特許文の複雑さに関する考察

A Consideration for the Complexity of Texts in Patent 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

一般に、特許文書におけるテキスト（以下、特許文）は複雑であり、非専門家にとっては内容の把握が困難であると言われている。新森ら [1] は、請求項の文章の特徴として、「文長がきわめて長い」、「記述スタイルが独特である」、「構文が複雑である」の 3 点を挙げている。また、一般財団法人日本特許情報機構では、実際の特許文を元に「文章の理解容易性・明晰性」と「機械翻訳の容易性」の観点から言い換え作業を行うことで、「短文化」をはじめとする文法的な言い換えルールを中心とした特許ライティングマニュアル（第 2 版）[2] を発行している。これらを踏まえると、特許文の複雑さの主な要因の一つとして、文法的な複雑さの存在は大きいと考えられる。

一方、現状では、特許文の複雑さにどのような傾向が見られているか、定量的に評価する研究は限られている。そこで本調査では、特許文の傾向分析の足がかりとして、2022 年 6 月に登録公報（B2）が発行された特許を対象に、文字数や係り受け距離を始めとする特許文の複雑さの傾向を調査した。

以下では、第 2 章で関連研究について述べた後、第 3 章で本調査の調査対象とした特許文の概要を述べる。その後、第 4 章にて特許文の複雑さについて多面的な調査の結果について述べ、第 5 章では特許と法令の傾向の差異に関する調査結果について述べる。

2 関連研究

特許文の文字数に関する報告として、新森ら [1] は、公開日が 1998 年 1 月から 3 月までの特許における請求項 1 の平均文字数を調査したところ、241.97 文字であったと述べている。また、通時的な調査として、長部 [3] は、特許明細書における平均文字数が、2001 年には約 7,600 字だったところ、2019 年には約 16,000 字と、2.1 倍に増加したと述べている。

特許文の文構造に関する研究として、横山 [4-5] は、第 1 請求項における文構造について、典型的なパターンを例示しつつその傾向について述べている。潮田 [6] は、特許文書における問題点の一つとして、長い名詞句表現を摘示してその特徴を述べている。橋本 [7] は、特許出願書類における、明細書と請求の範囲を対象として、特許文が複雑化する要因について述べている。鈴木 [8] は、特許文を対象とした自然言語処理の課題として、低頻度語、同一文脈の出現頻度の少なさ、表現の多様性による lexical gap を挙げている。

特許文の文字数の応用に関する研究として、岡田ら [9] は、請求項の文字数の逆数が、当該特許の価値と有意に関連することを示している。

3 調査対象の概要

本章では、本調査における調査対象とする特許文書と、その特許文書から抽出する特許文について述べる。併せて、外国から出願された特許について、制度等を補足的

に説明する。

3.1　調査対象

本調査の調査対象とする特許の母集団は、2022 年 6 月に登録公報（B2）が発行された特許とした。近年における特許出願等統計速報を見ると、特許の出願件数は例年 3 月に多い傾向が見られるものの、設定登録件数は月次の偏りは少ない傾向がある [10]。そこで本調査では、試験的に 2022 年 6 月という特定の年月を対象として抽出を実施した。より広範な通時的な変化を捉えた分析については、今後の研究に譲る。

また、本調査で調査の対象とする特許文は、特許の「要約」と「請求項 1」とした。これらはいずれも、当該特許の発明内容が端的に記述されており、特許文書の中でも重要な位置付けを占めていることを踏まえ、今回の調査の対象とした。

以上を踏まえ、本調査では、2022 年 6 月に登録公報（B2）が発行された特許を対象に、公開特許公報（A）または再公表特許（A1）が存在し、かつ、特許明細書の要約と、登録公報（B2）における請求項 1 が日本語のテキストとして抽出可能な特許を対象とした。そして、抽出対象の特許から、要約と請求項 1 を抽出することで、本データセットを構築した。抽出対象となった特許の条件等を【表 1】に示す。

なお、本データセットの作成においては、各テキストに対して、Unicode 正規化をはじめとする前処理を実施した。

表 1　本データセットの作成条件

項目	内容
特許の母集団	2022 年 6 月に登録公報（B2）が発行された特許
特許番号の範囲	7078812 ～ 7093906
抽出条件	以下を全て満たすこと ・公開特許公報（A）または再公表特許（A1）が存在 ・特許明細書の要約と登録公報（B2）における請求項 1 が、日本語のテキストとして抽出可能
抽出後の特許件数	14,262 件
抽出箇所	・特許明細書の要約 ・登録公報（B2）における請求項 1

3.2　「要約」

特許を出願する際には、要約書を願書に添付して特許庁に提出することが義務付けられており、要約書には要約と選択図を記載することとなっている。特許庁の Web サイト [11] によると、要約とは、「発明または考案の概要を平易な文章で簡潔に記載したものであり、一般の技術者が特許文献の調査の際に、その発明や考案の要点を速やかにかつ的確に判断できるように記載したもの」とされている。

また、特許法施行規則第二十五条の三及び様式第三十一において、要約は口語体で書くことが示されているほか、次のとおり様式が定められている。

- 原則として発明が解決しようとする課題、その解決手段等を平易かつ明りょうに記載する。この場合において、各記載事項の前には、「【課題】」、「【解決手段】」等の見出しを付す。
- 文字数は 400 字以内とし、簡潔に記載する。
- 要約の記載の内容を理解するため必要があるときは、選択図において使用した符号を使用する。

3.3　「請求項 1」

請求項とは、特許権が及ぶべき範囲を箇条書きで記したものであり、最初の請求項を「請求項 1」と呼ぶ。請求項には文字数の制限は無いものの、特許法施行規則第二十四条の四及び様式第 29 の 2 にて、要約と同様に口語体で記述することが示されている。

なお、「特許行政年次報告書 2022 年版」[12] によると、特許出願時における平均請求項の数は、2012 年以降逓増している傾向が見受けられる。このことから、時系列の推移と共に、請求項全体における請求項 1 の位置付け等が変化している可能性は否定できない。この点に関する調査は、今後の研究に譲る。

4　調査結果

本章では、前章で作成した本データセットを用いて、特許文書における要約と請求項 1 の複雑さに関する調査の結果について述べる。複雑さの調査に当たっては、文字列長、係り受け木の深さを採択した。文字列長は、テキストの複雑さを簡易的に測る指標として広く使われている [13]。係り受け木の深さは、文構造の複雑さを表す指標であり、テキストの読みやすさとの関連が報告されている [13-14]。

係り受け木の深さの算出には、GiNZA [1] による係り受け解析を行った。複数の文で構成されている特許文に対しては、文境界で分割し、各短文に対して係り受け解析を行い、係り受け木の深さの最大値を指標として用いた。

また、本調査に当たっては、主に次の観点から調査を実施した。

・要約と請求項 1 の間に何らかの相関は認められるか。
・見出しの構成にどのような傾向が見られるか。
・特許の分野ごとに異なる傾向が認められるか。

4.1　要約と請求項 1 の比較

【図 1】及び【図 2】にて、本データセット全体を対象とした、要約と請求項 1 の文字数の分布をそれぞれ示す。なお、要約の文字数の平均値は 283.5 字、中央値は 288 字であり、請求項 1 の文字数の平均値は 514.0 字、中央値は 491 字であった。

【図 1】より、要約の文字数の分布は、400 字を境に出現頻度が大きく下がっている。これは、3.2 節に述べたとおり、要約の文字数には 400 字の上限が定められていることが要因として考えられる。また、【図 2】より、請求項 1 の文字数は、300 字前後を中心として分布していると言える。また、請求項 1 の文字数の外れ値を見ると、請求項 1 が 4,000 字以上の特許文書が 5 件存在した。これらはいずれも、優先権主張国・地域が他国（中国または韓国）であるか、あるいは、特許協力条約（PCT）に基づいて出願された特許であった。

【図 3】に、二次元ヒートマップによる、要約と請求項 1 の文字数の分布を示す。なお、外れ値として、要約が 400 字超または請求項 1 が 1,000 字超のデータは削除した。また、要約と請求項 1 を対象に、文字数と係り受け木の深さのそれぞれについて相関分析を行ったところ、文字数の相関係数は 0.374（$p<0.001$）、係り受け木の深さの相関係数は 0.216（$p<0.001$）となった。このことから、要約と請求項 1 の複雑さには相関があると言える。

1　https://megagonlabs.github.io/ginza/

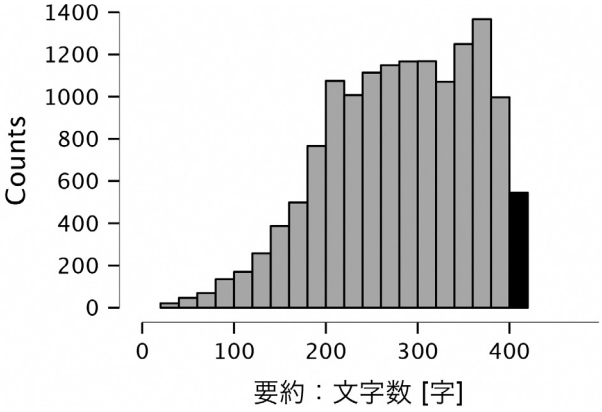

図 1　要約の文字数の分布（黒棒は 400 字超の外れ値）

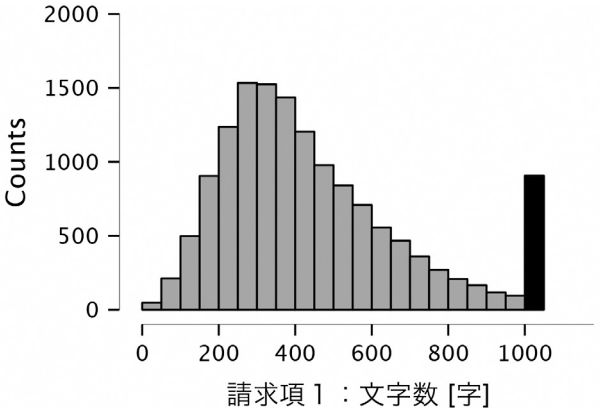

図 2　請求項 1 の文字数の分布（黒棒は 1,000 字超の外れ値）

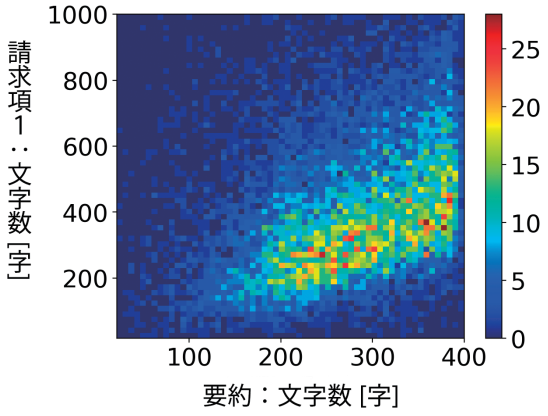

図 3　要約と請求項 1 の文字数の分布

4.2　要約の見出しの傾向

3.2 節に示したとおり、要約は、【課題】や【解決手段】等といった見出しに応じてテキストが分かれている。そこで、まずは【課題】と【解決手段】における各指標の傾向について分析した。なお、全ての特許にこれらの見出しが付与されているわけではなく、特に国外の特許においては見出しが付与されていないことが多い。また、これらの見出しには表記の揺れが散見されており、以下の分析は表記の揺れを包含して実施した。

【表 2】に見出し別の各指標を示す。まず、【課題】を

表２　見出し別の指標（いずれも平均値）

	【課題】	【解決手段】	（要約全文）
標本数	10,567	10,603	14,262
平均文数	1.02	2.02	2.94
文字数（全体）	52.76	234.18	283.54
文字数（１文当たり）	51.84	115.82	96.30
係り受け木の深さ	18.06	66.20	65.92

表３　クラス別出現頻度（上位３位）

クラス		出現頻度	割合
H01	基本的電気素子	1,892	13.3%
A61	医学または獣医学；衛生学	1,596	11.2%
G06	計算；計数	1,562	11.0%

表４　分野別の指標（標本数以外は平均値）

	要約				請求項１			
	H01	A61	G06	全分野	H01	A61	G06	全分野
標本数	1,892	1,596	1,562	14,262	1,892	1,596	1,562	14,262
文数	3.02	2.87	3.07	2.94	1.19	1.11	1.01	1.84
文字数（全体）	281.25	256.01	302.08	283.54	428.84	462.35	468.59	514.00
文字数（１文当たり）	93.18	89.13	98.49	96.30	359.65	418.08	462.66	279.38
係り受け木の深さ	67.51	56.12	63.28	65.92	156.22	137.55	156.34	160.78

見ると、１文当たり文字数が 51.84 文字と、他の書き言葉の文書でよく見られる程度の数値に収まっており、他分野と同程度の複雑さであると言える。一方で、【解決手段】は、１文当たりの文字数と係り受け木の深さが大きくなっており、【課題】よりも複雑な文となっていることが推察される。

4.3　分野別の比較

本データセットを対象に、国際特許分類のクラス別に出現頻度を調査した。出現頻度が上位３位のクラスの分布等を【表 3】に示す。次に、この上位３位のクラスを対象に、要約と請求項１の各指標の傾向を調査した。

本調査の結果を【表 4】に示す。H01 クラス（基本的電気素子）は、概ね全体の平均に似通った傾向を示している。一方で、A61（医学または獣医学;衛生学）は、全体と比較すると文字数が少なく、係り受け木の深さは浅くなっている。この理由として、A61 クラスに属する特許の要約及び請求項１では、化学式や特定の成分名の摘示に留まっているものが散見されており、結果として文字数等が少なくなりやすい傾向があるものと考えられる。また、G06（計算；計数）は、３分野の中で最も文字数が多いが、適切に文を区切っているためか、１文当たり文字数と係り受け木の深さは全体平均と同程度に収まっている。これらの結果から、特許の分野に応じて特許文の傾向が異なることが示唆される。

5　特許と法令との比較

前章までの調査を通じて、特許の分野による特許文の指標の傾向の差異が認められた。一方で、特許文そのものが、他の分野の文とどの程度異なるかは明らかにはなっていない。そこで本章では、試験的な調査として、特定の分野の特許における要約と法令のテキストを対象に、各指標にどのような特徴が見られるか、調査を実施した。

4.3 節の調査結果を踏まえると、特許の分野に応じて特許文の傾向が異なることが分かっている。そのため、本章の調査では、特許の分野を E04（建築物）に、法令を建築基準法及び建築基準法施行規則を対象とした。建築分野を採択した根拠としては、特許と関連する法令が明確であったことが挙げられる。法令文のデータセットは、建築基準法及び建築基準法施行規則を対象に、本則中の項文を対象に、末尾が句点（。）で終わっている文を抽出することで作成した。

本調査の結果を【表 5】に示す。今回、E06 クラスの特許が 89 件と限られているため、統計的な検定による有意差の確認は難しい。しかしながら、【表 5】を見ると、特許は平均文字数が少ないにもかかわらず、平均係り受け木の深さが深くなっていることから、特許文の

方が法令文よりも複雑であることが推察される。

表 5　建設分野を対象とした特許と法令の比較

	特許の要約	法令
標本数	89	1,228
総文数	239	2,220
平均文字数（1 文当たり）	104.2	124.4
平均係り受け木の深さ	73.0	41.6

6 おわりに

本調査では、2022 年 6 月に登録公報（B2）が発行された特許を対象に、特許文の複雑さの傾向について多面的に調査した。本調査の結果として、特許の分野によって、特許文の複雑さの指標に傾向の差異が認められた。また、建築関係の分野においては、特許文の方が法令文よりも複雑であることが推察される。

一方で、本調査で対象とした指標は限定的であり、また、調査対象の特許や比較対象の産業文書も限られている。そのため、本調査の結果を足がかりとして、引き続き特許文の複雑さについて考察を深めて参りたい。

参考文献

[1] 新森昭宏，奥村学，丸川雄三，岩山真．手がかり句を用いた特許請求項の構造解析．情報処理学会論文誌，Vol. 45, No. 3, pp. 891–905, 3 2004.

[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（第 2 版），第 2 版，3 2019.

[3] 長部喜幸．SDGs に貢献する特許戦略，6 2021. https://www.nikkei.com/article/DGKKZO72649530X00C21A6KE8000/（2022-08 閲覧）．

[4] 横山晶一．特許文請求項の構造に関する調査．Japio YEAR BOOK, pp. 242–245, 2016.

[5] 横山晶一．パターンによる特許文請求項の構造解析．Japio YEAR BOOK, pp. 298–301, 2017.

[6] 潮田明．特許文における長い名詞句表現の解析の問題について．Japio YEAR BOOK, pp. 284–285, 2011.

[7] 橋本康重．特許の日本語．専門日本語教育研究，Vol. 12, pp. 9–14, 2010.

[8] 鈴木祥子．機械による特許分析の課題とアプローチ．情報の科学と技術，Vol. 67, No. 7, pp. 355–359, 2017.

[9] Yoshimi OKADA, Yusuke NAITO, and Sadao NAGAOKA. Claim Length as a Value Predictor of a Patent. IIR Working Paper 16-04, Institute of Innovation Research, Hitotsubashi University, 5 2016.

[10] 特許庁．特許出願等統計速報．https://www.jpo.go.jp/resources/statistics/syutugan_toukei_sokuho/index.html（2022-08 閲覧）．

[11] 特許庁．要約書の概要．https://www.jpo.go.jp/system/patent/shutugan/sakusei/ygaiyo.html（2022-08 閲覧）．

[12] 特許庁．特許行政年次報告書 2022 年版，7 2022. https://www.jpo.go.jp/resources/report/nenji/2022/index.html（2022-08 閲覧）．

[13] 渡邉亮彦，村上聡一朗，宮澤彬，五島圭一，柳瀬利彦，高村大也，宮尾祐介．TRF: テキストの読みやすさ解析ツール．言語処理学会第 23 回年次大会発表論文集，pp. 477–480, 2017.

[14] Sarah Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics（ACL’05），pp. 523–530, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.