

日本の上場会社の知的財産関連文書におけるテキスト指標の分析

—知的財産報告書・有価証券報告書・特許出願書類の比較—

Analysis of Textual Metrics in Intellectual Property Related Documents of Japanese Listed Companies



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。株式会社東京証券取引所 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

企業価値の源泉は有形資産から知的財産や無形資産へ急速に移行している。この流れを受けて、2021 年 6 月のコーポレートガバナンス・コードの改訂により、上場会社は知的財産投資を経営戦略と整合的とし、かつ「分かりやすく具体的に」開示すること、実施しない場合にはその理由を説明すること（Comply or Explain）が求められている^[1]。さらに、内閣府は「知財投資・無形資産の投資・活用戦略の開示及びガバナンスに関するガイドライン Ver.2.0」^[2]を公表し、企業に対して、知的財産・無形資産に関する戦略が、競争優位やキャッシュフロー創出につながる因果関係を、ロジック／ストーリーとして定性的・定量的に開示し、企業と投資家との対話につなげることを推奨している。

こうした流れの中で、上場会社が知的財産に関する動態を表すための言語として、「産業日本語」の活用が期待される可能性がある。産業日本語研究会は、産業日本語を「産業・技術情報を人に理解しやすく、かつ、コンピュータ（機械）にも処理しやすく表現するための日本語」と定義する。特許出願書類等の専門性の高い技術文書は、長大な複合名詞列や曖昧な構文により読解が困難になりがちであり、これに対して産業日本語は「短文化」「省略しない」「係り受けを明確にする」「誰にでも伝わる表現への言い換え」等のルールを定めている^[3]。この考え方は特許出願書類だけでなく、知的財産に関連する他の開示文書にも適用可能であり、読者である投資家との対話を支援する可能性がある。

しかしながら、知的財産に関する文書について、産業日本語の考え方に即した記述がどの程度なされているかについては、定量的及び体系的な検証がほとんど行われていない。さらに、従来研究の多くは、分析対象は特許出願書類や有価証券報告書に限られ、知的財産報告書を同一のフレームワークで分析した研究は見当たらない。また、これらの文書は、読者層も目的も異なることから、言語表現の特徴は異なると考えられる。

そこで、本研究では、日本の上場会社が開示する、知的財産報告書、有価証券報告書における研究開発活動の項目（以下、有報 R&D）、特許出願の要約書における概要（以下、特許要約）及び特許請求の範囲における最初の請求項（以下、特許第一請求項）を対象に、複数のテキスト指標を用いて文書種別間の差異を明らかにすると共に、可読性の決定要因について考察する。これにより、知的財産に関する文書に適したテキストのあり方について示唆を得ることを目指す。

なお、本研究は、機械可読性の高い知的財産報告書がまだ限られているという現状の制約下で、必ずしも網羅的ではないテキスト指標を利用して実施した、試験的な取り組みである。そのため、結果の一般化には慎重な検討が必要であり、今後データと手法の拡充に伴い改めて検証を行う余地が残されている。

本稿の構成は以下のとおりである。第 2 章で関連研究を概観する。第 3 章ではリサーチデザインとして、分析対象の上場会社、データセット、テキスト指標の構築方法、推計モデルを説明する。第 4 章で分析結果を報告し、第 5 章で結論と今後の課題を示す。

② 関連研究

上場会社の知的財産に関する情報について、投資判断上の重要性を示唆する報告として、杉光ら^[4]が挙げられる。当該研究では、中型株（TOPIX Mid400の構成銘柄）について、登録特許件数が多い企業ほど7-10年後に成長株としての特徴が強くなることを報告した。ただし、当該研究では、知的財産開示のテキスト特性については言及されていない。

上場会社による日本語の開示資料を対象としたテキスト指標に関する研究としては、可読性とトーンの2つの論点が挙げられる。

第一に、可読性については、廣瀬ら^[5]が、日本の有価証券報告書におけるMD&Aの可読性と企業の将来の業績との関連性を指摘している。

第二に、トーンについては、加藤ら^[6]が、有価証券報告書におけるMD&Aのトーンが、将来の企業業績に対する予測力を有することを明らかにしている。

これらの研究は、有価証券報告書等を対象にテキスト指標と企業パフォーマンス等との関連を示してきた。しかしながら、知的財産報告書、有報R&D、特許出願書類という異なる文書を、同一フレームワークで比較した実証は見当たらない。本研究では、この文書横断的比較に焦点を当て、文書間のテキスト指標の傾向の差異を分析し、可読性の決定要因を分析する。

③ リサーチデザイン

3.1. 分析対象の上場会社

十分なテキスト量を確保するため、本研究では次の4要件を満たす企業を分析の対象とした。

- ・2025年6月末時点で、日本国内の証券取引所に上場していること
- ・2025年6月末までに開示された最新のコーポレート・ガバナンスに関する報告書にて、「知的財産報告書」の文字列が含まれていること
- ・2025年6月末時点で、自社のウェブサイト上で、知的財産報告書またはそれに類する書類を公開し、かつその文書においてテキストが埋め込まれていること（テキストを画像化していないこと）
- ・2023年内に出願し、2025年6月末時点で登録が

表1 データセットの概要

会社	文書種類	テキスト数	平均文字列長
A 社	知的財産報告書	40	67.91
	有報 R&D	31	94.30
	特許要約	53	93.73
	特許第一請求項	53	137.19
B 社	知的財産報告書	29	67.26
	有報 R&D	42	95.39
	特許要約	59	94.14
	特許第一請求項	59	197.01
C 社	知的財産報告書	58	65.26
	有報 R&D	32	93.05
	特許要約	30	111.06
	特許第一請求項	30	230.67

完了した日本国内の特許が、10件以上であること
この結果、3社の上場会社が全ての要件を満たした。本稿では、特定企業へのバイアスを避ける目的で、これらの企業名はA社、B社、C社と匿名化した。

3.2. データセット

本研究では、A社、B社、C社の3社を対象にして、知的財産報告書、有報R&D、特許要約及び特許第一請求項のテキストを基にしたデータセットを構築した。

知的財産報告書としては、2025年6月末時点でインターネットで公開されている、各社の最新版のPDFファイルを使用した。この時、各上場会社の知的財産報告書のPDFファイルから、表紙、前書き、目次、裏表紙のページを削除した。その後、知的財産報告書中の各項目ごとにテキストを抽出し、1つのテキストとして取り扱った。

有報R&Dは、2025年6月末時点でEDINETで公開されている、最新版の有価証券報告書から抽出した。上場会社は自社の研究開発活動について、有価証券報告書で開示することが義務付けられている。本研究では、この研究開発活動の欄のテキストについて、改行で区切ることで、複数のテキストを抽出した。

特許出願書類については、2023年内に出願され、2025年6月末時点で登録済みの日本国内の特許のうち、要約と第一請求項のテキストが含まれているものを対象とした。表1に、このデータセットの概要を示す。

3.3. テキスト指標の構築方法

本研究では、以下の4つのテキスト指標を使用した。

(A) NERatio :

固有表現数 ÷ 総単語数で計算される、テキストに含まれる固有表現の割合であり、具体性の代理指標である。固有表現とは、固有名詞や数値表現によって表される表現を意味する。この固有表現割合の値が大きいほど、固有表現の割合が多く、内容が具体的であると考えられており、会計分野のテキストでの利用が報告されている^[7]。なお、NERatioで表される具体性は、あくまで固有表現が用いられる程度に過ぎず、コーポレートガバナンス・コードで推奨されている具体的な開示とは異なる概念である可能性があることは留意すべきである。

(B) HD-D (Hypergeometric Distribution Diversity)^[8] :

語彙多様性の指標の一つであり、テキストの長さへの依存が少ないとされている。HD-Dは、次の式(1)のとおりに、テキスト内で登場する各単語について、「ある小規模サンプリングサイズ n の中に該当単語が含まれる確率」を、ハイパージオメトリック分布を用いて求め、それらの確率を総和することで算出される。

$$HD-D = \frac{1}{n} \sum_{i=1}^V \left(1 - \frac{\binom{n-f_i}{k}}{\binom{N}{k}} \right) \quad (1)$$

ここで、 N は文書における総単語数、 V は文書中の異なり形態素数、 f_i はそれぞれのユニークな単語 i の出現頻度、 n はサンプリングサイズを表す。また、 $\binom{x}{y}$ は二項係数を表し、「 x 個の中から y 個を取り出す組み合わせの総数」を意味する。なお、一般的には、サンプリングサイズは42が選択されている。HD-Dの値が大きいほど、多様な単語が含まれている、すなわち語彙多様性が高いテキストとみなされる。

(C) Depth :

テキスト中に含まれる文の構文木の深さの平均値である。この値が大きいほど、文が複雑であると考えられる。

(D) Grade :

柴崎・玉岡^[9]の多変量回帰式により日本語教科書の学年相当を推定する値である。具体的には、次の式(2)のとおりに算出される。

表2 テキスト指標の平均値の一覧

会社	文書種類	NE ratio	HD-D	Depth	Grade
A社	知的財産報告書	0.061	0.771	3.378	13.485
	有報R&D	0.106	0.796	3.946	14.994
	特許要約	0.170	0.765	4.384	14.520
	特許第一請求項	0.208	0.732	4.273	19.028
B社	知的財産報告書	0.049	0.779	3.304	12.706
	有報R&D	0.093	0.793	3.702	14.890
	特許要約	0.088	0.681	5.547	14.009
	特許第一請求項	0.074	0.637	7.157	22.571
C社	知的財産報告書	0.070	0.782	3.278	13.123
	有報R&D	0.126	0.802	3.513	14.588
	特許要約	0.162	0.733	5.213	15.172
	特許第一請求項	0.156	0.703	6.178	22.865
全体	知的財産報告書	0.062	0.778	3.316	13.142
	有報R&D	0.107	0.796	3.716	14.829
	特許要約	0.134	0.723	5.042	14.445
	特許第一請求項	0.141	0.686	5.874	21.311

$$Y = -0.145X_1 + 0.587X_2 + 14.016 \quad (2)$$

ここで X_1 は文の総文字数に対する平仮名の割合(%)を示し、 X_2 は平均述語数を示す。 Y が小さいほど、そのテキストの可読性が高いことを示す。

なお、固有表現抽出と構文木の深さの計算には GiNZA¹ を使用し、HD-D と Grade の計算には MeCab² と UniDic³ を使用した。

3.4. 仮説と推計モデル

本研究で対象とする文書は、読者層も目的も異なるため、言語表現の特徴が大きく異なると考えられる。これを検証するため、本研究では、次の仮説を設定した。

H1 (文書種類効果)

文書種類(知的財産報告書、有報 R&D、特許要約、第一請求項)は、固有表現割合 (NE ratio)、語彙多様性 (HD-D)、構文的複雑さ (Depth)、可読性 (Grade) の平均値に有意な差をもたらす。

一般的に、構文的複雑さ (Depth) が大きいほど、文は読みにくいとされている。また、分かりやすさを重視している文書では、専門性の高い用語の、より平易な表現への言い換えが行われていると考えられる。これらを踏まえ、さらに次の仮説を設定した。

H2 (可読性決定要因)

具体性 (NE ratio) と構文的複雑さ (Depth) は読みにくさを高め、語彙多様性 (HD-D) は読みにくさを緩和する。

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

2 <https://taku910.github.io/mecab/>

3 <https://clrd.ninjal.ac.jp/unidic/>

表 3 相関係数の一覧

		NE_ratio	HD-D	Depth	Grade
NE_ratio	相関係数	—			
	p値	—			
HD-D	相関係数	0.202	—		
	p値	< .001	—		
Depth	相関係数	-0.123	-0.559	—	
	p値	0.005	< .001	—	
Grade	相関係数	0.016	-0.438	0.603	—
	p値	0.709	< .001	< .001	—

H3（企業効果の頑健性）

企業固定効果を導入しても、H2 の係数符号は不変である。

これらの仮説の検証のために、式 (3) 及び式 (4) のとおり、推定モデルを設定した。

推定モデル 1

（文書種類効果の検証：一元配置 ANOVA）

$$y_{ij} = \mu + \tau_{d(j)} + \varepsilon_{ij} \quad (3)$$

ここで、 y_{ij} は各テキスト指標、 μ は母平均、 $\tau_{d(j)}$ は文書種類 d の効果、 ε_{ij} は残差項を示す。

推定モデル 2

（可読性の決定要因：重回帰分析）

$$Grade_{uj} = \alpha + \beta_1 NEratio_{ij} + \beta_2 HD-D_{ij} + \beta_3 Depth_{ij} + u_{c(j)} + \varepsilon_{ij} \quad (4)$$

ここで、 $u_{c(j)}$ は企業固定効果、 ε_{ij} は残差項を示す。

この推定には OLS を使用し、多重共線性を $VIF \leq 10$ で確認する。

本研究では、これらのモデルの分析を通じて、文書種類間の差異（H1）と、具体性・多様性・構文深度が可読性に与える影響（H2、H3）を検証する。

4 分析結果

4.1. 文書種類別の記述統計

表 2 に文書種類別のテキスト指標の平均値の一覧を、表 3 に各テキスト指標における相関係数を示す。また、図 1-1 から図 1-4 に、それぞれの分布を示す。

どの会社においても、知的財産報告書において、NE ratio が最小、Depth が最小、Grade が最小（難易度が最も低く、読みやすい）という傾向が得られた。

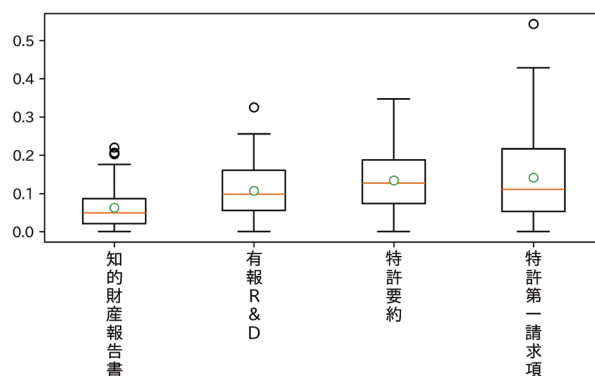


図 1-1 NEratio の分布

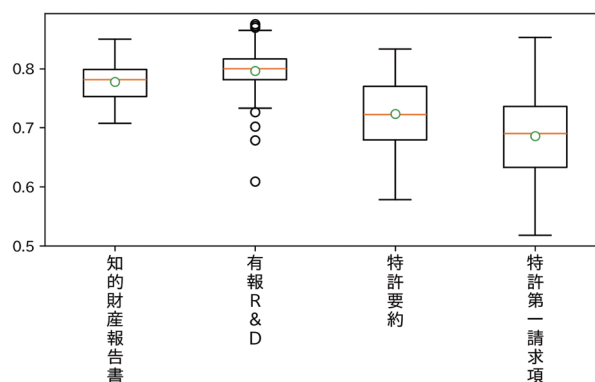


図 1-2 HD-D の分布

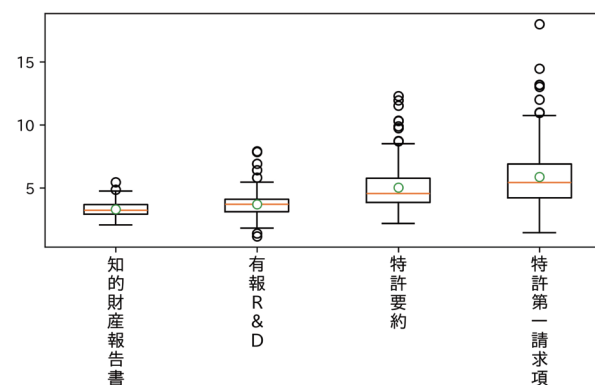


図 1-3 Depth の分布

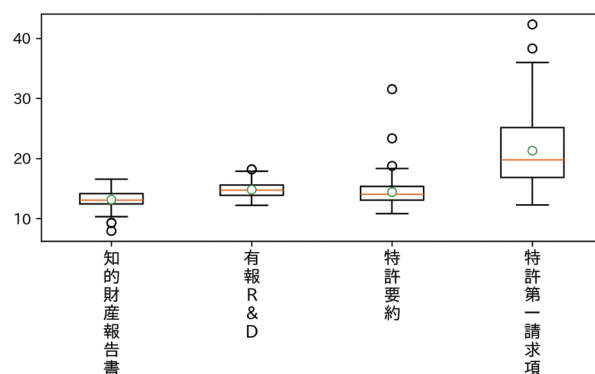


図 1-4 Grade の分布

また、どの会社においても、特許第一請求項において、HD-D が最小、Depth が最大、Grade が最大（難易度が最も高く、読みにくい）という傾向が得られた。

4.2. H1：文書種類効果の検証

一元配置 ANOVA の結果は、4 つのテキスト指標の全てにおいて、文書種類効果が 1% 水準で有意だった。Tukey HSD の結果は、知的財産報告書と有報 R&D の間において HD-D と Depth、特許要約と特許第一請求項の間において NE ratio、特許要約と有報 R&D の間において Grade で、それぞれ有意差は見られなかったものの、それ以外のペア差においては有意差が認められた。特に、知的財産報告書と特許要約、知的財産報告書と特許第一請求項、有報 R&D と特許第一請求項においては、全てのテキスト指標において有意差が認められた。

これらの結果は、H1 を概ね支持した。

4.3. H2：可読性の決定要因

重回帰分析の結果、自由度調整済み決定係数 (Adj. R^2) は 0.380 であり、次のとおり係数が得られた。

- ・ NEratio: $\beta = +6.27$ ($p < 0.01$)
- ・ HD-D: $\beta = -12.26$ ($p < 0.01$)
- ・ Depth: $\beta = +1.20$ ($p < 0.01$)

従って、高い固有表現割合と深い構文が読みにくさを高め、語彙多様性が読みにくさを緩和するという仮説どおりの符号で、H2 が支持された。

4.4. H3：企業固定効果を考慮した頑健性

企業ダミーを加えた固定効果モデルの結果、自由度調整済み決定係数 (Adj. R^2) は 0.385 とわずかに改善した。NEratio、HD-D、Depth の 3 つのテキスト指標の係数は、全て有意かつ符号不変だった。そのため、企業固有要因を制御しても、NEratio、HD-D、Depth の 3 つのテキスト指標が可読性を規定する関係は頑健であり、H3 も支持された。

4.5. 考察

分析の結果、固有表現が多く深い構文を持つテキストほど読みにくい一方、語彙多様性が可読性を緩和することが確認された。しかしながら、文書の種類ごとに、可

読性の改善に向けた方針は異なると考えられる。以下では、文書の特性の違いを踏まえて考察を行う。

まず、特許要約と特許第一請求項は、法的かつ技術的事実を忠実に記述する媒体であり、重要語をあえて言い換えたり固有名詞を削減したりする余地は小さい。従って、可読性向上の現実的手段は、語彙や固有表現の操作よりも、構文を浅く保つことにある。

次に、知的財産報告書では、特許出願書類で使用するような語彙は、想定される読者にとっては分かりにくいと考えることで、技術的な固有名詞を避けつつ、多様かつ一般的な語彙を用いて記述することによって、既に分かりやすく具体的な開示を実現している可能性がある。

最後に、有報 R&D では、特許出願書類ほどではないにせよ、知的財産報告書と比較すると固有表現が多く、すなわち、技術的な固有名詞が多用されている可能性がある。さらに、知的財産報告書と比較すると、語彙多様性が乏しく、構文が深い傾向にある。そのため、可読性の向上のためには、知的財産報告書と同様に、技術的な固有名詞の言い換えを行うと共に、構文の深さを浅く保つことが解決策の一つとして考えられる。ただし、有報 R&D は、制度上の理由から記載内容に固有表現を多用せざるを得ない可能性が否定できないため、より詳細な考察等は今後の研究に譲りたい。

以上を踏まえると、これらの文書の可読性の向上においては、「短文化」「省略しない」「係り受けを明確にする」「誰にでも伝わる表現への言い換え」をはじめとする、産業日本語のルールが準用できる可能性がある。今後、自然言語処理技術の発展により、産業日本語の考え方に即した機械校正ツールが出現すれば、分かりやすく具体的なテキストの作成が容易になると考えられる。

4.6. 本研究の限界

本研究の限界として、3 点挙げられる。

第一に、サンプル企業が 3 社に限定されており、結果の妥当性に限界がある。本研究では機械可読性の高い知的財産報告書を開示している上場会社に限ったことから、今後の研究として、分析対象を有報 R&D と特許出願書類に限ることでサンプル対象の上場会社を増やすことが考えられる。

第二に、本研究で使用した可読性指標はテキストに基

づくため、図表の効果を考慮できていない。今後の研究では、図表を考慮した分かりやすさについても考慮する必要があると考えられる。

第三に、本研究では、可読性と投資家反応を直接結び付けていない。今後は、株価や出来高等との関連を検証することで、知的財産関連文書におけるテキストの可読性と投資家反応の間の関係性を明らかにすることが求められる。

⑤ おわりに

本研究では、上場会社3社の知的財産報告書、有価証券要約、特許要約、特許第一請求項を対象として、複数のテキスト指標を用いて文書種別間の差異を明らかにすると共に、可読性の決定要因について検証した。実験の結果、知的財産報告書が最も可読性が高く、特許第一請求項が最も可読性が低いという傾向が明らかになった。また、固有表現の多さと構文の深さは読みにくさを高め、語彙多様性はそれを緩和することが示された。そのため、分かりやすく具体的な開示の実現に向けては、構文を浅くすることが全体的に有効であると共に、文書の種類に応じて適切に固有表現を使用し、多様かつ一般的な語彙を用いて記述することが有効であると考えられる。

今後の課題としては、サンプルの拡大に加え、知的財産関連文書における図表の効果の検証や、可読性と投資家反応との関連を探ることで、企業と投資家の対話を支援する、分かりやすく具体的な開示の実践策の検討が挙げられる。

参考文献

- [1] 株式会社東京証券取引所．コーポレートガバナンス・コード．2021．<https://www.jpx.co.jp/equities/listing/cg/tvdivq0000008jdy-att/nlsgeu000005lnul.pdf> (2025年6月30日閲覧)
- [2] 内閣府．知財・無形資産の投資・活用戦略の開示及びガバナンスに関するガイドライン Ver.2.0. 2023．https://www.kantei.go.jp/jp/singi/titek2/tyousakai/tousi_kentokai/pdf/v2_shiryo1.pdf (2025年6月30日閲覧)
- [3] 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．

- [4] 杉光一成，立本博文，波多野紅美，山内明，力久未知可：知財情報を活用した投資判断の可能性，証券アナリストジャーナル，Vol. 61, No. 10, pp. 16-26, 2023.
- [5] 廣瀬喜貴，平井裕久，新井康平．MD&A 情報の可読性が将来業績に及ぼす影響：テキストマイニングによる分析，経営分析研究，No. 33, pp. 87-101, 2017.
- [6] 加藤大輔，五島圭一．有価証券報告書のテキスト分析：経営者による将来見通しの開示と将来業績，金融研究，Vol. 40, No. 3, pp. 45-75, 2021.
- [7] Ole-Kristian Hope, Danqi Hu, and Hai Lu. The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, Vol. 21, No. 4, pp. 1005-1045, 2016.
- [8] P. M. McCarthy and S. Jarvis. Vocab: A theoretical and empirical evaluation. *Language Testing*, 24(4):459-488, 2007.
- [9] 柴崎秀子，玉岡賀津雄．国語科教科書を基にした小・中学校の文章難易学年判定式の構築，日本教育工学