Title	異なる文章ジャンルの判別可能性に関する調査 : ブログ本文、新聞社説、文学作品、論文を対象として
Sub Title	
Author	村田, 年(Murata, Minori) Lossa, Roman
Publisher	慶應義塾大学日本語・日本文化教育センター
Publication year	2014
Jtitle	日本語と日本語教育 No.42 (2014. 3) ,p.125- 135
JaLC DOI	
Abstract	
Notes	調査報告
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00189695-20140300-0125

慶應義塾大学　日本語・日本文化教育センター紀要
『日本語と日本語教育』42：125–135（2014.3）

異なる文章ジャンルの判別可能性に関する調査

―ブログ本文、新聞社説、文学作品、論文を対象として―

村田　年、ロマン・ロッサ

1. はじめに

異なるジャンルの文章を言語指標によって判別できるかどうかということについては、村田（2012）でまとめたように、接続語句、助詞相当句、複合動詞の後項動詞、慣用句などがその指標となり得ることをこれまで実証してきた。

今回は、自然言語処理の手法によって、ネット上から収集したブログの文章を使用する機会が得られたので、従来の文章コーパスの一部を加えて四つのジャンルの文章（ブログ、新聞社説、文学作品、論文）を対象として、新たな五つの指標によって、異なる文章ジャンルの判別の可能性を探るために調査を行った。コーパスとしては限られた小さなものではあるが、異なるアプローチによる最初の試みである。本稿ではその調査結果を報告する。

2. 調査方法

2–1　分析に用いた文章資料―個人作成コーパスとブログの文章資料

2–1–1　個人作成コーパス

村田（2012）などで従来、利用してきた文章コーパスである。このコーパスは、大学・大学院で日本語による学習・研究を志す日本語学習者が、論文やレポートを書く際に、その文章モデルになると考えられる、論理展開が明示的な専門分野の論文を中心とした、いわゆる書きことばの文章を

収集することを目的として作られたものである。また、その比較対象の資料としては、専門分野の論文に比べて論理の展開が非明示的だと考えられる新聞社説と文学作品の文章を収集し、加えてある。今回はその中から以下のものを選んだ。ここには新たに加えた資料も含まれている。

A. 物理学論文:『日本物理学会誌』(1997) 第 52 巻 No.1〜12 の各号の「最近の研究から」の掲載論文 2 編ずつ合計 24 編 (除　数式・記号)。

B. 文学論文: 学術雑誌『中古文学』『中世文学』『近世文藝』『日本近代文学』の四誌各 10 冊の掲載論文から単純無作為抽出による各 6 編ずつ計 24 編 (除引用部分)。

C. 経済学論文:『経済研究』一橋大学経済研究所編 (2000) VOL. 51 No. 1〜4 の全論文 19 編 (除数式)。

D. 理工学講演論文: 以下の三つの学会講演論文集から掲載順に 190 編。但し、図式が多数含まれているもの、英文のものは除いた。

(a) 情報処理学会講演論文

情報処理学会第 66 回全国大会 (2004) 論文集より 62 編

(b) 電子情報通信学会講演論文

2004 年電子情報通信学会総合大会 (2004) 論文集より 63 編

(c) 応用物理学会講演論文

第 51 回応用物理学会関係連合講演会 (2004) 論文集より 65 編

E. 文学作品 76 編: (1) 近代文学作品: 近代文学作家 12 名の作品計 26 編。(2a) 現代文学作品: 星新一編を泡坂妻夫の短編計 25 編。(2b) 現代文学作品: 星新一編と泡坂妻夫の短編計 50 編。

F. 新聞社説

1996 年 12 月 1 日から 31 日までの 1 か月分の日本経済新聞、朝日新聞、読売新聞、毎日新聞の 4 紙の全社説 (日経 51 編、朝日 55 編、読売 58 編、毎日 58 編) 222 編。さらに朝日新聞 1985 年から 2005 年までの毎月 9 日の社説 (休刊日の場合は 10 日) 合計 252 日分、489 編 [1)]。

重複を除いた総数 711 編。

2-1-2　ブログ資料

ドイツのマルティン・ルター大学ハレ・ヴィッテンベルグの情報学部と日本学研究所が共同で開発したコーパス（「原発ブログコーパス」）を使用する許可を得たので、それを用いる。このプロジェクトの目的は、東日本大震災に関する日本のブログ記事を収集し、その中でどのような内容が書かれているのかを調査することで、特に「原発」という語が含まれているブログ記事をダウンロードして調査が行われている。

本調査で対象とするブログ資料は、上記資料の中で、「牛」、「セシウム」という語が含まれた 1931 編のブログである。このブログ資料を調査すると、高い確率で同じブログ内に重複部分があることがわかった。このような重複部分は、新聞記事から引用されたもの、他のブログから取られたものなどいろいろである。この 1931 篇中、1379 編は重複部分がない資料であった。重複部分を抽出すると、その数は 867 編に上り、重複部分がない資料の数と合わせると 1931 編を超える。この事実から、一つのブログに重複部分がニ回以上あるものがあることがわかる。

本調査では、重複部分がない 1379 編の「ブログ本文」を利用する。

2-2　指標と方法

2-2-1　指標の選択

本調査では、文章ジャンルを判別するために、以下の五つの指標を選んだ。

a.　文の長さ（一文当たりの平均の文字数）

b.　文末の敬体の使用

c.　感嘆符の使用

d.　接続詞の使用

e.　指示語の使用

2-2-2　方法

調査には日本語テキストの分析のために新しく開発されたプログラムを利用した[2)]。指標ごとに具体的な調査方法を記す。

（a）一文当たりの平均文字数を計算するため、句点の場所、あるいは行末で中断したところを「文」とみなす。その各文の文字数を合計し、文数で割って一文当たりの平均文字数を計算する。

（b）文章を一文ずつに分け、形態素解析エンジンである MeCab[3)]を利用して、各文を分かち書きにし、文末動詞を分析して、「ます」「です」などの敬体が一文当たりどのぐらい含まれているかを調べる。

（c）感嘆符の使用は、すべてのブログ資料中の一文当たりの感嘆符の使用頻度を計算する。

（d）（b）の作業の延長で、指示語「この、その、あの、どの、これ、それ、あれ、どれ、こう、そう、ああ、どう」の一文当たりの使用頻度を調べる。

（e）（b）の作業の延長で、接続詞 35 語（また、しかし、だが、そして、すなわち、あるいは、なお、ただし、したがって、それで、つまり、すると、そうして、だから、ところが、もっとも、さて、けれども、要するに、または、それに、なぜなら、では、それでは、ところで、ちなみに、それなのに、それとも、こうして、ついで、ならびに、そのうえ、それなら、ともあれ、しかるに）の一文当たりの使用頻度を調べる。

3.　結　　果

調査結果を表 1 から表 3 に示す。

a.　文の長さ

文の長さについては、文字数を基準とする。文学作品が四つのジャンルの中では一番短く、論文がその約 2 倍で一番長い。ブログ本文と新聞社説の一文の平均文字数は、新聞社説の方が若干長いが、論文より約 10 字短いという結果になっている。論理的に書くことを目的とした論文が、一文

表 1　各指標ごとの調査結果

指標		ブログ本文	新聞社説	文学作品	論文
	総編数	1379	711	76	257
	総文数	66938	25316	6773	16437
1	文の長さ	36.0863	38.2812	23.5048	47.5341
2	文末動詞の形（敬体）	24.0148％（16075）	0.3437％（87）	16.3738％（1109）	0.2494％（41）
3	感嘆符	0.0614/sent.（4107）	0.0/sent.（8）	0.0109/sent.（74）	0.0/sent.（0）
4	指示語の頻度	0.1725/sent.（11546）	0.2256/sent.（5711）	0.2703/sent.（1831）	0.3176/sent.（5221）
5	接続詞の頻度	0.0543/sent.（3632）	0.083/sent.（2100）	0.0855/sent.（579）	0.1559/sent.（2563）

/sent.: 1 文当たりを意味する

表 2　指示語の頻度

指示語	ブログ本文	新聞社説	文学作品	論文
この	3.6616％（2451）	4.8546％（1229）	3.8683％（262）	11.8513％（1948）*
これ	2.6547％（1777）	3.4445％（872）	1.7274％（117）	3.9849％（655）*
こう	0.2853％（191）	0.2568％（65）	1.2255％（83）*	0.0548％（9）
その	3.1671％（2120）	5.8817％（1489）	7.0427％（477）	8.6877％（1428）*
それ	2.4201％（1620）	3.6341％（920）	5.2857％（358）*	3.8997％（641）
そう	2.0258％（1356）	1.1732％（297）	3.5878％（243）*	0.4015％（66）
あの	0.3272％（219）	0.1067％（27）	1.4174％（96）*	0.0304％（5）
あれ	0.1449％（97）	0.0237％（6）	0.3691％（25）*	0.0243％（4）
ああ	0.1001％（67）	0.0％　（0）	0.4282％（29）*	0.0％　（0）
どの	0.3959％（265）	0.4898％（124）	0.2362％（16）	1.5149％（249）*
どれ	0.2584％（173）	0.1975％（50）	0.2658％（18）*	0.1703％（28）
どう	1.4357％（961）	2.1409％（542）*	1.4469％（98）	0.7848％（129）

*: MAX

表 3　接続詞の頻度

接続詞	ブログ本文	新聞社説	文学作品	論文
また	1.0189%（682）	1.1179%（283）	1.0926%（74）	4.4595%（733）*
しかし	0.9113%（610）	2.69%　（681）*	1.7865%（121）	2.0563%（338）
だが	0.127%　（85）	1.8131%（459）*	0.502%　（34）	0.2494%（41）
そして	0.8949%（599）	0.3516%（89）	1.4469%（98）*	1.3628%（224）
すなわち	0.0299%（20）	0.0198%（5）	0.0443%（3）	1.0951%（180）*
あるいは	0.13%　（87）	0.1975%（50）	0.0148%（1）	0.8822%（145）*
なお	0.0956%（64）	0.3674%（93）	0.1034%（7）	0.8822%（145）*
ただし	0.1389%（93）	0.1422%（36）	0.0%　（0）	0.8213%（135）*
したがって	0.0403%（27）	0.0553%（14）	0.0443%（3）	0.8213%（135）*
それで	0.118%　（79）	0.079%　（20）	0.1919%（13）*	0.0122%（2）
つまり	0.2629%（176）	0.1343%（34）	0.1476%（10）	0.8578%（141）*
すると	0.0299%（20）	0.004%　（1）	0.6644%（45）*	0.0548%（9）
そうして	0.0105%（7）	0.0119%（3）	0.4134%（28）*	0.0061%（1）
だから	0.4078%（273）*	0.0948%（24）	0.2953%（20）	0.0487%（8）
ところが	0.1434%（96）	0.3555%（90）*	0.2067%（14）	0.2008%（33）
もっとも	0.0747%（50）	0.1817%（46）	0.1034%（7）	0.2069%（34）*
さて	0.1434%（96）	0.0079%（2）	0.1772%（12）	0.3164%（52）*
けれども	0.1165%（78）	0.1027%（26）	0.3839%（26）*	0.1338%（22）
要するに	0.0359%（24）*	0.0158%（4）	0.0295%（2）	0.0061%（1）
または	0.0508%（34）	0.0395%（10）	0.0295%（2）	0.2555%（42）*
それに	0.0598%（40）	0.0751%（19）	0.251%　（17）*	0.0304%（5）
なぜなら	0.0314%（21）	0.0119%（3）	0.0%　（0）	0.0913%（15）*
では	0.0941%（63）	0.0593%（15）	0.1329%（9）	0.2251%（37）*
それでは	0.0478%（32）	0.0751%（19）	0.1034%（7）*	0.1034%（17）*
ところで	0.0702%（47）	0.004%　（1）	0.0591%（4）	0.1521%（25）*
ちなみに	0.1464%（98）*	0.0%　（0）	0.0%　（0）	0.0548%（9）
それなのに	0.0314%（21）	0.1027%（26）*	0.0591%（4）	0.0122%（2）
それとも	0.0553%（37）	0.0514%（13）	0.0886%（6）*	0.0304%（5）
こうして	0.0299%（20）	0.0514%（13）	0.0738%（5）*	0.0669%（11）
ついで	0.0478%（32）*	0.004%　（1）	0.0443%（3）	0.0304%（5）
ならびに	0.0%　（0）	0.0%　（0）	0.0148%（1）	0.0426%（7）*
そのうえ	0.0%　（0）	0.0316%（8）*	0.0148%（1）	0.0061%（1）
それなら	0.0209%（14）	0.0316%（8）*	0.0295%（2）	0.0061%（1）
ともあれ	0.009%　（6）	0.0158%（4）*	0.0%　（0）	0.0061%（1）
しかるに	0.0015%（1）	0.0%　（0）	0.0%　（0）	0.0061%（1）*

*: MAX

が一番長いということは、目的を述べ、因果関係を論じ、具体的な例示をするなどの書き方によると考えられる。

b.　文末の動詞の形（敬体）

論文と新聞社説は、通常、文末は常体形であるため、ブログ本文、文学作品の二つのジャンルとは数値的に大きな差異が見られる。今回の調査では、文末に敬体が一番多く使用されていたジャンルはブログ本文である。この理由としては、ブログ本文がそもそも直接読者に向けて書かれたものであるということが一番大きいと考えられる。文学作品も敬体で書かれているものもあるが、現代作品のショートショートの内容を見ても、第三者の目で書いたもの、会話形式で話を成り立たせているものなど、表現方法は多様である。ブログ本文の方が文学作品より、読者との直接のコミュニケーションを目的としていると考えられ、今回のような結果になったと考えられる。

c.　感嘆符

感動、驚き、強調などの感情を表す感嘆符は、論理性の高い文章では使われにくいと考えられる。今回の調査でも、新聞社説、論文では出現せず、一番多く使われているのがブログ本文であった。この結果は上記の b. にもつながり、直接読者に話しかける、あるいは訴えかけることを目的とするブログでは、読者との共感を求める表現を用いることが多いと推測される。

d.　指示語

指示語としては「こ・そ・あ・ど」の使用頻度について調査した。その結果を表 2 に示す。

「この、その、どの」は、四つのジャンルの中で、論文に最も多く使われ、「こう、それ、そう、あの」は文学作品に最も多く、「どう」は新聞社説に最も多かった。「あれ」は、ブログ本文に最も多くなっている。指示語の中で、「この、その」が論文によく使われていることについては、連体詞としての用法のほか、「この結果」「その結果」「このため」「そのため」などのよ

うに、因果関係を表す接続詞としても用いられることが頻度を高くしていると考えられる。また、文学作品に多い「こう」「そう」については、副詞用法のほか、「こうして」「そうして」のように一つの接続詞として使われる場合も含まれている。「ああ」については、文学作品に最も多く出現しているが、MeCab ではすべて感動詞として検索されていた。そこで、連体詞用法「ああいう」について再検索すると、ブログ本文に 5 例、文学作品に 1 例見つかった。例は非常に少ないが、指示語「ああ」はブログ本文と文学作品のみで使用されていることがわかった。

「あの」は、文学作品に最も多く、「あれ」はブログ本文と文学作品で頻度が高い。「あ」の文脈指示用法[4)]を考えると、「あ」系の指示語は、論理性の高い論文や、客観性が求められる新聞社説には出現しにくいことが推測される。

「ど」系では、論文で「どの」が四つのジャンルの中で特に多く用いられている。「どう」は新聞社説で非常によく用いられている。「どれ」については各ジャンルとも低い割合で用いられ、特に際立った特徴はない。内容を見ていくと、論文では「どのように」「どのような」が非常によく用いられている。新聞社説では、「〜たらどうか」「〜みてはどうか」などの「どうか」で使われるもののほか、「どうすべきか」「どう違うか」「どうしたらよいか」「どう説明するのか」「どうしようもない」などがあった。

e. 接続詞

接続詞の使用頻度については表 3 の通りである。

四つのジャンルの中で、論文によく使われる接続詞を見ていくと以下のものがある。

また、すなわち、あるいは、なお、ただし、したがって、つまり、

さて、または、では、ところで、ならびに、しかるに

「さて」「では」「ところで」「しかるに」については、これまで調査対象としていないので比較ができないが、「すなわち」「なお」「ただし」「したがって」

「つまり」については、村田（2002）の結果を裏付ける結果となっている。

次に、日本語教育の視点から、書きことばと話しことばの差異に関して、興味深い点を以下に記す。

（1）同じ順接の意味の接続詞でも、「また」は論文によく用いられ、「そして」「そうして」「こうして」は文学作品によく用いられていることがわかる。

（2）逆接の意味の接続詞では、新聞社説、論文で「しかし」がよく用いられているのに対して、「けれども」は文学作品でよく使われている。また、「だが」は新聞社説に多用されている。

（3）因果の意味の「したがって」は、論文に非常に多く用いられているが、「だから」はブログ本文によく見られ、文学作品でも相対的に多く用いられていることがわかる。

（4）添加の意味の接続詞では、「そのうえ」は新聞社説に多く、「それに」は文学作品に多い。

（5）並列ならびに選択を意味する「また」「ならびに」「または」「あるいは」は論文に最も多く現れている。

（6）転換の意味を持つ「さて」「では」「ところで」の三つの接続詞は、論文ジャンルに最も多く現れている。内訳では「さて」が最多である。分野の使用状況を見ると、最もよく使われているのは文学論文で、つづいて経済学論文である。工学論文には使用例がなく、物理学論文では「さて」6 例、「では」4 例、「ところで」3 例となっている。今回の調査で、これらの転換の意の接続詞も指標として有効な可能性が示されたので、今後の課題としたい。

このほか、各ジャンルに特徴的な語としては、ブログ本文は「要するに」「ちなみに」「ついで」、新聞社説は、「ところが」「それなのに」「それなら」「ともあれ」、文学作品は「それで」「すると」「それとも」、論文は、「もっとも」「なぜなら」「しかるに」であった。

4. おわりに

今回は、自然言語処理の手法を用いて、ブログの文章を加えた四つの異なる文章ジャンルの判別を可能とする指標を探して調査を行った。限定的なコーパスを対象とした調査ではあったが、本稿で取り上げた五つの指標が有効である可能性が示された。今後は、大量のデータを対象として、より詳しい調査と分析を行いたい。

謝　辞

本調査ではドイツのマルティン・ルター大学ハレ・ヴィッテンベルグの情報学部と日本学研究所が共同で開発したコーパス（「原発ブログコーパス」）を使用しています。使用をご快諾いただいたヒンデンブルグ教授とオーバーレンダー教授に心から感謝いたします。

注

1）国立国語研究所　共同研究プロジェクト「テキストにおける語彙の分布と文章構造」（H22.10−H23.3）の共同研究者として参画した際に共同利用を許可された社説データである。データ元は朝日新聞テキストデータベース『聞蔵Ⅱ』である。
2）ロマン・ロッサ（Roman Lossa）開発によるプログラム「日本語テキストアナライザー」を用いて分析を行った。
3）MeCab はオープンリソースの形態素解析エンジンである。
4）「あ」の文脈指示の条件は、指示対象が話し手と聞き手の共通知識であることである（吉本啓「日本語の指示詞コソアの体系」『指示詞』P118）

参考文献

市川　孝（1978）『国語教育のための文章論概説』教育出版

金水　敏・田窪行則編著（1992）『指示詞』ひつじ書房

村田　年（2002）「論理展開を支える機能項目―接続項目、助詞相当句による文章のジャンル判別を通じて―」『計量国語学』vol. 23, No. 4, pp. 186−206

村田　年（2012）「文章のジャンル判別に寄与する指標の研究―専門日本語教育への応用―」石田・金編著『コーパスとテキストマイニング』共立出版 166−180

「日本語テキストアナライザー」開発時の参考文献

山崎　誠「語の平均使用度数に現れるテキストの特徴」特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（2010）

George Forman, Kave Eshghi, & Stephane Chiocchetti“Finding Similar Files in Large Document Repositories”Hewlett-Packard（2005）