一部のみが文アライメントスコア下位として 7 万文対に選ばれたものと考えらえる。なお、本調査においては、前述のとおり前者が修正されれば後者も自動的に改善されるため、結果的には後者も全件修正されている。

『対訳コーパス A』100 万文対における「文末の数字+ピリオド過分割」の発生予測数を「①-1 過分割」等と同じ論理で試算すると、総発生件数は 7,045 件、発生頻度 0.7%程度となる。7 万文対での不備発生数は対象選定時に主要なターゲットとされたことで突出したが、『対訳コーパス A』全体での推定発生頻度は特段高いものではないことがわかる。

6.3.2. 「② 文アライメントの不備」

インドネシア文、日本文とも文分割は妥当に行われている（一文単位に適当に区切られている）が、文対として双方の対応にずれがある場合は「② 文アライメントの不備」と見なした。まずは典型例を示す。

＜不備の典型例＞

［P00202303472_JP2023543993A SEQ:203］

修正前	ｲﾝﾄﾞﾈｼｱ	Gambar 6(A) menunjukkan letalitas anak mencit dan Gambar 6(B) menunjukkan perbedaan berat badan hewan antara hari ke-4 (P4) dan P7 pasca melahirkan.
		Semua batang galat mewakili SEM *p<0,05 ……h kontrol yang diinjeksi dengan bahan pembawa saja.
	日　本	図６Ａは仔の死亡率を示し、図６Ｂは、出生後４日目（Ｐ４）とＰ７の間での動物体重差を示す。全ての誤差バーはＳ．Ｅ．Ｍ．を表す。
		＊ｐ＜０．０５は、各群の平均値を、……ダネットの事後検定からの結果。
修正後	ｲﾝﾄﾞﾈｼｱ	Gambar 6(A) menunjukkan letalitas anak mencit dan Gambar 6(B) menunjukkan perbedaan berat badan hewan antara hari ke-4 (P4) dan P7 pasca melahirkan.
		Semua batang galat mewakili SEM *p<0,05 ……h kontrol yang diinjeksi dengan bahan pembawa saja.
	日　本	図６Ａは仔の死亡率を示し、図６Ｂは、出生後４日目（Ｐ４）とＰ７の間での動物体重差を示す。~~全ての誤差バーはＳ．Ｅ．Ｍ．を表す。~~
		全ての誤差バーはＳ．Ｅ．Ｍ．を表す。＊ｐ＜０．０５は、各群の平均値を、……ダネットの事後検定からの結果。

上例（修正前）は、インドネシア文１文が日本語２文に対応付けられ、１：２の文対となっている。しかしながら日本文２文目の「全ての誤差バーはＳ．Ｅ．Ｍ．を表す。」は、正しくは次行のインドネシア文と対応づけられるべきものである。

つまり本例は、文分割処理で適切な単位に区切られた「全ての誤差バーはＳ．Ｅ．Ｍ．を表す。」という文を、文アライメント処理で正しい相手と対応づけることができなかった結果と見なせる。このため「② 文アライメントの不備」と判定される。

修正は上例のようにカット＆ペーストで正しい相手と対応づける形で行う。こうすることで、前後行の文対の対応も同時に改善される。

「② 文アライメントの不備」に該当するもう１つの典型例は、文対の一方のみに見出し語が付随しているケースが挙げられる。前出の「①-2 見出し語の不適な連結」との区別がやや難しいが、前者が「文対の双方の文頭に見出し語が存在しており（つまり内容上の対応は取れているが)、かつ本文との区切りが不明瞭な場合」を指すのに対し、こちらは「文対の一方のみにしか見出し語が存在せず、内容が対応していない（つまり対応づけに失敗している)」ため、文アライメントの不備と見なすべきである[56]。一例を示す。

［P00202005006_JP2021524681A SEQ:15］

修正前	ｲﾝﾄﾞﾈｼｱ	Invensi ini berhubungan dengan mekanisme untuk mendistribusikan pesan peringatan publik ketika perangkat pengguna dilampirkan pada jaringan bukan-3GPP. Latar Belakang Invensi
	日　本	本発明は、ユーザ機器デバイスが非３ＧＰＰネットワークに接続しているときに、公共警報メッセージを配布する仕組みに関する。
修正後	ｲﾝﾄﾞﾈｼｱ	Invensi ini berhubungan dengan mekanisme untuk mendistribusikan pesan peringatan publik ketika perangkat pengguna dilampirkan pada jaringan bukan-3GPP. ~~Latar Belakang Invensi~~
	日　本	本発明は、ユーザ機器デバイスが非３ＧＰＰネットワークに接続しているときに、公共警報メッセージを配布する仕組みに関する。

上例は、インドネシア文（修正前）の文末に見出し語「Latar Belakang Invensi（発明の背景)」が付随しており、日本文と対応していない。この見出し語は、文分割の後処理で本文と強制的に分割されており（⇒3.4.6.<課題 1>参照)、文アライメントが適切に行われれば、本文と誤連結されることなく、日本の見出し語と単独で対応づけられるはずである。しかしながら上例はそうなっておらず、インドネシア見出し語のみが、文アライメント処理によって前文の末尾に再連結されている。このため、文アライメントの不正といえる。

[56] ただし、そもそも他方の文献（すなわち前後行）に対応する見出し語が存在しない場合もある。この場合は「③-2 様式の相違」と判定される。

＜発生頻度＞

人手確認対象 7 万文対における「文アライメントの不備」の発生頻度を下表に示す。

表 6-6　7 万文対における「② 文アライメントの不備」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	498	450	102
	0.7%	90.4%	20.5%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	373	359	51
	4.3%	96.3%	13.7%
文アライメントスコア下位のもの（61,321 文対）	125	91	51
	0.2%	72.8%	40.8%

「文アライメントの不備」は対訳コーパスの不備として想起される代表例であるが、7 万文対における実際の発生数は 498 件、発生率は 0.7%と比較的低頻度であった。文アラインメントに失敗した文対では、文単位の大規模な不一致が発生する。このため、重大なものは文長比や文アライメントスコアが『対訳コーパス A』の選定範囲を逸脱し、あらかじめ除外されたと考えられる。この結果は、本調査が採用した文長比×文アライメントスコアによる選別の妥当性に一定の裏付けを与えている。

なお、「文アライメントの不備」も「インドネシア文がピリオドで終わらない文対」での発生頻度が顕著に高い。これは、典型例に挙げたような「見出し語がインドネシア文末（のみ）に誤連結されている」ケースが多く該当したためである。換言すれば、文長比や文アライメントスコアで排除されない文アライメントの不備は、こうした比較的小規模な不一致が主であると考えられる。

『対訳コーパス A』100 万文対における「文アライメントの不備」を試算すると、総発生件数は 2,356 件、発生頻度 0.2%程度と推定される。

6.3.3.「③ 文内容の相違」

文分割や文アライメントが適切に行われていても、日本文とインドネシア文の原文データ自体に何らかの相違があることで、文対の内容は不一致となりうる。こうした原文データ自体の相違による不一致は「③ 文内容の相違」にカテゴライズし、5 種のサブカテゴリに細分化した。

6.3.3.1.「③-1. 内容の意図的な変更」

日本とインドネシアの文献で記載内容に相違があり、それが（誤記等ではなく）意図的なものであると見なせる場合、本カテゴリに分類した。

＜不備の典型例＞

「内容の意図的な変更」と判定された主な事例を以下に示す。

［P00202200767_JP2022540238A SEQ:3］

修正前	ｲﾝﾄﾞﾈｼｱ	Metode tersebut dijalankan oleh perangkat terminal dan mencakup: memilih sumber daya PUSCH dari sumber daya saluran bersama uplink fisik PUSCH target, di mana sumber daya PUSCH target adalah sumber daya PUSCH yang digunakan untuk mengirimkan Msg A dalam prosedur akses acak 2 langkah; dan mentransmisikan data EDT dengan menggunakan sumber daya PUSCH yang dipilih.
	日　本	前記方法は、２ステップランダムアクセスプロセスにおけるＭｓｇＡ伝送のためのＰＵＳＣＨリソースであるターゲットＰＵＳＣＨリソースから、ＰＵＳＣＨリソースを選択することと、選択されるＰＵＳＣＨリソースによってＥＤＴデータを送信することとを含む。
修正後	ｲﾝﾄﾞﾈｼｱ	Metode tersebut dijalankan oleh perangkat terminal dan mencakup: memilih sumber daya PUSCH dari sumber daya saluran bersama uplink fisik PUSCH target, di mana sumber daya PUSCH target adalah sumber daya PUSCH yang digunakan untuk mengirimkan Msg A dalam prosedur akses acak 2 langkah; dan mentransmisikan data EDT dengan menggunakan sumber daya PUSCH yang dipilih.
	日　本	前記方法は端末機器により実行され、２ステップランダムアクセスプロセスにおけるＭｓｇＡ伝送のためのＰＵＳＣＨリソースであるターゲットＰＵＳＣＨリソースから、ＰＵＳＣＨリソースを選択することと、選択されるＰＵＳＣＨリソースによってＥＤＴデータを送信することとを含む。

上例の修正前の文対は、インドネシア文の「tersebut dijalankan oleh perangkat terminal」に対応する内容が日本文に存在しない。このため日本文に「端末機器により実行され」と追加し、双方の内容を一致させた。この差異は文脈から自明な内容の省略によるものと考えられるが、状況的に（誤記等ではなく）意図的な変更の範ちゅうと見なされる。

「内容の意図的な変更」には、出願内容の変更など記載内容が全面的に変わるようなものも含まれる。しかしながら、『対訳コーパス A』に関しては、本調査で作成した日インドネシア対訳コーパスの中でも対応精度が最良と見なされる 100 万文対が選定されている。このため、日インドネシア間で極端に内容が異なる文対はあらかじめ排除されていると考えてよい。その結果、「③-1 内容の意図的な変更」と判定された文対の大半は、ごく小規模で局所的な相違が大半であった。下例もその典型例である。

［P00201503795_JP2016046246A SEQ:52］

修正前	ｲﾝﾄﾞﾈｼｱ	Setiap cip LED dihubungkan dengan resistor (R5, R6) untuk membatasi arus listrik yang dipasok ke cip LED sehingga LED itu bekerja tidak melebihi batasnya.
	日　本	各ＬＥＤチップは抵抗器（Ｒ１、Ｒ２）に接続され、ＬＥＤチップに供給される電流が限界値以内で流れるよう制限する。
修正後	ｲﾝﾄﾞﾈｼｱ	Setiap cip LED dihubungkan dengan resistor (R1, R2) untuk membatasi arus listrik yang dipasok ke cip LED sehingga LED itu bekerja tidak melebihi batasnya.
	日　本	各ＬＥＤチップは抵抗器（Ｒ１、Ｒ２）に接続され、ＬＥＤチップに供給される電流が限界値以内で流れるよう制限する。

上例のインドネシア文と日本文は実質的に同じ内容であるが、要素付番のみ「(R5, R6)」から「（Ｒ１、Ｒ２）」へ変化している。誤記とは考えにくく、「③-1 内容の意図的な変更」に該当する。このような場合は、どちらかの付番に合わせて内容を一致させた。上例ではインドネシア文を日本文に合わせて「(R1, R2)」と修正している。

他には、文対の一方のみにカッコ補記が付されていることによる不一致も多く見られた。これらも「内容の意図的な変更」の一種として、文対の内容を一致させる修正を施した。一例を示す。

［P00202305372_JP2023549540A SEQ:445］

修正前	ｲﾝﾄﾞﾈｼｱ	Senyawa-senyawa ini dapat ditandai dengan simbol “R” atau “S” bergantung pada konfigurasi substituen di sekitar atom karbon stereogenik, tetapi praktisi yang ahli akan mengenali bahwa suatu struktur dapat menunjukkan pusat kiral secara implisit.
	日　本	これらの化合物は、立体中心（ｓｔｅｒｅｏｇｅｎｉｃ）炭素原子の周りの置換基の構成に応じて、記号「Ｒ」または「Ｓ」によって指定され得るが、当業者は、構造がキラル中心を暗黙のうちに示しうることを認識するであろう。
修正後	ｲﾝﾄﾞﾈｼｱ	Senyawa-senyawa ini dapat ditandai dengan simbol “R” atau “S” bergantung pada konfigurasi substituen di sekitar atom karbon stereogenik (stereogenic), tetapi praktisi yang ahli akan mengenali bahwa suatu struktur dapat menunjukkan pusat kiral secara implisit.
	日　本	これらの化合物は、立体中心（ｓｔｅｒｅｏｇｅｎｉｃ）炭素原子の周りの置換基の構成に応じて、記号「Ｒ」または「Ｓ」によって指定され得るが、当業者は、構造がキラル中心を暗黙のうちに示しうることを認識するであろう。

上例はインドネシア文の「atom karbon stereogenik」に対応する日本語に、「立体中心（ｓｔｅｒｅｏｇｅｎｉｃ）炭素原子」と英訳のカッコ補記が付されていた。この状態でも実質的内容は同一といえるが、本例は「③-1 内容の意図的な変更」の一種と判定され、インドネシア文の当該箇所を「atom karbon stereogenik (stereogenic)」とカッコ補記を付加する修正が施された。

上掲の 3 例をはじめとする実際の不備事例及び『対訳コーパス A』の選定手段に鑑みて、『対訳コーパス A』における「③-1 内容の意図的な変更」は総じて、付番の変更や省略された語の補足など小規模な相違にとどまると考えられる。

＜発生頻度＞

人手確認対象 7 万文対における「内容の意図的な変更」の発生頻度を下表に示す。

表 6-7　7 万文対における「③-1 内容の意図的な変更」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	1,367	915	534
	2.0%	66.9%	39.1%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	196	172	52
	2.3%	87.8%	26.5%
文アライメントスコア下位のもの（61,321 文対）	1,171	743	482
	1.9%	63.5%	41.2%

7 万文対における「内容の意図的な変更」の発生数は 1,367 件、発生率は 2.0%であった。ここから推定される『対訳コーパス A』100 万文対での総発生件数は 19,031 件、発生頻度は 1.9%程度となる。これは全不備の中で第 3 位の発生頻度である。

ただし、前記したとおり、不備判定された事例には当初想定された「技術内容に関する根本的な変更」というべきものは見当たらず、典型例に挙げたように付番の変更やカッコ補記の有無などの小規模な相違が大半であった。したがって、『対訳コーパス A』全件においても不備の程度は総じて軽微であると見られる。

6.3.3.2.「③-2 文章や語句の誤り」

誤記やタイプミスによる内容の相違がある場合、「③-2 文章や語句の誤り」と判定した。

＜不備の典型例＞

［P00201601017_JP2016529825A SEQ:2］

修正前	ｲﾝﾄﾞﾈｼｱ	Suatu metode komunikasi nirkabel yang mencakup menerlma laporan pengukuran dari beberapa peranti nirkabel pertama yang beroperasi dalam daerah tertentu dan beroperasi pada spektrum yang berbeda dari spektrum ASA.
	日　本	ワイヤレス通信の方法は、特定のエリア中で動作するとともに、ＡＳＡスペクトルとは異なるスペクトル上で動作する複数の第１のワイヤレスデバイスから測定報告を受信することを含む。

修正後	ｲﾝﾄﾞﾈｼｱ	Suatu metode komunikasi nirkabel yang mencakup menerima laporan pengukuran dari beberapa peranti nirkabel pertama yang beroperasi dalam daerah tertentu dan beroperasi pada spektrum yang berbeda dari spektrum ASA.
	日　本	ワイヤレス通信の方法は、特定のエリア中で動作するとともに、ＡＳＡスペクトルとは異なるスペクトル上で動作する複数の第１のワイヤレスデバイスから測定報告を受信することを含む。

上例はインドネシア文にスペルミスがあり、正しくは「menerima（受信する）」と綴るべき語が「menerlma」と誤記されている。このため「③-2 文章や語句の誤り」と判定された。

本カテゴリには、主にインドネシア文における中間スペースやピリオド、カッコの欠落や重複なども多数指摘された。以下に例示する。

［P00202005200_JP2021507085A SEQ:22］

修正前	ｲﾝﾄﾞﾈｼｱ	Invensi ini lebih lanjut berhubungan dengan penggunaan komposisi dan polioktenamer yang mengandung polialkenamer
	日　本	さらに、本発明は、ポリアルケナマー含有組成物の使用、ならびにポリオクテナマーに関する。
修正後	ｲﾝﾄﾞﾈｼｱ	Invensi ini lebih lanjut berhubungan dengan penggunaan komposisi dan polioktenamer yang mengandung polialkenamer.
	日　本	さらに、本発明は、ポリアルケナマー含有組成物の使用、ならびにポリオクテナマーに関する。

上例のインドネシア文（修正前）は文末にピリオドが打たれておらず[57]、文として若干不自然である。実質的な内容の不一致とは言い難いが、日本語には句点が打たれており、これと対応させるべくピリオドを追加する修正が施された。

[57] 原文献におけるピリオド欠落が原因である。

＜発生頻度＞

人手確認対象 7 万文対における「文章や語句の誤り」の発生頻度を下表に示す。

表 6-8　7 万文対における「③-2 文章や語句の誤り」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	3,490	3,271	292
	5.0%	93.7%	8.4%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	631	622	27
	7.3%	98.6%	4.3%
文アライメントスコア下位のもの（61,321 文対）	2,859	2,649	265
	4.7%	92.7%	9.3%

7 万文対における「文章や語句の誤り」の発生数は 3,490 件、発生率は 5.0%と、全カテゴリで最多であった。典型例に示したとおり、その大半はピリオドの欠落やタイプミス、やや不自然な表現の改善など小規模なものである。こうした不備は、ごく小規模であるゆえに文長比や文アライメントスコアでの排除が難しく、他の不備に比べて『対訳コーパス A』に多数残留したものと見られる。

また、この不備に関しては「インドネシア文がピリオドで終わらない」文対と「文アライメントスコア下位」の文対とで発生率に大きな違いは見られなかった。誤記や不自然な表現はあらゆる文で発生する性質であるためと考えられる。ただしインドネシア文と日本文での発生比率は圧倒的に前者に偏っていた。

『対訳コーパス A』100 万文対における「文章や語句の誤り」の推定総発生件数は 47,223 件、発生頻度は 4.7%程度と試算され、全類型で最多となった。7 万文対の集計結果から、その 9 割程度がインドネシア文であると見られる。

本カテゴリに該当する不備は、基本的には原文献自体の品質の問題であり、対訳コーパスの作成手法にかかわらず不可避的に発生する性質のものである。誤記はあらゆるケースが想定されるため機械的な検出は困難であり、対訳コーパスへの一定量の混入は避けられない。4.7%という発生頻度は他の不備と比べて突出して高率であるが、前掲の典型例に示したように、不備の規模は総じて極めて小さく、内容理解への悪影響もほとんど無い場合が多いと考える。

6.3.3.3.「③-3 様式の相違」

「様式の相違」とは、国ごとの特許公報の記載ルールの違いなどによる内容の不一致を指す。

＜不備の典型例＞

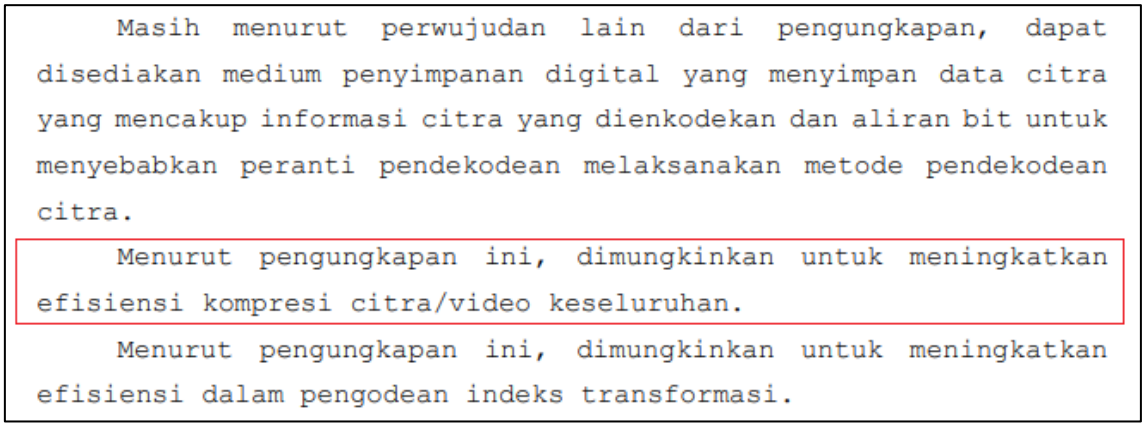
「様式の相違」に該当する文対の典型例として、ファミリー文献の一方のみに見出し語が使用されていることによる不一致が想定された。実際、人手確認対象 7 万文対においても該当事例が多数検出されている。一例を示す。

［P00202204438_JP2022551116A SEQ:39］

修正前	ｲﾝﾄﾞﾈｼｱ	Menurut pengungkapan ini, dimungkinkan untuk meningkatkan efisiensi kompresi citra/video keseluruhan.
	日　本	【発明の効果】本文書によると、全般的な映像／ビデオ圧縮効率を上げることができる。

上例は、日本文の「【発明の効果】」に対応する見出し語がインドネシア文に存在しない。こうしたケースでは、直前の文の末尾に見出し語が存在する場合が多く、そうであればカット＆ペーストで見出し語を移設し、不備判定は「② アライメントの不備」となる。しかしながら、本例に関してはそもそも当該インドネシア文献に対応する見出し語が存在していない（下図参照。赤枠が当該インドネシア文）。

図 6-2　PDF 公報 P00202204438 の該当行とその周辺の状況



このような場合は、文献の記載内容自体が異なっており、文アライメントの不備ではない。また、相違の原因も意図的な変更や誤記ではなく二国間の公報の様式の差と言うべきものであるため、不備判定は「③-3 様式の相違」となる。修正は下記のとおり、日本文から見出し語を削除して、双方の内容を一致させることとなる。

修正後	ｲﾝﾄﾞﾈｼｱ	Menurut pengungkapan ini, dimungkinkan untuk meningkatkan efisiensi kompresi citra/video keseluruhan.
	日　本	~~【発明の効果】~~本文書によると、全般的な映像／ビデオ圧縮効率を上げることができる。

ただし、実際に「③-3 様式の相違」と判定された文対の中で特に多かったのは、次例のように文対の一方のみに段落番号が存在するパターンであった。

［P00202003678_JP2021501300A SEQ:60］

修正前	ｲﾝﾄﾞﾈｼｱ	Masalah lain dengan sistem-sistem pengeringan semprot adalah potensi kerusakan terhadap produk jadi setelah selesai proses pengeringan.
	日　本	[0011]噴霧乾燥システムのもう１つの問題は、乾燥プロセスの完了後に完成品が損なわれる可能性があることである。
修正後	ｲﾝﾄﾞﾈｼｱ	Masalah lain dengan sistem-sistem pengeringan semprot adalah potensi kerusakan terhadap produk jadi setelah selesai proses pengeringan.
	日　本	~~[0011]~~噴霧乾燥システムのもう１つの問題は、乾燥プロセスの完了後に完成品が損なわれる可能性があることである。

上例では、日本文の冒頭に段落番号を示す「[0011]」が記載されており、インドネシア文との相違点となっている。通常、日本公報の段落番号は XML タグ化されているためテキストとしては抽出されないが、文献によっては本件のように段落番号がテキスト入力されている場合がある。

また、下例のようにインドネシア文のみに段落番号が存在するケースも、日本文ほどではないが一定数見られた。

［P00202303329_JP2023541434A SEQ:180］

修正前	ｲﾝﾄﾞﾈｼｱ	[0066] Istilah “pasien” atau “subjek” digunakan diseluruh spesifikasi untuk menjelaskan hewan, disukai manusia atau hewan domestik, yang diobati, mencakup pengobatan profilatik, dengan komposisi sesuai dengan pengungkapan ini disediakan.

	日　本	「患者」または「対象」という用語は本明細書全体を通して、本開示による組成物を用いた予防的治療を含む治療が提供される動物、好ましくはヒトまたは家畜を記載するために使用される。
修正後	ｲﾝﾄﾞﾈｼｱ	~~[0066]~~Istilah “pasien” atau “subjek” digunakan diseluruh spesifikasi untuk menjelaskan hewan, disukai manusia atau hewan domestik, yang diobati, mencakup pengobatan profilatik, dengan komposisi sesuai dengan pengungkapan ini disediakan.
	日　本	「患者」または「対象」という用語は本明細書全体を通して、本開示による組成物を用いた予防的治療を含む治療が提供される動物、好ましくはヒトまたは家畜を記載するために使用される。

こうした段落番号の不一致は、人手確認 7 万文対において多数検出されていることから、『対訳コーパス A』の選定に用いた文長比や文アライメントスコアでは特定し難い不備であるといえる。ただし、冒頭に段落番号がある文の検出自体は比較的容易であり、一括で機械的対応を行うことが可能である。このため本調査では、『対訳コーパス A＋』及び『対訳コーパス A-B＋』に対し、段落番号と見なされる文頭の文字列を一括除去する追加修正を実施した。詳細は 6.4.1.項で後述する

＜発生頻度＞

人手確認対象 7 万文対における「様式の相違」の発生頻度を下表に示す。

表 6-9　7 万文対における「③-3 様式の相違」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	1,797	582	1,353
	2.6%	32.4%	75.3%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	327	307	86
	3.8%	93.9%	26.3%
文アライメントスコア下位のもの（61,321 文対）	1,470	275	1,267
	2.4%	18.7%	86.2%

7 万文対における「様式の相違」の発生数は 1,797 件、発生率は 2.6%と、全カテゴリで第 3 位の発生頻度であった。典型例として示した見出し語の不一致、段落番号の不一致な

ど、「様式の相違」も総じて小規模な不備であり、このため文長比や文アライメントスコアでは排除されず、『対訳コーパス A』に多数残留したと考えられる。

「様式の相違」は全カテゴリで唯一、インドネシア文よりも日本文のほうが修正件数が多くなった。これは、本カテゴリの相違の主な原因である見出し語や段落番号が、日本文のみで使われているケースが多かったことを示している。

『対訳コーパス A』100 万文対における「様式の相違」の総発生件数は 24,119 件、発生頻度は 2.4%程度と試算される。

6.3.3.4.「③-4 ヘッダ、フッタの混入（テキスト抽出エラー）」

インドネシア文は、PDF 公報データ上のテキスト部分からヘッダやフッタなどの不要部分を機械的にカットして本文のみを抽出している。このため、イレギュラーな公報データが存在すると、ヘッダやフッタの除去が不完全となり、インドネシア文中に不要な文字列が残存する可能性がある。「③-4 ヘッダ、フッタの混入」は、こうしたテキスト抽出時のエラーを分類するためのカテゴリとして設けた。

＜不備の典型例＞

人手確認対象 7 万文対中、「ヘッダ、フッタの混入」と判定された文対は 5 件のみ存在した。いずれも、行番号と見られる数字が本文中に混入している事例であった。以下、一例を示す。

［P00202207045_JP2023504196A SEQ:384］

修正前	ｲﾝﾄﾞﾈｼｱ	Stimulasi ulang dari sel NT-NT dengan nanopartikel anti-CD3/CD28 yang berfungsi sebagai kontrol positif dan 20 menunjukkan bahwa kurangnya aloreaktivitas P-T NT-NT bukan karena sel T yang terganggu/tidak berfungsi.
	日　本	抗ＣＤ３／ＣＤ２８ナノ粒子によるＮＴ－ＮＴ細胞の再刺激は、陽性対照として機能し、Ｐ－ＴＮＴ－ＮＴ同種反応性の欠如が、損なわれた／機能しないＴ細胞に起因しないことが実証された。
修正後	ｲﾝﾄﾞﾈｼｱ	Stimulasi ulang dari sel NT-NT dengan nanopartikel anti-CD3/CD28 yang berfungsi sebagai kontrol positif dan ~~20~~ menunjukkan bahwa kurangnya aloreaktivitas P-T NT-NT bukan karena sel T yang terganggu/tidak berfungsi.

	日　本	抗ＣＤ３／ＣＤ２８ナノ粒子によるＮＴ－ＮＴ細胞の再刺激は、陽性対照として機能し、Ｐ－ＴＮＴ－ＮＴ同種反応性の欠如が、損なわれた／機能しないＴ細胞に起因しないことが実証された。

本例は、インドネシア文（修正前）に「20」という PDF 公報の行番号が混入したものである。不備の生じた理由や、同種の事例については 6.4.3.項で後述し、ここでは割愛する。

なお、行番号の混入は文献単位で発生する性質のものであるため、今回検出された 5 文対と同じ文献（全 3 文献）に由来する文対に対しては、全件を改めてチェックし、行番号が混入していた全ての文対（『対訳コーパス A＋』では 19 文対）の追加修正を実施した。これについても 6.4.3.項で詳述する。

＜発生頻度＞

人手確認対象 7 万文対における「ヘッダ・フッタの混入」の発生頻度を下表に示す。

表 6-10　7 万文対における「③-4 ヘッダ・フッタの混入」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	5	5	0
	0.01%	100.0%	0.0%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	1	1	0
	0.01%	100.0%	0.0%
文アライメントスコア下位のもの（61,321 文対）	4	4	0
	0.01%	100.0%	0%

7 万文対における「ヘッダ・フッタの混入」の発生数は 5 件（3 文献）、発生率は 0.01%と、全カテゴリ中最も少数であった。これら 5 件ともインドネシア PDF 公報のイレギュラーな記載が原因であり、予測不可能につき不可抗力といえる。今回の結果から、本調査で実施したテキスト抽出時のヘッダ・フッタ除去は、インドネシア公報が特殊な様式で記載されない限り、完璧に機能すると結論できる。

『対訳コーパス A』100 万文対における「ヘッダ・フッタの混入」の総発生件数は、上記カウント結果に基づけば 100 件程度と試算される。今回検出された 3 文献での該当文対

数は合計 24 文対（人手確認 5＋追加修正 19）であったが、これら以外にも、人手確認対象とならなかった文献で同様の不備が発生している可能性がある。

6.3.3.5.「③-5 データ形式に起因する相違」

「データ形式に起因する相違」とは、主に日本文献ではイメージデータとなっている配列表やテーブルがインドネシア文献ではテキストデータで記載されており、後者のみでテキスト抽出されることに起因する不一致を指す。

＜不備の典型例＞

「データ形式に起因する相違」は、通常インドネシア文のみが不当に長大となり、『対訳コーパス A』の選定条件となる一般的な文長比（日：イン＝1：2.3～3.1）を大きく逸脱することになる。このため『対訳コーパス A』に含まれる可能性は低いと想定された。

しかしながら、実際に人手確認を実施したところ、下例のように「文中の化学式が日本文はテキスト表題（【化学式 xx】等）を伴う（本体はイメージデータ）のに対し、インドネシア文はイメージデータのみでテキスト表記を伴わないため文中に不自然な抜けが生じているケースが「③-5 データ形式に起因する相違」の一種と見なされたことで、若干数のカウントが生じた。一例を示す。

［P00202302677_JP2023541404A SEQ:567］

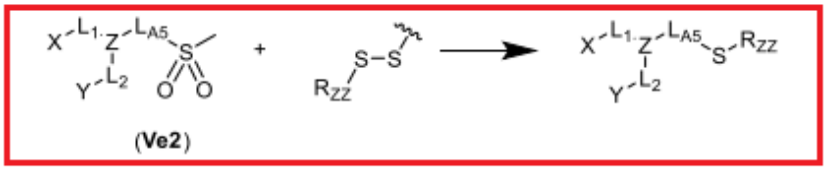
修正前	ｲﾝﾄﾞﾈｼｱ	Contoh skema reaksi ditunjukkan di bawah ini: （抜け１）dimana RZZ mencakup zat RNAi, dan （抜け２）menunjukkan titik koneksi ke setiap kelompok yang cocok yang dikenal dalam bidang ini.
	日　本	反応スキームの例を以下に示す。【化学式８】式中、ＲＺＺは、ＲＮＡｉ剤を含み、【化学式９】は、当該技術分野で既知の任意の好適な基への結合点を示す。

上例は、インドネシア文、日本文とも文中に化学式のイメージデータが挿入されており、この部分はテキスト化されていない。このためインドネシア文は化学式イメージデータ部分が抜け落ちており（ここでは「(抜け１)」「(抜け２)」と表現)、文法的に成立しない文となっている。

インドネシア公報原本の状況は次図のとおりである。赤枠部分が化学式のイメージデータであり、それぞれ「(抜け１)」「(抜け２)」に相当する。

図 6-3　PDF 公報 P00202302677 の該当行とその周辺の状況

Dalam beberapa perwujudan, prekursor modulator PK/PD dapat mencakup moitas sulfon dan dapat bereaksi dengan disulfida. Contoh skema reaksi ditunjukkan di bawah ini:



(Ve2)

dimana R_{ZZ} mencakup zat RNAi, dan  menunjukkan titik koneksi ke setiap kelompok yang cocok yang dikenal dalam bidang ini.

一方、日本文はイメージデータに添えて「【化学式 8】」「【化学式 9】」というテキスト表題が存在するため、ひとまず文として成立しており、インドネシア文との間に内容の不一致が発生している。このようなケースは、インドネシアと日本の文献のデータ形式に起因する相違の一種と見なした。

こうした場合の修正は、下に示したように、インドネシア文の所定の位置に「【化学式 8】」「【化学式 9】」に相当するインドネシア表題「[Rumus Kimia 8]」「[Rumus Kimia 9]」を補うことで、双方の内容を一致させた。

修正後	ｲﾝﾄﾞﾈｼｱ	Contoh skema reaksi ditunjukkan di bawah ini: [Rumus Kimia 8] dimana RZZ mencakup zat RNAi, dan [Rumus Kimia 9] menunjukkan titik koneksi ke setiap kelompok yang cocok yang dikenal dalam bidang ini.
	日　本	反応スキームの例を以下に示す。【化学式８】式中、ＲＺＺは、ＲＮＡｉ剤を含み、【化学式９】は、当該技術分野で既知の任意の好適な基への結合点を示す。

なお、文の構造によっては「[Rumus Kimia x]」を挿入するよりも日本文から「【化学式 X】」を除去するほうが自然な文対になるケースもあるため、どちらの対処を行うかは各文対の状況に応じて都度判断した。

＜発生頻度＞

人手確認対象 7 万文対における「データ形式に起因する相違」の発生頻度を下表に示す。

表 6-11　7 万文対における「③-5 データ形式に起因する相違」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	240	224	59
	0.3%	93.3%	24.6%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	116	115	12
	1.3%	99.1%	10.3%
文アライメントスコア下位のもの（61,321 文対）	124	109	47
	0.2%	87.9%	37.9%

7 万文対における「データ形式に起因する相違」の発生数は 240 件、発生率は 0.3%と比較的少量であった。典型例に示したとおり、文中にイメージデータが挿入されている文が主な対象となるため、絶対数自体が少ないためと考えられる。反面、この条件に合致する文では確実に発生する性質の不備であり、かつ、対訳コーパスのソースとして特許文献データを使用する以上、一定量の混入は避けられないタイプの不備である。

『対訳コーパス A』100 万文対における「データ形式に起因する相違」の総発生件数は、上記カウント結果に基づけば 2,099 件、発生率は 0.2%程度と試算される。全類型で第 2 位の発生頻度であるが、前述のとおり、その主な原因のひとつである段落番号に対してはコーパス全件を対象に一括除去する追加修正を実施するため、最終的な発生頻度は上記推定よりも顕著に少なくなると予測される。

6.3.4.「④ その他の相違」

ここまで上げた各類型に該当しない相違は、「④ その他の相違」に分類した。結果、事前に想定していなかった不備のパターンが 1 種類検出された。次項に示す。

＜不備の典型例＞

【① 英文／英単語の混入】

人手確認対象の 7 万文対のインドネシア文において、文全体が英文となっている文対が一定数検出され、「④ その他の相違」に分類された。特許文献では技術用語や固有名詞、引用文献名などで意図的にオリジナルの英語表記が使われる場合もあるが、人手確認で指摘された文の多くは文全体が英文であり、意図的な英語の使用とは異なる状況と見なせる。以下、一例を示す。

［P00202001672_JP2020533774A SEQ:2］

修正前	ｲﾝﾄﾞﾈｼｱ	The present invention discloses a lighting installation having an LED lamp (19), normally consisting of a series string of individual LED's (18), which is supplied by a rectifier (20, 200).
	日　本	本発明は、通常、整流器（２０、２００）により電力供給される、個々のＬＥＤ（１８）の一連のストリングからなる、ＬＥＤランプ（１９）を有する照明設備を開示する。
修正後	ｲﾝﾄﾞﾈｼｱ	Penemuan ini mengungkapkan perlengkapan pencahayaan dengan lampu LED (19), yang biasanya terdiri dari serangkaian LED individual (18) yang bertenaga penyearah (20, 200).
	日　本	本発明は、通常、整流器（２０、２００）により電力供給される、個々のＬＥＤ（１８）の一連のストリングからなる、ＬＥＤランプ（１９）を有する照明設備を開示する。

上例は要約部分の文対だが、インドネシア文が完全に英語で書かれている。このため全体をインドネシア文に再翻訳する修正を要した。

こうした文対では、インドネシア文ではなく英文が日本文と対応づけられているが、文アライメント処理では入力文が想定した言語であるかはチェックされず、想定と異なる言語であっても、内容的に一致していれば対応づけがなされ、文アライメントスコアも低値とはならない。このため、『対訳コーパス A』中にも英文が混入する結果となった。

上記状況から、『対訳コーパス A』には、人手確認対象 7 万文対以外の文対（93 万文対）

においても英文が含まれる可能性が非常に高い。

この状況を受け、本調査では、インドネシア文が英語である文対の候補を機械的に抽出し、目視で判定して英文であれば排除する追加対応を行った。対応内容の詳細は 6.4.2.項で後述する。

【② その他の事例】

「④ その他の相違」と判定された文対の大半は先に示した【①英文の混入】であり、他に特定のパターンの不備は見られなかった。英文混入以外の「その他の相違」はいずれも特殊な内容であり、多発することは考えにくい。以下、これらの一例を示す。

［P00202200824_JP2022543419A SEQ:420］

修正前	ｲﾝﾄﾞﾈｼｱ	MAIN¥VIYU¥35371881_1.docx Setelah agitasi isi selama 2-4 jam pada 20 ± 5° C, campuran reaksi (yang mengandung air dan fase organik) diambil sampelnya oleh IPC untuk analisis konversi (Senyawa D ≥99,6%).
	日　本	２０±５℃における２～４時間の内容物の撹拌後、変換分析のために（水及び有機相の両方を含む）反応混合物のＩＰＣサンプルを採取した（化合物D≧９９．６％）。
修正後	ｲﾝﾄﾞﾈｼｱ	~~MAIN¥VIYU¥35371881_1.docx~~ Setelah agitasi isi selama 2-4 jam pada 20 ± 5° C, campuran reaksi (yang mengandung air dan fase organik) diambil sampelnya oleh IPC untuk analisis konversi (Senyawa D ≥99,6%).
	日　本	２０±５℃における２～４時間の内容物の撹拌後、変換分析のために（水及び有機相の両方を含む）反応混合物のＩＰＣサンプルを採取した（化合物D≧９９．６％）。

上例では、インドネシア文の冒頭に「MAIN¥VIYU¥35371881_1.docx」という文字列が混入している。末尾の拡張子からは MS Word のファイル名と思われる。出願書類作成時の編集ミスと考えられるが、「③-2 文章や語句の誤り」とは性質が異なるため「その他の相違」と判定された。頻繁に発生する不備とは考えにくく、対処の優先順位は低い。

＜発生頻度＞

人手確認対象 7 万文対における「その他の相違」の発生頻度を下表に示す。

表 6-12　7 万文対における「④ その他の相違」の発生頻度

確認対象	発生数／発生率	修正	
		ｲﾝﾄﾞﾈｼｱ	日本
人手確認対象 7 万文対全件	61	58	13
	0.1%	95.1%	21.3%
［内訳］			
インドネシア文がピリオドで終わらないもの（8,679 文対）	14	14	4
	0.2%	100.0%	28.6%
文アライメントスコア下位のもの（61,321 文対）	47	44	9
	0.1%	93.6%	19.2%

7 万文対における「その他の相違」の発生数は 61 件、発生率は 0.1%とごく少量であった。このカウントからは、『対訳コーパス A』100 万文対におけるの総発生件数は 1,005 件程度と試算される。ただし、『対訳コーパス A』全件を対象に英文の検出・除外のための追加措置を講じることで、これらの大部分は除去される。

6.3.5 不備類型別『対訳コーパス A』全件での推定発生数

前項までに示した 10 種の不備類型の＜発生頻度＞の分析では、人手確認対象 7 万文対での発生頻度に基づき、『対訳コーパス A』全 100 万件での推定発生数を試算した。本項に試算結果の一覧表を示す。

表 6-13　不備類型別『対訳コーパス A』全件での推定発生数一覧

不備の類型	『対訳コーパス A』推定 発生数と発生頻度	（参考）7 万文対 の発生数と頻度
①-1 過分割	6,090	1,449
	0.6%	2.1%
①-2 見出し語の不適な連結	15,506	1,528
	1.5%	2.2%
①-3 文末の数字＋ピリオド過分割	7,045	3,313
	0.7%	4.7%
② 文アライメントの不備	2,356	498
	0.2%	0.7%
③-1 内容の意図的な変更	19,031	1,367
	1.9%	2.0%
③-2 文章や語句の誤り	47,223	3,490
	4.7%	5.0%
③-3 様式の相違	24,119	1,797
	2.4%	2.6%
③-4 ヘッダ、フッタ等の混入	100	5
	0.01%	0.01%
③-5 データ形式に起因する相違	2,099	240
	0.2%	0.3%
④その他の相違	1,005	61
	0.1%	0.1%
合　計	124,573	13,748
	12.4%	19.6%

類型別の試算に基づく『対訳コーパス A』全件での不備発生数は 124,573 件、発生率は 12.5%程度となった。5.6.3.項での全不備数による試算結果の 123,482 件、12.3%より若干増加したが、各類型の発生率を小数点第 1 位に丸めて試算したことによる誤差と考えられる。

各類型の『対訳コーパス A』全件での推定発生率を見ると、人手確認対象 7 万文対で「インドネシア文末がピリオド以外」の文対に偏って発生していた「①-1 過分割」と「①-3 文末の数字＋ピリオド過分割」の発生率がそれぞれ 2.1%→0.6%、4.7%→0.7%と大幅に低下した以外は、おおむね同程度の発生率となっている。7 万文対での発生率が 5.0%と最も高かった「③-2 文章や語句の誤り」は 100 万文対全件での推定発生率も 4.7%で最も高率であった。次いで「③-3 様式の相違」2.4%、「③-1 内容の意図的な変更」1.9%と続く。これら上位 3 種の不備は、いずれも文分割や文アライメント処理ではなくソースデータ自体の不一致であり、現状では不可避なものである。

上位 3 種の不備の推定発生数は合計 90,373 件であり、全類型合計の推定発生率 12.5%のうち 9.0%（約 7 割）を占める。ただし、『対訳コーパス A』におけるこれらの不備の大半は、各類型の典型例で見られたとおり、誤字脱字やピリオドの欠落（③-2）、段落番号の有無（③-3）、単語レベルのカッコ補記の有無（③-1）といった、文の実質的内容に影響しないごく軽微なものであると見られる。

これに対し、主にインドネシア文分割処理の失敗による「①-2 文末の数字＋ピリオド過分割」0.7%と「①-1 過分割」0.6%、文アライメント処理の失敗による「②文アライメントの不備」0.2%は、発生数は 3 種合計で 15,491 件、発生率 1.5%と少数であるが、文の実質的内容に影響するため、不備としてはより重大である。ソースデータの問題ではないため、使用するツールや前後処理の改善次第では抑制できる可能性がある。

「①-2 見出し語の不適な連結」は推定発生数 15,506 件で全類型中第 4 位である。これもテキストデータ抽出処理の過程で発生するものであり、対訳コーパス作成処理の失敗に含めるべきかもしれない。ただし、この不備は現状の（文分割処理を前提にした）対訳コーパス作成技術では必然的に生じるものであり（⇒3.5.1.4.）、解決は難しい。なお、この不備が生じた文対は、双方で見出し語と本文が連結され不自然な文となるが、文対の実質的内容には相違はないため、重大度は比較的低い。

6.4. 検出された不備パターンへの機械的対応

人手確認により検出された不備パターンの一部は、機械的処理により『対訳コーパス A』及び『対訳コーパス A-B』全体を対象に改善できるものが存在する。本調査では、以下の対応を実施した。

6.4.1. 文頭の段落番号の除去

「③-3 様式の相違」として、文対の一方のみに「[0001]」という様式の段落番号が書かれている文対の存在が多数確認された（⇒6.3.3.3.）。

日本公報には、段落ごとに【0001】のフォーマットで段落番号が表示される。ただし、この段落番号は XML タグで自動表示されるものであり、本文テキストデータには含まれない。このため通常は、段落番号を有さないインドネシア公報との不一致は発生しない。

しかしながら、日本公報の中には、本文中で各パラグラフ冒頭に段落番号を記載しているものが存在することが判明した。その一部が人手確認の対象となり、インドネシア文との不一致が指摘される結果となった。6.3.3.3.項で取り上げた実例を再掲する。

[P00202003678_JP2021501300A SEQ:60]

修正前	ｲﾝﾄﾞﾈｼｱ	Masalah lain dengan sistem-sistem pengeringan semprot adalah potensi kerusakan terhadap produk jadi setelah selesai proses pengeringan.
	日　本	[0011]噴霧乾燥システムのもう１つの問題は、乾燥プロセスの完了後に完成品が損なわれる可能性があることである。

上例では、日本文のみに段落番号「[0011]」が使用されており、インドネシア文との不一致の原因となっている。

なお、本項冒頭でインドネシア公報は通常は段落番号を有さないと述べたが、一部、段落番号が記載されているものも存在し、これによって日本文との不一致が生じているケースも検出された。一例を示す。

[P00202100591_JPWO2020004495A1 SEQ:181]

ｲﾝﾄﾞﾈｼｱ	[0041] Fungsi pembuatan kunci (624) adalah fungsi untuk membuat sepasang kunci pribadi (611) dan kunci publik (612).
日　本	鍵生成機能６２４は、秘密鍵６１１と公開鍵６１２の鍵ペアを生成する機能である。

このように人手確認対象 7 万文対において段落番号の不一致が多数見られたことから、『対訳コーパス A』や『対訳コーパス A-B』の全体においても、文対の一方のみに段落番号を有するケースは相当数存在すると見られた。

この不一致に網羅的に対応すべく、本調査では、『対訳コーパス A』及び『対訳コーパス A-B』の全文対を対象に、文頭に位置する「[xx…xx]」(※「xx…xx」は任意桁数の数字を示す。全角・半角は問わない[58])」を段落番号と特定し、一括で除去する修正を実施した。結果、『対訳コーパス A』では日本文 8,625 文とインドネシア文 2,379 文、『対訳コーパス A-B』では日本文 74,032 文とインドネシア文 15,959 文が一括修正された。

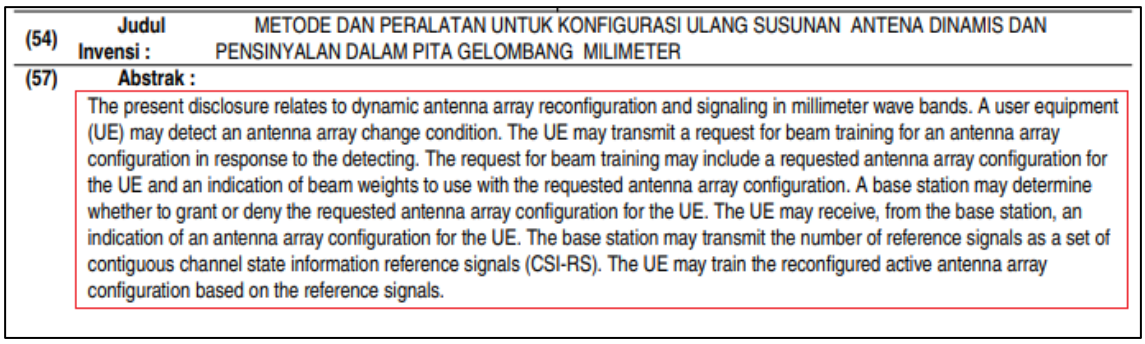
なお、上記修正件数から、段落番号は日本文献、インドネシア文献のいずれでも発生する性質であること、ただしその頻度は日本のほうが 3〜5 倍ほど多いことがわかった。

6.4.2. 英文の除去

6.4.2.1. インドネシア文への英文の混入

6.3.4.項で述べたとおり、人手確認・修正作業によりインドネシア文献への英文の混入が確認された。一例を示す。

図 6-4　インドネシア文献 P00202205012 の PDF 公報より抜粋



(54) **Judul Invensi :** METODE DAN PERALATAN UNTUK KONFIGURASI ULANG SUSUNAN ANTENA DINAMIS DAN PENSINYALAN DALAM PITA GELOMBANG MILIMETER

(57) **Abstrak :**

The present disclosure relates to dynamic antenna array reconfiguration and signaling in millimeter wave bands. A user equipment (UE) may detect an antenna array change condition. The UE may transmit a request for beam training for an antenna array configuration in response to the detecting. The request for beam training may include a requested antenna array configuration for the UE and an indication of beam weights to use with the requested antenna array configuration. A base station may determine whether to grant or deny the requested antenna array configuration for the UE. The UE may receive, from the base station, an indication of an antenna array configuration for the UE. The base station may transmit the number of reference signals as a set of contiguous channel state information reference signals (CSI-RS). The UE may train the reconfigured active antenna array configuration based on the reference signals.

この文献は発明の名称（図の上部）や明細書本文（図示せず）はインドネシア語で記述されているが、要約（図中赤枠）のみ英文が用いられている。

本調査で使用した文アライメント手法は入力言語を限定しないユニバーサルなものであるため、言語を問わず、内容が一致すれば文を対応付ける。このため、上例のようにイン

[58] 日本文、インドネシア文とも、段落番号は［0001］のフォーマットであるものが大半で、【】が用いられるケースは皆無であった。

ドネシア文に混入した英文も区別なく対応付けられ、文アライメントスコアも低スコアとならない。文長比も正常値となるため、同様の事例は『対訳コーパス A』及び『対訳コーパス A-B』中に少なからず混入していると予測される。本来インドネシア文であるべき文が異国語となっており、機械翻訳エンジンの学習に用いた際にも悪影響を及ぼす懸念が大きい。

人手確認対象のインドネシア文が英文であった場合は人手でインドネシア文に修正されるが、『対訳コーパス A』や『対訳コーパス A-B』の人手確認対象外の文対にも英文が混入している可能性が非常に高い。このため、『対訳コーパス A』及び『対訳コーパス A-B』の全件を対象に、以下の 2 条件でインドネシア文から英文の候補を特定した。

- 言語判定ツールで「英語」と判定された文
- 文中に「the」「in」「for」「are」を含む文

上記 2 条件のいずれかに該当する文対を目視で確認し、英文又は英語混じり文と判定された文を全て除去した。この措置により、『対訳コーパス A』から 229 文対、『対訳コーパス A-B』からは（これら 229 文対を含めて）1,084 文対が除去された。

次項より、上記 2 条件それぞれの実施内容を説明する。

6.4.2.2. 言語判定ツールで「英語」と判定された文の除去

インドネシア文に混入する英文を特定するための言語判定ツールには、fastText[59]を使用した。本ツールで『対訳コーパス A』及び『対訳コーパス A-B』の全インドネシア文(4,980,059文）の言語判定を実施し、英スコア 0.9（英文である可能性が 90%）以上のもの 212 文に対して目視判定で英文か否かを選別した。

結果、次図に示すとおり、スコア 0.9 以上となった文の大半は実際に英文であった。

図 6-5　言語判定ツール fastText で英語スコア 0.9 以上となった文のサンプル

スコア	文長	インドネシア文
0.9800	64	One section of bilateral kidneys from each animal was evaluated.
0.9026	65	A video data stream having a video encoded thereinto is provided.
0.9988	65	Ref Ab 2 and Ref Ab 3, as described above, were used as controls.
0.9405	67	This sorting process 706 can be performed manually or by a machine.
0.9505	68	The residue obtained was stirred in dietileter (30 mL) **dan** filtered.
0.9837	71	Each sample was analyzed by UHPLC, followed by calculation of the FLP%.
0.9396	71	Pharmaceutically acceptable salts include acid and base addition salts.
0.9349	71	The **Senyawa** plate was prepared by 3-fold **dan** 11-point serial dilutions.
0.9075	71	Tracking travel control with respect to vehicle ahead will be canceled.
0.9113	72	Present invention relates to a lift assembly (300) in an aerial vehicle.
0.9030	72	The grinder is fed by a hopper with a flow rate control flap (180 kg/h).
0.9298	74	The at least four inner frames are detachably disposed in the outer frame.
0.9851	75	As the ATP concentration increased, the ligasi reaction was inhibited more.
0.9771	75	Images were captured over 5 randomly selected fields at 400x magnification.
0.9710	76	The OH groups are reacted by acetylation with an excess of acetic anhydride.
0.9741	77	Now, pedestrians can be detected but vehicle detection accuracy has degraded.
0.9817	78	[0237] Each sample was analyzed by UHPLC, followed by calculation of the FLP%.
0.9376	80	Alcohol is reacted in countercurrent with the respective alkali metal hydroxide.
0.9104	80	It was shown that addition of PEG caused **suatu** increase in the **efisiensi ligasi**.
0.9854	82	[127] The compound as shown by (201) may be prepared as described in CN103380113A.
0.9413	82	Increased cytosine DNA-methyltransferase activity during colon cancer progression.
0.9498	82	The container is sealed and then mixed, either manually or by a mechanical device.
0.9759	84	The molecular biological reagents were used according to the manufacturer protocols.
0.9685	84	The vertical rotor (118) is operational during forward flight of the aerial vehicle.

図中赤字で示したように、英文中に若干のインドネシア語が存在する文もあったが、これらも対訳コーパスには不要であるため除去した。

このように fastText で非常に精度よく英文を検出することができたが、唯一、文が極端に短い文においてインドネシア文を英文と誤認したケースが若干数見られた。サンプルを下図に示す。

[59] https://fasttext.cc/blog/2017/10/02/blog-post.html

図 6-6　fastText で英文と誤認されたインドネシア文のサンプル

スコア	文長	インドネシア文
0.9293	16	**Reaksi disaring.**
0.9143	17	**Hasil = 7,0 gram.**
0.9078	17	**Larutan disaring.**
0.9205	33	**LCMS: kemurnian=96%, MH+ = 414,0.**
0.9612	33	Test results are reported in MPa.
0.9849	34	Other examples are also described.
0.9409	45	The spring balance is suspended in this hole.
0.9773	47	**Reaksi** was **dimula**i by switching on the UV lamp.

上図に青字で示した 4 行はいずれも英語スコア 0.9 以上であったが、実際にはインドネシア文であり除去する必要はない。fastText では、文長がごく短いものに限り、このような誤判定が発生する傾向が見られた。本調査では、これら誤認文（全 4 文）は目視判定により除去対象から外した。

fastText で英文として抽出した 212 文のうち、これら短文 4 文を除く 208 文は全て英文であった。スコア 0.9 以上という条件で抽出した場合の英文率は 98.1%と極めて高い。誤認された 4 文についても、極端な短文は学習データとしての価値が比較的低いことを考えると、今後の実用時には目視判定を省略し、スコア 0.9 以上の文は無条件で除外するほうが効率的と考える。

ただし、スコア 0.9 以上で英文率 98.1%という結果からは、この条件に合致しない（つまりスコア 0.9 未満の）英文も少なからず存在している可能性が高い。本調査では、こうした英文も極力除去すべく、もう 1 つの条件を併用した。次項に示す。

6.4.2.2. 頻出英単語を含む文

本調査では、fastText のスコアが 0.9 未満のインドネシア文のうち、頻出英単語「␣the␣」「␣of␣」「␣for␣」「␣is␣」「␣are␣」の 5 語のいずれかを含むものは英文の可能性があるとして目視チェックの対象とした。結果、12,150 文が抽出され、目視チェックされた。

しかしながら、抽出されたインドネシア文の大部分は、引用文献が原語（つまり英語）で記載されているものなど、意図的に英語が用いられたもので、除去対象とすべきでない文が大半であった。結果、本チェックで除去対象と判定されたインドネシア文は、12,150 文のうち 876 文（7.2%）にとどまった。以下、該当文をいくつか示す。

表 6-14　頻出英単語を含む文のうち除去対象のサンプル

No.	英ｽｺｱ	インドネシア文
1	0.8008	Untuk suatu practical implementatipada above matrix is not of interest as the constraints are typically applied in the time-domain.
2	0.7604	Molekul RNA berunting tunggal tusuk rambut including suatu sekuen penghambat ekspresi gen in the ligasi larutan reaksi in invensi may be purified by suatu metode known to those skilled in the art.
3	0.5904	Selanjutnya provide herein is metode tersebut, dimana sebelum pemberian virus onkolitik yang dimodifikasi, virus vaksinia onkolitik, atau komposisi farmasi subjek telah didiagnosa dengan kanker.
4	0.4690	The parfum element of said produk konsumen can be a combination of parfum mikrokapsul seperti yang didefinisikan di atas dan free atau non-encapsulated parfum, as well as other types of parfum mikrokapsul than those here- disclosed.
5	0.4118	Dalam perwujudan-perwujudan tertentu, the antibodies disclosed herein do not compete for the binding of ApoC3 with a lipid atau a lipoprotein.

本チェックで検出された文の大半は、上例に示したとおり、一文中にインドネシア語と英語が混在するタイプであった。参考に fastText の英語スコアを添付したが、例えば英語が 24 語中 3 語のみ英語である事例 3 のスコアが 0.5904 と比較的高いのに対し、英語が 22 語中 16 語を占める事例 5 のスコアは 0.4118 と低値であるなど、英語の含有率とスコアとの間に明瞭な相関性は見られない。このため、英語スコアの範囲指定でこの種の英語混じり文を効率よく特定することは困難である。

前記結果のとおり、特定の英単語を含む文の目視チェックは、大量のインドネシア文を目視チェックする作業負荷と、それによって除外できた文数とが見合わず、効率が非常に悪いことが分かった。この手法によって『対訳コーパス A』から 156 文（0.02%）、『対訳コーパス A-B』からはさらに 720 文、計 876 文（0.02%）の英文／英語混じり文を除去できたが、そのためには人手による大量のチェック作業が必要となり、相応のコストを要する。『対訳コーパス A』に混入する英文／英語混じり文はごく少量であることを考えると、今後対訳コーパスを作成する際は、英文混入への対処は言語判定ツールの使用のみと割り切るのが妥当と結論される。

6.4.3. 文中の行番号の削除

人手確認・修正作業により、インドネシア文献 3 件（P00202207045、P00202210351、P00202102343）から採取した文対において、下例のようにインドネシア文に行番号が混入しているものが存在することが判明した（⇒6.3.3.4.）。

［P00202207045_JP2023504196A SEQ:384］

修正前	ｲﾝﾄﾞﾈｼｱ	Stimulasi ulang dari sel NT-NT dengan nanopartikel anti-CD3/CD28 yang berfungsi sebagai kontrol positif dan 20 menunjukkan bahwa kurangnya aloreaktivitas P-T NT-NT bukan karena sel T yang terganggu/tidak berfungsi.
	日　本	抗ＣＤ３／ＣＤ２８ナノ粒子によるＮＴ－ＮＴ細胞の再刺激は、陽性対照として機能し、Ｐ－ＴＮＴ－ＮＴ同種反応性の欠如が、損なわれた／機能しないＴ細胞に起因しないことが実証された。

上例のインドネシア文中に存在する「20」は、PDF 公報の欄外に表示される行番号である。本調査では、インドネシア PDF 公報からテキストデータを抽出する際、行番号を対象外とする措置を講じていたが（⇒2.2.2.）、この文献（P00202207045）では、一部のページにおいて、本文中の表が行番号の位置よりも外側に突出して記載されていた（次ページ図 6-7 参照）。本調査で使用した除去ロジックでは、行番号を「本文よりも外側にある数字」と定義していたため、本文が行番号よりも外側にあると、正しく除去が行えない。

当該文献には下図と同様の状況にあるページが複数存在し、これらのページから取得した文対には、インドネシア文に行番号が混入したものがある。このため、この文献の該当ページから取得した全文対を対象に、行番号が混入しているものを目視で特定し[60]、修正を施した。

[60] この不備に関しては機械的対応が困難であったため、目視チェックを実施した。

他の 2 文献のうち、P00202210351 については上記 P00202207045 と同様の状況であったため、同じ対処を行った。一方、P00202102343 については、通常は 5 行ごとに付される行番号が 6 行ごとに付されていた（次ページの図 6-8 参照）。本調査で用いた行番号除去ロジックは 5 行ごとの付番であることを前提としていたため、この文献は全編にわたり行番号が除去されなかった。このため、全ての文に対して目視による特定・除外を実施した。

図 6-7　インドネシア PDF 公報 P00202207045 における本文と行番号の状況

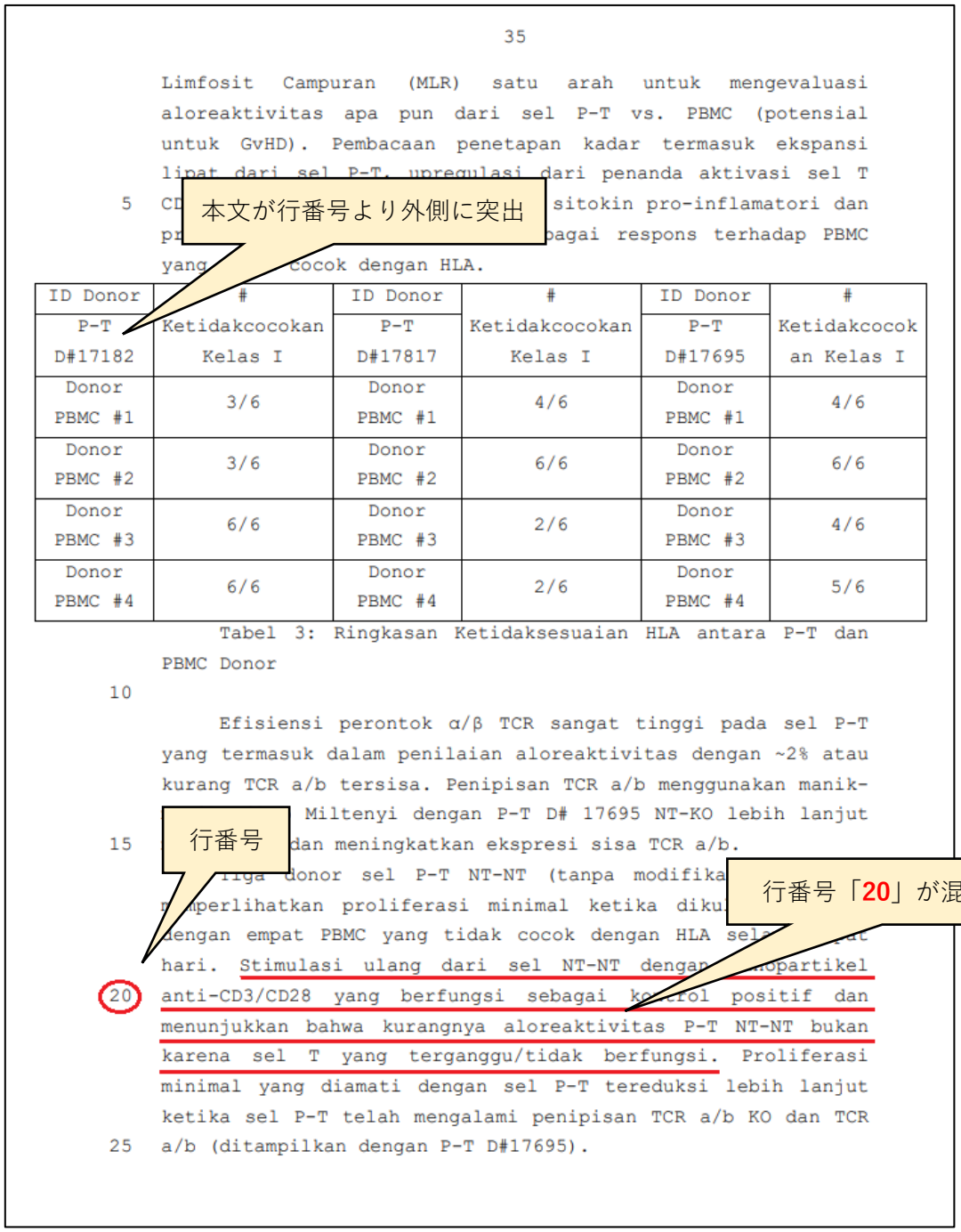
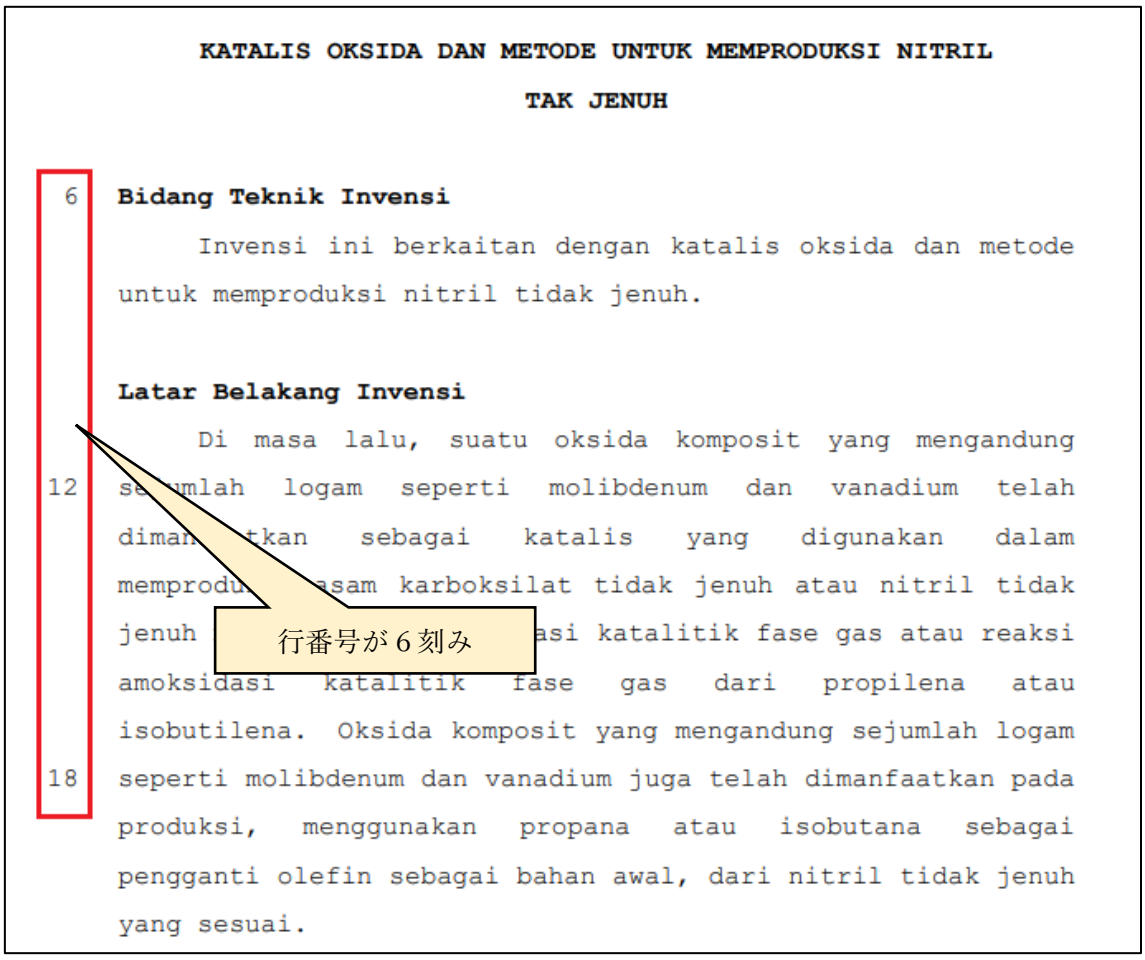


図 6-8　インドネシア PDF 公報 P00202102343 における行番号の状況



結果、これら 3 文献に由来する『対訳コーパス A』の 19 文対（人手確認・修正作業で修正されたものを除く）と、『対訳コーパス A-B』の 194 文対（『対訳コーパス A』と重複するものを含む）、そしてスコア C～D に属するもの 49 文対の計 243 文対が修正された。
下表に内訳を示す。

表 6-15　行番号混入 3 文献に対する修正結果

文献	コーパス A	コーパス A-B	スコア C～D	合計
P00202207045	4 文対	10 文対	4 文対	14 文対
P00202210351	0 文対	2 文対	1 文対	3 文対
P00202102343	15 文対	182 文対	44 文対	226 文対
合計	19 文対	194 文対	49 文対	243 文対

合計欄はコーパス A-B＋スコア C+D（コーパス A は A-B にも含まれるため）。

7．日インドネシア語公報に特化した対訳辞書の作成

本調査では、人手確認・修正作業の対象とした 7 万件の日インドネシア文対における頻出の用語・フレーズについて、約 2,000 語の対訳辞書を作成した。本章で詳細を示す。

7.1. 対訳辞書候補の選定

7.1.1. 対訳辞書候補の取得・選定方法

対訳辞書候補の取得および選定は、以下の手順で行った。

① 統計機械翻訳技術を用いて、日インドネシア文対 7 万件から翻訳モデルを取得する。この翻訳モデルには、日インドネシア文対 7 万件における用語・フレーズの出現状況を統計的に処理して得られる日インドネシア用語対の情報が含まれている。

② 並行して、日インドネシア文対の日本文 7 万文を形態素解析し、名詞句と判定された語句のみをピックアップし、出現頻度をカウントする。

③ 手順②で取得した日本語名詞句と手順①の翻訳モデルの用語対情報を照合し、日本語名詞句と対訳関係となるインドネシア語を取得する。

④ 手順③で取得した日インドネシア用語対について、7 万文対での出現頻度をカウントし、上位 2,000 語＋α[61]を対訳辞書候補とする。

上記処理の結果、出現頻度 1 位の対訳語は「実施形態／perwujudan」の 2,832 回であった。以下、「化合物／senyawa」933 回、「組成物／komposisi」808 回と続く。頻度上位 2,000 位では、出現頻度 6 回の対訳語の一部までカバーされた。

7.1.2. 対訳辞書候補の重複排除

上記手順③で取得される日インドネシア対訳語データは、一部、集約すべき重複が発生した。具体的には、インドネシア語における大文字／小文字の相違の集約である。

対訳語データは 7 万件の文対から抽出している。このため、同じインドネシア語句であっても、文対によって全て小文字のもの、先頭のみ大文字のもの（文頭に存在する場合

[61] 不採用語や修正により重複となる語など、2,000 語のカウント対象外となる語の発生を見越し、本調査では約 6%（122 語）を上乗せした。

等)、全て大文字のもの（発明の名称から抽出された場合等）など大文字／小文字の相違が存在する。

上記手順③では、対訳語データを作成する際、大文字／小文字の差異も別語彙として扱った。そのうえで、大文字／小文字の差異のみの語彙それぞれの出現頻度をカウントし（手順④)、最後に、最も出現頻度の多い語彙に集約した。

こうすることで、通常の語彙は全て小文字のものに集約される一方で、たとえば「ＤＮＡ配列／sekuens DNA」など、大文字の使用が一般的な語彙はそちらに集約されることになる。

7.2. 対訳辞書候補の目視チェック

7.2.1. 対訳辞書候補の目視チェックの手順

7.1.項で取得した対訳辞書候補に対し、日インドネシア語を解する作業者による確認と修正を実施した。確認・修正の手順は以下とした。

① 2,000 語＋α の対訳辞書候補について、対訳語（日本語×インドネシア語）と、その対訳語が用いられた実例を作業者に提示する。

② 作業者は、対訳語が対訳辞書の語彙として適切であるかを確認し、適切であれば対訳辞書に採用する。

③ 対訳語が適切でない場合、添付された実例に基づき[62]、対訳語を適切な形に修正する。

④ ただし、添付された実例から対訳語を適切な形に修正することが不可能である場合（つまり、適切な対訳語が実例で使用されていない場合）は、「不採用」とする。

⑤ 採用した対訳語が特許用語である場合は、フラグを立てて識別可能とする。

⑥ 採用した対訳語が名詞句以外である場合も、別のフラグを立てて識別可能とする。

[62] 実例で使用されていない語句への修正（つまり作業者の任意による語彙選択）は、特許文献由来の対訳語とならなくなるため、(大文字／小文字の置換を除き）すべて禁止とした。

⑦ 手順④で「不採用」となったもの、手順⑤で「特許用語」と判定されたもの、そして手順③での修正により別の対訳語と同一になったものを除外した結果、語数が2,000 に満たなかった場合は、対訳辞書候補を追加で選定し、上記①以降の手順で2,000 語を達成する。

7.2.2. 対訳辞書候補の目視チェックの結果

上記手順で目視チェックを実施した結果を以下に示す。

表 7-1　対訳辞書候補の目視チェック結果一覧

項　目	該当語数	辞書語数
［対訳辞書候補］	2,122 語	
［目視チェック結果］		
・不採用	-12 語	
・採用	2,110 語	2,110 語
・そのまま採用	1,948 語	
・修正して採用	162 語	
・インドネシア語修正	156 語	
・日本語修正	8 語	
・重複排除	-20 語	2,090 語
・特許用語	-90 語	2,000 語
［対訳辞書（技術／一般用語)］	2,000 語	
・うち名詞以外	6 語	

上表のとおり、機械的に作成した対訳辞書候補 2,122 語のうち語彙として採用（修正して採用を含む）されたのは 2,110 語で、採用率は 99.4%ときわめて効率であった。

採用された 2,110 語のうち 1,948 語（92.3%）は修正なしで採用されており、日本語、インドネシア語とも、語句がおおむね適切な単位で抽出されていることが示された。修正された 162 語はインドネシア語の修正が大部分（156 語。うち 2 語は日本語も修正）であり、本調査で使用した手法では日本語の切り出し精度のほうが優秀であったといえる。

なお、修正の結果、他の語句と重複（日本語・インドネシア語とも完全一致）した語句が 43 語と比較的多く見られた。その大部分は、インドネシア語の切り出しに過不足があり、修正された結果、正しく切り出された語句と一致したものであった。今後、本調査と同様の方式で辞書を作成する際は、目視チェック前に「インドネシア語が他の語句中に完

全に含まれ、かつ日本語が同一」の辞書候補を除外しておくことで、より効率性が高められる。

重複排除を経た対訳辞書候補 2,090 語には、特許用語が 90 語含まれた。特許用語は別途『特許用語辞書』として作成するため除外され、技術／一般用語からなる『対訳辞書』の総語彙数は 2,000 語となった。）

7.3. 特許用語辞書の作成

特許公報で通常使用される特許用語については、技術／一般用語からなる『対訳辞書』とは別に『特許用語辞書』として作成した。特許用語辞書には、以下の 3 種の用語、計 128 語を収録した。

① 7.2 で「特許用語」と判定された対訳語　……　90 語
② 文分割後処理[63]に用いた定番見出し語　……　27 語（①との重複排除後）
③ インドネシア公報の書誌事項の見出し語　……　11 語（①、②との重複排除後）

7.4. 対訳辞書における優先順位

本調査で作成した対訳辞書は、日インドネシア対訳語としては全語彙ユニークであるが、日本語、インドネシア語それぞれにおいては、同じ語彙が複数含まれることがある。例えば、「アミノ酸配列」という語彙は、インドネシア語の異なる下記 4 種の語彙が辞書に含まれている。

表 7-2　対訳辞書中の「アミノ酸配列」と 4 種の訳語

日本語	インドネシア語	出現頻度
アミノ酸配列	sekuens asam amino	79
アミノ酸配列	urutan asam amino	54
アミノ酸配列	sekuen asam amino	23
アミノ酸配列	rangkaian asam amino	16

本調査では、こうした一方の言語での重複は辞書から排除しなかった。これは、対訳辞書を双方向（日本語⇒インドネシア語、インドネシア語⇒日本語）の翻訳に対応させるためである。

[63] 3.4.6.項の表 3-6 参照。

対訳辞書を機械翻訳に使用する場合、上記「アミノ酸配列」のように複数の訳語が登録されていても、採用されるのは常に頻度最上位の訳語「sekuens asam amino」となる。したがって、日本語⇒インドネシア語の方向に関しては、それ以外の訳語は不要である。

だが、これら頻度最上位以外の訳語を辞書から排除してしまうと、同じ辞書をインドネシア語⇒日本語の方向で使用した場合に、「アミノ酸配列」と訳すことができるのは「sekuens asam amino」のみであり、「urutan asam amino」ほか下位の語には対応できなくなる。これを避けるため、本調査の辞書では一方の言語における重複は許容した。

8．日インドネシア語対訳コーパスによる機械翻訳エンジンの学習効果の評価

本調査で作成した日インドネシア語対訳コーパス『対訳コーパス A＋』及び『対訳コーパス A-B＋』をニューラル機械翻訳エンジンに段階的に学習させ、各段階の翻訳精度を評価することで、日インドネシア特許文献の機械翻訳に対する精度向上効果を検証した。本章にその結果をまとめる。

8.1. 使用エンジン

機械翻訳エンジンには、ニューラル機械翻訳エンジン「みんなの自動翻訳 TexxTra@」を使用した。日本語からインドネシア語への翻訳には「汎用 NT（日本語→インドネシア語)」を、インドネシア語から日本語への翻訳には「汎用 NT（インドネシア語→日本語)」を選択した。

8.2. エンジンの学習

エンジンの学習は、デフォルトの機械翻訳エンジンに対し、本調査で作成した対訳コーパスを追加学習（アダプテーション）させる形で実施した。

8.2.1. 段階的な追加学習の実施

機械翻訳エンジンへの対訳コーパスの追加学習は、日本語→インドネシア語、インドネシア語→日本語とも、下表に示したように『対訳コーパス A＋』及び『対訳コーパス A-B＋』の全件[64]を用いて段階的に行った。

表 8-1 『対訳コーパス A＋』及び『対訳コーバス A-B＋』の段階的学習の詳細

段階	追加学習データ	備考
⓪	なし	デフォルト（無学習）状態
①	『対訳コーパス A＋』より 6.8 万文対[65]	人手確認・修正対象文対のみ
②	〃 〃 〃 〃 より 19.8 万文対	
③	〃 〃 〃 〃 より 39.8 万文対	
④	〃 〃 〃 〃 より 59.8 万文対	
⑤	〃 〃 〃 〃 より 79.8 万文対	
⑥	〃 〃 〃 〃 より 99.8 万文対	『コーパス A＋』全件
⑦	『対訳コーパス A-B＋』より約 500 万文対	『コーパス A-B＋』全件

[64] ただし 2,000 文対は評価用データに使用するため学習データからは除外した。詳しくは 8.4.項参照。

[65] 7 万文対から評価用データ 2,000 文対を除外したため 6.8 万文対となっている。以下同じ。

本調査で作成した『対訳コーパス A＋』は 100 万文対、『対訳コーパス A-B＋』は約 500 万文対（4,978,974 文対）からなる。かつ、後者は前者を完全に包含している。

この状況を踏まえ、追加学習は、まずは 100 万文対からなる『対訳コーパス A＋』を約 20 万文対ずつ段階的に学習させ（段階②〜⑥）、その後、これと重複しない『対訳コーパス A-B＋』の全件を追加学習させた（段階⑦）。さらに、学習前のデフォルト状態のエンジン（⓪）、ならびに人手確認・修正作業の対象とした 7 万文対のみを学習させたエンジン（①）も用意した。なお、段階②〜⑥は文アライメントスコアの上位順に 20 万文対ずつを学習させた。

本調査では、これら 8 種の学習段階（⓪〜⑦）の機械翻訳エンジンで次項に示す評価用データを都度機械翻訳し、その翻訳精度の変化を追うことで、対訳コーパスの学習効果を評価した。

8.2.2. 学習時に除外されるデータ

「みんなの自動翻訳」では、文長が 300 語を超える学習データは自動的に除外される。ただし、学習させた対訳コーパスにおいて除外が発生したか否かや、具体的にどの文対が除外されたか等の詳細は示されない。また、ここでいう「語」の定義も公開されておらず、どのような基準で語数がカウントしているかも不明であるため、ユーザ側で除外された文対を特定することも難しい。

本調査で学習データに用いた『対訳コーパス A＋』及び『対訳コーパス A-B＋』に含まれる文対の中には、明らかに 300 語を超えるものも存在する。例えば下例は『対訳コーパス A-B＋』で最長レベルの文対のインドネシア文であり、文長は 8,128 文字である。

［P00202215176_JP2023527878A SEQ: 506］

Dalam beberapa, satu atau lebih agen terapeutik tambahan dipilih dari inhibitor enzim pengonversi angiotensin (ACE) a(n), agonis reseptor Adenosin A3, agonis reseptor Adiponektin, inhibitor protein kinase AKT, aktivator kinase AMP, aktivator protein kinase yang diaktifkan AMP (AMPK), agonis reseptor Amilin, antagonis reseptor Angiotensin II AT-1, agonis reseptor androgen, inhibitor kinase 1 pengatur sinyal apoptosis (ASK1), inhibitor sitrat liase ATP, antagonis Apolipoprotein C3 (APOC3), modulator protein Autofagi, inhibitor Autotaksin, inhibitor reseptor tirosin kinase Axl, stimulator protein Bax, agonis Kalsitonin, modulator reseptor Kanabinoid, inhibitor Kaspase, stimulator Kaspase-3, inhibitor Katepsin (misalnya, inhibitor katepsin B), inhibitor Kaveolin 1, antagonis kemokin CCR2, antagonis kemokin CCR3, antagonis

kemokin CCR5, antagonis CD3, stimulator kanal Klorida, pelarut kolesterol, inhibitor CNR1, inhibitor Siklin D1, inhibitor Sitokrom P450 7A1, inhibitor Sitokrom P450 2E1 (CYP2E1), inhibitor Diasilgliserol O asiltransferase 1 (DGAT1), inhibitor inhibitor Diasilgliserol O asiltransferase 2 (DGAT2), …（中間省略）… seperti verinurad (RDEA3170); Agonis reseptor VIP 1/VIP 2, seperti LBT-3627; dan Inhibitor xantin oksidase, seperti TMX-049, TMX-049DN.

このインドネシア文の文法的な語数（つまりスペースで区切られる文字列数）は 946 語であった。1 語あたり平均 8.59 文字であり、単純計算では 300 語で 2,577 文字と換算できる。各コーパス中、インドネシア文長が 2,577 文字を超える文対が学習データから自動的に除外されると仮定すると、『対訳コーパス A＋』では 702 文対、『対訳コーパス A-B＋』では 1,275 文対が該当するが、いずれも全体数に比してごく少量（0.07%、0.03%）であり、本調査で作成した対訳コーパスにおいては例外的な存在と見なせる。

8.3. 自動評価と人手評価

各学習段階における機械翻訳品質の評価は、自動評価と人手評価により実施した。自動評価指標には BLEU と RIBES の双方を使用した。人手評価は、特許庁が公表している「特許文献機械翻訳の品質評価手順（ver1.0）[66]」に示された各評価項目をベースに、本調査に必要な評価項目を追加する形で実施した。

8.4. 評価用データ

機械翻訳の評価を行うためには、まず機械翻訳に入力する原文データが必要である。また、機械翻訳結果と比較するための正解データ（参考訳文）に用いるため、原文データの人手翻訳文も必要となる。つまり、原文とその人手翻訳文の対訳データが一定量必要である。

本調査では、人手確認・修正を経た 7 万文対から 2,000 文対を選定して評価用データとした。日本→インドネシア方向の機械翻訳の評価では各文対の日本文が原文データ、インドネシア文が正解データとなり、インドネシア→日本方向では両者が逆になる。なお、自動評価用データには 2,000 文対の全件を用い、人手評価データはそこから 100 文を選定した。

また、評価用データに選定した文対は全て学習データから除外した。評価用データと同一の文対を学習データに用いると機械翻訳精度面で不当に有利となり、実用時の品質のシミュレーションとならないためである。

---

[66] https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html

以下、自動評価用／人手評価用データ選定の詳細を記す。

8.4.1. 自動評価用データ

自動評価用データには、『対訳コーパス A＋』のうち人手確認・修正作業の対象とした 7 万文対から 2,000 文対を選定した。選定に際しては、各項目について十分なサンプルを含めるため、発明の名称から 100 文対、要約と特許請求の範囲から各 500 文対、明細書から 900 文対を選定した。また、特定の技術内容に偏らぬよう、各項目とも一文献からの選定は 1 文対のみとした[67]。

技術分野については、極力、母集合 7 万文対における筆頭 IPC のセクションの構成率に沿った文対数となるよう留意した。ただし、特に明細書以外の 3 項目は 7 万文対での絶対数が少なく、かつ一文献から 1 件のみという方針を優先したことで、若干の偏りが生じた。

下表に自動評価用データ 2,000 件の選定結果を項目×IPC セクション別に示す。あわせて、母集合である 7 万文対の状況も参考に示す。

表 8-2　自動評価用データ 2,000 件の詳細

項目	文対数	筆頭 IPC のセクション							
		A	B	C	D	E	F	G	H
発明の名称	100 文対	17	8	38	1	0	8	10	18
要約	500 文対	110	40	150	18	12	20	55	95
特許請求の範囲	500 文対	130	19	237	3	2	3	27	79
明細書	900 文対	195	90	222	29	28	51	83	202
合計	2,000 文対	452	157	647	51	42	82	175	394

表 8-3　人手確認対象 7 万文対の詳細［参考］

項目	文対数	筆頭 IPC のセクション							
		A	B	C	D	E	F	G	H
発明の名称	215 文対	36	18	93	1	0	8	21	38
要約	1,820 文対	410	155	604	18	12	71	202	348
特許請求の範囲	902 文対	238	31	440	6	6	4	44	133
明細書	67,063 文対	14,489	5,275	22,360	812	436	1,893	6,764	15,034
合計	70,000 文対	15,173	5,479	23,497	837	454	1,976	7,031	15,553

[67] 項目間での重複は若干数発生した。

8.4.2. 人手評価用データ

人手評価用データは、自動評価用データの中から 100 文を目視で選定した。項目ごとの選定数も自動評価用データの比率を維持し、発明の名称から 5 文対、要約と特許請求の範囲から各 25 文対、明細書から 45 文対とした。選定にあたっては、文中に同一の技術用語が複数回用いられているものを選定した。ニューラル機械翻訳における大きな課題である「技術用語の訳ゆれ」の発生状況を各文対で把握するためである。

なお、この「同一の技術用語が複数回用いられている」という条件に合致する文対は非常に少なく、このため IPC セクションについては均等に選定することは不可能であった。下表に項目別、IPC セクション別の人手評価データ 100 件の内訳を示す。

表 8-4　人手評価用データ 100 件の詳細

項目	文対数	筆頭 IPC のセクション							
		A	B	C	D	E	F	G	H
発明の名称	5 文対	0	0	1	0	0	0	3	1
要約	25 文対	9	5	5	1	0	1	1	3
特許請求の範囲	25 文対	4	1	11	0	0	0	3	6
明細書	45 文対	13	3	9	0	0	6	4	10
合計	100 文対	26	9	26	1	0	7	11	20

8.5. 自動評価の結果

各学習段階の機械翻訳エンジンによる 2,000 文の自動評価用データの機械翻訳結果に対して、自動評価指標 BLEU 及び RIBES を用いて翻訳品質の評価スコアを取得した。本項にその結果を示す。

なお、BLEU、RIBES とも、評価には参照訳文が必要となる。どちらの手法も、参照訳文と機械翻訳結果とを比較し、その類似度によって評価スコアを算出するためである。したがって、自動評価用データの各文対は、一方が機械翻訳の対象文、もう一方が参照訳文として用いられる。日本→インドネシアの機械翻訳評価においては自動評価用データの各日本文が機械翻訳の対象文、各インドネシア文が参照訳文となり、インドネシア→日本の機械翻訳評価においては両者が逆となる。

8.5.1. インドネシア→日本の機械翻訳品質の自動評価結果（BLEU、RIBES）

自動評価用データのインドネシア文 2,000 文を各学習段階の機械翻訳エンジンで日本語に機械翻訳した結果に対し、BLEU 及び RIBES による自動評価を実施した。各段階の BLEU 及び RIBES スコアの全件平均値の推移を下表に示す。

表 8-5 「みんなの自動翻訳」の学習段階別 BLEU／RIBES スコア平均値（イン⇒日）

段階 (学習量)	⓪ 0	① 6.8 万	② 19.8 万	③ 39.8 万	④ 59.8 万	⑤ 79.8 万	⑥ 99.8 万	⑦ 500 万
BLEU	0.421	0.539	0.512	0.534	0.555	0.573	0.602	0.658
増減	－	+0.118	-0.027	+0.022	+0.021	+0.018	+0.029	+0.056
RIBES	0.829	0.874	0.864	0.875	0.880	0.885	0.894	0.908
増減	－	+0.045	-0.010	+0.011	+0.005	+0.005	+0.009	+0.014

上表のとおり、全体的には BLEU、RIBES スコアとも学習データが増加するごとに順調に良化している。BLEU、RIBES ともスコアが最も伸びたのが無学習状態のエンジンに『対訳コーパス A＋』を初めて学習させた段階①であり、その後もおおむね学習データ量に比例したスコアが伸びている。また、大規模な『対訳コーパス A-B＋』を学習させた段階⑦では段階①に次ぐスコアの向上が見られた。『対訳コーパス A-B＋』は『対訳コーパス A＋』に比べてやや品質が劣ると考えられるが、膨大なデータ量により、それを上回るメリットが生じたと結論される。

なお、両スコアとも段階②のみ一時的にスコアが下落している。考えられる理由としては、段階①の学習データ 6.8 万文対が全て人手確認・修正済みデータであり、最上級の品質であったのに対し、段階②の学習データは『対訳コーパス A＋』で文アライメントスコアが最上

位の 19.8 万文対であり、人手確認・修正済みデータはほとんど含まれていない。両段階の学習データに品質面で大きな差があるとは考えにくいが、データの内容はほぼ完全に異なるため、偶然の偏りによって学習効果の多寡が生じ、一時的なスコアの下落が生じた可能性がある。

全コーパスを学習させた段階⑧では、BLEU が 0.658、RIBES が 0.908 という高スコアに到達した。双方とも無学習時から 0.237 ポイント、0.079 ポイントと顕著に上昇しており、本調査で作成した対訳コーパスにより、特許文献のインドネシア語⇒日本語方向の機械翻訳において高い学習効果が得られることが実証された。

8.5.2. 日本→インドネシアの機械翻訳品質の自動評価結果（BLEU、RIBES）

続いて、自動評価用データの日本文 2,000 文を各学習段階の機械翻訳エンジンでインドネシア語に機械翻訳した結果に対する各段階の BLEU 及び RIBES スコアの全件平均値の推移を示す。

表 8-6 「みんなの自動翻訳」の学習段階別 BLEU／RIBES スコア平均値（日⇒イン）

段階 (学習量)	⓪ 0	① 6.8 万	② 19.8 万	③ 39.8 万	④ 59.8 万	⑤ 79.8 万	⑥ 99.8 万	⑦ 500 万
BLEU	0.371	0.467	0.447	0.462	0.479	0.506	0.533	0.583
増減	－	+0.096	-0.020	+0.016	+0.016	+0.028	+0.027	+0.050
RIBES	0.857	0.884	0.880	0.884	0.889	0.893	0.896	0.908
増減	－	+0.026	-0.004	+0.004	+0.005	+0.004	+0.003	+0.012

日本語→インドネシア語方向の自動評価結果には、インドネシア語→日本語方向と共通する点と、相違する点とが見られた。

共通する点としては、BLEU、RIBES とも段階②での一時的な下落を除き学習データ量に応じて順当にスコアが上昇していること、段階①で最も顕著な上昇が見られたこと、『対訳コーパス A-B＋』を学習させた段階⑦でもそれに次ぐ大幅な上昇が見られたことが挙げられる。それぞれの理由も、前項に記したインドネシア語→日本語についての考察がそのまま適用できる。

一方、相違点としては、日本語→インドネシア語方向では段階⑦で到達した BLEU スコアが 0.583 とやや低く、また無学習時からのスコアの伸びも BLEU で 0.212、RIBES で 0.051 とやや鈍い点が挙げられる。考え得る理由としては、本調査で作成した対訳コーパスでは総じてインドネシア文に不備が多いことが挙げられる。このことが、出力されるインドネシア

文の翻訳品質に影響した可能性がある。

とはいえ、BLEU、RIBES とも無学習時から大幅なスコアアップが果たされており、日本語→インドネシア語方向においても、本調査で作成した対訳コーパスが高い学習効果を有することは十分に示された。

8.5.3. 項目別スコア集計結果

自動評価用データ 2,000 文は、発明の名称由来の 100 文対、要約由来の 500 文対、特許請求の範囲由来の 500 文対、そして明細書由来の 900 文対で構成される。この 4 種の項目別に自動評価スコアの平均値を集計した結果を下 2 表に示す。

表 8-7　項目別 BLEU／RIBES スコア平均値の推移（イン⇒日）

項目		⓪	①	②	③	④	⑤	⑥	⑦
発明の名称 100	B	0.343	0.482	0.499	0.488	0.487	0.527	0.525	0.513
		－	+0.139	+0.017	-0.011	-0.001	+0.040	-0.002	-0.012
	R	0.812	0.871	0.883	0.893	0.892	0.905	0.901	0.886
		－	+0.059	+0.012	+0.010	-0.001	+0.013	-0.004	-0.015
要　約 500	B	0.382	0.457	0.451	0.463	0.472	0.482	0.495	0.533
		－	+0.075	-0.006	+0.012	+0.009	+0.010	+0.013	+0.038
	R	0.817	0.848	0.847	0.850	0.855	0.857	0.864	0.873
		－	+0.031	-0.001	+0.003	+0.005	+0.002	+0.007	+0.009
特許請求の範囲 500	B	0.530	0.657	0.625	0.659	0.683	0.704	0.744	0.810
		－	+0.127	-0.032	+0.034	+0.024	+0.021	+0.040	+0.066
	R	0.864	0.920	0.901	0.921	0.921	0.925	0.939	0.958
		－	+0.056	-0.019	+0.020	0.000	+0.004	+0.014	+0.019
明細書 900	B	0.392	0.526	0.484	0.508	0.537	0.557	0.592	0.659
		－	+0.134	-0.042	+0.024	+0.029	+0.020	+0.035	+0.067
	R	0.817	0.863	0.852	0.860	0.870	0.876	0.886	0.902
		－	+0.046	-0.011	+0.008	+0.010	+0.006	+0.010	+0.016

表 8-8　項目別 BLEU／RIBES スコア平均値の推移（日⇒イン）

項目		⓪	①	②	③	④	⑤	⑥	⑦
発明の名称 100	B	0.391	0.461	0.472	0.443	0.473	0.466	0.484	0.482
		－	+0.070	+0.011	-0.029	+0.030	-0.007	+0.018	-0.002
	R	0.817	0.880	0.871	0.876	0.879	0.886	0.882	0.886
		－	+0.063	-0.009	+0.005	+0.003	+0.007	-0.004	+0.004
要　約 500	B	0.328	0.397	0.385	0.395	0.399	0.408	0.420	0.439
		－	+0.069	-0.012	+0.010	+0.004	+0.009	+0.012	+0.019
	R	0.843	0.867	0.861	0.867	0.868	0.869	0.871	0.874
		－	+0.024	-0.006	+0.006	+0.001	+0.001	+0.002	+0.003
特許請求の範囲 500	B	0.410	0.511	0.492	0.504	0.531	0.588	0.638	0.722
		－	+0.101	-0.019	+0.012	+0.027	+0.057	+0.050	+0.084
	R	0.895	0.910	0.910	0.911	0.918	0.926	0.932	0.950
		－	+0.015	0.000	+0.001	+0.007	+0.008	+0.006	+0.018
明細書 900	B	0.371	0.482	0.454	0.478	0.494	0.519	0.542	0.597
		－	+0.111	-0.028	+0.024	+0.016	+0.025	+0.023	+0.055
	R	0.849	0.879	0.874	0.880	0.886	0.889	0.891	0.905
		－	+0.030	-0.005	+0.006	+0.006	+0.003	+0.002	+0.014

表 8-7 がインドネシア語→日本語方向、表 8-8 が日本語→インドネシア語方向のスコア平均値である。双方とも、おおむね同様の傾向が見て取れる。

まず段階間の平均スコアの推移に関しては、発明の名称のみ各段階でスコアがやや不規則に増減しているが、その他の 3 項目は、前項で示した全件平均と同じく、段階②で一時的にスコアが低下した以外は、段階を追うごとに順調にスコアが向上している。『対訳コーパス A＋』初回学習時（段階①）で最も大きくスコアが伸び、『対訳コーパス A-B＋』全件学習時（段階⑦）がそれに次ぐという状況も両方向、3 項目でおおむね一致している。

発明の名称のみは後期段階においてもスコアが低下するケースが多く見られ、他の 3 項目と印象が異なるが、これはサンプル数が他より顕著に少なく、平均スコアの粒度が粗いための誤差と見るべきであろう。全体的な推移を見れば、平均スコア最上位はイン→日方向では BLEU、RIBES とも段階⑤、日→イン方向では BLEU が段階⑥、RIBES が段階⑤と⑦が同値と、いずれも後期段階でスコアが最良となっており、学習データを増強することで翻訳品質が改善している状況は見て取れる。

一方、項目別のスコア水準の比較では、両方向、2つの評価指標とも、全段階において特許請求の範囲のスコアが他の3項目から突出して高くなった。他の3項目は翻訳方向や評価指標、段階によって順位の変動はあるものの、特許請求の範囲の突出度と比較すれば同程度の範ちゅうであった。

特定の項目のみ翻訳精度が顕著に高くなる理由としては、まず学習データの偏りが想起される。しかし、今回の学習データである『対訳コーパスA』及び『対訳コーパスA-B』における特許請求の範囲由来の文対の構成比はそれぞれ0.7%、4.0%とごく低く（⇒5.5.1.、5.5.2.)、この状況には該当しない。

学習データに偏りがないとなると、評価用データの翻訳難易度に項目間で差があったと考えられる。機械翻訳の難易度は概して文長に比例するため、各項目の評価用データの平均文長を比較した結果を下表に示す。

表8-9　評価用データの項目別平均文長

項目		インドネシア文	日本文
発明の名称	100文	75.9文字	28.6文字
要約	500文	262.6文字	98.0文字
特許請求の範囲	500文	182.5文字	70.4文字
明細書	900文	312.4文字	120.1文字
全　体	2,000文	255.7文字	97.6文字

上表のとおり、特許請求の範囲由来の文対の平均文長は、インドネシア、日本文とも全件平均を大きく下回っており、要約や明細書に比べれば翻訳難易度はやや低かったと見なせる（発明の名称の平均文長はさらに短いが、自然文でないため翻訳難易度の比較が難しく、ここでは考慮しない)。

特許請求の範囲に由来する文対、すなわち請求項は、長大で複雑な独立請求項と、短文かつ文構造がパターン化されている従属請求項とに大別される。前者は翻訳難易度が極めて高い一方、後者は短文であり、かつ文構造が文献間でパターン化されているため、特定のパターンを有さない要約や明細書に比べても翻訳難易度は低くなる。

評価用データの抽出元である人手確認対象7万文対のうち、特許請求の範囲由来の文対は902文対とごく少なく、このうち821文対が従属請求項であった[68]。評価用データ500

[68] 日本文に「請求項」が含まれる文対を従属請求項と見なした。

文はこの母集合から選定しており、必然的に従属請求項が 454 文対と大半を占めた。残る独立請求項由来の文対も、ほぼ全て、請求項の一部分（例えば一構成要素の説明部分）のみが分割された短文が大半であった。その結果、他の項目に比べて全体的な翻訳難易度が低くなり、評価スコアが突出したと考えられる。

他の 3 項目（発明の名称、要約、明細書）に関しては、平均スコアが極端に低いものはなく、おおむね同等の翻訳精度と評価された。学習データである『対訳コーパス A+』及び『対訳コーパス A-B+』は明細書由来の文対の比率が圧倒的（97.4%、93.3%）であり（⇒5.5.1.～2.）、このため他の 3 項目の翻訳精度が相対的に低くなる懸念があったが、今回の評価結果を見る限りそのような問題は生じていない。このことから、特許文献から対訳コーパスを作成する際、各項目の構成比の偏りが翻訳精度に悪影響を及ぼす懸念は小さく、構成比を無理に均等化させるなど過度に配慮する必要性は低いと結論される。

8.5.4. IPC セクション別スコア集計結果

自動評価用データ 2,000 件の自動評価スコアを筆頭 IPC のセクション別（A～H）に集計した結果を下表に示す。「向上」欄は段階⓪から段階⑦にかけてのスコアの伸びを示した。

表 8-10　IPC セクション別 BLEU／RIBES スコア平均値の推移（イン⇒日）

IPC		⓪	①	②	③	④	⑤	⑥	⑦	向上
A	B	0.418	0.531	0.503	0.528	0.541	0.560	0.596	0.655	0.237
	R	0.826	0.871	0.857	0.868	0.873	0.878	0.890	0.912	0.086
B	B	0.350	0.479	0.448	0.465	0.491	0.514	0.538	0.602	0.252
	R	0.815	0.856	0.852	0.853	0.860	0.860	0.874	0.887	0.072
C	B	0.466	0.567	0.542	0.561	0.583	0.600	0.628	0.678	0.212
	R	0.850	0.886	0.879	0.887	0.892	0.896	0.908	0.917	0.067
D	B	0.324	0.414	0.409	0.425	0.443	0.456	0.462	0.571	0.247
	R	0.795	0.827	0.837	0.845	0.849	0.852	0.853	0.881	0.086
E	B	0.371	0.485	0.461	0.435	0.461	0.462	0.523	0.584	0.213
	R	0.801	0.849	0.849	0.841	0.837	0.855	0.873	0.871	0.070
F	B	0.370	0.473	0.464	0.468	0.479	0.494	0.535	0.595	0.225
	R	0.819	0.872	0.860	0.868	0.863	0.859	0.871	0.880	0.061
G	B	0.401	0.503	0.490	0.500	0.531	0.549	0.565	0.631	0.230
	R	0.813	0.850	0.850	0.851	0.860	0.872	0.872	0.889	0.076
H	B	0.420	0.579	0.537	0.577	0.603	0.622	0.651	0.697	0.277
	R	0.818	0.884	0.867	0.891	0.897	0.902	0.909	0.917	0.099

表 8-11　IPC セクション別 BLEU／RIBES スコア平均値の推移（日⇒イン）

IPC		⓪	①	②	③	④	⑤	⑥	⑦	向上
A	B	0.359	0.453	0.438	0.443	0.463	0.492	0.511	0.575	0.216
	R	0.854	0.881	0.879	0.878	0.888	0.893	0.891	0.907	0.053
B	B	0.332	0.435	0.406	0.435	0.427	0.463	0.483	0.535	0.203
	R	0.847	0.875	0.865	0.860	0.865	0.876	0.873	0.894	0.047
C	B	0.389	0.484	0.465	0.477	0.498	0.527	0.560	0.599	0.210
	R	0.862	0.889	0.886	0.890	0.895	0.900	0.907	0.910	0.048
D	B	0.319	0.382	0.395	0.403	0.406	0.419	0.436	0.500	0.181
	R	0.832	0.859	0.856	0.863	0.865	0.862	0.866	0.884	0.052
E	B	0.321	0.422	0.372	0.398	0.391	0.420	0.458	0.501	0.180
	R	0.829	0.883	0.867	0.871	0.865	0.866	0.877	0.881	0.052
F	B	0.303	0.382	0.357	0.363	0.374	0.404	0.436	0.527	0.224
	R	0.841	0.854	0.846	0.859	0.848	0.845	0.856	0.880	0.039
G	B	0.391	0.482	0.457	0.474	0.494	0.508	0.536	0.580	0.189
	R	0.857	0.882	0.876	0.884	0.891	0.891	0.889	0.908	0.051
H	B	0.388	0.495	0.472	0.502	0.519	0.547	0.571	0.616	0.228
	R	0.868	0.894	0.890	0.901	0.905	0.907	0.909	0.921	0.053

上 2 表とも、各段階で最も高いスコアを青字、低いスコアを赤字で示している。全体の傾向として、インドネシア→日本語方向では、初期段階は C セクション、中盤以降は H セクションが最も高スコアで、D セクションがほぼ全段階で最も低スコアとなった。一方、日本→インドネシア語方向では、ほぼ全段階において H セクションが最も高スコア、F セクションが最も低スコアであった。

各セクションとも異なる文集合であり、そもそもの翻訳難易度が均一ではない。このため、必ずしもスコアが低いセクションの学習データが相対的に貧弱であるとは言えない。事実、上記傾向は対訳コーパス学習前の段階⓪の時点から見て取れる。

対訳コーパスの学習効果の分野間の差異は、むしろ、「向上」欄に示した段階⓪から段階⑦にかけてのスコアの伸びで比較するべきであろう。スコアの伸びの最上位と最下位は、インドネシア→日本語方向では最上位が H セクション、最下位が C セクション（BLEU）と F セクション（RIBES）であり、日本→インドネシア語方向も最上位は H セクション、最下位は E セクション（BLEU）と F セクション（RIBES）であり、両方向おおむね同傾向である。

ここで、『対訳コーパス A+』100 万文対ならびに『対訳コーパス A-B+』約 500 万文対の筆頭 IPC セクション別の構成比を下表に示す。

表 8-12 『対訳コーパス A』IPC セクション別構成比

IPC	対訳コーパス A+		対訳コーパス A-B+	
A	210,783	21.1%	1,009,068	20.3%
B	65,118	6.5%	512,257	10.3%
C	323,760	32.4%	1,531,730	30.8%
D	9,627	1.0%	71,379	1.4%
E	5,356	0.5%	45,826	0.9%
F	22,840	2.3%	211,048	4.2%
G	105,224	10.5%	465,591	9.4%
H	257,292	25.7%	1,132,075	22.7%
合計	1,000,000	100.0%	4,978,974	100.0%

上表のとおり、各 IPC セクションに属する文対数や構成比には大きな偏りがある。本調査では、入手できる限りのファミリー文献を対訳コーパスのソースに用いている。ファミリー文献の件数自体が技術分野によって大きく異なるため、そこから採取した対訳コーパスの文対数にもそれに準じた技術分野の偏りが生じるのは当然のことである。

スコアの伸びが良好であった H セクション、低調であった F、E セクションを上表に照らすと、対訳コーパス量（すなわち学習データ量）が 8 セクション第 2 位と多かった H セクションは学習効果が高く、コーパス量が少ない F、E セクションは学習効果が相対的に低かったことがわかる。したがって、学習データ量の差がスコアの伸び具合に表れたとも解釈できる。だが、各コーパスでの構成比が最大である C セクションはインドネシア→日本語方向の BLEU で最下位となるなどスコアの伸び具合は平凡であるという重大な矛盾も存在する。そもそも学習データ量の極端な差に比べると、各セクション間のスコア伸び率の差異は概して小さい。

これら IPC セクション別評価スコアの状況から判断して、本調査で作成した対訳コーパスは、技術分野ごとの構成比に顕著な偏りはあるものの、学習データに用いた際に各技術分野の機械翻訳品質に極端な優劣が生じる懸念は小さいと考えられる。

8.6. 人手評価の結果

人手評価文 100 文を各学習段階の機械翻訳エンジンで翻訳した結果に対して、人手評価を実施した。本項にその結果を示す。

8.6.1 人手評価の内容

人手評価は、特許庁が公表している「特許文献機械翻訳の品質評価手順（ver1.0）」の「内容伝達レベルの評価」「重要技術用語の評価」「流暢さの評価」の 3 種の評価と、これに加えて機械翻訳における典型的な 10 種の不備の有無をカウントする「誤訳のカテゴリ別チェック」、計 4 種の評価を実施した。以下、各評価の内容を示す。

8.6.1.1. 内容伝達レベルの評価

「内容の伝達レベルの評価」では、機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを、正解データ（参考訳文）の内容に照らして、下記 5 段階の評価基準で主観的に評価した。

> 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がほとんどない。（～20%）

ここでいう「重要情報」とは、原文に含まれる全ての実質情報（技術的要素とその相互関係）を指す。文に含まれる全ての重要情報を対象に、それぞれの重要度と翻訳精度を考慮して、機械翻訳結果の内容の伝達レベルを総合的・主観的に評価した。なお、各評点に付記したパーセンテージは、機械翻訳結果が原文の意味をどの程度正確に伝達しているかを示す大まかな目安であり、「重要情報」の個数から厳密に算出するものではない。

8.6.1.2. 重要技術用語の翻訳品質評価

人手評価対象文は、各文について 1 語、その文で使われている技術用語を「重要技術用語」として選定してある（そしてその対訳文である参考訳文で用いられている訳語が「正解訳語」となる）。この技術用語について、機械翻訳結果において適切に翻訳されているかを以下の評価基準 A～D で評価した。

A（適訳語）：正解訳語に照らし、技術的に同義かつ一般的に用いられる訳語である。
B（可訳語）：技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。
C（誤訳語）：誤訳である。
D（不訳語）：訳漏れ、もしくは原語のままである。

上掲の「適訳語（評価 A）」は、技術的に適切であれば必ずしも正解訳語と同一でなくてもよい。具体的には、「検索」用途に有用な訳語、すなわち検索の際に同義語展開の範疇に含められる語であれば「適訳語」と見なした。これに対して「可訳語（評価 B）」は、一般的な訳語でないため検索用途には向かないものの、意味はおおむね通じ、「照会（粗読）」用途には有用である訳語を指す。

8.6.1.3. 流暢さの評価

機械翻訳結果の「文としての読みやすさ、理解しやすさ」のみを、下記 5 段階の基準で評価した。つまり、機械翻訳結果を独立した文として扱い、原文や参考訳文との整合性を考慮せずに日本文／インドネシア文として自然かどうかのみを評価した。

5：文意が明解で、人間が書いた日本文／インドネシア文に近い。
4：日本文／インドネシア文として不自然な箇所を若干含むが、文意は明解である。
3：日本文／インドネシア文として不自然な箇所があり、文意がわかりにくい。
2：日本語／インドネシア語の文法規則に反する表現をかなり含む。
　文意がわからない。
1：日本文／インドネシア文として成立していない。

8.6.1.4. 誤訳のカテゴリ別チェック

機械翻訳結果に対し、下記①～⑨の各カテゴリに該当する誤訳の有無をカウントした。

表 8-13　機械翻訳における典型的な誤訳カテゴリ 9 種とその概要

カテゴリ	説明
① 技術用語の誤訳	技術用語が誤訳されている。 ※翻訳されていない場合は④又は⑤。
② 一般用語の誤訳	一般用語が誤訳されている。 ※翻訳されていない場合は④又は⑤。
③ 係り受けの誤り	文中の各要素の関係性（係り受け）が正しくない。
④ 訳抜け	原文中の語句が翻訳されず、欠落している。
⑤ 未知語	原文中の語句（またはその一部）が翻訳されず、原語のままである。
⑥ 湧き出し	原文に存在しない語句が出現している。
⑦ 数値・記号エラー	原文中の数値や記号が正しく反映されていない。
⑧ 訳ゆれ	文中に複数回出現する技術用語の訳語が統一されていない（異なる訳語で訳されている）。 ※訳語が適訳でない（①や②に該当）場合も、同じ訳語が使用されていれば「⑧訳ゆれ」とは判定しない。
⑨ その他	上記①～⑧に該当しない重大な不備が存在する。

8.6.2. インドネシア→日本の機械翻訳品質の人手評価

インドネシア→日本の翻訳方向の機械翻訳文の人手評価結果を本項にまとめる。

8.6.2.1. 内容伝達レベルの評価結果

人手評価文 100 文の各段階における内容伝達レベル（5〜1）の集計結果は下表のとおりであった。

表 8-14　内容伝達レベルの評価結果（イン⇒日）

	レベル 5	レベル 4	レベル 3	レベル 2	レベル 1	平均と増減
段階⓪	50	27	19	4	0	4.23
段階①	55	36	6	2	1	4.42(+0.19)
段階②	58	33	8	1	0	4.48(+0.06)
段階③	57	33	9	0	1	4.45(-0.03)
段階④	60	33	7	0	0	4.53(+0.08)
段階⑤	59	36	5	0	0	4.54(+0.01)
段階⑥	68	26	6	0	0	4.62(+0.08)
段階⑦	74	19	5	2	0	4.65(+0.03)

人手評価文 100 文の平均レベルは、無学習状態の段階⓪の 4.23 から学習データを増強するごとに向上していき、最終段階の段階⑦では 4.65 に到達した。段階③のみ一時的に前段階から下落しているものの、全体的には自動評価と同じ傾向を示しており、自動評価結果の妥当性を裏付ける結果となった。

翻訳品質の水準に関しても、無学習状態の段階⓪の時点から 100 件中 77 件がレベル 4 以上（レベル 5 が 50 件、レベル 4 が 27 件）と良好な品質であり、裏を返せばさらなる改善の余地が少ない状況であったが、学習データを増強する都度品質はさらに向上し、最終的には『対訳コーパス A-B＋』全件を学習させた段階⑦でレベル 4 以上が 93 件（レベル 5 が 74 件、レベル 4 が 19 件）と機械翻訳としては最上級といえるレベルに到達した。

段階⑦でレベル 2 と低評価であった 2 文（名 04、明 35）についても、前者に関しては段階⓪の時点での評価はレベル 5 であり、以降も段階①でレベル 2 となった以外は各段階ともレベル 5 であった。後者も段階⓪の時点でレベル 3 であり、段階③〜⑤ではレベル 4 となっている。つまり、両文とも、段階⑦よりも学習データが少ない状況で、すでにレベル 5 やレベル 4 の品質に到達している。こうした状況から、これら 2 文は学習データの不足によって恒常的に低品質となっているわけではなく、ニューラル機械翻訳に特徴的であ

る、学習データの全体構成が変化することに伴う突発的な訳文・訳語の変化により、一時的に低品質となったものと見られる。

ニューラル機械翻訳は、学習データの全体構成が変わる都度、多数の文の翻訳結果が変化することが知られている。その結果、上記2文のような一時的な品質低下が発生する場合がある。前記したとおり、最終段階で低評価であった2文もそれ以前の段階ではレベル5や4と高い評価を得ている。調査したところ、人手評価100文中、全段階を通じて恒常的にレベル3以下と評価された文は存在しておらず、したがって学習データ自体は、潜在的には100文全てを高品質に翻訳できるだけの質と量に達していると見なせる。

8.6.2.2. 重要技術用語の翻訳品質の評価結果

次に、人手評価文100文それぞれについて選定した重要技術用語の機械翻訳品質(A～D)の集計結果について本項に示す。

表 8-15　重要技術用語の内容伝達レベルの評価結果（イン⇒日）

	A	B	C	D
段階⓪	83	13	4	0
段階①	86	13	0	1
段階②	82	17	1	0
段階③	84	13	1	2
段階④	87	12	1	0
段階⑤	90	9	1	0
段階⑥	86	12	1	1
段階⑦	92	5	2	1

重要技術用語の評価結果も、段階⓪の時点から評価A（適訳語）が83%と高かったが、その後の段階的学習により適訳語率はおおむね順調に向上し、最終の段階⑦で92%ときわめて高い水準に達するなど、内容伝達レベルとほぼ同じ傾向を示した。

なお、段階⑦で評価C（誤訳語）～D（未訳語）と低評価であった3文（要07、要12、明35）とも、それ以前の段階では評価Aが多く(各文とも6回)、この点も内容伝達レベルの評価と同様、学習データの質量ではなくニューラル機械翻訳特有の突発的な訳文変化によるものと見なせる。

8.6.2.3. 流暢さの評価結果

原文との整合性等は不問とし、機械翻訳結果の日本文としての自然さのみを 5〜1 の 5 段階で評価した「流暢さの評価」の集計結果は下表のとおりであった。

表 8-16　流暢さの評価結果（イン⇒日）

	スコア 5	スコア 4	スコア 3	スコア 2	スコア 1	平均と増減
段階⓪	59	25	13	3	0	4.40
段階①	52	38	8	1	1	4.39(-0.01)
段階②	59	33	6	2	0	4.49(+0.10)
段階③	62	31	5	1	1	4.52(+0.03)
段階④	64	31	3	2	0	4.57(+0.05)
段階⑤	60	33	5	1	1	4.50(-0.07)
段階⑥	61	36	1	2	0	4.56(+0.06)
段階⑦	74	22	4	0	0	4.70(+0.14)

上表を見ると、各段階の評価 5 の件数やスコア平均値の増減など、内容伝達レベルや重要技術用語に比べると若干不安定な変動も見られるが、基本的には流暢さに関しても学習データが増強されるにつれて全体品質は向上しているといえる。その水準も、評価 4 以上の割合が段階⓪の時点で 84%、最終の段階⑦では 96%に達しており、きわめて流暢な日本文が出力されている。

8.6.2.4. 誤訳のカテゴリ別チェック結果

機械翻訳の典型的な 9 種の誤訳類型について、各文での発生の有無をカウントした。下表に集計結果を示す。

表 8-17　誤訳のカテゴリ別チェック結果（イン⇒日）

誤訳カテゴリ／学習段階	⓪	①	②	③	④	⑤	⑥	⑦
① 技術用語の誤訳	10	5	6	5	6	5	5	2
② 一般用語の誤訳	7	8	6	5	5	6	4	2
③ 係り受けの誤り	13	7	7	8	7	5	4	1
④ 訳抜け	13	5	0	6	8	4	3	5
⑤未知語	2	1	1	1	1	2	0	1
⑥ 湧き出し	2	6	5	5	5	5	7	11
⑦ 数値・記号エラー	8	8	7	6	8	8	9	10
⑧ 訳ゆれ	0	0	0	0	0	1	0	0
⑨ その他	0	1	0	1	0	0	0	0

各カテゴリの誤訳の段階別件数の推移を見ると、学習が進むにつれて発生数が低減していくタイプと、各段階で一定数発生するタイプとに大別される。前者は「①技術用語の誤訳」「②一般用語の誤訳」「③係り受けの誤り」が該当し、後者はそれ以外（④～⑨）が該当する。換言すれば、前者は「学習データの増強によって発生件数を抑制できる」タイプ、後者は「学習データの多寡にかかわらず一定数の発生が不可避」であるタイプとなる。以下、主な誤訳カテゴリについて、誤訳が発生した事例を示しつつ検討する。

＜①技術用語の誤訳＞

技術用語の誤訳は、無学習の時点から発生数は 100 文中 10 文（10%）と比較的少なく、「みんなの自動翻訳」がデフォルト状態からある程度のインドネシア技術用語をカバーしていることがうかがえた。その後、本調査で作成した特許文献由来の対訳コーパスを学習させていくにつれて誤訳の発生数は徐々に減少し、約 500 万文対の『対訳コーパス A-B＋』を学習させた段階⑦での発生数は 2 件のみとなった。下例はそのうちの 1 件である。

［要 07：P00201506694_JP2012500201A　段階⑦］

イン原文	Yang diungkap disini adalah metode pengobatan hiperglisemia yang lebih baik dengan suatu gabungan dari suatu insulin yang bekerja sangat cepat dan glargin insulin yang terdiri atas pemberian insulin ultra cepat prandial, dan pemberian suatu dosis pertama glargin insulin di dalam 6 jam bangun selama sehari.

参考訳文	超速効型インスリンの食事時投与、及び 1 日の朝目覚め後 6 時間以内の 1 回目のインスリングラルギンの投与を含む、超速効型インスリンとインスリングラルギンの組み合わせによる高血糖症の改善された治療方法を開示する。
機械翻訳	超速効型インスリンの食事時投与、及び 1 日の朝目覚め後 6 時間以内の 1 回目のインスリングラルギンの投与を含む、超速効型インスリンとインスリングラルギンの組み合わせによるインスリン投与の改善された治療方法を開示する。

機械翻訳はおおむね参考訳文のとおりであるが、唯一、「高血糖症（hiperglisemia)」と訳すべき箇所が「インスリン投与」となっている。このため技術用語の誤訳と判定された。

「高血糖症」と「インスリン投与」は意味が全く異なり、ナンセンスな誤訳であるが、ニューラル機械翻訳には、原文中の語句が翻訳できなかった場合、同じ文の他の語句を重複出力させて埋め合わせる傾向がある。本例もこれに該当すると見られる。また、今回誤訳された「hiperglisemia」であるが、それ以前の学習段階ではほぼ「高血糖症」と正しく訳されている（段階⑥でのみ「低血糖症」という正反対の意味に誤訳された)。このように、それまで正しく翻訳できていた語句が学習データの増強による全体構成の変化によって突発的に誤訳されるというのも、ニューラル機械翻訳によく見られる傾向である。

このように、(本来正しく訳せるはずの語句の）突発的な誤訳は学習データの多寡にかかわらず不可避的に発生する。ただし、今回の集計結果から、こうした突発的な誤訳であっても、学習データを増強させるにつれて全体としての発生頻度は徐々に減少させられることが示された。

こうした「①技術用語の誤訳」の状況は、同様の発生数の推移を見せた「②一般用語の誤訳」及び「③係り受けの誤り」にも等しく該当する。

＜④訳抜け＞

訳抜けの発生頻度は、段階⓪の 13 文が最多であるが、段階①以降も 0 文～8 文の間で不規則に変動しており、学習データの増強と相関した変動となっていない。以下、段階⑦で発生した一例を示す。

［要 12：P00201810393_JPWO2017217505A1　段階⑦］

イン原文	Suatu struktur akses (100) untuk mengakses suatu struktur laut terbuka mencakup sejumlah platform (21, 22, dan 23) yang dipasang pada suatu keliling luar suatu struktur laut terbuka (A) dan yang disusun berdampingan satu dengan lainnya, dan suatu unit pelat penambat (3) yang dipasang di bawah platform (21, 22, dan 23). Sejumlah platform (21, 22, dan 23) memiliki ketinggian berbeda yang saling menguntungkan, dan platform (21, 22, dan 23) dipasang sepanjang arah keliling struktur laut terbuka (A).
参考訳文	外洋構造物へのアクセス構造物（１００）は、外洋構造物（Ａ）の外周に設けられ、隣り合って配置される複数のプラットフォーム（２１，２２，２３）と、プラットフォーム（２１，２２，２３）の下方に設けられた接舷板部（３）と、を備え、複数のプラットフォーム（２１，２２，２３）の高さは互いに異なり、プラットフォーム（２１，２２，２３）は外洋構造物（Ａ）の周方向にわたって設けられている。
機械翻訳	オープンシー構造体(A)の外周に設置され、互いに隣接して配置された複数のプラットホーム(21、22、23)と、プラットホーム(21、22、23)の下方に設置されたアンカープレートユニット(3)と、を備え、複数のプラットホーム(21、22、23)は、互いに異なる高さを有し、プラットホーム(21、22、23)は、オープンシー構造体(A)の周方向に沿って設置される。

機械翻訳文には、参考訳文冒頭の「外洋構造物へのアクセス構造物（１００）」に相当する語句が存在しない。それ以降は過不足なく翻訳されており、「外洋構造物へのアクセス構造物（１００）」の訳抜けと判定された。

この語句は段階⑥までは毎回それなりに妥当な訳語で訳出されており、＜技術用語の誤訳＞で取り上げた「高血糖症」と同じく、何らかの理由で突発的に翻訳できなくなったものと考えられる。その際、文中の何らかの語句を重複出力して埋め合わせれば「用語の誤訳」と判定され、埋め合わせなければ「訳抜け」となるが、根本的な原因は共通していると推測される。

＜⑤未知語＞

各段階で 0～2 文と少数検出された。その大半は［要 18］において「kehilangan penguapan Noack」という語句が、参考訳文で「ノアック蒸発損失」となっていたのに対し、各段階の機械翻訳文では「Noack 蒸発損失」となっていたため、原語のまま出力されていると判断され、未知語と認定されたものであった。ただし、日本文献でも「Noack」という表記は使用例があり（J-Plat-Pat で 1,194 件ヒット）、「学習データに該当語が存在しないため、やむなく原語のまま出力した」という本来の意味での未知語に相当するかは判断が難しい。

なお、これ以外に未知語と判定された 2 文も、それぞれ、前後の段階では「EU インデックス」と訳されていた語が「EU Index」と訳されたもの（請 10）と、前後の段階では「高 Mn 鋼」と訳されていた語が「MN-High 鋼」と訳されたもの（請 09）であり、やはり判断は難しい。いずれにせよ、発生件数はごく僅かであり、内容理解上の悪影響も軽微であった。

＜⑥湧き出し＞

無学習時は 2 文と少量であったが、『対訳コーパス A＋』学習以降は 5～7 文とやや増加し、約 500 万文対の『対訳コーパス A-B＋』を学習させた段階⑦では 11 件と最多となった。湧き出し事例の多くは、下例のように、原文に無い特許特有の表現が湧き出すものであった。

［要 02：P00201501206_JP2015529717A　段階⑦］

イン原文	Khususnya diuraikan di sini komposisi karet yang mengandung campuran karet alam, nano karbon dan karbon hitam di mana jumlah relatif dalam bagian per seratus karet (pphr) dari nano karbon terhadap karbon hitam berkisar dari sekitar 1:40 hingga sekitar 1:2 dan jumlah relatif dalam bagian per seratus karet (pphr) dari nano karbon terhadap karet alam berkisar dari sekitar 1:100 hingga sekitar 10:100 dan di mana komponen nano karbon didispersi awal di dalam komponen karet alam.
参考訳文	特に、天然ゴム、ナノカーボンおよびカーボンブラックの混合物を含み、ゴム 100 部あたり部(pphr)でナノカーボン対カーボンブラックの相対量が約 1:40 から約 1:2 の範囲にあり、ゴム 100 部あたり部(pphr)でナノカーボン対天然ゴムの相対量が約 1:100 から約 10:100 の範囲にあり、ナノカーボン成分が天然ゴム成分内に予備分散されているゴム組成物が提供される。
機械翻訳	本明細書に特に記載されるのは、天然ゴム、ナノカーボン及びカーボンブラックの混合物を含むゴム組成物であって、カーボンブラックに対するナノカーボンのゴム 100 分率(pphr)での相対量が約 1:40～約 1:2 の範囲であり、天然ゴムに対するナノカーボンのゴム 100 分率(pphr)での相対量が約 1:100～約 10:100 の範囲であり、ナノカーボン成分が天然ゴム成分内に予

	め分散している、ゴム組成物である。

機械翻訳文には、インドネシア原文で明示的に記載されていない「本明細書～」という文言が存在する。インドネシア文もそのような文意であるものの、原文中に存在しない語句が訳出されており、湧き出しの一種といえる。他の事例も多くは原文に存在しない「前記」の存在などを指摘したものであり、内容理解上の悪影響は小さいものが大半と見られる。事実、上記事例に関しても、内容伝達レベルは 5 と判定されている。

特許文を模すことで生じる湧き出しは、特許文献由来のコーパスを学習することでむしろ発生しやすくなる性質とも考えられる。各段階のカウント結果もそれを裏付けている。また、総じて特定の文で継続的に発生する傾向が強く、上例に関しても段階①以降、全ての段階で同様の湧き出しが発生している。

＜⑧訳ゆれ＞

ニューラル機械翻訳は、同じ語句であっても周囲の語の状況などに応じて最適の訳語をその都度選定して出力する方式である。このため複数の訳語をもつ語句でも文脈に沿った訳し分けが可能である。その反面、同じ訳語を使うべき語句に異なる訳語を用い、内容理解を妨げる場合もある。「訳ゆれ」とは、この状況を指す。

本調査で用いた人手評価対象文は全件、文中に同一の技術用語が複数回使用されているものが選ばれている。機械翻訳文で、これらの語が異なる訳語で訳されている場合、「訳ゆれ」と判定した。ただし、全 100 文、全段階で発生した訳ゆれは 1 例のみ（段階⑤の［要 22］）であった。実例を示す。

［要 22：P00201911900_JP2020521064A　段階⑥］

イン原文	Helm yang meliputi lapisan penyerap energi, lapisan relatif keras yang lebih keras daripada lapisan penyerap energi dan dibentuk di luar lapisan penyerap energi dan sejumlah pelat luar yang dipasang di permukaan luar lapisan relatif keras; di mana pelat-pelat luar tersebut dipasang di lapisan relatif keras shingga, jika terjadi tumbukan pada pelat luar, pelat luar dapat bergeser di sepanjang lapisan relatif keras dan bergerak relatif terhadap pelat-pelat luar lain; dan antarmuka dengan gesekan rendah ditempatkan di antara permukaan luar lapisan relatif keras, dan sekurang-kurangnya sebagian permukaan pelat-pelat luar yang bersentuhan dengan permukaan luar lapisan relatif keras turut ditumbuk jika terjadi tumbukan pada pelat luar.
参考訳文	エネルギー吸収層と、エネルギー吸収層の外側に形成されエネルギー吸収

	層よりも硬い比較的硬い層と、比較的硬い層の外側表面に載置された複数の外側プレートとを備え、外側プレートが、ある外側プレートに対する衝撃下でその外側プレートが比較的硬い層に亘って摺動し他の外側プレートに対して相対的に移動できるように、比較的硬い層に載置され、比較的硬い層の外側表面と外側プレートの表面の少なくとも一部との間に、ある外側プレートに対する衝撃下で比較的硬い層の外側表面と接触する低摩擦接触部が設けられた、ヘルメット。
機械翻訳	ヘルメットであって、エネルギー吸収層と、エネルギー吸収層よりも硬く、エネルギー吸収層の外側に形成された比較的硬質の層と、比較的硬質の層の外面に固定された複数の外側プレートとを含み、外側プレートが、外部プレートに衝突した場合に、外部プレートが、比較的硬質の層に沿って移動し、他の外部プレートに対して移動することができるように、外部プレートが、比較的硬質の層に固定され、低摩擦の界面が、比較的硬質の層の外面の間に配置され、比較的硬質の層の外面と接触している外側プレートの表面の少なくとも一部は、外部プレートに衝突した場合に一緒に粉砕される、ヘルメット。

原文では、参考訳文の「外側プレート」に相当する「pelat luar」と「pelat-pelat luar」（複数形）が計 7 回出現する。全て同じ要素を指しており、同じ訳語で訳すべきである。

しかしながら、上例の機械翻訳では「外側プレート」（3 回）と「外部プレート」（5 回）という二種の異なる訳語に「訳ゆれ」している。「外部プレート」も不適訳ではないが、訳語が異なることによって異なる要素と誤認され、内容理解が困難となる。

数量的には単数形「pelat luar」が「外側プレート」、複数形「pelat-pelat luar」が「外部プレート」と訳されている可能性もあるが、後者の出現回数が異なるなど、文全体の翻訳精度も低く、確言は困難である。段階⑥以外の機械翻訳では単数形、複数形とも「外側プレート」で統一されており、段階⑥のみ何らかの理由で文全体の翻訳に失敗し、その影響でこの語句の訳語も不統一となった可能性もある。

このように、人手翻訳対象 100 文の範囲では、突発的に発生した一例を除き、同一文中での訳ゆれは検出されなかった。この結果から見て、同一文中の用語が頻繁に訳ゆれすることは考えにくい。一方、同一文献に属する複数の文間での訳ゆれについては本調査では検証しておらず、発生の多寡は未知数である。

＜⑨その他＞

その他の不備としてカウントされた 2 文（[請 09] の段階③、[明 35] の段階①）は、いずれも、「翻訳文が一切出力されていない」という重大な不備であった。

この不備の発生は、機械翻訳の実施中から認識されていた。その中には、機械翻訳を再実行することで翻訳文が得られるケースもあり、その場合はその機械翻訳文を使用した。一方、処理を繰り返しても翻訳文が一向に得られない場合もあり、本調査では 3 回試行して翻訳文が得られない場合は取得を断念した。機械評価 2,000 文においては、以下の件数がそれに該当する。

表 8-18　自動評価対象 2,000 文中、機械翻訳文が取得できなかった件数（イン⇒日）

学習段階	⓪	①	②	③	④	⑤	⑥	⑦
イン⇒日	0	3	0	1	0	1	3	7
（参考）日⇒イン	0	1	1	1	1	4	5	10

翻訳文の不出力は長文に偏る傾向にあり、このため同一の文で複数回発生するケースも見られる。また、対訳文は一方が長文であれば他方も長文となるため、文対の双方（すなわちインドネシア文→日本文の翻訳と日本文→インドネシア文の翻訳の双方向）で不出力が発生するケースが多い。事実、人手評価対象文でこの不備が検出された［請 09]、[明 35]とも、双方向で翻訳文の不出力が発生している（ただし発生のタイミングは双方向で異なり、前者は段階⑦、後者は段階③で発生）。

不出力が生じるのは、入力文が長大であると「みんなの自動翻訳」の利用状況によってはリソースが不足し、タイムアウトとなることによる。都度のリソース状況に左右されるため、再試行で翻訳文が得られることもあり、また同一の文であっても学習段階によって翻訳文が得られる場合と得られない場合がある。基本的には学習データ自体の問題ではないが、両方向とも学習データ量が多くなるにつれて発生数が増加していることから、膨大な学習データ量が「みんなの自動翻訳」全体のリソースを圧迫することで、長文のタイムアウトが生じやすくなったとも解釈できる。

翻訳文の不出力は機械翻訳として最も重大な不備である。本調査では 3 回の試行により発生数が抑えられているが、1 回のみの実施であれば発生数はさらに多くなる。本調査で作成した対訳コーパスの内容自体の問題ではないものの、学習データ量を増強するにつれて発生数が増えていることは事実であり、対訳コーパスのサイズが発生の一因である可能性がある。

ただし、全体的には学習データ量を増やすほど翻訳品質は向上しており、また学習データが少量の時点から若干数の不出力は発生している。このため不出力への対処策としては、学習データ量を抑制するのではなく、長文をあらかじめ分割したり、不出力時に翻訳を再実行する等の前後処理を導入することが有効と思料する。

8.6.3. 日本→インドネシアの機械翻訳品質の人手評価

続いて、日本→インドネシア方向の機械翻訳文の人手評価結果を本項にまとめる。

8.6.3.1. 内容伝達レベルの評価結果

人手評価文 100 文の各段階における内容伝達レベル（5～1）の集計結果は下表のとおりであった。

表 8-19　内容伝達レベルの評価結果（日⇒イン）

	レベル 5	レベル 4	レベル 3	レベル 2	レベル 1	平均と増減
段階⓪	46	38	14	1	1	4.28
段階①	70	23	6	0	1	4.54(+0.26)
段階②	70	24	4	1	1	4.55(+0.01)
段階③	74	23	3	0	0	4.70(+0.15)
段階④	72	21	6	0	1	4.54(-0.16)
段階⑤	79	16	3	0	2	4.68(+0.14)
段階⑥	76	19	3	0	2	4.62(-0.06)
段階⑦	78	15	3	1	3	4.60(-0.02)

日本語→インドネシア語方向の内容伝達レベルの評価結果は、段階③で平均レベル 4.70 とピーク値をマークしたが、その後は増減を繰り返し、最終段階⑦ではピーク値から 0.1 ポイント低い 4.60 となった。ただし、この数値自体はインドネシア語→日本語の最終段階での平均レベル 4.65 と同水準であり、どちらも機械翻訳として最上級の品質に達しているといえる。この結果から、本調査で作成した対訳コーパスが、日本語→インドネシア語方向においても学習データとして十分な翻訳品質改善効果を有することが確認された。

日本語→インドネシア語方向の機械翻訳において後期段階で平均レベルが低下傾向を示した理由は定かでないが、もともとサンプルデータが少量であり、ランダム的に発生する誤訳の数に段階ごとの偏りが生じることは十分に想定される。最終段階⑦で到達した品質水準が双方向ほぼ同等であったことに鑑みれば、むしろ日本語→インドネシア語方向では段階③、段階⑤において誤訳の発生が一時的に少量に偏ったとも解釈できる。

8.6.3.2. 重要技術用語の翻訳品質の評価結果

次に、日本語→インドネシア語方向の重要技術用語の機械翻訳品質（A～D）の集計結果について示す。

表 8-20　重要技術用語の内容伝達レベルの評価結果（日⇒イン）

	A	B	C	D
段階⓪	78	18	3	1
段階①	81	15	2	2
段階②	80	16	3	1
段階③	86	12	2	0
段階④	86	11	1	2
段階⑤	88	11	1	0
段階⑥	85	13	2	0
段階⑦	91	5	2	2

上表のとおり、段階⓪での評価 A（適訳語）は 78%であり、これが最終段階⑦では 91%に向上している。インドネシア語→日本語方向と比べると、段階⓪では 5 ポイント、段階⑦では 1 ポイント低いが、全体的にはおおむね同等の翻訳水準であり、コーパスの学習による改善度合いもほぼ同等といえる。段階⑦で C～D と低評価だった 3 語が、それ以前の段階で A 評価を得ている点も同じであり、全体的にインドネシア語→日本語方向における考察内容がそのまま当てはまる。

8.6.3.3. 流暢さの評価結果

日本語→インドネシア語方向の流暢さの評価結果は下表のとおりであった。

表 8-21　流暢さの評価結果（日⇒イン）

	スコア 5	スコア 4	スコア 3	スコア 2	スコア 1	平均と増減
段階⓪	46	38	14	1	1	4.27
段階①	70	23	6	0	1	4.61(+0.34)
段階②	70	24	4	1	1	4.61(±0.00)
段階③	74	23	3	0	0	4.71(+0.10)
段階④	72	21	6	0	1	4.63(-0.08)
段階⑤	79	16	3	0	2	4.70(+0.07)
段階⑥	76	19	3	0	2	4.67(-0.03)
段階⑦	74	22	4	0	0	4.64(-0.03)

内容伝達レベルの集計結果と同様、段階③がピークであり、最終段階ではそれより低下している。ただし段階⓪から段階⑦まで全段階を通じての平均レベルの改善度+0.37 ポイントはインドネシア語→日本語方向の+0.30 ポイントよりも高値であり、最終段階⑦での平均スコア 4.64 も、逆方向の 4.70 に比べて大きな遜色はない。段階⑦における評価 4 以上の割合も 96%ときわめて高く、インドネシア文としての流暢さは充分であると結論される。

8.6.3.4. 誤訳のカテゴリ別チェック結果

日本語→インドネシア語方向における誤訳カテゴリ別の発生頻度のカウント結果を下表にまとめる。す。

表 8-22　誤訳のカテゴリ別チェック結果（日⇒イン）

誤訳カテゴリ／学習段階	⓪	①	②	③	④	⑤	⑥	⑦
① 技術用語の誤訳	2	3	3	4	1	2	3	1
② 一般用語の誤訳	3	3	2	3	2	4	5	4
③ 係り受けの誤り	8	2	3	2	3	1	1	1
④ 訳抜け	12	7	4	3	11	6	6	6
⑤未知語	3	0	1	0	0	0	0	0
⑥ 湧き出し	6	5	2	1	4	3	6	8
⑦ 数値・記号エラー	3	3	5	5	5	5	5	8
⑧ 訳ゆれ	0	0	0	0	0	0	0	0
⑨ その他	0	0	1	0	0	0	0	2

インドネシア語→日本語方向のカウント結果と比べると、「①技術用語の誤訳」及び「②一般用語の誤訳」の発生数が無学習時点からごく少量である点が異なるが、その他の各カテゴリの傾向はおおむね同一であった。

唯一、「⑤未知語」のカウント結果について、段階⓪で 3 文（[明 02]［明 06］[明 43]）、段階②で 1 文（[明 43]）がカウントされたが、これは本来の未知語、すなわちインドネシア語の機械翻訳文中に、原文である日本語がそのまま出力されているというものではなく、全て「インドネシア語の機械翻訳文中に、英語が出力されている」というケースであった。本来は「⑨ その他」とすべき不備である（現状、⑨は翻訳文不出力のみカウントされている）。一例を示す。

［明 06：P00201910038_JP2020520196A　段階⓪］

日本原文	ＥＤＣアルゴリズムが非全０初期可変状態を用いて初期化されるので、誤り検出器３２５は、制御情報ベクトル４０５－ａ、４０５－ｂの異なるビット長により異なるＥＤＣ値を生成し、したがって、誤り検出器３２５は、異なる長さを有する制御情報ベクトル４０５－ａと制御情報ベクトル４０５－ｂとを区別することが可能であり得る。
参考訳文	Karena algoritma EDC diinisialisasi dengan keadaan variabel permulaan non-semua-nol, detektor kesalahan 325 menghasilkan nilai EDC yang berbeda karena panjang bit yang berbeda pada vektor informasi kontrol 405-a, 405-b, dan maka detektor kesalahan 325 bisa mampu untuk membedakan antara vektor informasi kontrol 405-a, 405-b yang memiliki panjang yang berbeda.
機械翻訳	Sebagai algoritma EDC diinisialisasi menggunakan non-semua 0 initial variable states, error detector 325 menghasilkan nilai EDC yang berbeda karena panjang bit yang berbeda dari vektor informasi kontrol 405 a, 405 b, sehingga error detector 325 dapat membedakan antara vektor informasi kontrol 405 a dan 405 b memiliki panjang yang berbeda.

上例では、原文の「誤り検出器３２５」がインドネシア語（detektor kesalahan 325）ではなく英語で「error detector 325」と訳されている。他の２文も、それぞれ「三量体化ドメイン（domain trimerisasi）」が「domain trimerization」（［明 02］）、「ソフトゲル（gel lunak）」が「softgels」（［明 43］）と英語が出力されている。

インドネシア公報にしばしば英文や英単語が混入する問題については、6.4.2.項で詳述した。本調査で作成した『対訳コーパス A＋』及び『対訳コーパス A-B＋』からは英文は極力除去したが、上記各ケースは無学習時点から発生しており、「みんなの自動翻訳」のデフォルトの学習データ自体に英文や英単語が混入している可能性が高い。そもそもインドネシア文であっても意図的に英語が使用される場合もあり、本調査で作成した対訳コーパスもそれらは除去していない。それでも、対訳コーパスの学習後は、段階②での１文を除き英語の出力は発生しておらず、本調査で実施した英文のクリーニングに一定の効果があったことが窺える。

8.6.2.5. 評価結果の総括

本章に示した各種評価結果により、「みんなの自動翻訳」を用いて実施したインドネシア語⇔日本語間の機械翻訳は、いずれの方向においても、本調査の対訳コーパスを学習させることで全体的な翻訳品質を向上させられることが明確に示された。品質改善の度合いもおおむね学習データ量に正比例しており、本調査で作成した大規模な日インドネシア対訳コーパスが機械翻訳の学習データとして有用であることが確認された。

また、各評価結果とも翻訳品質の水準が極めて高いことを示しており、本調査の対訳コーパスを学習させた機械翻訳エンジンが、インドネシア⇔日本の特許文献翻訳ツールとして十分な実用性を有することも検証された。

今回の調査で検出された主な課題としては、ニューラル機械翻訳特有の突発的な誤訳の発生、長文における翻訳文不出力、そしてエンジンのデフォルト学習データにおけるインドネシア文への英語の混入が挙げられる。ただし、これらはいずれも対訳コーパスが直接の発生原因ではなく、本調査で作成した対訳コーパス自体は、質、量のいずれにおいても特段の課題は見られなかった。

９．日インドネシア語対訳コーパスの整備に関する課題の分析と解決策の提言

日インドネシアのファミリー特許由来の対訳コーパスを実際に作成した結果に基づき、日インドネシア語対訳コーパスの整備に関する課題の分析と、解決策がある課題についてはその提言を行う。

9.1. 文分割手法に関する課題

本調査では、インドネシア文の文分割処理には pySBD、日本文には Stanza を使用した。いずれも全体的には良好な処理結果が得られたが、一部、改善すべき課題も判明した。文分割処理における主な課題として、以下が挙げられる。

9.1.1. インドネシア文における過分割

インドネシア文の文末記号であるピリオドは、他の用途で使用されることも多い。このため、句点が文末記号のみに用いられる日本文に比べ、文分割の難易度は高くなる。本調査では、インドネシア文に対応した文分割手法として現状で精度が最も良好と判断した pySBD を使用したが、この pySBD においても、改善すべき課題が見られた。具体的には、文末が数値＋ピリオドの場合、その直前で過分割され、次文冒頭に（あたかも箇条書き番号のように）連結されてしまう事象である。

この不備が発生した場合、後続処理である文アライメント処理でのリカバリ、すなわち次文冒頭の数値＋ピリオドを切り離して前文末尾に再連結することは不可能である。よって本調査で作成した対訳コーパスにも、この不備が発生している文が多数混入した。『対訳コーパス A』100 万文対においては、7,045 文程度発生していると推定された（⇒6.3.1.3.）。加えて、上記「文末の数字＋ピリオド過分割」と同じ原因で発生していると考えられる「文末のアルファベット＋ピリオド過分割」も同コーパスから多数検出された（⇒6.3.1.1.）。

『対訳コーパス A』においては、「インドネシア文末がピリオド以外」である文対全件を人手確認・修正作業の対象としたことで、上記の過分割は全て修正されたが、今後の対訳コーパス作成において同様の対処を行うには相応のコストが必要となる。より高性能なインドネシア文分割手法の登場が待たれる。

なお、人手作業を介さない簡易的な対応策としては、「インドネシア文末がピリオド以外、かつ直後の文の冒頭が数字＋ピリオドである場合、その数字＋ピリオドを当該インドネシア文末に強制的にカット＆ペーストする」という機械処理でもおおむね同等の改善結果が得られると思料する。

9.1.2. 日本文の分割処理

日本文の文分割処理には Stanza を使用した。日本語の文分割は難易度が低いと考えられ、Stanza の文分割処理においても、カッコ補記の中途での過分割などの突発的な発生は見られたものの、インドネシア文の分割のような顕著な課題は見られなかった。

9.2. 文アライメント手法に関する課題

文アライメント処理に関しては、日インドネシアのファミリー文献間の対応度は比較的良好であり、本調査で作成した全文対（7,797,351 件）の約 40.0%（3,118,196 件）が対応度スコア A、約 23.9%（1,861,863 件）がスコア B と判定された。

本調査で使用した文アライメント手法 BGE＋vecalign は N： 1 の対応づけにも無難に対応しており、処理精度面でも特段の課題は見られなかった。

9.3. テキストデータ抽出処理に関する課題

本調査では、対訳コーパス作成のための初手の作業として、ファミリー関係にあるインドネシア公報、日本公報のそれぞれから必要なテキストを抽出した。テキストデータ抽出時に実施した各種処理の課題と有効性は以下のとおりである。

9.3.1. テキスト抽出箇所の特定

公報データ中、テキスト抽出箇所の特定は、インドネシア公報は見出し語、日本公報は XML/SGML タグを手がかりとした (⇒2.2.)。このように特定したテキストデータから最終的には大量の良質な対訳コーパスが得られており、特段の課題は見られなかった。今後の対訳コーパス作成時も同様の手法が有用であると結論される。

9.3.2. テキストデータの全連結

PDF 文書からテキストデータを抽出すると、各行末で改行されているため文が細切れの状態となる。このため、抽出したテキストは全て連結することが必須となる。本調査においても、抽出したテキストデータは要約、明細書などの項目単位で全文を連結した。しかしながら、この必須の処理の結果、「見出し語と本文の連結」という課題が不可避的に生じることとなる。

＜見出し語と本文の連結＞

特許文献では、汎用のものから独自のものまで数多くの見出し語が使用されている。テキストデータの全連結によってこれらの見出し語と直後の本文とが連結されると、後続の文分割処理で分割できなくなる。

全文が連結されたテキストデータは、文分割処理で一文単位に分割される。具体的には、文末記号であるピリオドや句点を手がかりに文の区切りを判断して分割処理が行われる。しかしながら、見出し語は通常、末尾に文末記号を有さない。このため、テキスト抽出時に直後の本文と連結されると、文分割処理での再分割は事実上不可能となる。

見出し語がカッコで括られている場合は本文との境界は明確であるため、連結されたままでも対訳コーパスとして問題はないが、そうでない場合、本文との境界が不明瞭であり、使用性が悪くなる。また、学習データとして用いられる際も、全体が一文として解析されるため、学習精度に悪影響を及ぼす懸念がある。

この課題は、現状、PDF 文書をソースとし、文分割手法を用いて対訳コーパスを作成する方法論においては不可避といえる。

本調査では、この課題への部分的な対処策として、汎用の見出し語を収集し、文分割処理後の文中に該当する文字列が存在した場合、これを再分割する後処理を導入した。見出し語と本文の連結という課題の全てを解決するものではないが、少なくともその一部、すなわち汎用見出し語が文対双方の文頭に連結されているケースへの対処策として、今後の対訳コーパス作成においても有用と考える。

9.3.3. インドネシア文献からのページ番号・行数表示の自動除去

インドネシアの全文明細書は PDF データであり、各ページの末尾にページ番号、そして 5 行おきに行数表示がなされている。PDF データから単純にテキスト抽出を行うと、これらの文字列も本文の途中に混入してしまう（⇒2.2.2.）。このため、本調査ではテキスト抽出時にこれらを自動的に排除するロジックを導入した。

このロジックはきわめて精密に動作し、『対訳コーパス A』の 7 万文対を対象とした人手確認作業では、本文へのページ番号や行数表示の混入はごく僅か（3 文献 5 文対）であり、その原因も当該文献におけるイレギュラーなレイアウトに起因するものであった（⇒6.3.3.4.）。これらごく少数のイレギュラー文献を除けば、全てのインドネシア文献を適切に処理できており、今後も類似の PDF 公報をデータソースとする際には、本調査で用いた各手法を適用できると考える。

9.3.4. 請求項番号の除去

インドネシア PDF 公報の特許請求の範囲は、各請求項の冒頭に「数字＋ピリオド」の形式で請求項番号が付されている。テキストデータ抽出ではこれら請求項番号も抽出対象となるが、後続のインドネシア文分割処理では、この請求項番号と直後の請求項本文とが分割

される場合とされない場合があった。

これに対し、日本 XML 公報の特許請求の範囲では、請求項番号は XML タグとして存在しており、【請求項 XX】という本文と分割された状態で取得される。このため、インドネシア文において請求項番号と本文とが分割されない場合、対応の不一致が発生する。また、インドネシア文の請求項番号が本文と分割された場合も、インドネシア文が「XX.」、日本文が「【請求項 XX】」という、学習データとして無用、もしくは有害な対訳データが大量に生成される結果となる。

本調査では、これを避けるため、インドネシア文分割処理の実施後に、特許請求の範囲から得られた各文の冒頭から請求項番号、すなわち「数字＋ピリオド」を一括除去する処理を行った。あわせて、日本 XML 公報からはタグ化された請求項番号（【請求項 XX】）をテキスト抽出対象外とした。こうすることで、請求項本文のみを対訳コーパス化し、無用な請求項番号由来の対訳データを一掃できた。

今後も特許文献由来の対訳コーパスを作成する際は、対象文献の状況に応じた請求項番号への対処は常時考慮すべき事項となる。本調査で講じた措置は、この課題の解決策の一例として大いに参考となると考える。

9.4. データに関する課題

本調査で実施した対訳コーパス作成手法の問題ではないが、ソースに用いた日本とインドネシアの特許文献データに関しても数点の課題が判明した。以下に列挙する。

9.4.1. インドネシア文献における英語の混入

第 6 章にまとめた人手確認対象 7 万文対での各類型の不備発生頻度の調査分析の結果、『対訳コーパス A』で発生する不備の約 7 割は、対訳コーパス作成過程で生じるのではなく、それ以前のソースデータ自体に起因することがわかった（⇒6.3.5.）。さらに、その過半数を占める「文章や語句の誤り」の約 9 割がインドネシア文で生じていることが判明した。

なかでも、対訳コーパスの整備において大きな課題となるのが、インドネシア文献における英語の混入である。その結果、本調査で作成した対訳コーパスにも英文と日本文の文対が混入したため、作成後に機械的な検出を行う必要が生じた。日インドネシア対訳コーパスに英文が混入すると、これを学習させた日本語→インドネシア語の機械翻訳において、翻訳結果に英語が混じる懸念があるためである。

＜対処策＞

この課題に対し、本調査では以下の２つの手法でインドネシア文に混入する英文の検出・除去を試みた。

手法①：言語自動判定ツールの利用
手法②：英文に頻出する英単語による検索

各手法の詳細及び実施結果は 6.4.2.項に記載したとおりであり、手法①に関しては抽出精度はきわめて高い反面、インドネシア語と英語が一文中に混在する「英語混じり文」を抽出できず、手法②ではこうした英語混じり文を一定量抽出できたものの、ノイズ（除外する必要のない文）の比率が極めて高く（約 92.8%）、かつ全件を目視でスクリーニングする必要があり、作業量やコストの観点で実用性に乏しい。

このため、今後のコーパス作成作業においては、英文の抽出精度が高く、かつノイズが極めて少ない（約 1.9%）手法①のみを実施し、これらに対してのみ目視による取捨選択を行うか、もしくは取捨選択を省略して全件をコーパスから除去するのが現実的である。

9.4.2. 段落番号の有無による不一致

本調査で作成した対訳コーパスでは、正しく対応づけられた文対において、インドネシア文、日本文のいずれか一方のみに段落番号（主に[0001]というフォーマット）が付されており、その点で内容の不一致となるケースが数多く見られた。『対訳コーパス A』の選抜に用いた文長比や文アライメントスコアの限定ではこれらを排除できておらず、このため別途の対処が必要であった。

本調査では、この課題への対処策として、インドネシア文、日本文とも文頭が「[」＋数字＋「]」で始まる場合、この部分を段落番号と見なして除去する機械的処理を実施した。この処理により対訳コーパス中の段落番号の大半を除去することが可能である。効率的で現実的な解決策であり、今後の対訳コーパス作成においても等しく有用であると考える。

9.5 機械翻訳エンジンの学習データ用途に関する課題

ニューラル機械翻訳エンジン「みんなの自動翻訳」を用いた検証の結果、本調査で作成した『対訳コーパス A＋』及び『対訳コーパス A-B＋』は、日インドネシア特許文献の機械翻訳のための学習データとして高い翻訳品質改善効果を有することが確認された。

各種評価の結果から、全コーパスを学習させた状態での翻訳品質は機械翻訳として最上級の水準に達しており、本調査の対訳コーパスを学習させた機械翻訳エンジンが、インドネ

シア⇔日本の特許文献翻訳ツールとして十分な実用性を有することが示された。

検証では、ニューラル機械翻訳特有の突発的な誤訳の発生、長文における翻訳文不出力、そしてエンジンのデフォルト学習データにおけるインドネシア文への英語の混入などの課題も見られた。いずれも日インドネシア特許文献の機械翻訳における要改善事項である。ただし、これらの不備は直接の原因が学習データにあるとは考えにくく、本調査の対訳コーパス自体には、機械翻訳エンジンの学習データ用途における課題は見られなかった。

9.6. 総括

本調査で実施した日インドネシア対訳コーパス作成手法により、対応精度の良好な対訳コーパスを十分な件数、取得することができた。『対訳コーパス A＋』は目標とした 100 万文対、『対訳コーパス A-B＋』は想定を大幅に超える約 500 万文対を取得でき、これらのコーパスが機械翻訳の学習データとして高い品質改善効果を有することも検証された。文分割処理、文アライメント処理を中心とした各工程においても大きな問題は生じず、幾つかの工程で発生した課題に対しても総じて有効な対処策を講じることができた。こうした実施結果により、本調査で用いた手法がファミリー特許文献を用いた今後の日インドネシア語対訳コーパスならびに他国語の対訳コーパスにおいて十分な実施可能性と有効性を有することが確認された。

その一方で、現状では発生が不可避で、かつ十分な対処が困難な課題もいくつか存在した。具体的には、インドネシア文分割手法 pySBD における文末の数字・英文字＋ピリオドの直前での過分割（⇒9.1.1.）、テキストデータ抽出処理における見出し語と本文の連結（⇒9.3.2.）、インドネシア文献における英語混じり文の混入（⇒9.1.2.）の 3 点が挙げられる。ただし、本調査ではこれらの課題に対してもそれぞれ可能な限りの対処策を実施しており、同様の策を講じることで各課題の悪影響を低減させることは可能である。