青山経営論集
第56巻 第1号
2021年7月

論文

テキストマイニングを用いたMD&A、リスク、ガバナンス情報の分析

矢澤　憲一　　伊藤　健顕　　金　鉉玉

キーワード
テキストマイニング
MD&A
リスク
コーポレート・ガバナンス
有価証券報告書

目次

1．背景と問題意識

本研究の目的は、テキストマイニング技術を用いて、有価証券報告書における記述情報の情報特性とその変化に関する実証的な証拠を提示することである。

企業活動のグローバル化、複雑化が進み、その一方で企業を取り巻く不確実性が高まる中、財務情報を補完する記述情報の重要性が高まっている。有価証券報告書において経営者による財政状態・経営成績およびキャッシュ・フローの分析（MD&A）、リスクおよびコーポレート・ガバナンスに関する記述情報の開示が2004年に要請され、その後、数度にわたり記載内容の拡充が行われてきた。近年では、2019年3月に金融庁から「記述情報の開示に関する原則」が公表され、「経営方針・経営戦略等」「MD&A」「リスク情報」を中心に望ましい開示に向けた取組みについての原則が示された。また、2020年の新型コロナウイルス感染症の感染拡大が企業活動に大きな影響を与えていることを受けて、MD&Aやリスク情報において、より充実した開示を行うことが要請されている。このように記述情報の重要性が高まる一方、わが国では有価証券報告書における記述情報の情報特性や経済効果が十分に検討されてきたとは言い難い。

理由の一つとして、記述情報は、会計・財務情報のような構造化されたデータとは異なる非構造化データであり、こうしたデータを定量化、指標化し、統計的に解析することが難しかったことが挙げられる。一方、2000年以降、テキストマイニングやそれに関連した技術の進展により、そのような性格をもつ記述情報についても情報内容を定量的に把握することが可能となってきている。こうした背景のもと、本研究では、有価証券報告書に含まれる主要な記述情報であるMD&A、リスク情報およびガバナンス情報を対象として、テキストマイニング技術を用いて情報内容の定量化を行い、それら指標の時系列的な変化および企業属性との関連性を明らかにする。

本研究の特徴は、分析対象を記述情報の拡充が行われた2004年から2018年までの15年間にわたった上場全社を対象としていること、文字数に代表される量的指標だけでなく、具体性（数字表現、固有表現）、わかりやすさ（可読性）、定型性（ボイラープレート）、粘着性（スティッキネス）、そしてセンチメントなど質的指標を算出し、これらテキスト指標と企業のファンダメンタルズとの関連性を検証している点にある。

2．制度と先行研究

2.1　記述情報に関する開示制度の拡充

2000年以降、有価証券報告書における記述情報の拡充が進んでいる。金融審議会金融分科会第一部会に設置されたディスクロージャーワーキング・グループは、2001年10月からディスクロージャーに関する制度整備について審議を行い、2002年12月16日に「証券市場の改革促進」として公開された。この提言を受け、「企業内容等の開示に関する内閣府令（以下、開示府令）」が改正され、2003年4月1日以降に開始する事業年度に係る有価証券報告書から「事業等のリスク」「財政状態及び経営成績の分析[1]」「コーポレート・ガバナンスの状況」の記載が義務付けられた。リスク情報は、わかりやすく、かつ、簡潔に開示されること、MD&Aは、提出会社の経営成績と財政状態についての

経営陣による分析が出来る限り幅広く、かつ、具体的に記載されること、ガバナンス情報は、提出会社の自主的な判断に基づき出来る限り幅広く、かつ、具体的に記載されることが求められた。

その後、2010 年にコーポレート・ガバナンスの記載内容の充実（コーポレート・ガバナンス体制の概要および採用理由、財務・会計に関する知見を有する監査役の有無、社外役員と内部統制部門との関係、社外役員の設置状況など）が行われ、また 2015 年に役員の男女別人数および女性役員の比率の記載、その後も役員報酬、政策保有株式の開示拡充などが行われてきた。また、2018 年にこれまで「事業の状況」において記載されていた「業績等の概要」及び「生産、受注及び販売の状況」が MD&A に統合されるとともに、MD&A の記載内容の充実（経営成績等に重要な影響を与えた要因についての経営者の認識・分析および経営方針・経営戦略等の中期的な目標に照らした際の経営成績等の経営者の分析・評価）が行われた。

記述情報の開示充実の流れをさらに加速させたものが、2018 年 6 月の金融庁金融審議会ディスクロージャーワーキング・グループによる報告書、および同年 12 月に金融庁による「記述情報の開示に関する原則（案）」（以下、開示原則）の公表である。開示原則では、投資家による適切な投資判断を可能とし、投資家と企業との深度ある建設的な対話を促進するため、経営方針・経営戦略等、経営成績等の分析、リスク情報を中心に、プリンシプルベースによる開示の考え方が述べられた。これを受けて、2019 年 1 月に開示府令の改正によって記述情報の大幅な拡充が実施され、2020 年 3 月以降から適用されている。

2.2　先行研究

会計情報の有用性低下、コンピュータ技術の発達により、2000 年以降、テキストマイニングを用いた財務報告に関するテキストデータの研究が増えつつある（Türegün 2019）。本節では、法定開示書類である米国の Form 10-K および日本の有価証券報告書に含まれる記述情報を対象とした分析を概観する [2]。

これまでの研究では、MD&A の可読性とセンチメント（トーン）に着目した分析が多く実施されてきた（首藤 2019）。例えば、Li（2008）は、1994 年から 2004 年にかけての米国企業を分析対象とし、アニュアルリポートの可読性と企業業績や収益の持続性との関係を分析している。分析の結果、収益の低い企業のアニュアルリポートの可読性が低く、可読性の高いアニュアルリポートを持つ企業の方が収益の持続性が高いことが明らかになっている。また、Feldman et al.（2010）は、アクルーアルやアーニングスサプライズをコントロールしたうえで、MD&A のトーンの変化が市場の反応と関連することを報告している。リスク情報では、情報量や情報の具体性に関する研究が行われている。例えば、Campbell et al.（2015）は、規模が大きく負債が多い企業、そして市場ベータが高い企業ほど、リスク情報の文字数が多いことを報告している。また、Hope et al.（2016）は、10-K が長くわかりやすい、アクルーアルが小さい、パフォーマンスが高い、そして規模が小さいほど、リスク情報の具体性が高いこと、そしてリスク情報の具体性が市場の反応と関連することを明らかにしている。このように MD&A やリスク情報の研究が進められている一方、ガバナンス情報の記述情報に関する研究

はほとんど見当たらない。

これまでの研究の多くが、特定の記述情報を対象に特定のテキスト指標を分析していたのに対して、Dyer et al.（2017）はForm 10-K全体を対象に多くのテキスト指標を分析している。分析の結果、1996年から2013年にかけてForm 10-Kの文長、ボイラープレート（定型性）、スティッキネス（粘着性）、冗長性が増加し、具体性、可読性、および数字字現の相対的な量が減少していることが報告されている。さらに、潜在的ディリクレ配分法（Latent Dirichlet Allocation、以下LDA）を用いた分析では、米国財務会計基準審議会（Financial Accounting Standards Board、以下FASB）と米国証券取引委員会（Securities Exchange Commission、以下SEC）による新たな要求事項が文長の増加の大部分を説明し、150のトピックのうち3つのトピック（公正価値、内部統制、リスク要因の開示）が事実上の増加のすべてを占めていることを明らかにしている。

日本企業を分析対象とした先行研究として、首藤・緒方（2009）は2006年度における上場企業300社のMD&Aを分析した結果、新規上場企業がMD&Aの開示に積極的であること、開示項目については多少の偏りがあること、MD&Aの記述において他の箇所の参照を指示している企業が少なくないことが明らかにされている。また、中野（2010）はリスク情報とMD&A情報を分析した結果、事業リスクの高い企業、大規模企業、市場からの注目度の高い企業および事業構造が複雑な企業ほど、積極的に開示する傾向があるということを報告している。Kim and Yasuda（2018）によると、このようなリスク情報は投資家のリスク認識に影響している可能性がある[3]。

野田（2016）は、2004年から2012年にかけての金融業を除く東証一部上場企業1,200社について有価証券報告書の記述情報（対処すべき課題、MD&A、リスク情報、ガバナンス情報）を分析している。分析の結果、定性情報の開示の積極性は社外取締役比率の高さなどコーポレート・ガバナンス関連の数値と関連がある一方、負債比率が高い企業や安定持株比率が高い企業は開示に消極的であり、パフォーマンスと定性情報の開示にはマイナスの関係があるという結果が示されている。またMD&Aにおける定性情報の開示は、業績が悪化している状況において、積極的に開示を行う点がマーケットからプラスに評価されているという結果が示されている。そして事業等のリスクの開示量はアナリスト予想精度に影響を与え、その開示内容によってアナリスト予想精度に与える効果が異なるということが明らかにされている。矢澤（2020）は、野田（2016）のフォローアップとして、2015年から2019年までの有価証券報告書の記述情報を分析し、対処すべき課題、MD&A、リスク情報、ガバナンス情報の文字数が増加していることを報告している。

このように、日米において法定開示書類に含まれるテキストデータの分析が蓄積されつつあるものの、こと日本に関しては有価証券報告書における記述情報の分析は十分であるとはいえない。本研究では、こうした背景を踏まえ、記述情報の拡充が行われた2004年から2018年までの15年間を対象として、米国の先行研究で検証されたテキスト指標をローカライズすることによって、時系列および国別で比較できる実証的証拠の提示を行う。

3．テキスト指標

テキストマイニングでは、自然言語で記述された非構造化データから何をどのように抽出するかが重要となる。本研究では情報の量（quantity）と質（quality）という 2 つの側面から定量化する。テキストデータの量を指数化する指標として本研究では、文字数、単語数、文章数の 3 つを使用する。当該指標の基本的前提は、テキストデータの量が多ければそれだけデータのもつ価値が高いということである。一方、いくらテキストデータの量が多くとも、それに質が伴わなければノイズとなり、情報の価値が高まるばかりか、価値を棄損させてしまうことにもつながる。そこでテキストデータの質を測る指標として本研究では、数字表現、固有表現、可読性、センチメント、ボイラープレート、スティッキネスの 6 つの指標を使用する（表 1 参照）。先述のように、有価証券報告書における記述情報の記載にあたり、わかりやすく、かつ具体的に記載することが求められている。数字表現と固有表現はテキストデータの具体性、すなわち記述が数字や固有の表現を用いて説明されているかどうかを指標化したものである。可読性は、文章が読み手に取って読みやすく書かれているかどうかを指標化したものである。次に、センチメント（トーン）はポジティブかネガティブかという記述情報の傾向を指標化したものである。そして、紋切り型の表現が問題となっていることから、ボイラープレートとスティッキネスを分析する。ボイラープレートとスティッキネスはテキストデータの定型性、粘着性をみるもので、他企業や前年度と同様の表現を用いているかどうかを指標化したものである。これらの指標が置いている前提は、テキストデータが具体的であるほど、読みやすいほど、そして定型性、粘着性が低いほど、データの質が高まるというものである。

表 1　分析指標の定義

分析指標	定　　義
文字数	各セクションの文字数（単位：文字）
単語数	各セクションの単語数（単位：語）
文章数	各セクションの文章数（単位：文）
数字表現	各セクションにおける数字表現（金額、割合、日付、回数、人数）の個数（単位：個）
固有表現	各セクションにおける固有表現（場所、人物、組織）の個数（単位：個） ※固有表現の抽出には NEologd 辞書を使用。
可読性	李（2016）による可読性指標。可読性＝平均文長 *−0.056＋漢語率 *−0.126＋和語率 *−0.042＋動詞率 *−0.145＋助詞率 *−0.044＋11.724
センチメント	ポジティブワードの個数／（ネガティブワードの個数＋ポジティブワードの個数）※ポジティブワードとネガティブワードは、L&M ワードリストを日本語に訳して使用
ボイラーワード	ボイラーワードの個数（単位：個） ※ボイラーワードは 8 単語から構成されるフレーズのうち、当該年度の調査対象企業の 3 割以上において用いれているフレーズ
スティッキーワード	スティッキーワードの個数（単位：個） ※スティッキーワードは、8 単語から構成されるフレーズのうち、前年度と同じフレーズ

テキストマイニングではテキストデータの分析にあたり、テキストを形態素といわれる分析単位に分割して解析する。本研究ではプログラミング言語として Python（Ver.3.7）、形態素解析システムと

してMeCab[4]を使用している。また、形態素解析用の辞書としてMeCabにバインディングされているipadic[5]に加えて、分析指標に応じてUniDic[6]、NEologd[7]および筆者らが独自に作成したユーザー辞書（UserDic）を使用している。ユーザー辞書は、有価証券報告書や財務諸表にのみ用いられる単語を抽出するために作成されたものである。たとえば、「利益剰余金」はipadic辞書では「利益」と「剰余金」の2つの要素に分割されてしまう。有価証券報告書の形態素を正しく分析するためには「利益剰余金」を一つの単語として認識することが重要で、そのためユーザー辞書が必要となる。本研究では、有価証券報告書から抽出したキーワードおよび日経NEEDS Financial QUESTの財務データ項目を利用してユーザー辞書を作成した。具体的には、業界および年度特性を考慮するため、2004年から2018年までの期間において年別・業種別（東証業種）に最もボリュームの大きい（ファイルの容量が大きい）有価証券報告書を選んだ（合計495社の有価証券報告書）。続いて、KH Coder[8]を利用して複合語を抽出し、抽出された複合語の中でスコアが100,000を超えるものを選んだ。最後に、FQ項目との重複などを調整し、2,180単語のユーザー辞書を作成した。

4. 実態分析

4.1 データ

本研究の分析対象は、日本の証券取引所に上場する企業であり、分析期間は、有価証券報告書に「財政状態、経営成績及びキャッシュ・フローの状況の分析（以下、MD&A）」「事業等のリスク（以下、RISK）」そして「コーポレート・ガバナンスの状況等（以下、CG）」が新設された2004年から2018年まで15年間である。プロネクサス社の提供するeolデータベースから各年度の有価証券報告書の各セクションをHTML形式でダウンロードした結果、55,701社・年が入手できた[9]。有価証券報告書における各セクションは、見出し、本文、図および表から構成される。そこで入手したHTMLデータを解析し、ここから句点で終わる文章を抽出できるサンプルに限定した結果、47,930社・年となった[10]。さらに、金融業（銀行、保険、証券、その他金融）を除き、分析に必要な企業属性・財務データが入手できる企業に限定した結果、分析対象サンプルとして44,710社・年が抽出された[11]。なお、内閣府令の改正は3月31日以降に終了する事業年度を適用の基準としているため、2004年は2004年3月31日から2005年3月30日に終了する事業年度の有価証券報告書から構成され、これ以降の年度も同様の区切り方で対象企業を抽出している。

表2　サンプルセレクション

上場企業2004-2018（3セクションのデータがHTML形式でダウンロード可）	55,701
（控除）	
3セクションのテキストデータが抽出できない	−7,771
金融業（銀行、保険、証券、その他金融）	−2,567
決算期変更、個別決算のみ	−450
企業属性、財務データが利用できない	−203
分析対象サンプル	44,710

表 3 はセクションごとに 44,710 社・年の記述統計を示している。MD&A は、文字数 2,284 字（平均、以下同）、単語数 1,224 語、文章数 33 文、数字表現 45 個、固有表現 5 個、可読性 0.43、センチメント指数 0.52、ボイラーワード 17 個、スティッキーワード 290 個となっている。RISK は、文字数 2,202 字（平均、以下同）、単語数 1,161 語、文章数 27 文、数字表現 3 個、固有表現 7 個、可読性 0.57、センチメント指数 0.21、ボイラーワード 31 個、スティッキーワード 776 個となっている。CG は、文

表 3　記述統計

MD&A	平均	標準偏差	10%	50%	90%
文字数	2,284	2,022	880	1,843	4,038
単語数	1,224	1,038	486	998	2,149
文章数	33.84	28.46	14	28	59
数字表現	45.69	30.88	16	41	78
固有表現	5.59	13.53	0	2	13
可読性	0.43	0.65	−0.376	0.483	1.17
センチメント指数	0.52	0.41	0.60	0.56	0.50
ボイラーワード	17.28	20.71	0	12	35
スティッキーワード	290.39	380.40	51	199	599
RISK	平均	標準偏差	10%	50%	90%
文字数	2,202	2,116	616	1,560	4,551
単語数	1,161	1,095	332	826	2,392
文章数	27.25	23.80	9	20	54
数字表現	3.66	6.31	0	1	10
固有表現	7.38	14.22	0	4	17
可読性	0.57	0.54	−0.082	0.582	1.2
センチメント指数	0.21	0.25	0.00	0.19	0.23
ボイラーワード	31.12	23.15	4	28	61
スティッキーワード	776.66	726.70	202	559	1,610
CG	平均	標準偏差	10%	50%	90%
文字数	4,106	2,288	1,278	3,912	6,887
単語数	2,173	1,195	679	2,080	3,626
文章数	61.60	32.80	20	60	102
数字表現	19.92	10.17	9	19	32
固有表現	20.11	16.97	2	17	42
可読性	0.66	0.37	0.237	0.65	1.08
センチメント指数	0.49	0.51	0.54	0.50	0.48
ボイラーワード	99.33	57.67	0	114	160
スティッキーワード	1379.46	781.04	373	1,328	2,381
企業属性	平均	標準偏差	10%	50%	90%
lnASSETS	10.36	1.71	8.33	10.21	12.61
LEVERAGE	2.75	10.55	1.26	1.98	4.32
ROA	4.70	10.76	−0.30	4.56	12.28
LOSS	0.15	0.36	0	0	1
BIG4	0.74	0.44	0	1	1
TSE1	0.75	0.43	0	1	1
JASDAQ	0.15	0.35	0	0	1
NJGAAP	0.02	0.15	0	0	0

字数4,106字（平均、以下同）、単語数2,173語、文章数61文、数字表現19個、固有表現20個、可読性0.66、センチメント指数0.49、ボイラーワード99個、スティッキーワード1,379個となっている。テキスト指標の側面からとらえると、文量はMD&AとRISKはほぼ同数、数字表現はMD&Aが多く、固有表現はCGが比較的多く、可読性はいずれも0.5下限値前後、センチメントはRISKがネガティブ傾向、ボイラープレートワードとスティッキーワードはCGとRISKが多くなっている。なお、企業属性を見ると資産規模（平均、以下同）は315億70百万円、ROAは4.7%、レバレッジは2.75倍、損失計上企業は15%、Big4による監査は74%、東証一部上場企業は75%、JASDAQ上場企業は15%、国際会計基準あるいは米国基準採用企業は2%となっている。

4.2 分析結果

本節ではテキスト指標の経年推移および構成要素の推移を分析する。

4.2.1 文量

図1パネルAからCは、抽出されたテキストデータの文字数、単語数、文章数を計算し、15年間の推移を示したものである。図1パネルAで示された各セクションの平均単語数をみると、2004年時点ではMD&Aが最も多く1,079語、それにRISK 777語、CG 457語と続く形であった。その後、CGの単語数が大きく増加し、2018年に3,097語と3セクションのなかで最も多くなっている。MD&Aは2004年からほぼ横ばいで推移していたが、2018年には「業績等の概要」と「生産、受注及び販売の状況」が統合されたため、前年比2倍超の2,489語となっている。RISKは経年的に増加傾向にあり、2018年2004年比2倍の1,421語となっている。同様に図1パネルBで文字数、パネルCで文章数を示している。いずれも単語数と同様の傾向にあり、2018年において文字数はMD&A 4,681語、RISK 2,699語、CG 5,876語、文章数はMD&A 70文、RISK 32文、CG 87文となっている。

4.2.2 数字表現

数字表現は、金額、割合、日付、回数、人数の各表現の使用個数から構成される。各セクションにおける数字表現の使用例を挙げると以下の通りである。

MD&A

> 当社グループは、（…中略…）企業価値の向上と持続的な成長を図るため、2017年度をスタートとする4ヶ年の第3次中期経営計画（2020年度目標：売上高3,400億円、営業利益170億円、ROE 8%以上）を策定し、（…中略…）（下線著者）

RISK

> 2016年12月には、道内で初めて鳥インフルエンザが発生しましたが、特段の消費減退はみられませんでした。（下線著者）」「広告・マーケティング事業の売上高のうち、当連結会計年度における100分の10以上の販売先は株式会社K社であり、その金額は当該セグメント売上高の12.6%を占めております。（下線著者、一部修正）

CG

2017 年度は、社外取締役ミーティングを 6 回開催（社外取締役出席率 100%）し、以下のテーマについて取り組みました。（下線著者）

図 1　文量

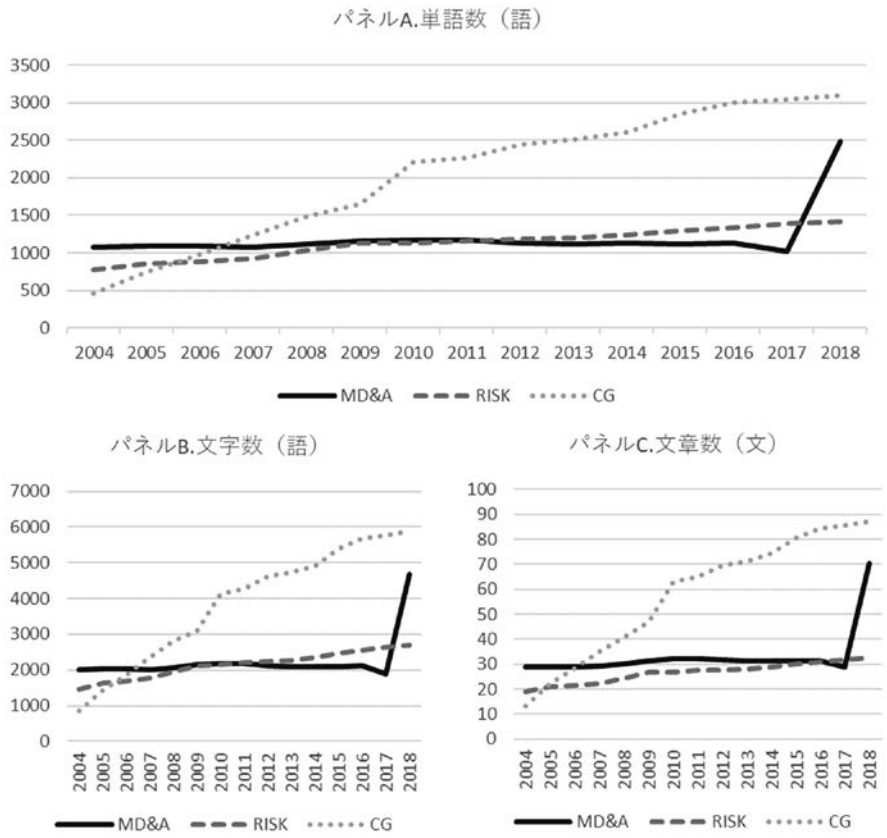
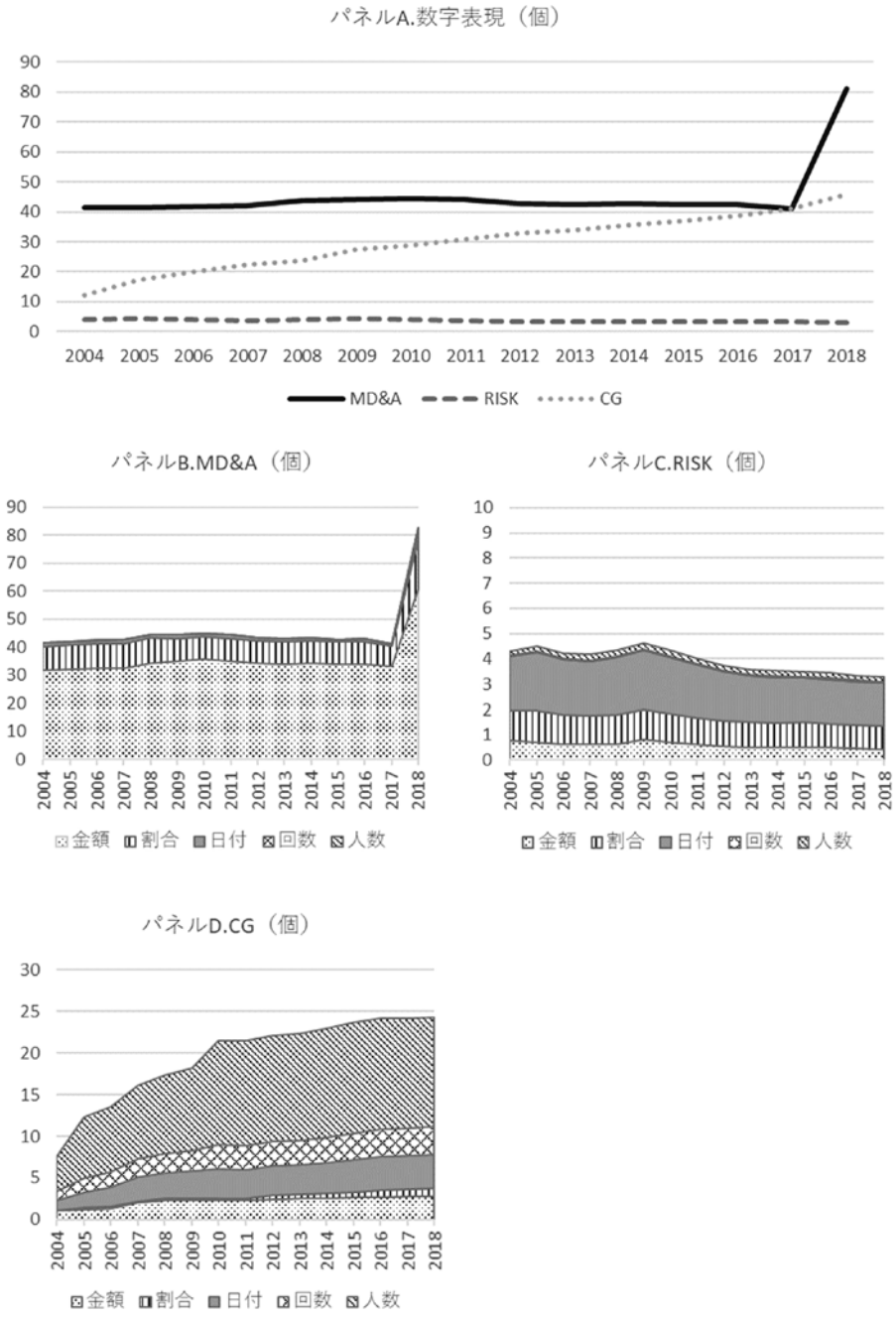


図 2 パネル A で示された数字表現の使用個数をみてみると、MD&A は 2004 年時点で 41 個、RISK は 4 個、CG は 7 個となっている。MD&A はその後ほぼ横ばいの傾向が続き、2018 年に単語数と同様に前年比 2 倍の 81 個となっている。CG も同様に経年的に増加傾向にあり 2018 年に 45 個となっている。一方、RISK は経年的にわずかながら減少傾向にあり、2018 年には 3 個となっている。まとめると、MD&A と CG は単語数の増加に比例して数字表現も増えている一方、RISK は単語数が増えているにも関わらず、数字表現はわずかに減少している点に特徴がある。

それではなぜこのような傾向が見られるのだろうか。図 2 パネル B から C で示された数字表現の内訳をみると、各セクションの特徴として MD&A は 41 個の数字表現のうち金額表現が 30 個を占め、残りはほぼ割合表現となっている。RISK は 4 個の数字表現のうち日付表現が 2 個、割合表現が 1 個となっている。CG は 7 個の数字表現のうち人数が 4 個、金額、日付、回数がそれぞれ約 1 個ずつを占めている。そのうえで、セクションごとの推移を見ていくと、MD&A では金額表現が経年的に増加傾向にあり、これが全体の増加をけん引していることがわまる。自社の財政状態と経営成績を財務

図 2　数字表現



数値に基づいてより具体的に開示する動きが進んでいることの現れであると考えられる。続いて RISK はもともと 2004 年において数字表現が 4 個と少なかったが、そのうち金額表現が 2004 年 0.7 個から 2018 年 0.4 個、割合表現が同 1.2 個から 0.9 個、日付表現が 2.1 個から 1.6 個といずれも減少している。すなわち自社を取り巻くリスクが複雑化・多様化するなかで、文量を割いてより自社の抱えるリスクを丁寧に記載しようとする傾向にある一方、それが企業経営に与える影響を予測することが

困難になり表現の抽象度が上がっているのかもしれない。最後に、CG は、人数表現が 2004 年 4 個から 2018 年 13 個、回数表現が同 1 個から 3.4 個、日付表現が同 1.2 個から 4 個とほぼ 3 倍に増加している。これらの傾向は、企業の採用するコーポレート・ガバナンスの体制、例えば、役員構成、各種組織の構成員の人数、活動状況などをより具体的に記載するようになってきていることを反映していると考えられる。

4.2.3　固有表現

固有表現は、地域、人名、組織から構成される。各セクションにおける数字表現の使用例を挙げると以下の通りである。

MD&A

> 当グループは北米、東南アジア、豪州などにおいて、天然ガス・石油の生産・開発事業、液化天然ガス（LNG）事業を行っており、石油・ガス価格は当グループの業績に少なからぬ影響を与えます。（下線著者）

RISK

> NIDEC の継続的な成功は主に NIDEC の創業者であり代表取締役会長（最高経営責任者）の永守重信氏の能力と手腕に依存しております。（下線著者）

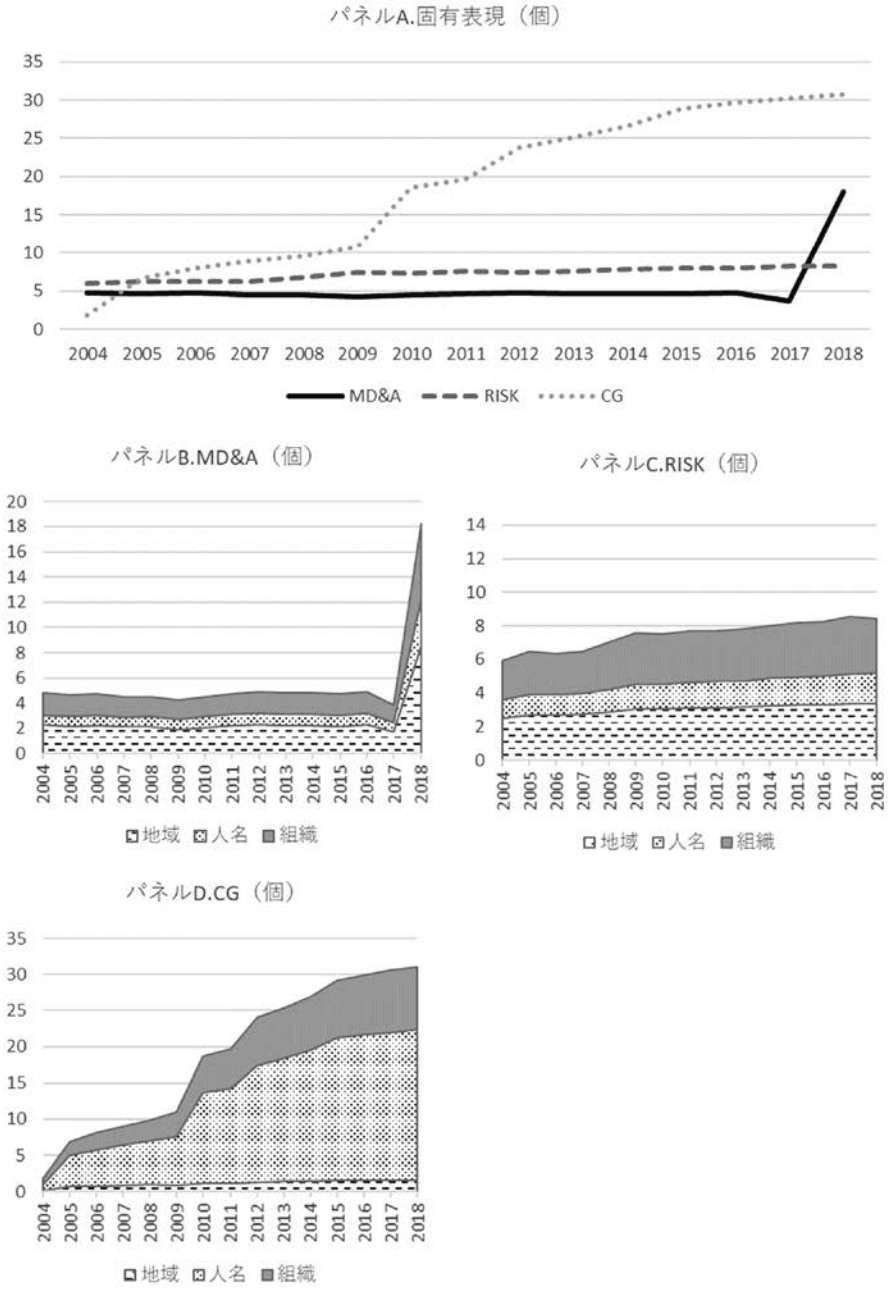
CG

> 小嶋一美氏は、株式会社パルコにて会計・財務に関する豊富な実務経験と専門知識を有し、会社経営を統括する充分な見識を有するものであります。（下線著者）

図 3 パネル A は数字表現の経年推移を示している。2004 年段階における固有表現は、MD&A 4.7 個、RISK　5 個、CG　1.8 個と金額表現と比較して全体的に少ない。MD&A はその後横ばいで推移し、数字表現と同様に 2018 年に前年比約 4 倍の 17 個となっている。RISK は、経年的に増加傾向にあり、2018 年に 8 個となっている。CG は、2009 年から 10 年にかけて前年比 2 倍に増加し、その後も増加傾向を維持し、2018 年には 30 個と 3 つのセクションのなかで突出して多くなっている。

数字表現と同様にそれぞれの変動要因を検討しよう。図 3 パネル B から D で示されたた固有表現の内訳をみると、各セクションの特徴として、MD&A は 4.7 個のうち地域表現が 2.2 個、組織表現が 1.7 個、人名表現が 0.7 個、RISK は 5.9 個のうち地域表現が 2.5 個、組織表現が 2.2 個、人名表現が 1 個、CG は 1.8 個のうち組織表現が 0.8 個、人名表現が 0.7 個、地域表現が 0.2 個となっている。個別にみていくと、MD&A は各指標ともほぼ横ばいで推移しているが、2018 年に地域が前年比 5 倍、組織と人名も約 4 倍と大幅に増加している。統合に伴い、情報内容が変わっている点が影響していると考えられる。RISK は地域、人名、組織とも経年的になだらかな増回傾向にある。先ほど見た数字表現は減少傾向にあるが、その一方で固有表現という点ではリスクをより具体的に記載する傾向にあると解釈できそうである。CG は 2010 年から 12 年にかけて人名と組織に関する記載が顕著に増加し、その後は緩やかな増加傾向にある。2010 年の内閣府令改正によるコーポレート・ガバナンスの記載充実の影響が表れているといえるだろう。

図 3　固有表現



4.2.4　可読性

本研究で使用する可読性指標は李（2016）により開発された指標であり、可読性は、平均文長、漢語率、和語率、動詞率、助詞率に定数を乗じて計算される。各セクションにおいて比較的可読性の高い表現を例示してみると、以下のようなものが挙げられる。

MD&A（4.53 初級後半やさしい）

このような経営環境の中、トヨタは、お客様の期待を超える「もっといいクルマ」づくりに取り組んできました。「トヨタのグローバルミッドサイズセダン」である「カムリ」を、TNGA（トヨタ・ニュー・グローバル・アーキテクチャー）に基づくプラットフォームやパワートレーンなどにより一新し、意のままの走りと美しいデザインを実現しました。

RISK（1.86 上級前半むずかしい）

当社は、地域に密着したスーパーとして埼玉県下に店舗を拡充しておりますが、各店の商圏内の同業他社との競合（オーバーストア）状況にあります。今後更に新規競合店舗が多数参入した場合、当社の業績に影響を与える可能性があります。

CG（1.63 上級前半むずかしい）

取締役会は、5 名の取締役で構成しており、うち 3 名は監査等委員である取締役です。法令及び取締役会規程に定める経営及び業務執行に関する重要事項を審議・決定しています。

図 4 パネル A をみると、2004 年における可読性スコアは、MD&A が 0.23、RISK が 0.69、CG が 0.75 となり、いずれの指標も 6 段階の判定基準の下限に位置していることがわかる。経年的な変化をみると、RISK は緩やかな低下傾向、CG は増減しつつも横ばい傾向にあるのに対して、MD&A は一貫して上昇傾向にあり、特に 2018 年に 0.8 と 3 つのセクションの中で最も可読性が高い水準に達している。

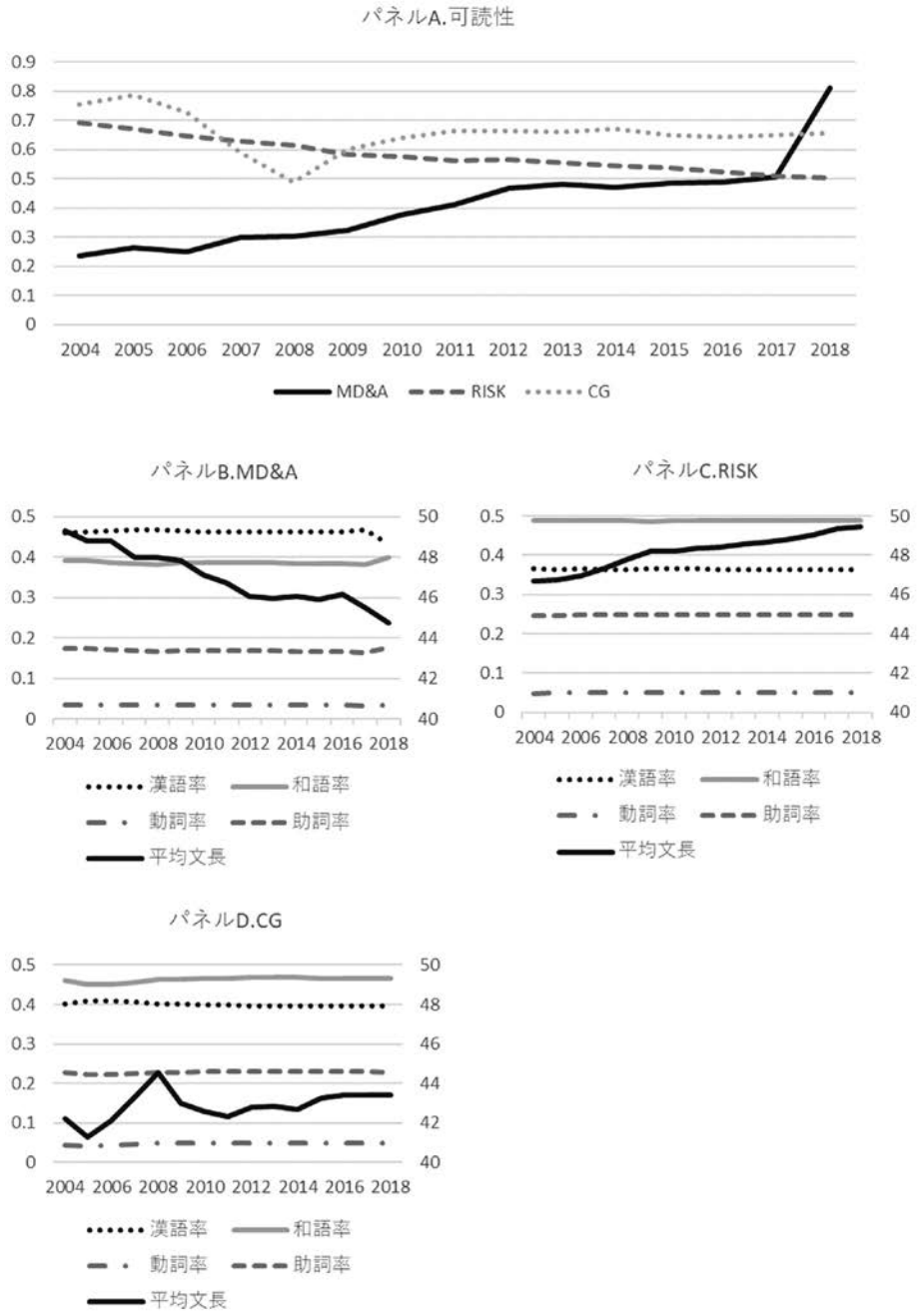
ではなぜこのような可読性の変化が生じたのだろうか。可読性スコアを構成する 5 つの要素のうち平均分長の寄与度が最も大きい（李 2016）。図 4 パネル B から D をみると、MD&A は 5 つの要素のなかで平均文長が継続的に下落傾向にあり（すなわち短文化が進んでいる）、これが可読性スコアを押し上げていることがわかる。加えて、2018 年は平均文長だけでなく、同時に、漢語率が低下し、和語率が上昇していることによって全体の可読性が上がっている点が特徴的である。MD&A とは反対に、RISK は平均文長の経年的な増加傾向が可読性を低下させている（すなわち、長文化が進んでいる）。CG も同様に、平均文長の変動が可読性の変動と重なっていることがわかる。以上から、MD&A では短文化が進む一方で、RISK では長文化が進んでいるという興味深い傾向がみてとれる。

4.2.5　センチメント指数

センチメント指数は、ネガティブワードとポジティブワードの個数の合計を分母にとり、分子にポジティブワードの個数をとる[12]。つまり、当該指数が高いほど、ポジティブな表現の割合が高いことを意味する。

図 5 パネル A で示されるように、2004 年時点では、MD&A は 52%、RISK は 23%、CG は 58% となっている。3 つのセクションのなかで、MD&A と CG はポジティブな表現とネガティブな表現がほぼ同割合使用されている一方、RISK はかなりネガティブな表現が多く用いられていることが伺える。経年的にみると、MD&A は 2009 年に前年の 51%から 46%へと 1 割ほどネガティブに変化している。これはリーマンショックに伴う景気悪化の影響を受け業績が悪化したことによるものと考え

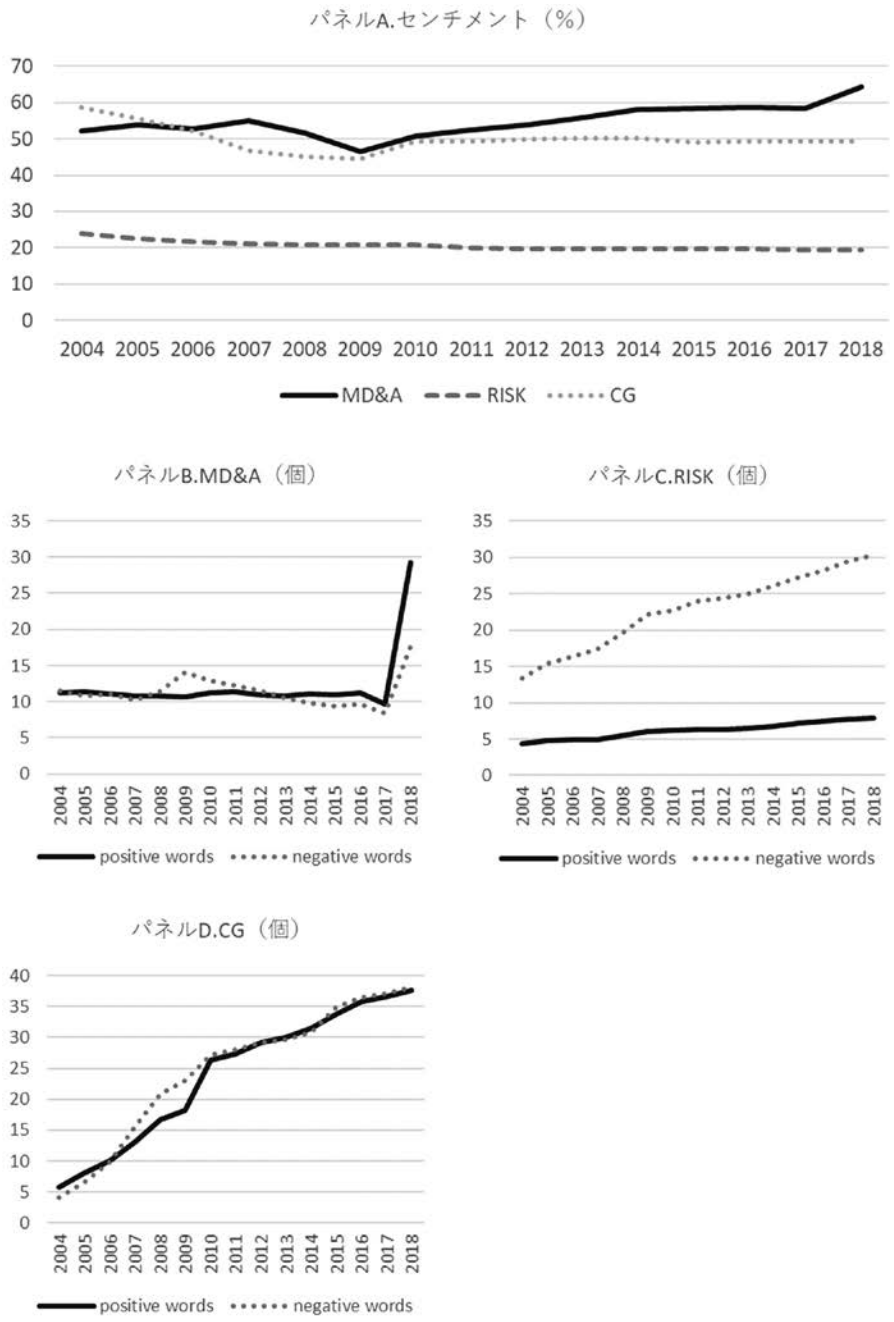
図4　可読性



られる。RISKは、経年的に指数が低下傾向にあり、2018年には19%まで低下している。CGは2009年まで低下傾向にあるものの、それ以降10年間は安定して推移している。

次に、センチメントの内訳をみていこう。図5パネルBからCはセンチメントをポジティブワードとネガティブワードに分け、それぞれの推移を示している。図5パネルBで示されたMD&Aをみると、ポジティブワードとネガティブワードの個数は約10個と経年的にほぼ同数のワードが使用さ

図 5　センチメント



れている。一方その推移をみると、ポジティブワードの個数は 2018 年を除いてほぼ横ばいであるのに対し、ネガティブワードの使用回数は 2009 年に前年の 11 個から 14 個に増加しており、これがセンチメント指数の低下をもたらしていることがわかる。当該傾向は、景気動向や企業業績に連動するのは主にネガティブワードであり、ポジティブワードはそれほど関連しない可能性を示唆している。図 5 パネル C で示された RISK も特徴的であり、2004 年においてポジティブワードが 4 回であるの

に対し、ネガティブワードは 13 回使用されている。この差は経年的に広がり、2018 年にはポジティブワードが 7 回に対して、ネガティブワードは 30 回となっている。すなわち、RISK においてもやはりネガティブワードがセンチメント指数の主な変動要因となっていることが伺える。最後に、図 5 パネル D の CG は、ポジティブワードとネガティブワードがほぼ同水準であり、経年的な増加傾向もほぼ同様であり、結果としてセンチメント指数が 50%水準で一定となっている。

4.2.6 ボイラープレートワード

本研究では 8 単語から構成されるフレーズをすべて抽出し、各年度において調査対象企業のうち 3 割以上の企業が用いているフレーズをボイラープレートワードとしている。ボイラーワードプレートワードとして抽出されたフレーズの個数は、2018 年において MD&A が 90、RISK が 63、CG が 184 となっている。ボイラーワードフレーズの具体例と会社数は以下の通りである。

MD&A

‘消費税’, ‘等’, ‘は’, ‘含ま’, ‘れ’, ‘て’, ‘おり’, ‘ませ’	2,941 社
‘文中’, ‘の’, ‘将来’, ‘に関する’, ‘事項’, ‘は’, ‘、’, ‘当’	2,321 社
‘わが国’, ‘において’, ‘一般’, ‘に’, ‘公正’, ‘妥当’, ‘と’, ‘認め’	2,201 社
‘に関する’, ‘認識’, ‘及び’, ‘分析’, ‘・’, ‘検討’, ‘内容’, ‘は’	2,120 社

RISK

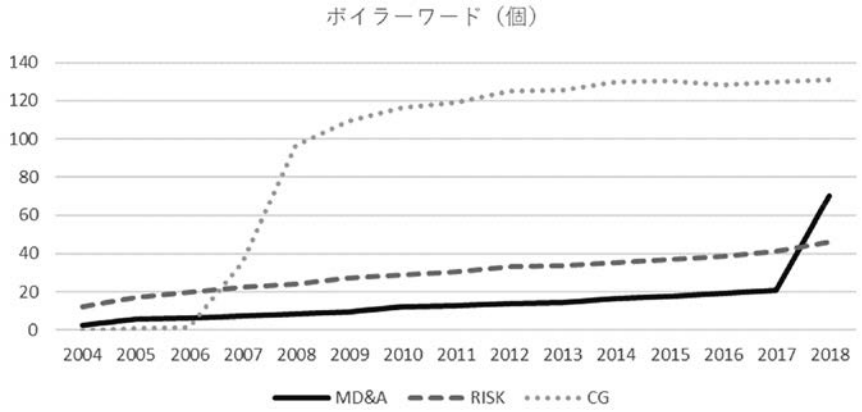
‘に’, ‘影響’, ‘を’, ‘及ぼす’, ‘可能性’, ‘が’, ‘あり’, ‘ます’	2,597 社
‘投資’, ‘者’, ‘の’, ‘判断’, ‘に’, ‘重要’, ‘な’, ‘影響’	1,836 社
‘経理’, ‘の’, ‘状況’, ‘等’, ‘に関する’, ‘事項’, ‘の’, ‘うち’	1,648 社

CG（2018 年）

‘議決権’, ‘の’, ‘3 分の 1’, ‘以上’, ‘を’, ‘有する’, ‘株主’, ‘が’	3,629 社
‘会社法’, ‘第’, ‘309’, ‘条’, ‘第’, ‘2’, ‘項’, ‘に’	3,160 社
‘株主総会’, ‘の’, ‘円滑’, ‘な’, ‘運営’, ‘を’, ‘行う’, ‘こと’	2,604 社
‘等’, ‘の’, ‘総額’, ‘が’, ‘1 億円’, ‘以上’, ‘で’, ‘ある’	1,946 社
‘株主’, ‘へ’, ‘の’, ‘機動的’, ‘な’, ‘利益還元’, ‘を’, ‘行う’	1,504 社

図 6 は各セクションのボイラープレートワードの個数の推移を示している。ボイラープレートワードの 1 社当たり平均語数をみると、2004 年では MD&A は平均 2 語、RISK は平均 33 語、CG は 0 となっている。RISK は国際情勢や景気動向など複数産業および企業に影響を与える要因もあることから同様のフレーズが使用されやすいものと考えられる。またすべてのセクションにおいて経年的にボイラープレートワードが増える傾向があり、上場企業間において段々と開示の「型」が形成されてきたといえる。CG は 2007 年から 2008 年にかけて、ボイラーワードが突出して増えている。この理由は、監査報酬の開示において「公認会計士法第 2 条第 1 項に規定する業務に基づく報酬」というフレーズ

図 6　ボイラーワード



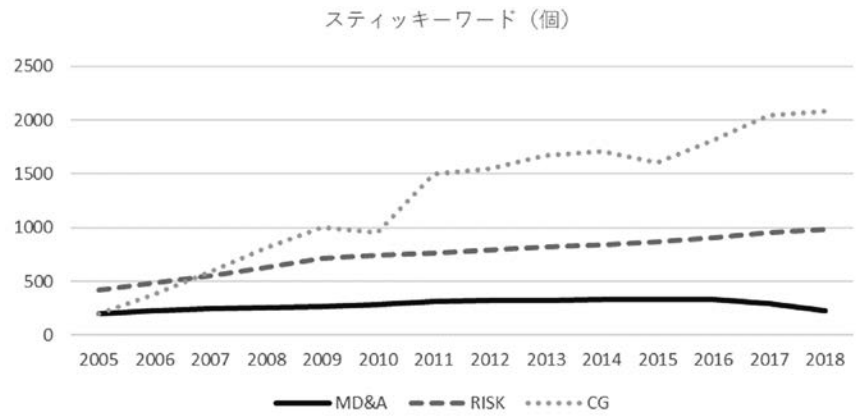
が多くの企業において統一的に使用されるようになったこと、および 2007 年 3 月期から株主総会の特別決議要件を変更した場合にはその内容及びその理由を記載することが要請され、多くの企業において定型的な説明「会社法第 309 条第 2 項の定めによる決議は、…」が用いられるようになったことに起因する。また、MD&A も 2018 年に統合を受けて前年比約 2 倍に増加している。この結果、2018 年段階では、CG が 195 語と最も多く、MD&A が 94 語、RISK が 61 語となっている。

4.2.7　スティッキーワード

本研究では、スティッキーワードは同一企業で使用される前年度と同じフレーズと定義される。指標の計算にあたり、ボイラープレートと同様に 8 単語から構成されるフレーズを抽出し、同一企業において前年度と同じフレーズの個数をカウントしている。

図 7 は各セクションのスティッキーワードの個数の推移を示している。スティッキーワードの 1 社当たり平均語数をみると、2005 年では MD&A は平均 201 語、RISK は平均 423 語、CG は平均 200 語となり、RISK におけるスティッキーワードが最も多いことがわかる。その後の経年変化は、ボイラーワードのそれとほぼ同様に右肩上がりとなっている。ただし、RISK の増加率に比べて、MD&A の増加率はそれほど高くなく、一方、CG はいくつかの山と谷が交互にくるという特徴がある。おそらく、MD&A は年度ごとに業績およびその影響要因が異なる可能性が高く、その説明もおのずと変化する一方、RISK は年度ごとに大きく入れ替わることはなく、累積的に増えていく傾向があるものと考えられる。また、ガバナンス情報は新たな開示項目が増える時点で一時的にスティッキネスが低下するものの、その後は前年と同じ表現が用いられるため、その後の期間で増加するものと考えらえる。2018 年段階では、CG が 2,082 語と突出して多く、RISK979 語、MD&A は 230 語と大きな差が表れている。なお、MD&A は 2018 年の改正により今までの開示形式からの変更が生じたため、スティッキーワードが減少した一方、多くの企業で同形式の開示がなされたためボイラープレートワードが大きく増加している点が特徴的である。

図7　スティッキーワード



4.2.8　本節のまとめ

以上の結果をセクションごとにまとめると、次のようになる。まず MD&A は、2018 年を除き、文量、数字表現、固有表現はほぼ横ばい、短文化が進むことにより可読性は増加、センチメントはほぼニュートラルであるが、全体的な業績下降期にはネガティブワードが増える傾向にある、ボイラープレートワードとスティッキーワードは経年的に一貫して増加傾向にある点が明らかになった。続いて、RISK は、文量は 15 年間を通して増加傾向にあり、数字表現（日付、割合、金額）が減少する一方で、固有表現（地域、人名、組織）は増加傾向にある、また長文化が進むことにより可読性は低下、センチメントはネガティブな表現をより多く使い、その割合が増えている、ボイラーとスティッキーはともに増加傾向にある。最後に、CG は、3 つのセクションのなかで最も顕著に文量が増加している、それに伴い数字表現（日付、回数、人数）が増加し、固有表現（人名、組織名）も 2010 年以降増加傾向にある、可読性は 2010 年以降ほぼ横ばい、センチメントも 2010 年以降ほぼ一貫してニュートラル、ボ 2008 年以降急激にボイラーワードが増加し、スティッキーワードは記載内容の変更時に減少するものの傾向として右肩上がりにある。

5. 実証分析

5.1　リサーチ・デザイン

本節では MD&A、RISK、CG から抽出されたテキスト指標と企業属性との関連性を分析する。そのために、前節で紹介した指標を次の通り調整する。まず、文量分析のために、各セクションの文字数を自然対数化した変数を用いる（*lnLength*）。数字表現および固有表現分析においては、それぞれの単語数をセクションの総単語数で基準化した値を用いる（*Hardratio* と *Specificity*）。続いて可読性については、前節で推定した指標をそのまま各セクションの可読性変数として分析する（*Readability*）。センチメント指数は、ポジティブワードの個数をネガティブワードとポジティブワードの個数の合計で割った値である（*Sentiment*）。最後にボイラープレートおよびティッキー表現としては、前節のプロセスで抽出したボイラープレートおよびティッキーワードの自然対数を用いる

（*lnBoiler* と *lnSticky*）。

テキスト指標と関係するファンダメンタル要因として、Dyer et al.（2017）に倣い、次の変数を取りあげる。まずは、多くの先行研究でディスクロージャーに影響を与えることが明らかになっている規模、負債依存度そして収益性を分析に含める（Verrechia, 1983; Lang and Lundholm, 1993; Kallapur and Trombley, 1999; Healy and Palepu, 2001; Miller, 2002）。規模は総資産の自然対数（*lnASSETS*）、レバレッジは総資産を純資産で割った値（*LEVERAGE*）を用いる。収益性としては経常利益を総資産で割った *ROA* と当期純利益が赤字であれば 1 を取る当期純損失ダミー（*LOSS*）の 2 つの変数を用いる。さらに、大手監査法人や上場市場および会計基準が記述情報に与える影響も合わせて分析する。BIG4 ダミー（*BIG4*）は KPMG あずさ、EY 新日本、トーマツ及び PwC あらたのいずれかの監査を受けている場合は 1 を取るダミー変数である [13]。東証 1 部ダミー（*TSE1*）と JASDAQ ダミー（*JASDAQ*）はそれぞれ東証 1 部上場企業そして JASDAQ 上場企業であれば 1 を取るダミー変数である。最後に非日本基準ダミー（*NJGAAP*）はアメリカ会計基準あるいは国際財務報告基準を採用している場合は 1 を取るダミー変数である。なお、すべての連続変数については上下 1%でウィンソライズしている。

推定においては欠落変数バイアスに対処するため企業固定効果を、マクロ経済的要因をコントロールするために年度の固定効果を加えている。また、記述情報の量が各指標に与える影響をコントロールするために、文量そのものを用いる文字数分析と、文字数で基準化した数字表現や固有表現を除いては、文字数の自然対数をコントロールする。係数の有意水準の計算には頑健な標準誤差を用いている。なお、変数間で大きな相関関係はみられなかった。

5.2　分析結果

5.2.1　MD&A

表 4 は MD&A についてテキスト指標と企業属性の関連をまとめたものである。まず、レバレッジと ROA および当期純損失ダミーの結果に注目すると、レバレッジが高く当期純損失を出すほど MD&A セクションの文量が多いことがわかる。これは財政状況が悪く収益性が低い企業ほど説明を多く行っていると解釈でき、そもそも MD&A が経営者による経営成績等の分析に関する記述が求められていることと整合的な結果であるといえる。一方、レバレッジが高いほど数字表現は減るが、収益性が高いほど数字表現が増える。以上から業績が好調な場合には具体的に開示し、そうではない場合には具体的な開示を控えている可能性が考えられる。また、収益性が高いほど固有表現が減るものの、その可読性は高くなる。また、レバレッジが高く収益性が低い企業ほど、記述情報がネガティブになる傾向にある。これらの結果は、MD&A セクションの記述情報に、企業の財政状態と収益性が大きく関連していることを意味する。

他の変数に目を向けると、規模が大きいほど固有表現が増えるものの可読性は低くなっている。これは規模が大きい企業ほど多角化または事業が複雑化している場合が考えられるため、より具体的な説明が必要となるためだと考えられる。また、大手監査法人の監査を受ける企業ほどボイラープレー

表 4　MD&A テキスト指標と企業属性

	lnLength	*Hardratio*	*Specificity*	*Readability*	*Sentiment*	*lnBoiler*	*lnSticky*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. *lnASSETS*	−0.010	−0.048	0.043***	−0.035***	−0.359	0.016	−0.013
	(−0.92)	(−1.46)	(4.66)	(−2.61)	(−0.88)	(0.79)	(−0.62)
2. *LEVERAGE*	0.008***	−0.026***	0.002	−0.006*	−0.542***	−0.001	0.002
	(3.04)	(−3.71)	(1.12)	(−1.87)	(−5.77)	(−0.15)	(0.39)
3. *ROA*	−0.001*	0.009***	−0.001***	0.003***	0.511***	0.004***	0.001
	(−1.86)	(5.87)	(−2.68)	(4.74)	(19.27)	(3.64)	(1.51)
4. *LOSS*	0.025***	−0.090***	0.002	−0.022***	−14.288***	−0.001	−0.018*
	(4.25)	(−5.17)	(0.45)	(−2.94)	(−44.45)	(−0.11)	(−1.68)
5. *BIG4*	−0.022	0.026	−0.002	0.036**	0.373	−0.082***	−0.055**
	(−1.45)	(0.64)	(−0.13)	(2.01)	(0.66)	(−2.95)	(−2.30)
6. *TSE1*	0.035**	−0.051	0.023*	−0.021	0.832	0.063**	0.016
	(2.22)	(−1.17)	(1.86)	(−1.12)	(1.34)	(2.13)	(0.62)
7. *JASDAQ*	0.021	−0.016	0.035***	0.004	0.852	0.060*	−0.053*
	(1.28)	(−0.32)	(2.70)	(0.21)	(1.27)	(1.93)	(−1.93)
8. *NJGAAP*	−0.053	−0.128*	0.103***	0.058*	−0.400	−0.353***	−0.250***
	(−1.30)	(−1.70)	(3.82)	(1.89)	(−0.37)	(−4.57)	(−5.34)
9. *InLength*				−0.102***	−0.291	0.555***	0.416***
				(−7.80)	(−0.64)	(30.12)	(24.00)
constant	7.482***	4.547***	−0.093	1.376***	57.552***	−3.489***	1.792***
	(64.74)	(13.61)	(−1.00)	(8.12)	(10.87)	(−14.40)	(7.08)
Observations	44,710	44,710	44,710	44,710	44,710	44,710	41,027
Adj R-squared	0.339	0.0552	0.0905	0.0955	0.224	0.604	0.174
Year & firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

注）係数の有意水準の計算には頑健な標準誤差を用いている．***1％水準で有意．**5％水準で有意．*10％水準で有意。

トあるいはスティッキー表現があまり用いられない傾向にある。興味深いことにアメリカ会計基準や国際財務報告基準を採用する企業の MD&A セクションは日本会計基準を採用する企業のそれより固有表現が多い一方でボイラープレートあるいはスティッキー表現が使われていない点である。アメリカでは 1960 年代から MD&A が開示されていることに加えて監査（あるいはレビュー）の対象でもあり、国際会計基準において MD&A に相当する MC（Management Commentary）も同様である。したがって、多くの企業が使うフレーズや昨年度用いた表現はなるべく避け、具体的な記述に力を入れている可能性がある。

5.2.2　RISK

表 5 は RISK についてテキスト指標と企業属性の関連をまとめたものである。分析結果によると、規模が大きくレバレッジが高い企業ほど、そして赤字企業ほどリスク情報の開示量が多い一方で東証 1 部上場企業ほど開示量が少ないことがわかる。規模が大きく負債依存度が高い企業そして収益性が

表 5　RISK テキスト指標と企業属性

	lnLength	*Hardratio*	*Specificity*	*Readability*	*Sentiment*	*lnBoiler*	*lnSticky*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. *lnASSETS*	0.028**	−0.059***	0.001	0.023**	−0.195	0.027	−0.011
	(2.19)	(−6.26)	(0.08)	(2.14)	(−0.64)	(1.38)	(−1.04)
2. *LEVERAGE*	0.005**	0.010***	−0.002	−0.003	−0.075	0.003	−0.009***
	(2.07)	(4.58)	(−0.97)	(−1.29)	(−1.22)	(0.58)	(−4.16)
3. *ROA*	0.000	−0.001**	0.001**	0.001***	0.016	0.001	0.001
	(0.08)	(−2.53)	(2.52)	(2.59)	(1.20)	(1.57)	(1.38)
4. *LOSS*	0.019***	0.017***	0.004	0.000	−0.104	−0.015*	−0.015**
	(3.30)	(3.53)	(0.73)	(0.07)	(−0.67)	(−1.66)	(−2.34)
5. *BIG4*	0.057***	0.009	0.010	0.016	0.249	−0.030	0.051***
	(3.42)	(0.66)	(0.73)	(1.18)	(0.58)	(−1.14)	(4.18)
6. *TSE1*	−0.037**	−0.010	0.010	−0.000	−0.017	0.021	0.008
	(−2.30)	(−0.75)	(0.59)	(−0.03)	(−0.04)	(0.86)	(0.64)
7. *JASDAQ*	−0.023	0.013	0.005	−0.003	−0.371	−0.006	−0.006
	(−1.32)	(0.89)	(0.28)	(−0.23)	(−0.85)	(−0.22)	(−0.44)
8. *NJGAAP*	0.057**	0.008	−0.023	−0.028	−0.017	−0.053	−0.026
	(2.31)	(0.45)	(−0.90)	(−1.33)	(−0.03)	(−1.22)	(−1.02)
9. *InLength*				−0.226***	1.599***	0.525***	0.790***
				(−14.58)	(3.78)	(19.80)	(53.96)
constant	6.735***	1.127***	0.662***	2.006***	15.183***	−2.046***	0.146
	(51.66)	(11.60)	(5.98)	(12.86)	(3.60)	(−7.33)	(1.00)
Observations	44,710	44,710	44,710	44,710	44,710	44,710	41,033
Adj R-squared	0.136	0.137	0.023	0.080	0.039	0.425	0.416
Year & firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

注）係数の有意水準の計算には頑健な標準誤差を用いている．***1％水準で有意．**5％水準で有意．*10％水準で有意。

低い企業ほど、ビジネスリスクが多いことを踏まえると、リスクの開示量が多くなっていると考えられる。さらに、東証 1 部上場企業はそうでない企業より比較的に安定した成熟した企業が多く、その結果その結果リスクの開示量が少なくなっている可能性がある。続いてレバレッジが高く赤字企業ほど数字表現が高くスティッキー表現が少ないこともわかる。財政状態が悪く収益性が低い企業はリスク要因を更新しつつ数字を用いながら具体的に記述している可能性がある。このように RISK セクションの記述情報にも、MD&A セクションと同じく、企業の財政状態と収益性が影響しているようである。

5.2.3　CG

表 6 は CG についてテキスト指標と企業属性の関連をまとめたものである。企業の財政状態と収益性が MD&A や RISK セクションの記述情報に影響を与える可能性がある一方で、CG についてはそのような傾向は強く見られない。むしろ他の要因が影響しているように見える。たとえば、規模が大

表 6　CG テキスト指標と企業属性

	lnLength	*Hardratio*	*Specificity*	*Readability*	*Sentiment*	*lnBoiler*	*lnSticky*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. *lnASSETS*	0.023***	−0.006	0.045***	−0.016*	−0.153	−0.026*	−0.030***
	(3.01)	(−0.53)	(3.42)	(−1.92)	(−0.56)	(−1.94)	(−3.24)
2. *LEVERAGE*	0.000	0.003	−0.001	−0.003	−0.041	0.007**	−0.002
	(0.27)	(1.17)	(−0.29)	(−1.36)	(−0.79)	(2.28)	(−1.09)
3. *ROA*	0.001	0.000	−0.001	−0.000	0.045***	−0.000	−0.000
	(1.52)	(0.11)	(−1.19)	(−1.26)	(3.63)	(−0.22)	(−0.64)
4. *LOSS*	−0.005	0.012*	0.003	−0.004	−0.026	0.001	0.006
	(−1.24)	(1.72)	(0.42)	(−0.81)	(−0.18)	(0.06)	(0.88)
5. *BIG4*	−0.006	−0.007	−0.003	0.011	1.816***	−0.059**	0.021
	(−0.54)	(−0.39)	(−0.16)	(0.87)	(4.39)	(−2.55)	(1.50)
6. *TSE1*	0.008	0.043**	0.020	−0.032**	−0.617	0.056***	0.024**
	(0.65)	(2.56)	(0.96)	(−2.41)	(−1.55)	(3.14)	(2.16)
7. *JASDAQ*	0.052***	0.034*	0.005	0.013	0.390	−0.014	0.003
	(4.17)	(1.93)	(0.20)	(0.94)	(0.94)	(−0.73)	(0.28)
8. *NJGAAP*	0.007	0.032	−0.072**	−0.034	1.161*	−0.022	−0.014
	(0.38)	(1.36)	(−2.21)	(−1.54)	(1.76)	(−0.86)	(−0.69)
9. *InLength*				−0.147***	−4.805***	0.411***	0.587***
				(−11.74)	(−12.46)	(17.18)	(39.58)
constant	6.420***	1.639***	−0.197	1.915***	91.110***	−2.509***	1.153***
	(81.25)	(13.99)	(−1.47)	(16.33)	(23.86)	(−12.76)	(8.51)
Observations	44,710	44,710	44,710	44,710	44,710	44,710	41,022
Adj R-squared	0.853	0.370	0.210	0.075	0.163	0.914	0.794
Year & firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

注）係数の有意水準の計算には頑健な標準誤差を用いている．***1％水準で有意．**5％水準で有意．*10％水準で有意。

きい企業ほど文量が多く具体的な表現を用いている一方でスティッキー表現が少ない。このセクションでは取締役に関する情報など具体的な表現が多く使われるが、規模が大きい企業ほど関連する情報量が多いことから、文字数および具体表現が多いと思われる。そして、取締役の移動等により開示内容が変化することからスティッキー表現が少なくなっているとも考えられる。

その他の要因については東証 1 部上場企業ほど CG セクションの可読性が低く、ボイラープレートおよびスティッキー表現が多くなるのは興味深い。

5.2.4　本節のまとめ

以上のように、企業が開示するテキスト情報にはいくつかの企業属性が関係することがわかった。たとえば、MD&A や RISK は業績や財政状況がテキスト情報の量やその質と関係している。特に、財政状況や業績の悪い企業ほど開示量が多く、業績の良い企業の可読性は高いという結果は、先行研究とも整合的である。さらに、業績の良い企業ほど MD&A の記述情報がポジティブになることも興

味深い。また、企業の規模は MD&A や RISK だけではなく CG とも関係している。その他、監査を受ける監査法人の大きさや企業の上場マーケット、そして会計基準などが企業の記述情報と関連しているようである。

6．結論と今後の課題

本研究は、テキストマイニング技術を用いて日本の上場企業の有価証券報告書における記述情報を長期的かつ包括的に分析した初めての研究であるといえる。具体的にいえば、本研究の特徴として、記述情報の開示要請がされた 2004 年から 2018 年までの上場企業全社を分析対象としていること、断片的な記述情報ではなく MD&A、リスク、ガバナンス情報を分析対象としていること、文量をはじめとして数字表現、固有表現、可読性といった複数の指標を計算、分析していることが挙げられる。以上から、本研究で用いられたテキストマイニング手法およびその分析結果は、学術的および実務的に重要な貢献を果たすものと考えられる。

本研究から得られた知見は以下の通りである。上場企業を対象に有価証券報告書の記述情報の実態を分析した結果、① 2004 年から 2018 年まで 15 年間の傾向を見ると、有価証券報告書の記述情報（MD&A、リスク情報、ガバナンス情報）は、内閣府令による企業情報の開示拡大を受けて全体的に増加傾向にあり、近年では特に MD&A とガバナンス情報において顕著な増加傾向が見て取れた。②数字表現については、2004 年時点において MD&A が最も使用頻度が高く、リスク情報とガバナンス情報では少なかった。その後、MD&A とガバナンス情報で数字表現の使用が増加する一方、リスク情報は微減傾向が観察された。③固有表現では、2004 年時点において 3 セクションともそれほど多くないものの、2009 年から 12 年にかけてガバナンス情報が、そして 2018 年に MD&A 情報が大幅な固有表現の増加を見せている。④可読性指標の分析結果から、2004 年時点においては 3 つのセクションともっとも難しい水準であることがわかった。その後、リスク情報は緩やかに低下、ガバナンス情報は増減しつつもほぼ横ばいであるのに対して、MD&A では一般して可読性スコアが上昇傾向にあることが観察された。

続いて、テキストマイニング手法を用いて算出した分析指標を用いて、企業の属性と記述情報の関連性を実証的に分析した結果、MD&A やリスク情報は業績や財政状況がテキスト情報の量やその質と関係していることがわかった。特に、財政状況や業績の悪い企業ほど開示量が多く、業績の良い企業の可読性が高いとの結果は、先行研究とも整合的である。さらに、業績の良い企業ほど MD&A の記述情報がポジティブになっていた。また、企業の規模は MD&A やリスク情報だけではなくガバナンス情報とも関係していることがわかった。その他、監査を受ける監査法人の大きさや企業の上場マーケット、そして会計基準なども企業の記述情報と関連していた。

本研究の課題は以下の通りである。分析の質は分析対象および分析手法の影響を受けるため、データベースおよび分析モデルの精緻化が求められる。また、本研究では記述情報と企業ファンダメンタルズとの関連性を分析しているが、記述情報の解釈方法や将来の企業および株価パフォーマンスとの関連性は分析できておらず、これらは今後の課題として残されている。

注

1 「財政状態及び経営成績の分析」は、名称変更が何度か行われており、2021 年現在では「経営者による財政状態、経営成績及びキャッシュ・フローの分析」と記載される。
2 米国における近年の研究動向については金（2015a）、金（2015b）および首藤（2019）を参照。またレビュー論文として Loughran and McDonald（2016）および Elshandidy et al.（2018）がある。
3 さらに、Fukukawa and Kim（2017）では日本企業のリスク情報の開示に監査人が影響している可能性を指摘している。
4 MeCab は、京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究で開発された形態素解析エンジンであり、研究、実務を問わず広く使用されている形態素解析システムである（https://taku910.github.io/mecab/）。
5 ipadic は、奈良先端科学技術大学院大学により公開されている形態素解析用の辞書であり、情報処理振興事業協会（IPA）で定義された IPA 品詞体系に基づいている。
6 UniDic は、国立国語研究所で規定された短単位と呼ばれる揺れのない斉一な単位で設計された形態素解析用の辞書である。
7 NEologd は、佐藤敏紀氏（東京工業大学大学院総合理工学研究科）によって開発が続けられている ipadic および UniDic の拡張辞書である（https://github.com/neologd/mecab-ipadic-neologd）。
8 KH Coder は形態素解析用のアプリケーションソフトである。KH Coder の概要と利用状況については樋口（2017）を参照されたい。
9 2018 年の MD&A はデータベースから正常な形で入手できない企業が多かったため、EDINET から XBRL 形式でダウンロードした。
10 HTML 形式のデータは、Python の html.parser で読み込まれ、p タグで囲まれた記述のうち、句点（。）で終わる文章を抽出した。ただし、2004 年から 2010 年は約半数の企業が p タグではなく、div タグや span タグを使用して文章を記述しているため、当該企業は分析対象から除外されている。
11 分析に必要なデータ（監査法人データを除く）は Astra Manager より入手し、監査法人関連の変数は、NEEDS 企業基本データ監査法人・監査意見データから入手した。
12 分析にあたり、本研究ではノートルダム大学のティム・ローラン教授とビル・マクドナルド教授により作成されたワードリスト（Loughran and McDonald Sentiment Word Lists [12]、以下 L&M ワードリスト）におけるポジティブワード（positive word）とネガティブワード（negative word）を日本語に翻訳したリストを用いて各企業のセンチメントを測定する。
13 なお、2007 年度までは中央青山に監査を受けている場合においても *BIG4* は 1 を取る。

参考文献

Bryan, S. H. (1997). Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review*, 72 (2), 285-301.

Campbell, J.L., H. Chen, D. S. Dhaliwal, H. Lu, and L. B. Steele. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19 (1), 396-455.

Dyer, T., M. Lang, and L. Stice-Lawrence. (2017). The evolution of 10-K textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64, 221-245.

Elshandidy, T., P. J. Shrives, M. Bamber, and S. Abraham. (2018) Risk reporting: A review of the literature and implications for future research. *Journal of Accounting Literature*, 40, 54-82.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. (2010) Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies,* 15(4), 915-953.

Fukukawa, H. and H. Kim. (2017). Effects of audit partners on clients' business risk disclosure. *Accounting and Business Research*, 47 (7), 780-809.

Healy, P. M. and K. G. Palepu. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1-3), 405-440.

Hope, O., D. Hu, and H. Lu. (2016). The benefits of specific risk-factor disclosures. *Review of Accounting Studies*. 21, 1005-1045.

Kallapur, S. and M. A. Trombley. (1999). The association between investment opportunity set proxies and realized growth. *Journal of Business Finance and Accounting*, 26 (3&4), 505-519.

Kim, H. and Y. Yasuda. (2018). Business risk disclosure and firm risk: Evidence from Japan. *Research in International Business and Finance*, 45, 413-426.

Lang, M. and R. Lundholm. (1993). Cross-sectional determinants of analyst ratings of corporate disclosures. *Journal of Accounting Research*, 31(2), 246-271.

Lev, B. and F. Gu. (2016). The End of Accounting and the Path Forward for Investors and Managers. Wiley.（バルーク・レブ、フェン・グー著、伊藤邦雄監訳『会計の再生』中央経済社 2018 年）

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45, 221-247.

Loughran, T. and B. Mcdonald. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69 (4), 1643-1671.

Loughran, T. and B. Mcdonald. (2016). Textual analysis in accounting and finance : A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Miller, G. S. (2002). Earnings performance and discretionary disclosure. *Journal of Accounting Research*, 40 (1), 173-204.

Verrehia, R. (1983). Discretionary disclosure, *Jarnal of Accounting and Economics,* 5(3), 179-194.

Türegün, N. (2019). Text mining in financial information. *Current Analysis on Economics & Finance*, 1, 18-26.

大谷潤・上利悟史・堀内隼・岡村健史（2018）「企業内容等の開示に関する内閣 府令等の改正」週刊経営財務、第 3351 号、2018 年 3 月 19 日.

金鉉玉（2015a）「定性情報に焦点を当てた研究動向（1）」企業会計、67（1）6-7.

金鉉玉（2015b）「定性情報に焦点を当てた研究動向（2）」企業会計、67（2）6-7.

金融庁（2019）「記述情報の開示に関する原則」2019 年 3 月.

工藤拓（2018）『形態素解析の理論と実装』近代科学社、2018 年.

財務会計基準機構（2018）「有価証券報告書の開示に関する事項―『一体的開示をより行いやすくするための環境整備に向けた対応について』を踏まえた取り組み」、2018 年 3 月.

首藤昭信（2019）「テキスト分析と会計学研究」『情報センサー』、143、8-10.

首藤昭信・緒方英明（2009）「実務研究 有価証券報告書における「財政状態及び経営成績の分析（MD&A）」について」『研究所レポート』プロネクサス総合研究所、3.

須田一幸（2004）「ディスクロージャー・レベルの決定要因」『ディスクロージャー戦略と効果』森山書店、107-122.

中野貴之（2010）「財務諸表外情報の開示実態―事業等のリスクおよび MD&A の分析―」『財務諸表外情報の開示と保証―ナラティブ・リポーティングの保証―』同文舘出版.

野田健太郎（2016）「有価証券報告書における定性情報の分析と活用―リスクの多様化にともなう望ましい対話のあり方―」『経済経営研究』37（1）、1-51.

樋口耕一（2017）「計量テキスト分析および KH Coder の利用状況と展望」『社会学評論』68（3）、334-350.

矢澤憲一（2020）「テキストマイニングを用いた会計、監査、ガバナンス研究の新たな潮流、そして二〇三〇年の監査研究」『會計』197（3）、297-308.

李在鎬（2016）「日本語教育のための文章難易度に関する研究」早稲田日本語教育学第 21 号、1-16.

本研究は JSPS 科研費（課題番号 JP20H01561）の助成を受けたものです。