翻訳小説を資料とした品詞比率と文書間類似度による明治中期口語文体分析

小西 光（国立国語研究所コーパス開発センター）†

The Colloquial *Genbun Itchi* Style Analysis on Translated Novels in Mid-Meiji Era by Part-of-Speech Rate and Document Similarity

Hikari KONISHI (National Institute for Japanese Language and Linguistics)

要旨

明治期の文体を論じる際、多様な文体から言文一致による口語体書き言葉成立へという変遷は指摘されているものの、その具体的な実態と詳細が明らかになっているとはいえない。本発表では明治中期に口語体で翻訳された翻訳小説を対象に「近代口語文翻訳小説コーパス」を構築し、明治 40 年代に成立したとされる口語体書き言葉への萌芽を観察する。

特徴量として名詞率に対する MVR の分布、全体の品詞比率および品詞・語彙素・出現書字形・品詞バイグラムの分布による文書間類似度を用い、『太陽コーパス』『近代女性雑誌コーパス』で「口語」とアノテーションされたデータとの比較を行った。その結果、名詞率と MVR の二次元グラフでは、『太陽』と『女性雑誌』の全データセットが翻訳小説五作品よりも近い位置にまとまって分布し、翻訳小説五作品とは異なることが明らかになった。一方、文書間類似度においては、翻訳小説五作品すべてに対して 1909（明治 42）年発行の『太陽』コアデータセットの距離が最も近いことが明らかとなった。

１．はじめに

国立国語研究所にて現在も近代語のコーパス整備が行われている。田中ほか(2012)では明治から昭和までをおよそ 15 年ごとに区切り、各時代のジャンルや文体など幅を持たせたコーパスの方向性を示している。国立国語研究所にて現在公開されているものは『明六雑誌コーパス』（明治前期）『国民之友コーパス』（明治中期）『太陽コーパス』『近代女性雑誌コーパス』（明治中期～大正期）の四つである。

一方、「近代口語文翻訳小説コーパスの構築と計量的文体研究」（研究課題番号：25770178）にて収録対象資料とした明治中期（特に明治 20 年代）の口語体翻訳小説とは、当時の文学界において初期言文一致体を試みた作家たちと密接不可分なものであり、新文体の獲得に無関係とは言えない[1]ものの、あまりその特徴が明らかにされることはなかった。口語体翻訳小説は、明治 40 年代に口語体としての書き言葉が統合・成立するその過程を捉える上で、押さえるべき資料と考える。

そこで、本発表では明治中期に口語体で翻訳された小説五作品を資料とし、その概要および品詞比率をまとめ、明治中期から大正期のコーパスである『太陽コーパス』『近代女性雑誌コーパス』（以下、『太陽』『女性雑誌』）の品詞・語彙素・出現書字形の情報を用いて文書間類似度の比較を行った。以下、２節では分析データをまとめ、３節では品詞比率と MVR、４節では各コーパスの年代別文書間類似度を比較し、５節でまとめとする。

† hkonishi@ninjal.ac.jp

1 加藤(2012)「（明治時代、）小説家は、自己の創作活動のために必要とする形式と内実を、彼の翻訳作業を通じて探索していたのだ」(pp. iv - v)

２．分析データ

２．１『太陽コーパス』『近代女性雑誌コーパス』について

2005 年に公開された『太陽コーパス』は、総合雑誌『太陽』（博文館刊）1895（明治 28）年、1901（明治 34）年、1909（明治 42）年、1917（大正 6）年、1925（大正 14）年発行の通常号全文をデータとするタグ付きコーパスである。含まれる記事数や文字数の基礎統計量については森(2014)にまとめられており、「１記事文字数、出版年ごと記事数・文字数・ジャンルにばらつきがあり（中略）非常に不均衡なコーパスである」との指摘があるなど取り扱いには注意を要する。本発表では特別な配慮は行わなかった。現在整備中の『太陽コーパス』にはコアデータと非コアデータという二種類のデータセットがあり、コアデータについては精緻な人手修正が行われ、精度の高いデータとなっている。今回の調査では発行年ごとにコアデータ（TC）と非コアデータ（TNC）を区別した。

また続いて 2006 年に公開された『近代女性雑誌コーパス』は、1894（明治 27）・1895（明治 28）年発行の『女学雑誌』31 冊（女学雑誌社）、1909（明治 42）年発行の『女学世界』6 冊（博文館）、1925（大正 14）年発行の『婦人倶楽部』3 冊（講談社）の全文をデータとするタグ付きコーパスである。『女性雑誌』には『太陽』のようなデータの区別が行われていないため、発行年ごとのデータセット（JC）としている。

両コーパスには、サンプル単位と形態素単位の両方に口語・文語（・漢文ほか）の情報が付与されており、本分析ではサンプル単位で「口語」と認定されたサンプルを利用する。サンプル単位の口語文にも、形態素単位には口語要素だけでなく文語要素（典拠・手紙ほか）が含まれるがこれらについては排除していない。

２．２「近代口語文翻訳小説コーパス」について

現在構築を進めている「近代口語文翻訳小説コーパス」の公開予定データは、表１の五作品である。このほかに現在修正中のものもあるが、資料の成立年代としては明治 20 年代を中心とした常体・口語体翻訳小説からなる形態素情報付きコーパス[2]となっている。なお、敬体の翻訳小説については、収録を予定していない。

口語体・文語体の判定については、『太陽』の文体情報付与基準と同様に「文末辞が「なり」「たり」「き」「けり」などで終わる文体は文語体、「だ」「である」「た」「です」「ます」などで終わる文体は口語体」(田中ほか 2012)とし、資料を選定した。「近代口語文翻訳小説コーパス」は基本的に全文口語文で構成されているが、『罪と罰』以外は地の文・会話文等をすべて含んだデータとなっており、『罪と罰』のみ当初地の文を分析対象としていたため、会話文や書簡文（第三回の大部分を書簡文が占める）を含んでいない（今後、品詞・形態素情報整備完了後、収録予定）。

表１に出典情報、表２に文の数、短単位の数、文の長さ、MVR[3]、名詞率[4]の値をまとめた。

『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を対象とした山崎(2014)の調査では、37 短単位数以下の文で全体の 90%をカバーしているという報告があり、五作品の文の長さが極端に長過ぎるということはなさそうではあるが、BCCWJ の文の長さの平均値よりはやや長いといえる。

また MVR については次節でも取り上げるが、小磯ほか(2010)の調査[5]によると BCCWJ 中

---

[2] 言語単位は BCCWJ を踏襲した「短単位」を採用し、品詞体系についても UniDic 品詞体系を用いた。（小椋ほか 2011）

[3] 樺島・寿岳(1965)　MVR=100*形容詞・形状詞・副詞・連体詞の数／動詞の数

[4] 樺島・寿岳(1954)では機能語を除いて名詞率を算出しているため、本稿でも同様の方法で算出した。

[5] 小磯(2010)では、分析に言語単位「長単位」を用いている。

の小説の MVR は 25～70 の間に収まり、これも文の長さ同様に大きな差異は見られず、{「玉を懐いて罪あり」「綠葉歎」} と {「洪水」『罪と罰』} の二組は近い値を示している。

表 1　「近代口語文翻訳小説コーパス」出典情報

作品名	原作者	訳者	原語	初出・刊行年	初出
あひゞき	ツルゲーネフ	二葉亭四迷	露語	明治 21(1888)年	国民之友
玉を懐(いだ)いて罪あり	ホフマン	森鴎外	独語	明治 22(1889)年	読売新聞
洪水	ブレツト、ハアト	森鴎外	独語	明治 22(1889)年	『柵草子』
綠葉歎	ドオデエ	森鴎外	独語	明治 22(1889)年	読売新聞
罪と罰	ドストエフスキー	内田魯庵	英語	明治 25(1892)年	単行本

表 2　「近代口語文体翻訳小説コーパス」文数・短単位数・文の長さ・MVR・名詞率

作品名	文数	短単位数[6]	文の長さ（短単位数／文数）	MVR	名詞率
あひゞき	159	5,557	34.95	68.46	43.06
玉を懐いて罪あり	892	25,636	28.74	47.63	53.55
洪水	124	4,429	35.72	55.90	47.51
綠葉歎	88	2,296	26.01	54.94	53.58
罪と罰	1,097	30,472	27.77	57.81	48.52
計	2,360	68,390	29.40	54.73	50.08

３.『太陽コーパス』『女性雑誌コーパス』と「近代口語文翻訳小説コーパス」の品詞比率

本節では品詞比率と MVR を用いた比較を行う。樺島・寿岳(1965)では、名詞率（以下、N 率）と MVR の関係から文章の特徴が明らかになるとした。本分析データについても、同様の手法で比較することとする。

３.１　名詞率と MVR

図 1 に「近代口語文翻訳小説コーパス」五作品と『太陽』『女性雑誌』における N 率に対する MVR の分布を示す。問題となる N 率については、「あひゞき」のみ他の四作品や『太陽』『女性雑誌』よりも値が小さく、MVR が「極めて大」とされる 56 以上の 68.5 という点から、樺島・寿岳(1965)で分類された「ありさま描写的」と言える。たしかに「あひゞき」は、語り手の視点が物陰から男女の逢引の一場面を描写するという短編であり、動作性の描写という点で、他の四作品とは異なっている。

他の四作品については、N 率は小から普通(45~54)の範囲にあり、MVR は「玉を懐いて罪あり」が普通(48～54)、「綠葉歎」「洪水」は大(54～56)、『罪と罰』は極めて大(56～)に位置している。また、「洪水」と『罪と罰』については、N＜MVR となっている。このことより、上記四作品の中では、「洪水」『罪と罰』は「ありさま描写的」、「玉を懐いて罪あり」「綠葉歎」は「動き描写的」な傾向性を持つものと考えられる。

一方、『太陽』『女性雑誌』のデータと比較をすると、N 率と MVR の関係において「あひゞき」「洪水」『罪と罰』は異なる傾向性があると言える。当然『太陽』『女性雑誌』は雑誌という性質上、小説以外の記事が含まれ、単純な比較はできない。一方で、『太陽』と『女性雑誌』というサンプリングした年代の異なるデータで、いずれも近しい値となったという点は、注目に値する。

[6] 空白・補助記号は除いた。

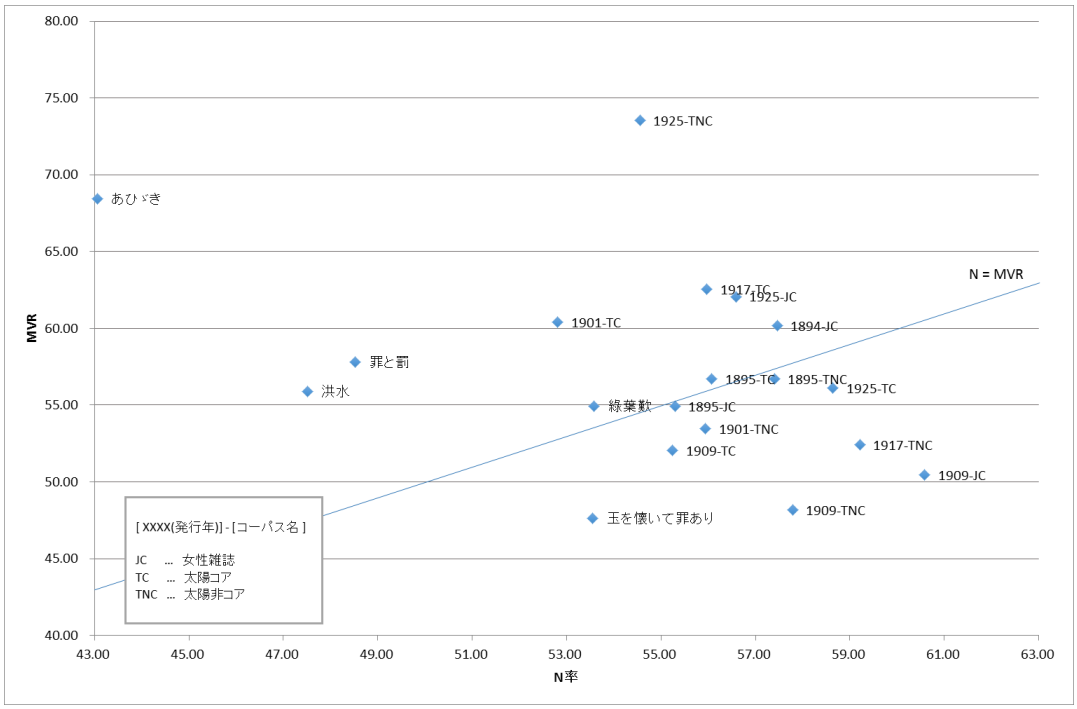

図 1　名詞率に対する MVR の分布

【例１】
此難に逢うて飾は取られたが、不思議と命を拾つた人の話に、何心なく道を行くと、突然頭を強く打たれ、其儘仆れて氣を失ひ、暫くして心付いて見れば、遙か離れた町に居て飾はなかつたといふ。家の中で殺されたものも、途で殺されたものも、撿屍の時に見ると、皆んな唯つた一つの突創が胸に在るばかり。解剖して見れば、心の臓が差し貫ぬかれてある。

（N 率：53.55　MVR:47.63 「玉を懐いて罪あり」）

【例２】
取分け自分の氣に入ツたはその面ざし、まことに柔和でしとやかで、取繕ろツた氣色は微塵もなく、さも憂はしさうで、そしてまた愛度氣なく途方に暮れた趣きも有ツた。たれをか待合はせてゐるのと見えて、何か幽かに物音がしたかと思ふと、少女はあわてゝ頭を擡げて、振り反つて見て、その大方の涼しい眼、牝鹿のものゝやうにをど／＼したのをば、薄暗い木蔭でひからせた。

（N 率：43.06　MVR:68.46 「あひゞき」）

【例３】
暫らくすると戸が少し開いて其隙間から部屋の主人が小さな眼を暗黒の中に燦つかせながら慥に猜疑の心をもて訪問者を吟味すると、溜段の上には多勢人がゐたから、やッと安神したらしく戸を排放した。少年は薄暗い前房に入ッた。壁一ト重を距てゝ奥は狭い臺所であッた。其部屋の中に黙然として屹立し不審しげにきッと少年を凝視めたは年配六十位の皺枯れて癯せこけた老婆で、鼻準透ッて鋭く尖り、陰険な色を帯びた眼光はギラ／＼人を射る様である。

（N 率：48.52　MVR:57.81 『罪と罰』）

表 3　「近代口語文翻訳小説コーパス」の品詞比率（機能語も含む）

	P	N	V	M	I	O
あひゞき	45.20	23.60	17.68	12.10	1.43	0.00
玉を懐いて罪あり	45.49	29.14	16.65	7.93	0.69	0.10
洪水	45.53	25.86	17.84	9.97	0.76	0.03
綠葉歎	45.60	29.15	15.77	8.66	0.83	0.00
罪と罰	43.55	27.39	17.87	10.33	0.85	0.00

３.２　機能語を含む作品全体の品詞比率

次に表３に「近代口語文翻訳小説コーパス」の助詞や助動詞といった機能語も含む全体の品詞比率[7]を示す。山崎(2014)の BCCWJ における品詞比率の調査（「。．！？」で終わる「通常の文」を対象とし、短単位を基準としたもの）に比べ、N の比率が 10 前後小さくなり、それ以外の V、M、I、P の値がいずれも高くなっている。山崎(2014)では「句点で終わる文に比べて疑問符、かぎ括弧で終わる文で、N の割合が低く、P の割合が多くなっているのは話し言葉的な要因が関係している可能性がある。」と指摘されている。現代語の品詞比率や考察を単純に近代語に対して適用することはできないが、BCCWJ の書籍データのうちの文学にデータを絞り、比較することを今後の課題としたい。

図２に「近代口語文翻訳小説コーパス」五作品と『太陽コーパス』『近代女性雑誌コーパス』におけるすべての品詞を対象とした品詞比率を図示する。「近代口語文体翻訳小説コーパス」では、樺島(1965)の示す通り V・M 率と N 率との間にやや相関が見られるが、『太陽』『女性雑誌』では、N 率と P 率の間に相関が見られる。これはテキストの内容（小説か評論か等）の問題と推察されるが、今後より詳細に調査していきたい。

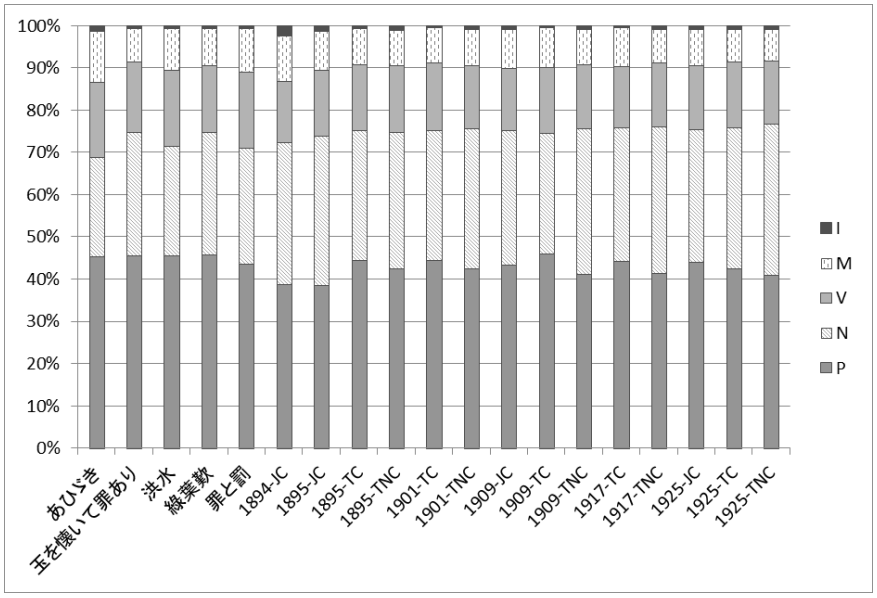

図 2　品詞比率の比較（機能語を含む）

[7] N（名詞類）：名詞、代名詞、接尾辞-名詞的、記号　V（動詞類）：動詞、接尾辞-動詞的　M（形容詞・形状詞・副詞類）：形容詞、形状詞、副詞、連体詞、接頭辞、接尾辞-形容詞的、接尾辞-形状詞的　I（接続詞・感動詞類）：接続詞、感動詞　P（助詞・助動詞類）：助詞、助動詞　O（その他）：未知語、漢文、英単語ほか（山崎 2014）

４.『太陽コーパス』『女性雑誌コーパス』と「近代口語体翻訳小説コーパス」の類似度

４.１　分析手法

以下では品詞分布・語彙素分布・出現書字形分布・品詞バイグラム分布の四種類の文体特徴量を用いた文書間類似度について検討する。各分布は頻度ベクトルの形式で保持し、頻度ベクトルのコサイン類似度を検討する。仮に比較する文書のベクトルを$\vec{s}$とし、比較される文書のベクトルを$\vec{t}$とすると、コサイン類似度は以下の式で表される：

$$\cos(\vec{s},\vec{t}) = \frac{\vec{s}\cdot\vec{t}}{|\vec{s}|\cdot|\vec{t}|}$$

通常、０から１の値をとり、文書間距離が近い（似ている）場合１に近い値を、最も文書間距離が遠い（似ていない）場合に０に近い値を取る。

品詞情報を用いた分布取得において、品詞「空白」と「補助記号-*」を排除した。UniDicの品詞体系には「名詞-普通名詞-一般」のように「[大分類]-[中分類]-[小分類]」と分類されているが、小分類まで用いている。品詞バイグラム分布において、文の先頭要素には “BOS” と当該品詞の対を特徴量として用いるが、バイグラムの前件・後件のいずれかが「空白」もしくは「補助記号-*」の場合は特徴量空間から排除してコサイン類似度の算出を行った。

４.２　各種分布による文書間類似度

表４～表７に「近代口語体翻訳小説コーパス」五作品それぞれと『太陽』『女性雑誌』の発行年別データセット（『太陽』のみコア・非コア区別あり）との文書間類似度をまとめた。

まず全体を通して共通する点を三点挙げる。一つ目は、どの特徴量においても 1894 年の『女性雑誌』データは、五作品のいずれに対しても文書間距離の値が小さく、また値の差分が、上位の値同士のそれと比較して大きい。原因を明らかにするべきであるが、次稿に

表 4　品詞分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.982	1909-TC	0.992	1909-TC	0.988	1901-TC	0.990	1909-TC	0.988	1909-TC
2	0.977	1925-JC	0.991	1895-TC	0.987	1909-TC	0.988	1917-TC	0.983	1909-JC
3	0.975	1901-TC	0.990	1901-TNC	0.985	1895-TC	0.988	1895-TC	0.983	1925-JC
4	0.971	1895-TC	0.990	1917-TC	0.984	1901-TNC	0.985	1901-TC	0.982	1895-TNC
5	0.970	1909-JC	0.989	1895-TNC	0.984	1895-TNC	0.985	1901-TNC	0.982	1895-TC
6	0.967	1901-TNC	0.989	1901-TC	0.982	1909-JC	0.985	1925-JC	0.980	1901-TC
7	0.965	1917-TC	0.988	1909-JC	0.981	1917-TC	0.984	1895-TNC	0.980	1901-TNC
8	0.964	1895-TNC	0.987	1925-JC	0.979	1909-TNC	0.983	1909-JC	0.979	1917-TC
9	0.962	1925-TC	0.986	1909-TNC	0.979	1925-JC	0.981	1909-TNC	0.979	1925-TC
10	0.961	1909-TNC	0.983	1925-TC	0.975	1895-JC	0.979	1925-TC	0.976	1909-TNC
11	0.954	1925-TNC	0.982	1917-TNC	0.971	1925-TC	0.978	1917-TNC	0.975	1917-TNC
12	0.951	1917-TNC	0.980	1925-TNC	0.970	1917-TNC	0.976	1925-TNC	0.973	1925-TNC
13	0.951	1895-JC	0.976	1895-JC	0.969	1925-TNC	0.968	1895-JC	0.970	1895-JC
14	0.917	1894-JC	0.900	1894-JC	0.897	1894-JC	0.898	1894-JC	0.897	1894-JC

表 5　語彙素分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.948	1901-TC	0.972	1925-JC	0.966	1901-TC	0.959	1909-TC	0.940	1917-TNC
2	0.945	1909-TC	0.964	1909-TC	0.960	1909-TC	0.957	1917-TC	0.938	1895-TC
3	0.933	1925-TC	0.960	1901-TC	0.956	1917-TC	0.957	1901-TC	0.932	1925-JC
4	0.928	1925-JC	0.958	1909-JC	0.950	1925-TC	0.953	1925-TC	0.931	1901-TC
5	0.925	1917-TC	0.952	1925-TC	0.948	1925-JC	0.948	1925-JC	0.924	1909-JC
6	0.916	1925-TNC	0.952	1917-TC	0.944	1925-TNC	0.946	1925-TNC	0.920	1895-JC
7	0.905	1909-JC	0.951	1901-TNC	0.937	1917-TNC	0.933	1909-JC	0.920	1909-TNC
8	0.903	1917-TNC	0.948	1925-TNC	0.933	1909-JC	0.933	1917-TNC	0.912	1925-TC
9	0.902	1895-TC	0.947	1917-TNC	0.931	1901-TNC	0.928	1895-TC	0.904	1917-TC
10	0.894	1901-TNC	0.943	1895-TC	0.929	1895-TC	0.925	1901-TNC	0.902	1901-TNC
11	0.882	1909-TNC	0.941	1895-JC	0.922	1909-TNC	0.916	1909-TNC	0.900	1895-TNC
12	0.867	1894-JC	0.937	1909-TNC	0.916	1895-JC	0.893	1895-JC	0.885	1909-TC
13	0.857	1895-TNC	0.936	1895-TNC	0.902	1895-TNC	0.891	1895-TNC	0.875	1925-TNC
14	0.845	1895-JC	0.913	1894-JC	0.881	1894-JC	0.879	1894-JC	0.858	1894-JC

表 6　出現書字形分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.925	1901-TC	0.979	1925-JC	0.964	1909-TC	0.956	1909-TC	0.947	1925-TC
2	0.925	1909-TC	0.977	1909-TC	0.964	1901-TC	0.954	1925-JC	0.946	1901-TC
3	0.916	1925-JC	0.971	1901-TC	0.959	1925-JC	0.953	1925-TC	0.944	1925-JC
4	0.907	1925-TC	0.969	1925-TC	0.957	1917-TC	0.953	1901-TC	0.942	1909-TC
5	0.905	1917-TC	0.965	1909-JC	0.956	1925-TC	0.952	1917-TC	0.931	1925-TNC
6	0.893	1925-TNC	0.964	1925-TNC	0.951	1925-TNC	0.947	1925-TNC	0.925	1917-TNC
7	0.890	1917-TNC	0.963	1901-TNC	0.945	1917-TNC	0.940	1901-TNC	0.924	1917-TC
8	0.884	1909-JC	0.963	1917-TC	0.943	1909-JC	0.940	1909-JC	0.922	1901-TNC
9	0.880	1901-TNC	0.963	1917-TNC	0.940	1901-TNC	0.940	1917-TNC	0.915	1909-JC
10	0.869	1909-TNC	0.954	1909-TNC	0.933	1909-TNC	0.932	1895-TC	0.913	1895-TC
11	0.868	1895-TC	0.953	1895-TC	0.932	1895-TC	0.931	1909-TNC	0.910	1909-TNC
12	0.865	1895-JC	0.947	1895-JC	0.925	1895-JC	0.916	1895-JC	0.896	1895-JC
13	0.849	1895-TNC	0.944	1895-TNC	0.914	1895-TNC	0.914	1895-TNC	0.893	1895-TNC
14	0.847	1894-JC	0.905	1894-JC	0.887	1894-JC	0.878	1894-JC	0.867	1894-JC

表 7　品詞バイグラムによる文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.956	1909-TC	0.983	1909-TC	0.973	1909-TC	0.971	1909-TC	0.967	1909-TC
2	0.949	1901-TC	0.973	1901-TC	0.968	1901-TC	0.962	1917-TC	0.955	1917-TC
3	0.931	1917-TC	0.972	1917-TC	0.958	1917-TC	0.955	1901-TC	0.952	1925-TC
4	0.930	1895-TC	0.968	1895-TC	0.952	1895-TC	0.949	1895-TC	0.951	1895-TC
5	0.917	1925-TC	0.956	1925-TC	0.933	1925-TC	0.940	1925-TC	0.948	1901-TC
6	0.910	1925-JC	0.949	1901-TNC	0.929	1901-TNC	0.933	1925-JC	0.940	1909-JC
7	0.907	1909-JC	0.948	1909-JC	0.927	1909-JC	0.931	1909-JC	0.939	1925-JC
8	0.902	1901-TNC	0.946	1895-TNC	0.923	1895-TNC	0.931	1901-TNC	0.936	1917-TNC
9	0.897	1895-TNC	0.946	1925-JC	0.921	1925-JC	0.925	1895-TNC	0.934	1895-TNC
10	0.887	1917-TNC	0.942	1917-TNC	0.914	1917-TNC	0.924	1917-TNC	0.932	1901-TNC
11	0.883	1895-JC	0.934	1909-TNC	0.913	1909-TNC	0.916	1925-TNC	0.929	1895-JC
12	0.883	1909-TNC	0.933	1895-JC	0.912	1895-JC	0.916	1895-JC	0.926	1925-TNC
13	0.881	1925-TNC	0.932	1925-TNC	0.905	1925-TNC	0.916	1909-TNC	0.921	1909-TNC
14	0.737	1894-JC	0.759	1894-JC	0.732	1894-JC	0.740	1894-JC	0.765	1894-JC

譲りたい。二つ目は、どの特徴量においても非コアデータである「*-TNC」の文書間距離の値が相対的に小さい[8]。これは自動解析誤りが文書間距離に影響を与えているものだと推察される。このことから、自動解析によるデータを大量に準備するよりも、少量の人手修正された翻訳小説・雑誌コーパス双方で準備することが信頼性の高い分析のためには重要であると考える。三つ目は、五作品が発表もしくは発刊された 1888 年（明治 21 年）から 1892 年（明治 25 年）に最も近いデータである『太陽』「1895-TC/TNC」と『女性雑誌』「1894/1895-JC」（明治 27、28 年）よりも 1901（明治 34）年、1909（明治 42）年との文書間距離の方が近い、つまり今回調査した特徴量においては 1894・95 年の文体よりも 1901 年・1909 年の文体の方に類似していることが読み取れる。

次に、表４～７の各分布について見ていく。

表４の品詞分布では、「洪水」以外で「1909-TC」との文書間距離が最も近い。「1909-TC」の次に文書間距離が近いデータセットは五作品すべてで異なっている。また、文書間距離の値の差分が「1894-JC」を除くと高々0.031 で抑えられ、ほぼ差がないといえる。次の表５と表６では、五作品それぞれ最も文書間距離の近いデータセットが異なっている。『罪と罰』のみ語彙素分布と出現書字形分布の文書間距離結果に差があり、他の四作品よりも値の小さい P 率（特に語彙素と出現書字形が一致する助詞）が影響しているものと推察される。最後に表７のバイグラム品詞分布だが、五作品すべてで「1909-TC」の文書間距離が最も近い。『罪と罰』と「綠葉歎」以外の三作品については、上位五データセットの文書間距離の近さが「1909-TC ＞ 1901-TC ＞ 1917-TC ＞ 1895-TC ＞ 1925-TC」の順で同じとなっている。『罪と罰』と「綠葉歎」については、上位二データセット「1909-TC ＞ 1917-TC」の順が同一である。また、他の表と比べて、文書間距離の差分が大きいことから、品詞バイ

[8] 表５「語彙素分布」の『罪と罰』のみ「1917-TNC」データセットの文書間距離が最も１に近いものとなっている。

グラム（2,495 次元）の特徴量が、データの分布を調べるのに最も適した粒度であったことが伺える（品詞・64 次元、語彙素・69,556 次元、出現書字形・106,609 次元）。

「1909-TC」にどのような記事が含まれているかというと、八サンプルすべて「文芸」の記事であり、一記事は中原青蕪による短編の翻訳である。このことから、「文芸」「小説」「文学」等のレジスタによる結果なのか、発行年代の文体による結果なのか、明確なことは指摘できないが、「翻訳小説」を「文芸」「小説」「文学」等のレジスタに含めるとすると、単純に 1909（明治 42）年前後に著された同レジスタのものに類似するという結果を重視する。

５．まとめ

本稿では、明治 20 年代の口語体翻訳小説五作品と『太陽』『女性雑誌』コーパスとの品詞比率、文書間類似度の比較を行った。３．１節では、樺島・寿岳(1965)の研究をもとに、N 率と MVR を図示化し、「あひゞき」「洪水」『罪と罰』は「ありさま描写的」、「玉を懐いて罪あり」「緑葉歎」は「動き描写的」な傾向性があることを明らかにし、『太陽』『女性雑誌』との関係があまり見られないことを示した。３．２節では、機能語を含んだ全体の品詞比率を示し、これまでの先行研究との関連性を確認したが、一方で『太陽』『女性雑誌』では N 率と P 率に相関が見られ、より詳細な調査は今後の課題とした。文書間類似度については、４．２節で五作品とも 1901 年・1909 年のデータと文書間距離が近く、品詞バイグラム分布においては五作品すべてで 1909 年のデータが最も似ているという結果が観察された。品詞の構成比率による文体的特徴（「ありさま描写的」「動き描写的」等）と文書間類似度との関連は見られなかった。

今後は、より具体的に言語現象と今回得られた結果との関連性を明らかにし、近代口語文の文体的特徴を明確に位置づけていくこととする。

謝　辞

本研究は、文部科学省科学研究費補助金若手研究(B)「近代口語文翻訳小説コーパスの構築と計量的文体研究」（平成 25～26 年度、領域代表者：小西光）による補助を得ています。

文　献

樺島忠夫(1955)「類別した品詞の比率に見られる規則性」『国語国文』24(6)、pp55-57

樺島忠夫・寿岳章子(1965)『文体の科学』綜芸舎

加藤百合(2012)『明治期露西亜文学翻訳論攷』東洋書店

小磯花絵・小椋秀樹・小木曽智信・宮内佐夜香(2010)「長単位情報に基づくジャンル間の文体に関する分析」『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（研究成果報告会）予稿集』、pp.183-190、国立国語研究所

森秀明(2014)「均衡性と代表性に配慮した『太陽コーパス』の分析法試論」『第６回コーパス日本語学ワークショップ予稿集』、pp.73-82、国立国語研究所

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規定集第４版（上）（下）』，特定領域研究「日本語コーパス」平成 22 年度研究成果報告書，国立国語研究所.

田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢(2012)『近代語コーパス設計のための文献言語研究成果報告書』国立国語研究所

山崎誠(2014)「言語単位と文の長さが品詞比率に与える影響」『第５回コーパス日本語学ワークショップ予稿集』、pp.233-242、国立国語研究所