特許出願書類からの日本語構文平易化コーパスの作成可能性の検討

An Examination of the Feasibility of Creating a Japanese Syntactic Simplification Corpus from Patent Application Documents

株式会社日本取引所グループ　総合企画部　主任研究員
土井　惟成

2015 年株式会社日本取引所グループに入社。東京証券取引所 IT 開発部などを経て、2020 年より現職。2022 年より株式会社 JPX 総研インデックスビジネス部を兼務。

n-doi@jpx.co.jp

1　はじめに

世界的かつ急速な社会情勢の変化に伴い、行政や企業では、正確かつ分かりやすいテキストで構成された情報発信が求められている。このようなテキストの執筆を支援するための取組みとしては、ライティングマニュアルの作成や、自然言語処理技術の利用が挙げられる。具体的には、2022 年 1 月に文化庁が取りまとめた、新しい「公用文作成の考え方」[1] や、2018 年 3 月に一般財団法人日本特許情報機構が発行した「特許ライティングマニュアル」[2] の改訂版がそれぞれ、ライティングマニュアルの例として挙げられる。また、執筆支援のための自然言語処理技術の一つに、テキスト平易化（Text Simplification）がある。

テキスト平易化とは、文の意味を保ちながら、読みやすい文へ変換する自然言語処理技術を指す。テキスト平易化のアプローチには、難解な語彙を平易な語彙に変換する語彙平易化（Lexical Simplification）と、複雑な構造の文を平易な構造の文に変換する構文平易化（Syntactic Simplification）がある[3]。日本語の語彙平易化に関する研究は、コーパスの構築を始めとして、幅広く行われているものの、日本語の構文平易化に関する研究は限られている。この要因として、構文平易化のための言語資源が乏しいことが考えられる。

そこで、本稿では、特許出願書類による日本語の構文平易化コーパスの作成可能性について検討し、これを通じて日本語の構文平易化に関する研究への貢献を目指す。特許出願書類における記載項目の一つである、特許請求の範囲は、特許法 36 条第 6 項第 2 号により、その記載が明確でなければならないこと（明確性要件）が定められている。一般的に、構造が複雑な文はその内容が明確ではないことが多い点を踏まえると、明確性要件を手がかりにすることで、特許の請求範囲をもとに文構造の複雑さを改善した文を収集できることが期待される。そのため、本稿では、明確性要件に反していたが後に修正が施された特許請求の範囲をもとに、構文平易化コーパスの作成可能性を検討する。このような言語資源を確立することは、特許文の明確性の向上を支援するための技術の実現に繋がるものと考えられる。なお、本稿は、明確性要件に関する情報をまとめつつ、限定的な事例をもとに可能性を論じるものであり、詳細な分析は今後の研究に譲る。

本稿の構成は次のとおりである。まず、本稿の関連研究について述べる（2 章）。次に、特許出願書類の審査における明確性要件について概観し（3 章）、明確性要件と構文平易化の関連性について述べる（4 章）。これらを踏まえ、構文平易化コーパスの作成方法を提案し、この可能性について考察する（5 章）。

2　関連研究

本稿に関する関連研究として、日本語のテキスト平易化の言語資源に関する研究、特許文を対象とした言語資源に関する研究、特許文の明確性を対象とした自然言語処理に関する研究について述べる。

2.1　日本語のテキスト平易化の言語資源

日本語のテキスト平易化の言語資源として、田中コーパス[3]を用いて人手によって作成した、「SNOWT15: やさしい日本語コーパス」[4]や、稲岡らによる日本語文法平易化のための対訳コーパス[5]が挙げられる。田中コーパスは、日本の大学生が教科書等の文を翻訳することで作成した日英対訳コーパスであり、その多くは短文で構成されている。また、これらの他に、現代日本語書き言葉均衡コーパス[6]等を用いて作成した、日本語の語彙平易化システムの性能を自動評価のためのデータセット[7]がある。また、長井ら[8]は、毎日新聞と毎日小学生新聞を用いて、文書単位で、語彙や文法を含む網羅的な平易化操作を対象とした平易化コーパスを作成している。これらに対して、本研究で提案する言語資源は、特許文という、複雑な文構造を持つことが多い文を用いて構築することを検討する。

2.2　特許文を含む言語資源

日本語の特許文を含む言語資源として、日英対訳コーパスが挙げられる。Utiyama らは、機械的にアライメントされた約 200 万文対の特許文による日英対訳コーパスを作成している[9]。また、英語の特許文を含むコーパスとして、The Harvard USPTO Patent Dataset（HUPD）が挙げられる[10]。HUPD は、2004 年 1 月から 2014 年 12 月の間に米国特許商標庁に提出された、英語の実用特許の出願書類をもとに作成した、構造化済みの大規模な多目的コーパスである。これらの言語資源は、機械翻訳や特許審査に関する分析のために利用されており、日本語の特許文を対象としたテキスト平易化のための言語資源は、本稿執筆時点では無いものと考えられる。

2.3　特許文の明確性を対象とした自然言語処理

特許文の明確性に関する研究として、明確性のモデリングに係る研究が挙げられる。Ashtor は、機械学習によって特許の請求範囲の明確性をモデル化することで、テキストの冗長性や反復性が、明確性を低下させること等を示している[11]。これに対して、本稿で提案するような、特許文の明確性の向上へそのものを目的とした研究は限定的である。

3　明確性要件

本章では、特許出願書類の審査における明確性要件の理解のため、明確性要件の位置付けを示すとともに、明確性要件違反の類型について述べる。

3.1　明確性要件の位置付け

特許法第 36 条第 6 項では、特許請求の範囲の記載に対する要件を定めている。この内容として、第 1 号は「特許を受けようとする発明の詳細な説明に記載したものであること（サポート要件）」、第 2 号は「特許を受けようとする発明が明確でなければならないこと（明確性要件）」、第 3 号は「請求項ごとの記載が簡潔であること（簡潔性要件）」を定めている。すなわち、特許請求の範囲の記載に対する明確性要件は法規範の一つである。

特許の審査においては、特許請求の範囲に基づいて新規性や進歩性等が判断される。そのため、特許請求の範囲の記載は、特許発明の技術的範囲が決定されるという点で重要な意義を有しており、その請求項から発明の内容が明確に把握できることが求められている。明確性要件は、このような特許請求の範囲の機能を担保する上で、重要な規定と位置付けられている。

特許要件の審査に当たる審査官にとって基本的な考え方を示す「特許・実用新案審査基準」では、明確性要件に関する判断や考え方についても述べられている。なお、「特許・実用新案審査基準」は、あくまで判断基準に過ぎず、法規範ではないことには留意すべきである。

3.2　明確性要件違反の類型

「特許・実用新案審査基準」では、特許請求の範囲の記載が明確性要件を満たさない場合の例として、5 種類の類型を示している。本稿では、「特許・実用新案審査基準」の記載を元に、これらの 5 類型を次のとおり呼称する。

（1）不適切な表現
（2）技術的な不備
（3）不明確なカテゴリー
（4）不明確な選択肢表現
（5）曖昧な範囲表現

以下では、これらの 5 類型の概要について述べる。

（1）不適切な表現

この類型には、請求項に日本語として不適切な表現がある場合や、出願時の技術常識を考慮しても、記載された用語の意味内容を理解できない場合に当てはまる。

（2）技術的な不備

この類型には、発明特定事項の内容に技術的な欠陥がある場合を始めとして、記載事項に技術的な誤りや理解が困難な記載を含んでいる場合に当てはまる。

（3）不明確なカテゴリー

この類型には、特許を受けようとする発明の属するカテゴリーが不明確な場合に当てはまる。ここで言うカテゴリーとは、物の発明なのか、方法の発明なのか、物を生産する方法の発明なのか、といった区分のことを指す。

（4）不明確な選択肢表現

この類型には、発明特定事項が選択肢で表現されているものの、その選択肢同士の関連性が低い場合に当てはまる。このような例として、「特定の電源を有する送信機又は受信機」といった表現が挙げられる。

（5）曖昧な範囲表現

この類型には、範囲を曖昧にし得る表現が含まれている場合に当てはまる。このような表現の例として、否定的表現（「～を除く」、「～でない」等）によって曖昧となる表現、上限又は下限だけを示すような数値範囲限定（「～以上」、「～以下」等）によって曖昧となる表現、定量的ではない表現（「やや」、「高温」等）、範囲を不確定とさせる表現（「約」、「およそ」等）が挙げられる。

4 明確性要件と構文平易化の関連性

明確性要件は、請求項から発明が明確に把握できることを求めており、言い換えると、記載された内容が誤解なく一意に解釈できる表現を求めている。これは、構文平易化の目的にも関連するものである。従って、明確性要件を満たすようなテキストの修正は、構文平易化と共通する部分がある。

一方で、明確性要件の全てが構文平易化と関連するとは言えないと考えられる。そこで、以下では、明確性要件と構文平易化の関連について、明確性要件違反の類型を踏まえて分析する。その後、構文平易化に関連する事例について述べる。

4.1 明確性要件違反の類型における「日本語として不適切な表現」

前述の明確性要件違反の類型には、日本語の表現上の問題だけではなく、特許請求範囲に特有な問題が含まれている。まず、類型（1）不適切な表現には、「日本語として不適切な表現」に加えて、「出願時の技術常識を考慮しても、記載された用語の意味内容を理解できない場合」が含まれている。後者は、構文平易化で解決すべき問題とは言い難く、特許請求範囲に特有な問題である。同様に、類型（2）-（5）も、特許請求範囲に特有な問題であり、これらは構文平易化の範囲を超える。

以上から、明確性要件が構文平易化と関連する点としては、明確性要件違反の類型のうち、類型（1）の中の「日本語として不適切な表現」に限られる。

4.2 構文平易化に関連する「日本語として不適切な表現」の事例

「日本語として不適切な表現」の修正過程を収集する方法として、そのような表現が含まれていることを理由に特許が拒絶され、その後に修正のうえ登録されている特許を収集することが考えられる。拒絶理由通知書には、当該出願がどの条項に違反したかが記載されており、明確性要件に違反した場合は当該表現の箇所とその理由が記載されている。しかしながら、明確性要件のうち、どの類型に該当するかまでは明記されておらず、拒絶理由通知書の記載内容から判断しなければならない。更に、本稿執筆時現在、J-PlatPat には、拒絶理由通知書を対象とした検索機能は備わっていない。

そこで、以下では、J-PlatPat の審決検索を利用して、請求項に「日本語として不適切な表現」を含む特許出願書類を、アドホックに探索する。J-PlatPat の審決検索では、特許審査の拒絶査定に対する不服の申し立てに関する手続きである、拒絶査定不服審判を対象にした、全文検索が可能である。これを踏まえ、次の条件で審決検索を実行した。

- 文献種別：拒絶査定不服審判
- 四法：特許
- 全文：第 36 条第 6 項第 2 号
- 全文：日本語として
- 審決結論：WY：特許（登録）

表１「日本語として」を含む明確性要件違反の類型

大区分	類型名	該当条件	例（太線下線部は該当箇所を示す）	例に関する備考	出典
文章表現に起因する問題点	誤記	誤字、脱字、衍字等が含まれている場合	動作可能に結合され前記液晶マトリクス内の前記行の**全てをの各々が**駆動する行ドライバ	「全てを各々が」の誤記	拒絶 2017-015930
	主語・目的語の不足	主語や目的語が不足している場合	上記核分裂爆燃波原子炉の第１領域から上記核分裂爆燃波原子炉の第２領域まで、**燃料を通って燃焼する工程**	何が「通って」何を「燃焼」するのか日本語として不明確	拒絶 2015-014624
	不明確な係り受け	文節の係り受け先の候補が複数考えられる場合	金属加工物、**特に**管の横欠陥を非破壊検査する	「特に」が「管」にのみ係っているのか、「管の横欠陥」に係っているのかが不明確	拒絶 2015-013519
	受動態の使用	受動態による記載のために文意が曖昧となっている場合	前記プロセッサが、…前記第１時間が閾値以上である場合、…表示する領域が**変更され**	審判中に、不明確である原因が受動態であることが明記されている	拒絶 2020-016176
特許請求範囲に特有な問題点	修飾部の不足	技術内容の明確化のために記載すべき表現が不足している場合	前記圧力面（D）が前記接合面（V）の< **80%である**	面の何が 80%未満であることを指しているかが不明確	拒絶 2015-016721
	不明確な代名詞	「その」や「前記」といった記載が、何を指しているのかが曖昧な場合	前記複数の抵抗発熱体の**前記抵抗**に基づいて前記ヒータの温度を決定し、	本請求項と本請求項が引用している請求項において、抵抗に係る記載が無い	拒絶 2021-011819
	不適切な単語の用法	記載されている用語の用法が不適切であることに起因して、文意が不明確となっている場合	…インストラクションを**備えている**コンピュータ読み取り可能な記録媒体	後に「記憶している」に補正	拒絶 2015-014780
	意味を理解できない用語	明細書及び図面の記載並びに出願時の技術常識を考慮しても、請求項に記載された用語の意味内容を理解できない場合	（審判の指摘箇所を編集） 請求項１の「感心のある入力信号」との記載は、日本語としてその意味するところが明らかでない…	－	拒絶 2016-000445
	意味不明	文意が不明であることが拒絶理由通知書等に記載されている場合	（審判の指摘箇所を編集） （7）請求項６に記載の「…」は、…、日本語としても技術的にも意味不明である。	－	拒絶 2020-011030

表２「日本語として」を含む明確性要件違反の類型分布

大区分	類型名	件数	割合
文章表現に起因する問題点	誤記	11	12.4%
	主語・目的語の不足	2	2.2%
	不明確な係り受け	14	15.7%
	受動態の使用	1	1.1%
特許請求範囲に特有な問題点	修飾部の不足	5	5.6%
	不明確な代名詞	2	2.2%
	不適切な単語の用法	17	19.1%
	意味を理解できない用語	26	29.2%
	意味不明	11	12.4%
合計		89	100%

2023 年 6 月 1 日に前記の検索を実行した結果、72 件の審判が該当した。この 72 件の審判から、日本語として不適切だと判断された表現を含むテキストとして、89 件のテキストを抽出した。そして、これらを注意深く観察することで、表 1 のとおり「日本語として」を含む明確性要件違反の類型を定義した。また、この定義に従い、89 件のテキストを人手で分類した結果を、表 2 に示す。

表 1 のとおり、「日本語として」を含む明確性要件違反の類型は、「文章表現に起因する問題点」と「特許請求範囲に特有な問題点」に大別されると考えられる。「文章表現に起因する問題点」は、テキストの曖昧性に関する一般的な問題点である。一方で、「特許請求範囲に特有な問題点」は、特許請求範囲に特化した問題点である。そのため、構文平易化について考えると、「文章表現に起因する問題点」について検討するべきである。

表 2 のとおり、試験的な分類の結果、「文章表現に起因する問題点」の割合は約 31.5% だった。この中には、文法誤り訂正に役立つと考えられる、「誤記」や「主語・目的語の不足」といった事例も含まれている。そして、構文平易化に特に関連すると考えられる類型は「不明確な係り受け」であり、その割合は約 15.7% だった。

以上を踏まえると、「日本語として不適切な表現」において、特許文に限らず、構文上の問題を含む事例として、「不明確な係り受け」が挙げられる。そして、拒絶

表３　不明確な係り受けの前後比較の例

例	補正前	補正後	出典
1	ユーザ装置がセルに位置し、かつ、複数のリモート無線ユニットが前記セルを共有する前記ユーザ装置のためのリソース割り当て方法	ユーザ装置がセルに位置し、かつ、複数のリモート無線ユニットが前記セルを共有する、前記ユーザ装置のためのリソース割り当て方法	拒絶 2015-022121
2	連続スペースとしてマシンのネットワーク化クラスタの少なくとも1つのノードにメモリの指定された部分を確保するよう構成されたキャッシュアダプタ	マシンのネットワーク化クラスタの少なくとも１つのノードにメモリの指定された部分を連続スペースとして確保するよう構成されたキャッシュアダプタ	拒絶 2017-014156
3	前記アンテナの動作周波数及び通信モードは，リソース及び性能が，予め設定された基準又はユーザの好み及び選択に基づいて，装置において最大限必要とされている場合に適応可能である	前記アンテナの動作周波数及び通信モードは，予め設定された基準，又は，ユーザの好み及び選択に基づいて，装置のリソース及び性能が最適化されるように適応可能である	拒絶 2016-002524
4	金属加工物、特に管の横欠陥を非破壊検査する	金属加工物である管の横欠陥を非破壊検査する	拒絶 2015-013519
5	第１の公開鍵証明書の有効期限まで，該第１の公開鍵証明書を用いた通信が可能な状態で	（削除）	拒絶 2018-006820

理由通知書を対象としてこのような事例を収集することが、構文平易化コーパスの構築に繋がると考える。

5　構文平易化コーパスの作成及び課題

本章では、前章までの分析を踏まえ、「不明確な係り受け」の事例を元に、構文平易化コーパスを試験的に人手で作成し、本コーパスの作成に係る課題を整理する。

本コーパスの作成に当たっては、「不明確な係り受け」の事例となる特許出願を対象に、これに関する補正の前後を比較した。この時、請求項全体を抽出すると。文字列長が過大となることから、「不明確な係り受け」に関する箇所のみを抽出した。

表３に、「不明確な係り受け」の前後比較の例を示す。例１と例２はそれぞれ、読点の追加と文節の並べ替えにより文構造を明確化したものであり、構文平易化の事例として利用可能である。例３と例４は、補正の過程で用語も明確化されており、これらを構文平易化の事例として利用するには判断を要する。また、請求項の補正においては、例５に示すように、曖昧な記載が削除されることや、あるいは請求項そのものを削除することも一般的に行われる。そのため、補正の前後比較には一定の作業負荷が生じる。更に、拒絶理由として「不明確な係り受け」が指摘されたとしても、その補正前後の比較により得られる構文平易化コーパスの規模は限定的になり得ると言える。そのため、より効率的な手法として、「不明確な係り受け」の事例を収集し、その修正例は人手による書き換えによって作成するという手法が考えられる。

6　おわりに

本稿では、特許出願書類を用いた日本語の構文平易化コーパスの作成可能性を検討した。明確性要件に基づく特許請求範囲の補正事例を利用し、構文平易化の事例を収集する手法は、一定の可能性があることを示した。しかしながら、具体的な作業手順や期待されるコーパスの規模を始めとする課題が残ることを確認した。この課題を解決する一つの方法として、構文上の問題を含む文を抽出するモデルを開発し、その抽出結果を人手で明確な文へと修正するといった手法が考えられる。

今後の研究では、「不明確な係り受け」の事例を収集し、それに基づく構文平易化コーパスの作成を行うことが考えられる。また、こうしたコーパスをもとに、構文平易化の自動化に関する技術の開発や評価も行われることが期待される。普遍的な構文平易化を狙ったコーパスの構築のみならず、特許出願書類に係る効率化を目指した自然言語処理技術の応用についても検討を進めたい。

参考文献

[1] 文化審議会、公用文作成の考え方（建議）（付）「公用文作成の考え方（文化審議会建議）」解説 , 2022. https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/pdf/93651301_01.pdf

[2] 一般財団法人日本特許情報機構特許情報研究所 . 特許ライティングマニュアル「産業日本語」第２版 . 2019.

[3] Shardlow, M. A Survey of Automated Text Simplification. International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing, pp.58–70, 2014.

[3] Tanaka Y. Compilation of a multilingual parallel corpus. In Proceedings of PACLING 2001, pp.265-268, 2001.

[4] Katsuta A., Yamamoto K. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. The 11th International Conference on Language Resources and Evaluation（LREC 2018）, pp. 461-466, 2018.

[5] 稲岡夢人、山本和英 . 日本語文法平易化コーパスの構築 . 言語処理学会第 25 回年次大会、pp. 375-378, 2019.

[6] Kikuo Maekawa. Balanced Corpus of Contemporary Written Japanese. In Proceedings of the 6th Workshop on Asian Language Resources. 2008.

[7] Kodaira T., Kajiwara T., Komachi M. Controlled and balanced dataset for Japanese lexical simplification, in Proceedings of the ACL 2016 Student Research Workshop（Berlin: Association for Computational Linguistics;）, pp.1–7, 2016.

[8] 長井慶成、岡照晃、小町守 . 文書単位の日本語テキスト平易化コーパスの構築に向けて . 言語処理学会第 29 回年次大会、pp. 1112-1117, 2023.

[9] Masao Utiyama and Hitoshi Isahara. A Japanese-English Patent Parallel Corpus. In MT Summit XI, pp. 474–482, 2007.

[10] Suzgun, Mirac and Melas-Kyriazi, Luke and Sarkar, Suproteem K. and Kominers, Scott Duke and Shieber, Stuart M. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications. arXiv preprint arXiv:2207.04043, 2022.

[11] Ashtor, Jonathan H. Modeling patent clarity, Research Policy, Vol. 51, pp. 104415, 2022.