高崎経済大学論集　第59巻　第２・３・４合併号　2017　61～73頁

公会計情報の言語的特徴：東京、ニューヨーク、ロンドンを対象とした分析

廣　瀬　喜　貴

Linguistic features of public accounting information: Analysis for Tokyo, New York, and London

Hirose Yoshitaka

Abstract

This paper clarifies the linguistic features of public accounting information disclosed by certain cities. The paper focuses on the reports disclosed by large cities（Tokyo, New York, and London）with many stakeholders, and analyzes the linguistic features of the disclosure content using the textual analysis. The results of the analysis clarified the following: 1）Traditional accounting terminology is a term that represents the reports containing public accounting information of cities; 2）On comparing important terms between cities, explanation terms emphasized by cities seem to differ; and 3）Readability of urban reports varies, but on average, it is about graduation degree of undergraduate university. The results indicate that 1）Even if it performs accrual-based accounting, it still faithfully expresses the character of the public accounting report based on cash-based accounting, 2）The fact that Tokyo focuses on assets and costs rather than other cities and does not explain fair value at all is helpful for Japanese practice, and 3）London's report is closest to corporate accounting and a benchmark for readability of public accounting information. The contribution of this paper is two-fold: it is the first study to clarify the actual situation of public accounting information of major cities, and their implications for Japanese practice.

1．はじめに：問題の所在と先行研究

本稿の目的は、都市[1]が開示している公会計情報の言語的特徴を明らかにすることである。

1　都市という表現は各国で異なっており、どの程度の大きさを都市と呼ぶのかが問題となるが、本稿では都市経済学及び都市工学で定義されているものとほぼ同義で都市という用語を使用している。具体的には、総務省統計局の定義に従い、各国の首都及び人口100万人以上（インドと中国は上位20位以上）を都市と呼んでいる。詳細は総務省統計局（2016）; United Nations（2015）を参照。

市、州、都道府県等、公会計を担う主体は、利害関係者に対しアカウンタビリティーを果たすため多くの公会計情報を開示している。公会計の分野では、欧州では2000年以降に実証分析（archival / empirical）の手法を用いた研究が行われている（Anessi-Pessina and Steccolini, 2007; Steccolini, 2004）。これらの先行研究は、財務諸表の財務データや開示されている情報内容（コンテンツ）を分析したものである。しかし、都市が公表している報告書は財務諸表のみで構成されるわけではなく、言語情報も多く含まれている。また、開示されている報告書の中で財務諸表が大部分を占めている場合であっても、勘定科目、注記、補足説明等は言語情報としての価値を有している。会計の分野では、財務会計や先端技術（Emerging technologies in accounting）、情報システム（Accounting information systems）等の学術雑誌で、テキストマイニングの手法を用いた研究が数多く行われている（Li, 2008, 2010; Bozanic, Z. and Thevenot, M., 2015; Muehlmann, Chiu, and Steve, 2015）が、公会計の分野では、言語情報の分析はこれまでなされてこなかったという背景がある。しかし、世界的な会計研究の文脈において財務情報とともに言語情報の重要性が注目されている現状を鑑みると、公会計情報の言語的特徴を明らかにすることは、公会計研究及び実務にとって重要な意義があるといえるだろう。

そこで、本稿では、他都市よりも相対的に利害関係者が多い大都市（東京、ニューヨーク、ロンドン）が開示している報告書に焦点を当て、開示内容がどのような言語的特徴を有しているのかを、テキストマイニングの手法を用いて探索した。

形態素解析、*TF-IDF*, *DP*, *N-gram*, *Fog index*, *TTR*の測定等の分析の結果、1）資産、収益、キャッシュ等の伝統的な会計の専門用語が、都市の公会計情報の報告書を代表する用語であること、2）重要な用語の都市間比較によると、都市によって重視している説明項目が異なるということ、3）都市の報告書の難易度は、小学6年程度から大学学部卒業を大幅に超える程度まで、かなりのばらつきがあるが、平均的には概ね大学学部卒業程度であること、という3点が明らかになった。

これらの結果は、1）発生主義会計を取り入れつつも現金主義会計を基礎とする公会計の報告書の性格を忠実に表していること、2）東京が他都市よりも資産と費用を重視している点や東京のみが公正価値に全く触れていないという点は、新公会計基準の導入を控えた日本の実務にとって参考になること、3）可読性が3都市の平均値に最も近く、企業会計の先行研究の結果に最も近いのはロンドンであることから、ロンドンの報告書は公会計情報の可読性のベンチマークとなること、を示唆している。

本稿の貢献は、1）主要都市の公会計情報の言語的特徴を明らかにすることで、日本の統一的な公会計基準で作成される財務情報の公表時の参考情報を提供することができるという実務へのインプリケーションと、2）主要都市の公会計情報の実態を解明した最初の研究であるという学術的貢献、の2点である。

本稿の構成は次のとおりである。まず2節では、サンプルの選択方法と報告書の基礎情報について記述する。次に3節で分析結果を示し、その解釈を検討する。最後に4節で結論と今後の検討課

題を論じる。

2．サンプルの選択と報告書の基礎情報

本稿では、市民をはじめとする各種利害関係者が多いと予想される大都市の報告書を分析対象とした[2]。具体的なサンプルの選択手順は次のとおりである。まず、総務省統計局が Web上で公表している「世界の統計2016」から、世界の主要都市の人口情報を取得し、人口ランキング上位20都市を確認した。次に、各都市の公式 Website にアクセスし、英語で情報公開がなされているか否かを確認した。そして、英語の情報が公開されている場合、サイトマップとサイト内検索があるか否かを確認した。サイトマップがある場合は、都市の概要（Overview）、情報公開（IR; Information）、経済（economy）等、公会計情報に関する報告書が公表されている可能性が高いページにアクセスした。サイト内検索がある都市は、併せて “annual report”, “financial statement”, “year book”, “annual”, “report”, “financial”, “statement”, “IR”, “budget”, “accounting” を検索した。このような手順でサンプルを抽出した結果、東京、ニューヨーク、ロンドンの３都市の報告書が分析対象となった[3]。これらの情報を示したものが図表１である。

図表１　主要都市人口の上位20と財務諸表の開示の有無

	都市	人口	開示の有無	公式 web site のURL
1	北京（ペキン）	19,610	-	http://www.beijing.gov.cn/
2	上海（シャンハイ）	14,349	-	http://www.shanghai.gov.cn/
3	イスタンブール	14,160	-	http://istanbul.gov.tr/tr
4	ブエノスアイレス	13,339	-	http://www.buenosaires.gob.ar/
5	ムンバイ	11,978	-	http://www.mcgm.gov.in/
6	モスクワ	11,918	-	https://www.mos.ru/en/
7	サンパウロ	11,153	-	http://www.prefeitura.sp.gov.br/
8	ソウル	10,008	-	http://www.seoul.go.kr/
9	デリー	9,879	-	http://delhi.gov.in/
10	リマ	9,736	-	http://www.munlima.gob.pe/
11	重慶（チョンチン）	9,692	-	http://www.cq.gov.cn/
12	ジャカルタ	9,608	-	http://www.jakarta.go.id/
13	カラチ	9,339	-	http://www.kmc.gos.pk/
14	東京都	8,946	✓	http://www.metro.tokyo.jp/
15	メキシコシティ	8,875	-	http://www.cdmx.gob.mx/
16	広州（クワンチョウ）	8,525	-	http://www.gz.gov.cn/
17	ニューヨーク	8,337	✓	http://www1.nyc.gov/
18	武漢（ウーハン）	8,313	-	http://english.wh.gov.cn/
19	バンコク	8,305	-	http://www.bangkok.go.th/
20	ロンドン	8,278	✓	http://www.london.gov.uk/

人口単位は1,000人。都市名及び人口は総務省統計局（2016）; United Nations（2015）を参照した。開示の有無の列の「✓」は、英語で記述された財務諸表を一組の電子媒体（pdfファイル等）の形式で開示している都市である。また、「-」は筆者が公式 Website 上から英語で記述された財務諸表を発見することができなかった都市である。なお、英語以外の財務諸表を開示している都市、予算（現金主義会計）の開示のみを行っている都市、決算をhtml上で簡易的に開示している都市も「-」に含まれている。URLの列に記載した都市の公式 Website の閲覧日は、いずれも2016年12月１日である。

2　市民以外の利害関係者は、公債に利害関係を持つ投資家も想定されるが、本稿では、市民が報告書を読むという前提に立脚して分析を行なっている。東京は東京都会計基準、ニューヨークはGovernmental Accounting Standards Board（GASB）のGenerally Accepted Accounting Principles（GAAP）、ロンドンは2014/15 Code of Practice on Local Authority Accounting および International Financial Reporting Standards（IFRS）にもとづいて作成されており、どの利害関係者を最も意識しているのかは都市によって異なっている。ニューヨークは、Popular Annual Financial Report（PAFR）という概要版の書類の開示も行っていることから、こちらの概要版が市民向けの開示なのかもしれない。

3　本稿では、注１の都市の定義に従って分析対象を特定しているが、東京は都道府県であり、ニューヨークとロンドンは市であることから、分析単位が揃っているのかという懸念は残る。この点については今後の検討課題である。

テキストデータの取得年度については、都市によって決算月が異なるため、本稿では、2015年の1月から12月に決算期を迎える期の報告書を分析対象とした。これらの手順を経て、分析対象となった報告書の一覧を示したものが図表2である。

図表2　分析対象となった報告書の一覧

都市	決算期	報告書の名称
東京	2015年3月期	FISCAL YEAR 2014 REFERENCE MATERIAL TO ANNUAL ACCOUNTS OF TOKYO METROPOLITAN GOVERNMENT FINANCIAL STATEMENTS
ニューヨーク	2015年6月期	THE CITY OF NEW YORK NEW YORK COMPREHENSIVE ANNUAL FINANCIAL REPORT OF THE COMPTROLLER FOR THE FISCAL YEAR ENDED JUNE 30, 2015
ロンドン	2015年3月期	GREATER LONDON AUTHORITY Statement of Accounts 2014/15 AUDITED

都市によって決算月が異なるため、本稿では、2015年の1月から12月に決算期を迎える期の報告書を分析対象とした。ニューヨークはPopular Annual Financial Report（PAFR）という概要版の書類の開示も行っている。

分析対象となった3都市の 報告書の基礎データを得るため、ページ数、ページサイズ（cm）、ファイルサイズ（bytes）を確認した。次に、Rを使用し、単語数（Word）、単語割合、数字の数、数字割合、不明文字数、不明文字割合、略語数、コンマの数、ドットの数、その他の句読点の数を測定した[4]。これらの基礎データを示したものが図表3である。

図表3　報告書の基礎データ

	東京	ニューヨーク	ロンドン
ページ数	128	422	140
ページサイズ（cm）	21×29.7	21.59×27.94	21.01×29.71
ファイルサイズ（bytes）	4,538,987	7,595,043	1,937,418
単語数（Word）	29,293	100,799	41,586
単語割合	53%	56%	79%
数字の数	24,000	74,855	9,726
数字割合	44%	42%	19%
不明文字数	1,489	4,517	1,084
不明文字割合	3%	2%	2%
略語数	24	172	28
コンマの数	10,686	39,448	4,091
ドットの数	1,038	67,220	1,630
その他の句読点の数	2,461	199,481	6,621

不明文字とは、フォントが正しく埋め込まれていない、文字化け等の理由により、認識することができなかった文字のことである。

報告書のページ数は、東京の128ページ、ロンドンの140ページに対し、ニューヨークは422ページであり、ニューヨークの報告書の情報量が多いことがわかる。また、東京のファイルサイズは約4.5MBであり、ロンドンの約1.9MBの2倍以上のサイズであった。しかし、単語数は、東京よりもロンドンの方が多く、同様に単語割合も東京よりもロンドンの方が高かった。企業を対象とした先行研究であるLoughran and McDonald（2014）は、企業が積極的に情報開示をしているというトー

4　Rとは、R言語および統計解析ソフトのことである。本稿の分析は、Mac OS Xにて、R version 3.3.2を使用した。

ンの代理変数としてファイルサイズを用いている。しかし、公会計の文脈においては、ファイルサイズは積極的開示のトーンの代理変数を表していないのかもしれない。また、東京は単語割合が53%、数字割合が44%であり、ニューヨークは単語割合が56%、数字割合が42%であり、単語と数字の割合がほぼ同様である。しかし、ロンドンは、単語割合が79%、数字割合が19%であり、東京とニューヨークとは異なる特徴を持っているということがわかった。

3. 分析結果とその解釈

3.1. 単語の出現頻度の測定と重要単語の特定

本節では、計算言語学の分野で開発されてきた一般的な手法（Gries, 2009; 小林、2015）、およびそれらを適用した会計の先行研究で用いられている手法（Loughran and Mcdonald, 2016）を用いて分析を行なった結果を示し、その解釈を論じる[5]。なお、分析にあたって、Mac OS X Version 10.11.6, R version 3.3.2, TreeTagger 3.2を使用した（Schmid, 1994; 1995）。

まず、文章を品詞ごとに分解し、名詞のみを抽出し、出現回数を測定するという形態素解析を行なった。各都市の名詞の出現回数をワードクラウドの形で表したものが図表４である。

図表４　各都市の名詞のワードクラウド

（A）東京

（B）ニューヨーク

（C）ロンドン

（D）合計

5　本稿では、pdfファイルから抽出したテキストデータをtxtファイルに格納し分析した。また、その分析対象のファイルに対し前処理を行った。テキストデータの前処理には、目的に応じて様々な方法が提唱されている（Gries, 2009; 小林, 2015）。前処理は、1）大文字を小文字へ変換、2）単語の前後のスペースを削除、3）句読点の削除、4）数字の削除、という処理を行った。

これは、出現回数が多い名詞を大きいフォントで表示し、出現回数が少ない名詞を小さいフォントで表示したものである。名詞の出現回数を単純にカウントすることは、言語処理の分野では一般的な手法であるが、さらに文書の持っている特徴を浮き彫りにするために、様々な重み付けの手法が提唱されている（Manning, Raghavan, and Schütze, 2008; Gries, 2009）。

そこで、これらの名詞の中から重要な用語である特徴語を抽出するために、この全名詞の出現頻度（Term frequency: *TF*）に、何個の文書に当該名詞が出現したのかを示す文書頻度（Document frequency: *DF*）の逆数（Inverse）を掛けて、*TF-IDF*（Term frequency inverse document frequency）を測定した。*TF-IDF*の計算方法は、Shannon（1948a; 1948b）においてその着想が示されて以来、様々な計算方法が提唱されており（Manning et al., 2008）、会計におけるテキストマイニング研究においても一般的な分析手法として確立されている（Loughran and McDonald, 2016）。本稿では、Dumais（1992）; Nakov, Popova, and Mateev（2001）で示されている（1）式を用いて全名詞の*TF-IDF*を測定した。*tf*は、ある単語の出現回数を全単語数で除して求めた割合（%）であり、ある文書における当該単語の出現頻度を表している。また、*ndocs*は総文書数を、*df*は当該名詞を含む文書数を表している。

$$TF-IDF=\log(tf+1)*\log(ndocs/df)+1 \quad (1)$$

各都市および全都市合計の*TF-IDF*の上位20語を示したものが図表5である。

図表5　*TF-IDF* の上位20

	東京	ニューヨーク	ロンドン	合計
1	assets	city	authority	assets
2	activities	services	london	accounts
3	accounts	year	income	revenue
4	cash	funds	statement	capital
5	transfer	fund	expenditure	fund
6	period	york	group	cash
7	revenue	new	gla	year
8	account	june	accounts	activities
9	expenses	personal	year	liabilities
10	liabilities	department	capital	statement
11	investment	comptroller	assets	investment
12	general	debt	greater	account
13	yen	community	march	general
14	subsidies	pension	fund	property
15	capital	budget	value	balance
16	surplus	total	balance	amount
17	property	board	rates	payments
18	fund	expenditures	account	period
19	payments	capital	audited	expenses
20	flows	tax	comprehensive	interest

Dumais (1992) ; Nakov, Popova, and Mateev (2001) にもとづき*TF-IDF*を測定した。

各都市の上位の単語は、出現回数と概ね似た傾向を持っている。各都市の上位20語のうち、東京の補助金（subsidies）、ニューヨークの年金（pension）、税（tax）、ロンドンの監査対象

（audited）、という単語は、他の都市にはみられない特徴語となっており、それぞれの都市が何を重視しているのかがわかる。

また、合計の列に着目すると、会計分野の専門用語が上位の特徴語になっているということがわかる。この結果は、報告書のテキストデータの全体傾向を表している。近年、CSR（Corporatesocial responsibility）情報やe-governmentに関する情報の開示等、公会計のディスクロージャーも企業会計と同様に多角化しているが（Ball, 2005; Pina, 2010）、依然として伝統的な会計の専門用語が報告書を代表する単語であるという事実が明らかになった。

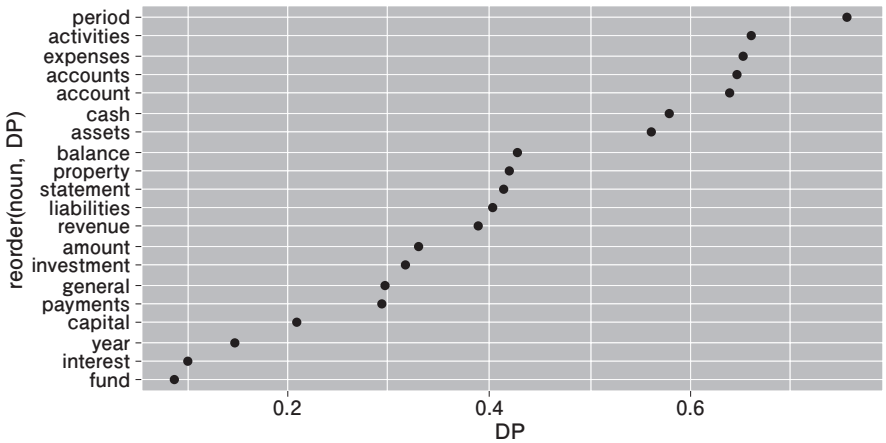
3.2. 重要単語の都市間比較

次に、これらの重要単語の出現頻度が、どの程度偏っているのかを明らかにするために、重要単語の割合の偏差（Deviation of Proportion: *DP*）を計算した 。*DP*の計算については、Gries（2008）; Lijffijt and Gries（2012）で提示されている次の３ステップに従った。まず、各都市の報告書の総単語数を全都市の総単語数で除して期待割合を求めた。次に、重要単語の各都市の実際の観測頻度を当該単語の全都市の合計頻度で除して観測割合を求めた。そして、期待割合と観測割合の差の絶対値を足し２で割ることによって*DP*を計算し、*DP*を次の（２）式で正規化し、０から１の数値で表した。*DPnoun* は正規化された*DP*、*min*（*s*）は最も小さいサブコーパス（都市のテキストデータ）の期待割合のことである。

$$DPnoun = DP / 1 - min(s) \qquad (2)$$

正規化された*DP*は、０から１の値を取り、０は観測割合が期待割合に近いことを、１は観測割合が期待割合から乖離していることを表している。観測割合が期待割合から乖離している程度が大きい単語を上に、小さい単語が下になるように並べたものが図表６である。

図表６　重要単語上位20語の標準の偏差 (Deviation of Proportion: *DP*)



観測割合が期待割合から乖離している程度が大きい単語を上に、小さい単語が下になるように並べた。
Gries（2008）; Lijffijt and Gries（2012）にもとづき計算した。

図表6. から、ファンド（fund）、利息（interest）、年（year）等の単語は、ほとんど偏りなく散らばっているのに対し、費用（expenses）、会計（account）、キャッシュ（cash）、資産（assets）等の単語は、偏って散らばっていることが明らかになった。とりわけ、財務諸表の構成要素である資産（assets）、負債（liabilities）、収益（revenue）、費用（expenses）という単語は、DPが中程度あるいは高程度であることがわかった。

そこで、これらの重要単語のうち、財務諸表の構成要素である資産、負債、収益、費用が、どの都市に多く／少なく出現しているのかを探索するために、それらの単語の出現頻度を集計したクロス表についてχ^2検定を行った結果、χ^2=101.349, df=6, p<0.01 で有意差が認められた。標準化残差を計算し、調整済み残差を計算するという残差分析を行なった結果を示したものが図表７である。

図表７　財務諸表の構成要素の残差分析の結果: 調整済み残差

	東京	ニューヨーク	ロンドン
assets	4.484***	-5.791***	0.964
liabilities	-4.571***	4.236***	1.086
revenue	-5.840***	4.254***	2.825***
expenses	5.613***	-1.554	-5.860***

***p<0.01

残差分析の結果、東京は、借方項目（資産と費用）の出現回数が多く、貸方項目（負債と収益）の出現回数が少なかった。ニューヨークは、資産の出現回数が少なく、負債と収益の出現回数が多かった。ロンドンは、収益の出現回数が多く、費用の出現回数が少なかった。すなわち、東京は資産と費用について、ニューヨークは負債と収益について、ロンドンは収益について、多くの説明を割いていることがわかった。また、東京は、負債と収益について、ニューヨークは資産について、ロンドンは費用について、説明が少ないということがわかった。このことは、都市によって、財務諸表のどの構成要素を重視して説明しているのかが異なっていることを意味している。新公会計制度が導入される日本においては、固定資産台帳の作成およびそれに伴う減価償却費の計上という資産と費用の管理が改革の柱となっており（総務省、2015）、以前から公会計情報の開示に対する意識が高く、日本の模範ともいえる東京が、資産と費用についてより多くの説明を割いていることは、直感とも整合的な結果といえるだろう。

3.3. *N-gram*

前項までは、１つの単語の分析を行なったが、本項では複数単語の組み合わせである*N-gram*を測定する。*N-gram*とは、ある文字列の中に*N*個の単語の組み合わせが出現する回数のことである（Shannon, 1948a; 1948b）。全文書における2-grams上位20を示したものが図表８である。

まず、財務諸表の構成要素である純資産（net assets）は133回出現し19位、包括利益（comprehensive income）は126回出現し20位であった。これらの用語は、前項までの分析では明らかにすることができなかった重要用語であり、ここに*N-gram*を測定する意義がある。また、前

項の分析では、資産（assets）が最も重要な単語であることがわかったが、2-grams を測定すると390回出現した固定資産（fixed assets）が３位であった。流動資産（current assets）は89回、無形資産（intangible assets）は20回しか出現していないことから、資産の中でも、特に固定資産を多く説明しているということが明らかになった。このことは、都市にとっての固定資産の管理の重要性を物語っている。また、固定資産が計上されている貸借対照表（balance sheet）は151回出現し17位となっている。キャッシュフロー計算書に関連する用語である、キャッシュフロー（cash flows）は219回出現し10位、財務活動（financing activities）は180回出現し13位、営業活動（operating activities）は170回出現し15位、投資活動（investing activities）は163回出現し16位であった。この結果は、キャッシュ（cash）や予算（budget）が重要単語であるという前項の結果と併せて解釈すると、現金主義会計で予算を統制している公会計の文脈において、キャッシュが重要な説明事項として捉えられていることを表している。また、18位にランクインした公正価値（fair value）は141回出現しているが、取得原価を意味するhistorical cost は７回、acquisition costs は３回であり、取得原価よりも公正価値という用語が多く出現しているということが明らかになった。なお、東京は公正価値という用語が１度も出現していなかった。日本の新公会計基準は、原則として資産を取得原価で評価することを要求していることから、この結果は注目に値する。今後、公会計における公正価値と取得原価の意義について、さらなる国際間比較の分析を行う必要があるだろう。

図表８　全報告書における2-gramsの出現回数の上位20

	2-grams	出現回数（回）
1	fiscal year	668
2	personal services	460
3	fixed assets	390
4	financial statements	336
5	net position	269
6	other accounts	268
7	community board	267
8	general fund	263
9	general account	237
10	cash flows	219
11	capital investment	218
12	long term	181
13	financing activities	180
14	debt service	177
15	operating activities	170
16	investing activities	163
17	balance sheet	151
18	fair value	141
19	net assets	133
20	comprehensive income	126

固有名詞とノイズを削除している。

3.4. 可読性と語彙の豊富さ

本項では、報告書の可読性（readability）と語彙の豊富さ（lexical diversity）を測定する。財務会計、先端技術（Emerging technologies in accounting）、情報システム（Accounting information systems）等の学術雑誌では、テキストマイニングの手法を用いた可読性（readability）の研究が積極的に行われている（Li, 2008, 2010; Bozanic, Z. and Thevenot, M., 2015; Muehlmann, Chiu, and Steve, 2015）。また、計算言語学、自然言語処理、計量国語学の分野では、語彙の豊富さを測定することも一般的である（Gries, 2009; 小林, 2015）。計算言語学の先行研究では、可読性と語彙の豊富さ

について様々な計算方法が提唱されているが、本稿では、会計の先行研究で一般的に用いられている可読性指標である*Fog index*と語彙の豊富さの指標である*Type-token ratio*（*TTR*）を計算した。

*Fog index*は、1文の平均単語数と全語数に対する3音節以上の単語の割合を用いて計算する可読性の指標であり、6から17の数値が求められる。6が小学6年程度の難易度、17が大学学部卒業程度の難易度を表している。*Fog index*は（3）式で計算することができる。

Fogindex＝0.4＊（1文の平均単語数+100＊全語数に対する3音節以上の単語の割合）　（3）

また、*TTR*は、異なり語数（types）を延べ語数（tokens）で割ったものである。異なり語数とは、あるテキストの中で、同一の単語が何度用いられていてもこれを一語とし、全体で異なる単語がいくつあるかをかぞえた数である（大辞泉第3版、2006）。*TTR*は（4）式で計算することができる。

TTR＝異なり語数／延べ語数　（4）

各都市の報告書の*Fog index*と*TTR*を示したものが図表9である[6]。

図表9　可読性と語彙の豊富さ

	東京	ニューヨーク	ロンドン	平均値
Fog index	27.67	6.91	19.93	18.2
TTR	0.03	0.04	0.08	0.05

分析の結果、東京は大学学部卒業レベルを大幅に超える難易度（27.67）、ニューヨークは中学2年程度の難易度（6.91）、ロンドンは大学学部卒業を上回る程度の難易度（19.93）、3都市の平均値は大学学部卒業を上回る程度の難易度（18.2）であることがわかった。ニューヨーク以外の*Fog index*は、大学学部卒業程度とされている最高難易度である17を上回っており、モデルが想定している範囲外の難易度である。もっとも関連する先行研究の分析結果も、モデルの範囲外の難易度であることが一般的である。たとえば、財務会計の先行研究であるLi（2008）では、アメリカの1994年から2004年までの55,719企業年のアニュアルレポートの*Fog index*は19.39であると報告されている。また、Lang and Stice-Lawrence（2015）では、世界の1998年から2011年までの85,793企業年のアニュアルレポートの*Fog index*は19.520であると報告されている。これらの企業会計の先行研究の結果に最も近いのはロンドンの報告書の難易度であることが明らかになった。なお、東京の難易度が極端に高い理由については、東京の報告書は他都市と比較して報告書全体における財務諸表の割合が高いことが考えられる。

また、*TTR*は、それぞれ、東京が0.03、ニューヨークが0.04、ロンドンが0.08であることがわかった。会計の分野における語彙の豊富さの先行研究では、Bozanic and Thevenot（2015）で

6　分析の前処理にて、テキストの本文以外の部分を削除した方が正確な値が得られるが、本稿ではその処理は行なっていない。その理由は、pdfファイルを分析対象としており、本文の抽出が技術的に困難であるからである。もっとも、可読性に関する企業会計分野の先行研究も同様の問題を抱えている。今後、iXBRLが普及すれば、本文のみを抽出し分析を行うことが技術的に可能になるだろう。

は、2004年から2012年までの1,837サンプルの利益のプレスリリース（earnings press releases）の*TTR*は、0.346であると報告されている。また、Humpherys, Moffitt, Burns, Burgoon, and Felix（2011）では、1995年から2004年の101サンプルのアニュアルレポートのMD&Aセクションの*TTR*は、不正をしていない企業は0.250であり、不正をしている企業は0.202であると報告されている。公会計の報告書には、財務諸表も含まれていることから、文章のみで構成されているプレスリリースやMD&Aセクションよりも語彙が豊富ではないという結果は直感的である。

これらの結果は、可読性と語彙の豊富さに関連する先行研究の知見を拡張しており、その点は本稿の貢献である。

4. おわりに：結論と今後の検討課題

本稿では、東京、ニューヨーク、ロンドンが開示している報告書に焦点を当て、開示内容がどのような言語的特徴を有しているのかを、テキストマイニングの手法を用いて探索した。

形態素解析、*TF-IDF*, *DP*, *N-gram*, *Fog index*, *TTR*の測定等の分析の結果、本稿で得られた結論は次の３点である。まず第１に、資産、収益、キャッシュ等の伝統的な会計の専門用語が、都市の公会計情報の報告書を代表する用語であるという事実が明らかになった。とりわけ、キャッシュや予算を重視しているという点は、発生主義会計を取り入れつつも現金主義会計を基礎とする公会計の報告書の性格を忠実に表している。次に第２に、重要な用語の都市間比較によると、都市によって重視している説明項目が異なるということが明らかになった。とりわけ、東京が他都市よりも資産と費用を重視している点、東京のみが公正価値に全く触れていないという点は、新公会計基準の導入を2016年に控えた日本の実務にとって参考になるといえるだろう。そして第３に、 都市の報告書の難易度は、小学６年程度から大学学部卒業を大幅に超える程度まで、かなりのばらつきがあるが、平均的には概ね大学学部卒業程度であることが明らかになった。可読性が３都市の平均値に最も近く、企業会計の先行研究の結果にも最も近いのはロンドンであることから、ロンドンの報告書は公会計情報の可読性のベンチマークとなるだろう。

本稿の貢献は、１）主要都市の公会計情報の言語的特徴を明らかにすることで、日本の統一的な公会計基準で作成される財務情報の公表時の参考情報を提供することができるという実務へのインプリケーションと、２）主要都市の公会計情報の実態を解明した最初の研究であるという学術的貢献、の２点である。

なお、本稿の限界として、サンプルである３都市の全てもしくはいずれかが特殊なケースであるという可能性は排除できない。また、単年度のみのデータの重要な用語のみを分析しているという点も限界である。したがって、より多くの都市の報告書の入手、単年度のみではなく時系列での比較、より多くの単語の分析を行うことが、今後の検討課題である。

情報技術の発展とともにe-governmentが進展し、Web上に都市の財務情報がアーカイブされる

ようになっているが（Caba Pérez, Pedro Rodríguez Bolívar, and López Hernández, 2008）、本稿で分析したような言語情報もWeb上に徐々にアーカイブされはじめている。したがって、公会計分野においても企業会計分野のように財務情報と言語情報を総合的に分析することが可能になっているといえるだろう。また、財務会計や計量国語学の分野では可読性の実験研究が行われていることから（Tan, Wang, and Zhou, 2015; 柴崎・原、2010; 柴崎・玉岡、2010）、今後、公会計分野においても言語情報に関する実験研究が進展する可能性が期待できる。

（ひろせ　よしたか　本学経済学部非常勤講師・高崎商科大学短期大学部講師）

参考文献

Anessi-Pessina, E., and Steccolini, I.（2007）. Effects of budgetary and accruals accounting coexistence: evidence from Italian local governments. *Financial Accountability & Management,* 23（2）, 113-131.

Ball, A.（2005）. Environmental accounting and change in UK local government, *Accounting, Auditing & Accountability Journal,* 18（3）, 346 - 373.

Bozanic, Z. and Thevenot, M.（2015）. Qualitative Disclosure and Changes in Sell-Side Financial Analysts' Information Environment. *Contemporary Accounting Research,* 32（4）, 1595-1616.

Caba Pérez, C., Pedro Rodríguez Bolívar, M., and López Hernández, A. M.（2008）. e-Government process and incentives for online public financial information. *Online Information Review,* 32（3）, 379-400.

Dumais, S.（1992）. Enhancing Performance in Latent Semantic Indexing（LSI）Retrieval. *Technical Report, Bellcore*.

Gries, S. Th.（2008）. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics,* 13（4）, 403-437.

Gries, S. Th.（2009）. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F.（2011）. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems,* 50（3）, 585-594.

小林雄一郎.（2015）.「R による英文テキスト解析」『東洋大学社会学部紀要』、53（1）、51-64.

Lang, M., and Stice-Lawrence, L.（2015）. Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics,* 60（2）, 110-135.

Li, F.（2008）. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics,* 45（2）, 221-247.

Li, F.（2010）. Textual analysis of corporate disclosures: A survey of the literature. *Journal of accounting literature,* 29, 143-165.

Lijffijt, J., and Gries, S. Th.（2012）. Correction to "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics,* 17（1）. 147-149.

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

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

Manning, C. D., Raghavan, P., and Schütze, H.（2008）. Scoring, term weighting and the vector space model. *Introduction to Information Retrieval,* 109-133.

Muehlmann, B. W., Chiu, V., and Liu, Q.（2015）. Emerging Technologies Research in Accounting: JETA's First Decade. *Journal of Emerging Technologies in Accounting,* 12（1）, 17-50.

Nakov, P., Popova, A., and Mateev, P.（2001）. Weight functions impact on LSA performance. *Proceedings of the Recent Advances in Natural language processing,* Bulgaria, 187-193.

Pina, V., Torres, L., and Royo, S.（2010）. Is e-government leading to more accountable and transparent local governments? An overall view. *Financial Accountability & Management,* 26（1）, 3-20.

Schmid, H.（1994）. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK.

Schmid, H.（1995）. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of*

the ACL SIGDAT-Workshop. Dublin, Ireland.
Shannon, C. E.（1948a）. A Mathematical Theory of Communication. *Bell System Technical Journal*. 27（3）, 379–423.
Shannon, C. E.（1948b）. A Mathematical Theory of Communication. *Bell System Technical Journal*. 27（4）, 623–666.
柴崎秀子・原信一郎(2010)「12 学年を難易尺度とする日本語リーダビリティー判定式」『計量国語学』第27巻第6号、215-232頁。
柴崎秀子・玉岡賀津雄（2010）「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」『日本教育工学会論文誌』第33巻第4号、449-458頁。
総務省（2015）「統一的な基準による地方公会計マニュアル」
総務省統計局（2016）「世界の統計2016」、http://www.stat.go.jp/data/sekai/0116.htm（閲覧日: 2016年12月11日）
Steccolini, I.（2004）. Is the annual report an accountability medium? An empirical investigation into Italian local governments. *Financial Accountability & Management,* 20（3）, 327-350.
Tan, H. T., Wang, E. Y., and Zhou, B.（2014）. How does readability influence investors' judgments? Consistency of benchmark performance matters. *The Accounting Review,* 90（1）, 371-393.
東京都（2015）「東京都会計基準」
United Nations（2015）*Demographic Yearbook 2014*.