

# 近代口語文翻訳小説コーパス構築の概要と計量的分析

著者	小西 光
雑誌名	国立国語研究所論集
号	11
ページ	37-61
発行年	2016-07
URL	<a href="http://doi.org/10.15084/00000840">http://doi.org/10.15084/00000840</a>

## 近代口語文翻訳小説コーパス構築の概要と計量的分析

小西 光

国立国語研究所 コーパス開発センター 非常勤研究員

### 要旨

現代日本語書き言葉の原型となる近代日本語の口語文体は、言文一致運動という意識的な文体変革から大きく動き出し、第二期国定教科書の完成・普及によって確立したとされる。近年、この近代口語文体の確立において、欧米文学の翻訳行為が大きな影響を及ぼしたと具体的に論じられ始めた。より詳細にその影響を明らかにするため、明治中期（明治16～30年）に発表された翻訳小説6作品と明治後期（同31～44年）・大正期に発表された創作小説2作品のコーパス構築を行った。

本稿では、まず構築したコーパスの概要と、それらの文の長さや文書間類似度の調査結果を示し、明治中期から後期にかけての文体の類似性について指摘する。本稿の調査では、文の長さ（一文における文節数の平均）や品詞比率、MVR（Modifier Verb Ratio）では近似の値を示し、文書間類似度では特徴的な結果は表れなかった。ただし、一文に含まれる接続助詞数のばらつきを調査すると、時代が下るにしたがってばらつきが小さくなるため、やはり時代による差があることは明らかとなった。

本稿で用いた手法によると、明治中期の翻訳小説と近代口語文体確立期の創作小説とに類似性が見出せることを示すことができた。これは、明治中期の段階で近代口語文体に近い文が産出され、それが読み手の目に触れていたことを意味しており、欧米の小説を翻訳することによる日本語への影響を示すこととなる\*。

**キーワード**：近代語、文体、口語文、翻訳小説、接続助詞

### 1. はじめに

『日本語大事典』（佐藤・前田2014）の見出し「言文一致体」は「明治初年から末期にかけて試みられた、規範文体の一つ。近代口語文体成立の母体（※下線筆者）となった文体である。言（話し言葉）と文（書き言葉）との懸隔をなくし、両者を近接させようとする言語改革のなかで生まれた規範文体」と定義されており、「口語体」は「現代一般に文章を書く際に用いられるスタイルで、文末に「である」「だ」「ます」「です」といった語尾をもつ。（中略）文語文に対立する概念であり、歴史的に言文一致運動の結果として定着（※下線筆者）したもの」と定義されている。これによると、明治時代の文体を時間軸上で考えた場合、始点としての「言文一致体」と終点としての「口語体」というものが明らかとなる。よって本稿では、明治前期（明治元～15年）に発生した言文一致運動の流れの中で試みられ、「近代口語文体成立の母体となった」文体を「言文一

\*本稿の内容は、第133回NINJALサロン（平成27年12月22日開催）での発表をもととしている。またその際いただいたご意見をもとに、修正を行った。

本研究は、JSPS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25～27年度、研究代表者：小西光、課題番号：25770178)による補助を得ています。また、本稿執筆にあたり、データ解析・分析に協力くださった浅原正幸准教授および加藤祥PDフェローに心より感謝申し上げます。そして、なにより素晴らしい研究環境を提供くださった国立国語研究所コーパス開発センターならびに所属する方々に対し、厚く御礼を申し上げます。

致体」とし、その「言文一致運動の結果として」明治 40 年代に確立されたものを「近代口語文体」とすることとする。当然ながらこれらはいつまでが言文一致体で、いつからが近代口語文体であるというふうに明確に線引きできるものではなく、むしろその変化を捉えることそれ自体が筆者の研究目的である。

現代日本語の書き言葉を考える際、その前身となった「近代日本語」がどのように変化・収束・定着していったのかを把握する必要がある。

国立国語研究所では現在も近代語のコーパス整備が続けられている。2015 年時点で公開されている「近代語のコーパス」は以下の四つである。収録データの発表順に①『明六雑誌コーパス』（明治 7～8 年、約 18 万語、[http://pj.ninjal.ac.jp/corpus\\_center/cmj/meiroku/](http://pj.ninjal.ac.jp/corpus_center/cmj/meiroku/)）、②『国民之友コーパス』（明治 20～21 年、約 101 万語、[http://pj.ninjal.ac.jp/corpus\\_center/cmj/kokumin/](http://pj.ninjal.ac.jp/corpus_center/cmj/kokumin/)）、③『太陽コーパス』（明治 20 年代後半～大正、約 1450 万字、博文館新社、国立国語研究所編 2005）、④『近代女性雑誌コーパス』（明治 20 年代後半～大正、約 210 万字、[http://pj.ninjal.ac.jp/corpus\\_center/cmj/woman-mag/](http://pj.ninjal.ac.jp/corpus_center/cmj/woman-mag/)）となっている。言語資料のサンプリングは、当時の書き言葉を代表するような「雑誌」を対象とし、対象となった「号」は全文がコーパス化されている。これらは、「近代語のコーパス」全体として代表性を志向するコーパスとなっている（田中 2012）。

近代日本を論じる際、欧米から流入した各種知識や技術をさし置いて論じることは不可能であり、それは文章・文体についても同じである。川戸（2014: 12-13）は「欧米の言語・作品をモデルとする新文体の創造ということが近代文学・近代文章語の形成を促した最大の要因であり、それこそが近代文学・近代文章語の成立の歴史をふりかえる上での最大の眼目となってしかるべきものである」と述べ、「欧米文学の直訳に基づく日本語の再構成」が起こったことを指摘している。しかしながら、これまで翻訳小説は日本語史の中で重点的に取り上げられることはなかった。そのため、その実態は明らかになっていない。「近代語のコーパス」の内にも翻訳小説は含まれているが、『国民之友コーパス』ではほぼ文語体による翻訳であり、口語体の翻訳小説は二葉亭四迷訳の「あいびき」のみである。川戸（2014）が指摘する欧米文学の影響を明らかにするためには、「近代語のコーパス」内のサンプルでは不十分であると考え、「近代口語文翻訳小説コーパス」を構築することとした。本コーパスを用いて、新文体の原型となった「言文一致体」を計量的に捉えていきたい。

以下では、2 節で近代日本語と文体、文体と翻訳文学について概観し、3 節で構築したコーパスの概要を示す。4 節で構築したコーパスおよび他のコーパスを用いた文体分析の試みと結果について述べ、5 節で全体のまとめと将来の展望について触れる。

## 2. 明治期の文章と翻訳文学

### 2.1 明治期と文体

江戸時代に入ると、書き手と読み手の幅が大きく拡大していき、日本語の文体は加速度的にその多様性を増してきた。自分と相手との身分や性別、関係性、状況・目的に応じた文体の選択肢は多岐にわたり、それらを踏襲した明治時代初期においても同様の様相であった。しかし主流と

なる漢文が確固として存在していたため、文体に迷うことはなかった。ところが明治時代に入ると、西洋文物の影響から自国の言語を見直すようになり、試行錯誤の末、現代書き言葉の基礎となる文体が確立され、広く日本中に定着していくこととなる。森岡（1991）では、明治時代に用いられていた文体を実用文系統と文学系統とに二分し、その変遷を辿りながら文体の多様性を表1のようにまとめている。これによると、明治初期では8種に分類された文体が、明治40年代に至って「言文一致体」（本稿の「近代口語文体」）へと収束していく流れが見られる。

表1 明治期における文体変遷（森岡 1991: 19 の系統図より）

文の種類		明治初期	明治10年代	明治20年代	明治30年代	明治40年代
実用文系統	文語	漢文訓読体	和漢折衷体	明治普通文		言文一致体
		和漢折衷体				
		候文				
	口語	問答体	演説体	初期言文一致体		
		講述体				
		談話体				
文学系統	口語	俗文体	講釈体	初期口語体	初期言文一致体	
	文語	和漢折衷体	雅俗折衷体		(雅俗混合体)	

森岡（1991）は、明治20年代の文学系統－口語体では「初期口語体」が主に用いられ、明治30年代では「初期言文一致体」が主に用いられていたとする。「初期口語体」「言文一致体」という用語が本稿とは逆となっているが、その指し示すものに違いはないと考える。

共通語としての日本語書き言葉がいつ成立したかという問題について森岡（1991: 64）は「言文一致体（東京語・標準語）の成立・完成という時点で、日本語というラングが変化したと見たい。（中略）正確には室町期から文法体系の変化が現れ始め、明治四十二年に至ってそれが完成したと言うべきだと思う。明治四十二年というのは、山本正秀氏が小説系統の文体から推して言文一致体の成立と見た年であり、私も第二期国定教科書がこの年に出たことをもって言文一致体の成立と考えるものである」と、それまでは各種の言語変化についてあくまでパロールの変化であるとしていた姿勢から一歩踏み込み、口語文法の確立を経た明治43年（第二期国定教科書の使用開始）時点で日本語のラングが変化したと位置づけている。

このように明治時代は、日本語の変化にとって大きな変化の仕上げとなった時である。

### 2.1.1 口語系文体と文語系文体

明治時代に入ると時代が「共通語」を要請するようになった。万民に伝わる共通性と平易さを求め、卑近な議題から抽象概念まで欧米語のように自由に論じられる日本語のモデルを探り、口語系「言文一致体」と文語系「明治普通文」の二大潮流が生まれていく。

口語文体の成立について森岡（1991: 32）は、「書き言葉の言文一致体（現代の口語文）も話し言葉の標準語（現代の全国共通語）も、次に述べる小説系の口語文よりも、この種の一対多のコミュ

ニケーションの伝統から発達し、形成されてきた」と指摘している。一対一の〈話すまま書く〉という談話体から一対多の演説体へ、演説体から論文や叙事文といった〈文章にも耐えうる〉口語文体へと、そのモデルは推移していった。しかし談話体も演説体も、あくまで基本は話し言葉であり、文章の言葉とするにはそこから待遇表現を省く必要があった。そこで待遇表現を取り除くと、その文末は四迷の言う「結尾の聞き苦しき」「嫌悪せらるべき」ものとなってしまう、書き手にしても読み手にしても受け入れがたいものだったというのが当時の位置づけである。森岡(1991: 43)においても「言文一致体の成立は、これまで文学系統の資料によって裏付けることが試みられたが、口語小説は遅れて出発した（※下線筆者）し、作者の文学意図によって初期には個人ごとに文体が異なり、どうやら口語小説の文体に共通性が生じてくるのは、明治三十年末もしくは四十年を過ぎてから（※下線筆者）である。しかも、その共通性は『吾輩は猫である』に象徴されるように演説の文体を取り入れることによって成立したと見られる」とあるように口語文体は、結果的に耳馴染むまでに20年程度を要したのである。

一方、文語文体については口語文体に先んじて「明治普通文」が発達し、漢文系の流れも汲んだその思考・発想や論理は、形而下から形而上まで現代的論説に耐えうるものであった。つまり、文末辞を替えれば言文一致体となり、その内容は現代語と相違ない論理的な展開が可能となっていた。それは当時言論の場で活躍していた人々の思考の成熟を意味していると言える。森岡(1991: 27)は「言語の背後にある思考法は、明治普通文と言文一致体とで著しく共通して」いるとし、明治時代の日本語は、先んじて文語文体の枠の中で、現代語と同様の文の構造・文章の構造を獲得していったと言えよう。

### 2.1.2 文学系統と実用文系統

文学系統の文体と実用文系統の文体は、その文章を書く目的自体が異なるように、文体にも差異が生じる。当時、文学の目的は風景・心理の「写実・描写」であり、実用文の目的は情報や主張の「伝達」であった。実用文系統の文語系文体も口語系文体も、また文学系統の文語系文体も、前時代から引き継いできた言語モデルがあったため、そのモデルを土台としていけばよかったが、文学系統の口語系文体については、目指すべき理想のモデルがすぐには見つからなかった。2.1.1節で一対多の演説体が口語文体のモデルとなったと述べたが、口語文体の発展については、芸術としての文学と文学者たちの功績が外せない。文学と文体については、次節で述べることとする。

ただ、明治時代に直面した文体の問題は、文学系統のものも実用文系統のものも共通しており、それは前時代の措辞やレトリック、文の接続構造といった強力な「縛り」をいかに脱していくかというものであった。特に文学は、いわゆる「美文」的な枠組みからの脱却を迫られることとなった。坪内逍遙は1886(明治19)年に「文章新論」と題して、「無用の雅言を廃すべき」「文章の音調の滑かならん事を主とせず」「古人の糟粕をやきなほし若くは其儘に切抜き来りて巧に新しく綴り合はすことを以来さつぱりと廃すべき」との具体的な三点を挙げて、目指すべき新しい文章を提唱している。

## 2.2 文体と文学

2.1 節では、明治時代と文体という枠でその流れを確認したが、次はそこから「文学」に絞って明治時代の文体をまとめたい。山本（1965）では、言文一致運動の流れを文学系統の資料により7区分に分けている。今回構築したコーパスは、以下の「第二期 第一自覚期」に発表された資料が中心となる。この第一自覚期は、明治19年に初めて二葉亭四迷と山田美妙によって言文一致体の小説が試みられ、それを皮切りに30名もの文筆家が言文一致体に挑むが、一方でその筆力の低さと読み手の馴染みの薄さから反言文一致の気運も高まり、言文一致論争ひいては未来の日本語文章についての論議が盛り上がった時期である。当時の言文一致体は、「洗練が足らず文章として未熟だった」（山本1965）ため、「第三期 停滞期」には、反動から文語系の雅俗折衷体が主流を占めるようになった。そんな流れの中でも「である体創始者」として名高い尾崎紅葉が言文一致体に着手し始め、「第四期 第二自覚期」に、そのこなれ始めた言文一致体を新聞小説として普及させていくのである。

第一期	発生期	慶応 2 年 - 明治 16 年	1866-1883
第二期	第一自覚期	明治 17 年 - 明治 22 年	1884-1889
第三期	停滞期	明治 23 年 - 明治 27 年	1890-1894
第四期	第二自覚期	明治 28 年 - 明治 32 年	1895-1899
第五期	確立期	明治 33 年 - 明治 42 年	1900-1909
第六期	成長・完成前期	明治 43 年 - 大正 11 年	1910-1922
第七期	成長・完成後期	大正 12 年 - 昭和 21 年	1923-1946

（言文一致運動の時期区分（山本1965: 33）より）

そして「第五期 確立期」に入ると、自然主義文学者たちの「内面的必要」から言文一致体が不可欠となり、その中で近代口語文体が確立されていくのである。

このように明治20年代には、「近代「文学」の発生、文章についての価値観の変化があり、近代小説という制度が出来上が」（齊藤2012: 224）り、「写実と描写に重きを置いた近代小説の文体上の必要から言文一致が求められ、以後外国の場合同様、言文一致の文章は、主に小説家の努力によって推進され発達を見た」（山本1965: 41）のである。

以上の流れから、近代口語文体成立の過程を把握する上で、やはり言文一致体を発達させた文学の実相をひとつ大きな軸とすべきと考える。

## 2.3 文学と翻訳

これまで2.1 節では、口語文体へ文体が統一されていく流れとともに、その過程で①演説体から待遇表現を除くこと、②論理的な思考を獲得すること、③前時代的な措辞やレトリック、文の接続構造にとらわれなくすることが不可避な要請であったことを確認し、2.2 節では、文体改良と文学の実践が密接であったことを確認した。そして本節では、これらすべてを可能にしたのが、欧米文学の「翻訳」という行為であったと述べる。

柳父ほか (2010: 38) も「近代文章史は欧文脈摂取の歴史 (中略)。この点で従来の文学史は翻訳の役割を過小評価している」と言うように、これまで「翻訳文学」というものは、近代文体史の傍流とされてきたと言ってよいだろう。欧米言語の影響により言文一致体の模索が起り、口語文体の完成が促されたことは、論を俟たない。「漱石や鷗外や、花袋・藤村など自然派のやや年輩の作家の文章には、文章語的口語体といった明治的限界があった。(中略) 前近代的な漢文式措辞法からの脱出や俗談的冗長性の克服、一方細叙にたけた欧文脈の一層のとりこみによる補強などが、近代口語文体の完成のために必要であった」(山本 1981: 49) とあるとおりである。

言文一致運動の中で言文一致体を試みた主流の作家たちは、いずれも欧米語に通じ、小説家である前に翻訳家であった<sup>1</sup>。

明治時代の文学界について、水野 (2007: 4) は「まず「翻訳-創作文学」という大きなシステム間の対立関係があり、翻訳システムの内部では「直訳-意識」という規範の競合が存在した。(中略) 文学的多元システムの中心にあった翻訳システムが相対的に未成熟な創作文学システムに影響を及ぼしていると想定」しているとする (図 1)。水野 (2007) によると、当時の翻訳システムは創作システムよりも優位 (中心) にあったのである。

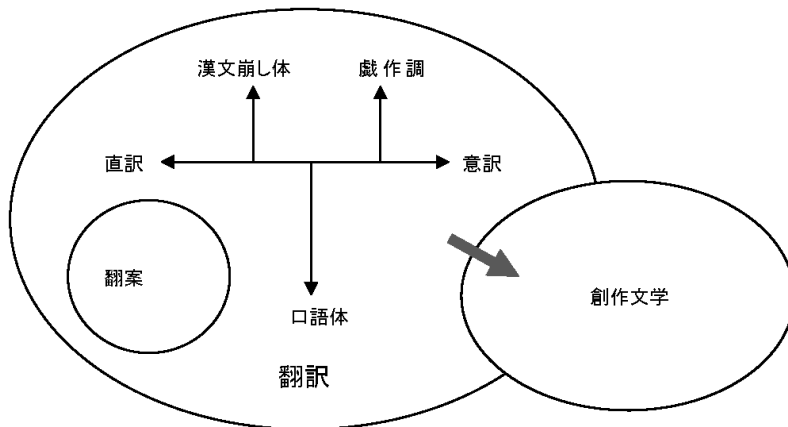


図 1 明治中期の文学的多元システム (水野 2007: 5 より)

欧米の小説の構造や、何を描き何を描かないのかといった描写法、レトリックに及ぶまで、読者として受けた趣致を日本語に落としこむ作業から、近代文学も言文一致体も創造されていったと言えるだろう。しかし一方で、まだ当時の作家たちには欧米作家たちの「文体を自由に再現するだけの「筆力」がなかったので、従来どおり逐語訳するよりなかった」(柳父ほか 2010) ので

<sup>1</sup> ①二葉亭四迷：東京外国語学校露語科 ②山田美妙：紅葉と幼友だち。硯友社メンバー。明治 21 (1888) 年に「正本はむれっと」、明治 24 (1891) 年に「魔王来」とシェークスピアを邦訳 ③尾崎紅葉：魯庵に「罪と罰」を教えた本人。明治 22 (1889) 年に「恋山賤 (こいのやまがつ)」(ゾラ) を翻訳 ④嵯峨の屋おむろ：明治 21 (1888) 年に「薄命のすゝ子」を連載。四迷の同級生。東京外国語学校露語科 ⑤森鷗外：4 年弱のドイツ留学。明治 21 (1888) 年に短編を邦訳 ⑥内田魯庵：明治 25 (1892) 年に『罪と罰』初邦訳 ⑦若松賤子：明治 23 (1890) 年に『小公子』初邦訳

ある。ただ、逐語訳しただけでは日本語としての完成度は低くなり、文学としては不十分である。翻訳文学の場合は、原作の雰囲気崩さず、その趣までを訳出する必要性があり、当時の翻訳者が《自在に》訳文（日本語）をコントロールするのは困難だったであろう。それゆえに、言文一致体翻訳小説は粗も目立つ。その点、4年弱ドイツに留学していた鷗外は、ドイツ語が内在化されているためか、翻訳文の日本語化に「余裕が感じられる」（加藤 2012）ものの、やはりぎこちなさがないわけではない。まして、欧米文学と同様の「詩想」や「インプレッション」を有する日本語文章を早々自由に生み出せるわけがないのである。

しかし、「直訳による不透明な翻訳表現は、目標言語の表現規範と衝突し、その文学言語システムのあり方いかんでは規範を揺り動かし、言語の表現可能性を拡大する可能性がある」と水野（2007）が指摘するように、直訳的な《ぎこちない文体》は日本語としての破格を許容し、清新な文体の誕生を促すものでもある。

以上見てきたように、文学の翻訳行為は、未来への文章彫琢の方向性を示し、日本語としての破格を広める役割を大いに担っていた重要な資料と考えられるのである。

### 3. 近代口語文翻訳小説コーパス構築の概要

筆者が「近代口語文翻訳小説コーパス」として作成したコーパスの概要を以下に示す。

#### 3.1 選定資料

コーパスに収録する翻訳小説作品は、表2の6作品とした。資料選定にあたり、「言文一致体」の初期段階を捉えるという基本姿勢から、言文一致運動が文学界で盛り上がった「第一自覚期」（～明治22年）に「言文一致体」で「翻訳」（翻案ではなく）された作品を中心とした。「翻訳」とは、原作の筋・物語世界の人物や事物・文体を改変せずに訳しているという意味である。これらは、厳密な逐語訳に限らず、やや意識を含むものも含んでいる。また、本コーパスにおいては、代表性よりも資料性を優先した。

表2 収録作品 - 1（言文一致体翻訳小説作品）

作品	原作者	訳者	原語	初出年	初出媒体
あいびき	ツルゲーネフ	二葉亭四迷	露語	明治21（1888）年7/8月	『国民之友』25-26
めぐりあひ	ツルゲーネフ	二葉亭四迷	露語	明治21（1888）～22（1889）年	『都の花』1,3-6
玉を懐いて罪あり	ホフマン	森鷗外	独語	明治22（1889）年	『読売新聞』
洪水	ブレット・ハート	森鷗外	独語	明治22（1889）年	『柵草子』
緑葉歎	ドオデー	森鷗外	独語	明治22（1889）年	『読売新聞』
小説罪と罰 卷一	ドストエフスキー	内田魯庵	英語	明治25（1892）年11月	単行本

次に、上記翻訳小説作品と対照するための創作小説作品は、表3の2作品とした。



表3 収録作品 - 2 (近代口語文体創作小説作品)

作品	原作者	初出年	初出媒体
破戒	島崎藤村	明治 39 (1906) 年	自費出版単行本
高瀬舟	森鷗外	大正 5 (1916) 年	『中央公論』

次に、電子テキスト化した際に使用した底本を表4に示す。

表4 使用底本

作品	種類	底本
あいびき	翻訳	『明治文學全集 17 二葉亭四迷・嵯峨の屋おむろ集』(筑摩書房)
めぐりあひ	翻訳	
洪水	翻訳	『水沫集』(初版)〈近代デジタルライブラリー〉
緑葉歎	翻訳	
玉を懐いて罪あり	翻訳	
小説 罪と罰	翻訳	『小説 罪と罰 卷一』(初版)〈近代デジタルライブラリー〉
破戒	創作	『現代日本文学大系 13 島崎藤村集 (一)』(筑摩書房)
高瀬舟	創作	『山椒大夫・高瀬舟』(岩波文庫)※新字新仮名

### 3.2 選定理由

3.1 節に挙げた8作品の選定理由について、以下に示す。

二葉亭四迷による翻訳2作品「あいびき」「めぐりあひ」は、言文一致運動の「第一自覚期」(明治17～22年)に発表され、当時文壇では言文一致体逐語訳による画期的な試みとして受け入れられ、言文一致文体の火付け役的存在となった。翻訳・創作行為において意識的に「言文一致」を実現させようとする苦心については、四迷本人による手記等が残っている。また「あいびき」「めぐりあひ」の翻訳は、「言文一致による新しい散文の世界を、花袋や独歩・有明らの当時の文学青少年の前に開いて見せた」(山本1965)と評価され、言文一致体近代小説の端緒とされる『浮雲』第三篇執筆への影響も指摘されている(加藤2012)。文学における初期言文一致体の翻訳作品として外せないものである。

森鷗外による『水沫集』は、鷗外が4年弱のドイツ留学からの帰国後、翻訳家として文学活動を開始した際に翻訳また創作したアンソロジー集(短編作品全20編。附録含む)である。初出は表2に示したとおりだが、翻訳が17編、「舞姫」を含む創作が3編と翻訳作品中心の作品となる。翻訳17編の作品種類としては、散文が14編、韻文1編(附録「於母影」)、戯曲2編という構成である。その翻訳散文作品14編のうち、文体の種別は、口語常体3編、口語敬体1編、文語体10編となっている。つまり、当時の鷗外の翻訳姿勢は、基本的には文語文体である。しかも加藤(2012)によると、作品の趣致を乱すような訳になってしまう箇所は省筆するという、柔軟な「意識」を行っていた。鷗外の翻訳姿勢について加藤(2012: 117-118)は、「鷗外は、個々の単語の対応に拘ることなく、翻訳をいわばスタイルの習得に役立つと考えて実験的に翻訳した」

とし、その証拠として「原典の文体の特色の違いによって、鷗外の翻訳の文体がくっきりと違ってきている」と翻訳の対象を文体の問題として捉えていたことを指摘している。このことから、文学作品の翻訳を通して、意識的に近代小説・文体を取り入れようとしていた点では四迷と同じであり、「た」調常体の完全な言文一致訳で、(中略)当時の翻訳文体中出色の出来ばえ(山本 1965: 582)との評価も高く、その日本語としての〈こなれ度〉という点では四迷より優っていると言えるだろう。鷗外は、明治 23 年以降言文一致体を用いることがなくなったものの、明治 42 年から再び言文一致体を採用して執筆活動を続け、「作家」として大正期まで長期的に活躍する重要な人物である。

内田魯庵による『小説 罪と罰 卷一』(以下、『罪と罰』)は、第 1 章から第 10 章で構成されている。これは、ドストエフスキーの初邦訳作品であり、『罪と罰』の冒頭からの一部分にあたる。また魯庵は、その「緒言」に英訳版『Crime and Punishment』からの重訳であること、友人・二葉亭四迷の協力を得た(つまりロシア語原文も参照していた)ことを記している。それによって、基本的には原文に忠実な訳となっている。上梓された時期は山本(1965)によって「停滞期」(明治 23～27 年)とされた時期にあたるものの、文体としては「である体」を積極的に用いており、「明治の翻訳小説中、これほど深大な感化影響を與えた作はない。それは二葉亭の『あひびき』と並んでまさに双璧である」(木村 1972: 401)とされ、北村透谷や島崎藤村等に影響を与えているという点では、重要な作品である。

「近代口語文体」の代表作として選んだ 2 作品についても、その選定理由を述べる。

島崎藤村著『破戒』は、日本自然主義文学の先駆けと位置づけられ、欧文脈を意識的に取り込んだ文体が特徴的であり、魯庵訳『罪と罰』との影響関係も指摘される(木村 1972)。

森鷗外著「高瀬舟」は、作品の舞台は江戸時代に置かれているが、『破戒』発表より 10 年が経過した大正期の短編作品である。鷗外が藤村や花袋といった自然主義文学の流れとは異なる独自の文体・作風を展開していた点、また本コーパス収録の翻訳 3 作品の発表より 27 年経過した作品という点で収録対象とした。

表 5 に、8 作品が何人称小説であるかと、あらすじをまとめた。

表 5 作品別の人称とあらすじ

作品	人称	あらすじ
あひびき	1	「自分」が目撃した男と女が忍び会う様を語る
めぐりあひ	1	「自分」がかつて一目惚れした女との再会
洪水	1	冒頭の自然描写と「私し」が女から聞いた話を語る
緑葉歎	3	かつて恋仲にあった少女と少年の出逢いと再会
玉を懐いて罪あり	3	ある殺人事件をめぐる女学士の働きとその人間模様
罪と罰	3	老婆殺害事件をめぐる主人公の苦悩
破戒	3	出生を隠し続ける主人公の苦悩
高瀬舟	3	流罪中の船上で交わされる罪人と同心の二人語り

### 3.3 構築手順

コーパス構築の手順を以下に示す。

未電子テキスト化の資料は、全文を国立国語研究所開発「近代語のコーパス」と同仕様（近藤 2014）で電子テキスト化し、それらテキストに形態素解析（MeCab-0.996）をかけ、人手にて解析結果の修正を行った。単位の認定・品詞判定については、小椋ほか（2011）を参考にした。翻訳小説 6 作品については中・長単位解析器「Comainu-0.71」で自動長単位解析を行い、係り受け解析器「CaboCha-0.69」を用いて自動係り受け解析を行った（いずれも人手修正は行っていない）。表 6 にテキスト情報（ルビ情報・文字種）と文字入力の仕様、表 7 に使用した形態素解析辞書をまとめた（以下、「玉を懐いて罪あり」を「玉を懐いて」と示す）。

表 6 ルビ情報・使用文字種情報・文字入力仕様

作品	ルビ	文字種	文字入力の仕様
あいびき	パラ	漢字・平仮名混じり	文字集合：JIS X 0213 のうち、(1) 康熙字典掲字、(2) UCS 互換字、(3) CJK 統合漢字拡張 B に符号位置が割り当てられる文字、を除外した範囲とする。この範囲にない文字は外字として「■」で入力する。 包摂規準：JIS X 0213 に準拠
めぐりあひ			
洪水	なし		
緑葉歎			
玉を懐いて	総ルビ		青空文庫のテキスト入力基準に従う (青空文庫「作業員手帳」 <a href="http://eunheui.sakura.ne.jp/aozora/">http://eunheui.sakura.ne.jp/aozora/</a> )
罪と罰	パラ		
破戒	パラ		
高瀬舟	パラ		

表 7 使用した形態素解析辞書

作品	形態素解析辞書
あいびき	旧仮名口語 UniDic (2014 年 8 月 内部公開版)
めぐりあひ	
洪水	
緑葉歎	
玉を懐いて	
罪と罰	近代文語 UniDic-1.3(旧仮名口語 UniDic 開発前の 2009 年解析のため)
破戒	解析済みデータを人手修正 (2014 年 5 月)
高瀬舟	

### 3.4 基礎統計量

最後に 8 作品における地の文のみの基礎統計量を表 8～11 で示す。ただし、『破戒』については現在データ修正中のため、四分の一の分量にあたる人手修正が完了している箇所までのデータを示した（平成 28 年 3 月修正完了予定）。

#### 3.4.1 文の長さ

服部（2008, 2011, 2012, 2013）では、明治 20 年代前半発表の言文一致体創作小説 4 作品におい

る「文の長さ(文長)」(文節数/文数)が算出されている。それによると各作品の文の長さは、『浮雲』第一篇 20.2, 第二篇 14.8, 第三篇 15.6, 「武蔵野」(山田美妙・初出) 14.3, 「武蔵野」(同左・単行本) 14.6, 「薄命のすゞ子」(嵯峨の屋おむろ) 14.0 となっており, 全体の平均は 15.6 となる(『浮雲』第一篇のみ値が大きいので, これを除外すると 14.7 となる)。一方, 表 8 の小計で計算すると, 翻訳小説 6 作品の平均文長(文節数/文数)が 11.9, 『破戒』と「高瀬舟」の平均文長が 10.4 と, 翻訳小説と同時期の言文一致体創作小説より 3~4 文節程度短く, 翻訳小説および近代口語文体小説の方が, 文が短いという結果になる。

表 8 文数・文節数・短単位数および平均文長(短単位数/文数)

作品	文数	文節数	短単位数 <sup>2</sup>		平均文長 (延べ短単位数/文数)
			延べ語数	異なり語数	
あいびき	123	1,692	3,634	1,059	29.5
めぐりあひ	655	6,587	14,792	2,413	22.6
洪水	124	1,875	3,926	904	31.7
緑葉歎	81	912	1,967	560	24.3
玉を懐いて	340	4,431	9,372	1,718	27.6
罪と罰	1,097	13,320	27,487	3,955	25.1
小計	<b>2,420</b>	<b>28,817</b>	<b>61,178</b>	<b>10,609</b>	<b>25.3</b>
破戒	1,136	11,642	24,965	3,324	22.0
高瀬舟	104	1,281	2,943	687	28.3
小計	<b>1,240</b>	<b>12,923</b>	<b>27,908</b>	<b>4,011</b>	<b>22.5</b>
計	3,660	41,886	89,086	14,620	24.3

表 9 文の長さ(文節数/文数)と文節の長さ(短単位数/文節数, 文字数/文節数)の平均・標準偏差

作品	文節数/文数		延べ短単位数/文節数		文字数/文節数	
	平均	標準偏差	平均	標準偏差	平均	標準偏差
あいびき	13.76	11.62	2.15	0.87	3.60	1.69
めぐりあひ	10.06	12.37	2.25	1.09	3.52	1.97
洪水	15.12	15.70	2.09	0.90	3.16	1.48
緑葉歎	11.26	7.76	2.16	0.85	3.29	1.55
玉を懐いて	13.03	9.03	2.12	0.86	3.25	1.60
罪と罰	12.14	8.64	2.06	0.91	3.28	1.83
破戒	10.25	6.45	2.14	0.94	3.25	1.48
高瀬舟	12.32	7.82	2.30	1.11	3.66	1.74
BCCWJ LBa	6.36	4.65	2.34	1.13	3.92	1.92

表 9 および次頁の図 2 では, 参考に BCCWJ (『現代日本語書き言葉均衡コーパス』) 非コア・

<sup>2</sup> 空白・補助記号は除く。

図書館サブコーパス (LB) の一部データ (LBa データの内、小説のみ。以下「BCCWJ LBa」) を参考に付与した。これは会話文も含んだデータとなっているため、文の長さが短く出ているものと考えられるが、それを勘案しても明治期の小説よりは文の長さが短くなっていると言えるだろう。また、文節の長さに差はないため、1 文節の中の語数は固定的であると言えることができる。

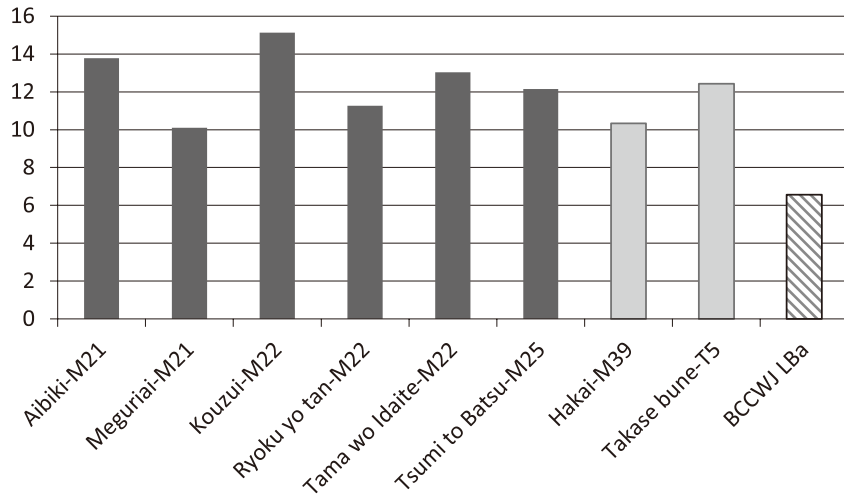


図2 平均文長 (文節数 / 文数)

### 3.4.2 品詞分布

次に、8 作品の品詞分布 (延べ語数) とその比率を示す (表 10・11)。品詞分布は、翻訳小説 6 作品についても、創作小説 2 作品についても、また表には入れていないが「BCCWJ LBa」についても、ほぼ同様の分布となっており、比率において有意な差を観察することはできない。公開されている品詞構成表で BCCWJ のレジスタ別 (白書・新聞・書籍・ブログなど) に品詞比率を見ると、レジスタ特有の差異を確認することができる。これより、類似した資料性をもつ文書での差異はよほど差のある文体でないかぎり、差が表れにくいと考えられる。

表 10 品詞分布 (延べ語数)

作品	名詞	動詞	副詞	形容詞	形状詞	接続詞	助詞	助動詞	その他	計
あいびき	830	669	152	115	99	40	1,212	437	80	3,634
めぐりあひ	3,560	2,574	537	353	267	134	4,807	2,006	610	14,848
洪水	993	700	120	117	55	28	1,391	399	123	3,926
緑葉歎	551	312	43	56	33	6	678	225	63	1,967
玉を懐いて	2,707	1,529	207	183	111	28	3,315	914	378	9,372
罪と罰	7,132	4,907	1,203	362	451	234	8,709	3,261	1,041	27,300
破戒	7,150	4,142	628	561	390	77	8,679	2,576	750	24,953
高瀬舟	742	547	74	50	35	41	1,063	343	73	2,968

表 11 品詞比率

(%)

作品	名詞	動詞	副詞	形容詞	形状詞	接統詞	助詞	助動詞	その他	計
あいびき	22.8	18.4	4.2	3.2	2.7	1.1	33.4	12.0	2.2	100
めぐりあひ	24.0	17.3	3.6	2.4	1.8	0.9	32.4	13.5	4.1	100
洪水	25.3	17.8	3.1	3.0	1.4	0.7	35.4	10.2	3.1	100
緑葉歎	28.0	15.9	2.2	2.8	1.7	0.3	34.5	11.4	3.2	100
玉を懐いて	28.9	16.3	2.2	2.0	1.2	0.3	35.4	9.8	4.0	100
罪と罰	26.1	18.0	4.4	1.3	1.7	0.9	31.9	11.9	3.8	100
破戒	28.7	16.6	2.5	2.2	1.6	0.3	34.8	10.3	3.0	100
高瀬舟	25.0	18.4	2.5	1.7	1.2	1.4	35.8	11.6	2.5	100

#### 4. 計量的文体分析の試み

本節では、構築したコーパスの短単位データを元に、MVR (Modifier Verb Ratio)、文書間類似度、接統助詞の配列について調査結果をまとめる。

##### 4.1 MVR 比較

樺島・寿岳 (1965) は文書の特徴を表す指標として「相の類 (形容詞・形容動詞<sup>3</sup>・副詞・連体詞) の語数 ÷ 用の類 (動詞) の語数 × 100」の式で表される「MVR」という値を提案し、MVR と名詞率 (N 率) の対比によって文体的な特徴を捉えようとしている。この MVR の値が高いと、様態描写が多く、低ければ動作描写が多いということになる。

N 率 (大, 54-56)	・	MVR (小, 34-41)	… 要約的文章
N 率 (小, 45-48)	・	MVR (大, 55-65)	… ありさま描写的文章
N 率 (小, 45-48)	・	MVR (小, 34-41)	… 動き描写的文章

(樺島・寿岳 1965: 36, 130 より作成)

表 12 および図 3 にその結果をまとめた。

3.4.2 節の品詞比率であまり有意な差が観察されなかった以上、名詞率に対する MVR の分布においても同様の結果となる。よって、品詞の分布という点においても、8 作品は前節同様似たような特徴を持つ文書であると言える。

<sup>3</sup> 国語研の開発した形態素解析辞書 UniDic には、「形容動詞」に相当するものとして「形状詞」が認定されている。本調査では、この「形状詞」を用いた。

表 12 N 率に対する MVR の分布

作品	N 率	MVR
あいびき	22.84	48.19
めぐりあひ	23.98	53.42
洪水	25.29	54.00
緑葉歎	28.01	54.49
玉を懐いて	28.88	42.45
罪と罰	26.12	51.01
破戒	28.65	46.04
高瀬舟	25.00	36.75

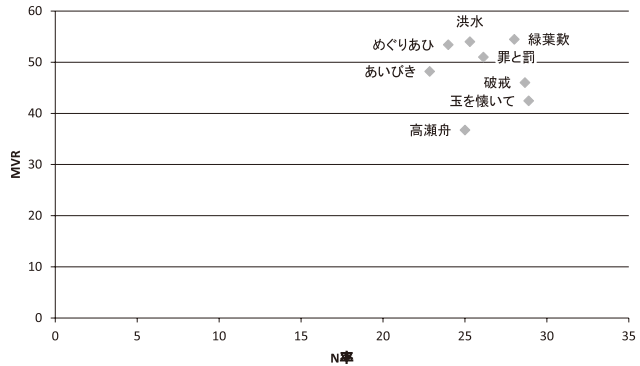


図 3 N 率に対する MVR の分布

#### 4.2 文書間類似度 (コサイン類似度)

次に、8 作品同士の文書間類似度を確認する。小西 (2015) では、同一の手法を用いて『太陽コーパス』『女性雑誌コーパス』(いずれも近代語のコーパス)と翻訳小説 5 作品(「めぐりあひ」以外。会話文も含む)の文書間類似度を見た。特徴量の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特徴量ベクトルを  $\vec{s}$  とし、比較される文書の特徴量ベクトルを  $\vec{t}$  とすると、コサイン類似度は以下の式で表される。

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

文書間類似度 (コサイン類似度) は、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 に近い値をとる<sup>4</sup>。

用いた特徴量は品詞・語彙素・書字形出現形 (以下、書字形)・品詞バイグラム<sup>5</sup>の分布である。小西 (2015) ではいずれの特徴量においても、大きな差が見られなかったものの、1900 (明治 33) 年以降の『太陽コーパス』コアデータ (人手修正済みの高精度データ。口語文のみ) が上位に来る結果となった。品詞バイグラム分布では、5 作品すべてで 1909 (明治 42) 年の『太陽コーパス』コアデータとの類似度が最も高く、品詞バイグラム分布の粒度がデータを比較する際に適した粒度ではないかと述べた。しかし、1909 年の『太陽コーパス』コアデータは、サンプルすべてが「文芸」の記事というデータ自体の問題があり、同レジスタの文書ゆえに類似したと考えられた。本節では 8 作品に「BCCWJ LBa」の一部データを加えて作品間のコサイン類似度を算出した。

用いた特徴量は、品詞、品詞バイグラム (2-gram)、語彙素、語彙素バイグラム、書字形、書字形バイグラム、接続助詞のみのバイグラム・トライグラム (3-gram) の八つである。バイグラムとは隣接した二つ組のことで、「蓮華寺では下宿を兼ねた」とあった場合、語彙素バイグラム

<sup>4</sup> 品詞分布では、「空白」と「補助記号-\*」を排除した。UniDic 品詞体系の分類「[大分類]-[中分類]-[小分類]」のうち、小分類まで用いた。

<sup>5</sup> 品詞バイグラム分布では、文の先頭要素「BOS」を含む対は特徴量としては排除した。

では「蓮華 - 寺」「寺 - で」「で - は」「は - 下宿」「下宿 - を」「を - 兼ねる」「兼ねる - た」という二つ組を作り、その組み合わせ同士の類似度を比較するというものである。またトライグラムは同様にして作った三つ組みのことである。

#### 4.2.1 品詞・語彙素・書字形分布の文書間類似度

文書間類似度の最高値と最低値の差に着目する。品詞分布の類似度の差は 0.022 (Max. 0.994「あいびき」対『罪と罰』, Min. 0.972「めぐりあひ」対「玉を懐いて」), 語彙素分布の類似度の差は 0.053 (Max. 0.967「洪水」対「玉を懐いて」, Min. 0.914「あいびき」対「BCCWJ LBa」), 書字形分布の類似度の差は 0.087 (Max. 0.969「緑葉歎」対「玉を懐いて」, Min. 0.882「めぐりあひ」対「高瀬舟」) となっている。このように全体で類似度に差が出ないということは、いずれの特徴量においても全 8 作品が類似しているとの結果でもある。

#### 4.2.2 品詞・語彙素・書字形バイグラム分布の文書間類似度

4.2.1 節と同じように文書間類似度の最高値と最低値の差に着目する。品詞分布の類似度の差は 0.048 (Max. 0.982「あいびき」対『罪と罰』, Min. 0.934「あいびき」対「BCCWJ LBa」), 語彙素分布の類似度の差は 0.263 (Max. 0.880「洪水」対「玉を懐いて」, Min. 0.617「あいびき」対「BCCWJ LBa」), 書字形分布の類似度の差は 0.268 (Max. 0.849『罪と罰』対『破戒』, Min. 0.581「あいびき」対「BCCWJ LBa」) となっている。バイグラムになると、どの特徴量においても「あいびき」と「BCCWJ LBa」の文書間類似度が低いことが分かる。

小西 (2015) で類似度を比較する際に適した粒度とされた品詞バイグラム分布の文書間類似度を表 13 に示した。ここからは、特に特徴的な傾向を指摘することはできない。

表 13 品詞バイグラム分布による文書間類似度

	あいびき		めぐりあひ		洪水		緑葉歎		玉を懐いて		罪と罰	
1	罪と罰	0.982	罪と罰	0.976	玉を懐いて	0.982	玉を懐いて	0.980	洪水	0.982	あいびき	0.982
2	洪水	0.978	洪水	0.973	破戒	0.978	洪水	0.972	罪と罰	0.980	玉を懐いて	0.980
3	玉を懐いて	0.973	あいびき	0.970	あいびき	0.978	破戒	0.971	緑葉歎	0.980	めぐりあひ	0.976

	破戒		高瀬舟		BCCWJ LBa	
	洪水	0.978	破戒	0.974	めぐりあひ	0.969
	玉を懐いて	0.978	洪水	0.968	破戒	0.965
	高瀬舟	0.974	めぐりあひ	0.963	罪と罰	0.960

#### 4.2.3 接続助詞 (語彙素) バイグラム・トライグラム分布の文書間類似度

接続助詞のバイグラム・トライグラムは、一文の中に含まれる接続助詞の組み合わせについて文書間類似度を見ることになる。接続助詞は複文を構成し、文の長短を左右する。また、文書の文体的特徴を示す有効な指標としても指摘されている (宮内 2012)。しかし、小説という同一の資料性においても有効な指標か否かを示すような具体的な調査は未だなされていない。



表 14 に接続助詞バイグラム分布の文書間類似度を、表 15 に接続助詞トライグラム分布の文書間類似度をまとめ、上位 10 位を墨地白抜きにし、下位 10 位を網掛けにした。これによると、接続助詞バイグラム・トライグラムどちらにおいても『罪と罰』は「BCCWJ LBa」以外の作品との類似度が低い。また接続助詞バイグラム・トライグラムどちらにおいても、「高瀬舟」は「あいびき」「めぐりあひ」「BCCWJ LBa」（バイグラムでは「緑葉歎」も）との類似度が高く、『破戒』は「あいびき」「めぐりあひ」との類似度が高い。しかし、これは実際に読んだ際の印象とは異なる。これについては次節で述べることとする。

品詞・語彙素・書字形のユニグラム・バイグラム同様、一文に含まれる接続助詞の組み合わせでは、有意と言える差や傾向は確認できなかった。

表 14 接続助詞バイグラム分布による文書間類似度

	BCCWJ LBa	あいびき	めぐりあひ	洪水	緑葉歎	玉を懐いて	罪と罰	破戒	高瀬舟
BCCWJ LBa		0.978	0.989	0.978	0.990	0.986	0.984	0.985	0.990
あいびき			0.994	0.984	<b>0.974</b>	<b>0.973</b>	<b>0.942</b>	0.993	0.988
めぐりあひ				0.991	0.985	0.985	<b>0.964</b>	0.988	0.992
洪水					0.984	0.983	<b>0.944</b>	<b>0.976</b>	0.986
緑葉歎						0.988	<b>0.962</b>	0.977	0.993
玉を懐いて							<b>0.964</b>	0.976	0.984
罪と罰								<b>0.955</b>	<b>0.955</b>
破戒									0.988
高瀬舟									

表 15 接続助詞トライグラム分布による文書間類似度

	BCCWJ LBa	あいびき	めぐりあひ	洪水	緑葉歎	玉を懐いて	罪と罰	破戒	高瀬舟
BCCWJ LBa		0.937	0.967	<b>0.921</b>	0.929	0.954	0.971	0.949	0.961
あいびき			0.985	0.956	0.930	0.925	<b>0.876</b>	0.966	0.967
めぐりあひ				0.962	0.948	0.948	<b>0.918</b>	0.964	0.976
洪水					0.934	<b>0.918</b>	<b>0.860</b>	0.924	0.946
緑葉歎						0.945	<b>0.870</b>	<b>0.909</b>	0.952
玉を懐いて							<b>0.922</b>	0.924	0.943
罪と罰								<b>0.912</b>	<b>0.896</b>
破戒									0.943
高瀬舟									

#### 4.3 接続助詞の配列

4.2.3 節では、接続助詞のバイグラム・トライグラムの文書間類似度を調査したが、指摘できるような結果は得られなかった。そこで本節では、一文に含まれる接続助詞の配列（並び）という点から調査を行う。服部（2008）では「節の運用に着目した調査は、明治期における近代文体の成立を考えるにあたり、文（Sentence）に対する意識や文章構造を考える指標となる可能性を

持つ」とし、「一文の長さは節そのものの長さよりは、節をどのように複数連結するかという運用の面に依るところが大きい」（服部 2008）と副詞節を中心とした節を計量的に調査しているが、本稿では文から切り離された節の頻度だけでは不十分と考え、次のようなデータを作成した。

係り受け解析を行ったデータより、文節の右端にある「接続助詞」を抽出する。この際、文を越えることはない。そして、文頭から文末までの接続助詞を順番に並べる。接続助詞をひとつでも含めば、「接続助詞配列タイプ」と考えることとする。例を以下に示す。（例文中の || は文節境界を示す）

- (1) 見れば || 二三の || 青年が || 店頭に || 立つて、 || 何か || 新しい || 雑誌でも || 猟つて || 居るらしい。（『破戒』）

この場合、抽出される接続助詞は「見れば」の「ば」、「立つて」の「て」、そして「猟つて」の「て」の 3 個となり、接続助詞の配列は「ば：て：て」となる。これをひとつの接続助詞配列タイプと考える。「猟つて居る」の「て居る」は、BCCWJ の長単位では複合辞と認定されているものの、現時点において人手修正済みの長単位が存在しないため、自動解析に委ねている。よって、複合辞に含まれるような接続助詞も含まれている。

表 16 に、この「接続助詞配列タイプ」が作品内に何種類（D） 現れるかをまとめた。

表 16 接続助詞配列タイプ

A：作品名	B：文数	C：文節数	D：接続助詞配列タイプ	D/B
あいびき	123	1,694	44	0.36
めぐりあひ	655	6,617	111	0.17
洪水	124	1,875	52	0.42
緑葉歎	81	912	24	0.30
玉を懐いて	340	4,431	77	0.23
罪と罰	1,097	13,322	175	0.36
小計	<b>2,420</b>	<b>28,851</b>		
破戒	1,136	11,742	96	0.08
高瀬舟	104	1,293	34	0.33
小計	<b>1,240</b>	<b>13,035</b>		
計	3,660	41,886		
BCCWJ LBa	14,918	97,758	303	0.02

「D/B」列には、文全体に占める接続助詞配列タイプの比率を示した。特に『破戒』は、その接続助詞配列タイプの比率が低く、多様な配列は用いられていないこととなり、「BCCWJ LBa」の 0.02 に値が最も近い。一方、「洪水」や「あいびき」『罪と罰』は相対的に多様な配列が用いられていることが分かる。

具体的にどのような配列があるかを見ていくと、8 作品と「BCCWJ LBa」すべてに共通して出現する接続助詞配列タイプは次の 6 種のみであった。

「て：て：て」「て：て」「て：ば」「から」「が」「て」

次に、接続助詞の配列がどのように分布しているかを見ていく。表 17 および図 4 には、例えば「接続助詞をひとつ含む文が何文あるか」のように、文中に含まれる接続助詞数別に文の数をまとめた。作品は、発表年順ではなく、右に行くほど文中に含まれる接続助詞数の「ばらつきが大きい」作品へと並べ替えてある。つまり、右に行くほど、文中の接続助詞数が極端に多い文が現れることを示している。このことから、確かに接続助詞の多さが文の長さに関係していると言いうことができる。これは、木坂（1976:313）で指摘されていた「初期の二葉亭翻訳文章の文脈を支える最も特徴的な」《累加的表現》を捉えたと言える。

また、作品の 8 割は接続助詞を 0～3 個含んだ文で構成されている。「めぐりあひ」以外の翻訳小説作品については、接続助詞 0 個の文があまり使われていないという点で、現代語より長文化の傾向が見られ、接続助詞を 9 個以上含むような長い文については、初期の言文一致体らしさを表していると言える。そういった意味で、「特徴的な文」というものに焦点を当てて分析を行っていくことが有効だと考えられる。

表 17 1 文中の接続助詞数別にまとめた文の数（単位：文）

1 文中に含まれる接続助詞の数	BCCWJ LBa	破戒	高瀬舟	緑葉歎	罪と罰	洪水	玉を懐いて	あいびき	めぐりあひ
1	4,709	355	26	19	330	30	113	39	146
2	1,866	248	22	17	239	30	81	17	90
3	615	92	16	10	131	12	42	19	41
4	160	51	10	3	92	12	21	8	32
5	49	14	5	2	37	5	10	6	15
6	17	7	1	1	12	6	4	3	13
7	7	2	0	0	7	2	4	1	7
8	0	0	0	1	1	0	2	2	10
9	1	0	0	0	3	1	2	1	2
10	0	0	0	0	1	0	0	2	2
11	0	0	0	0	1	1	0	0	2
12	0	0	0	0	0	0	1	1	0
13	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	1	1	1
21	0	0	0	0	0	1	0	0	0
24	0	0	0	0	0	0	0	0	1
0	7,494	367	24	28	243	24	59	23	292
総計	14,918	1,136	104	81	1,097	124	340	123	655

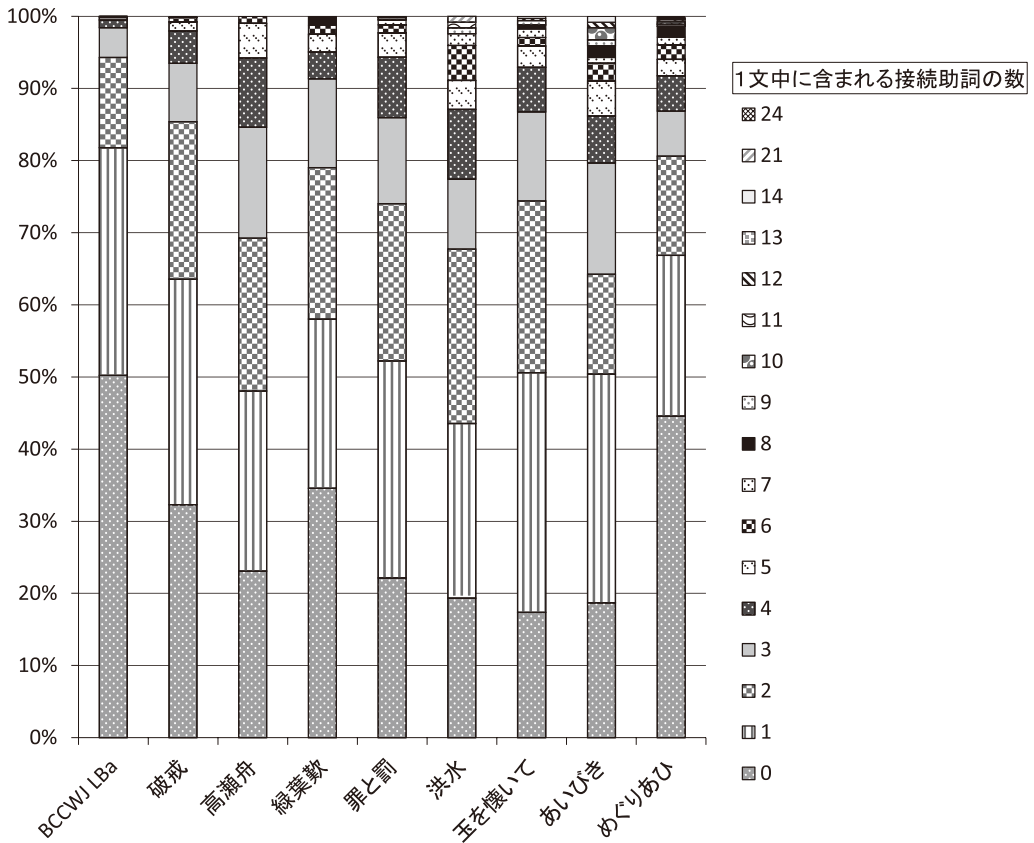


図4 1文中の接続助詞数別の文数比率

次の表18は、「1文中に含まれる接続助詞の数」と「1文中に含まれる「接続助詞テ」の数」の相関をまとめた。これを見ると、1文中に含まれる接続助詞の数が増える(文の長さが長くなる)と、それに応じて「接続助詞テ」の数も増えていくという正の相関が観察される。

特に多くの接続助詞を含む文が現れる作品は、「めぐりあひ」(13個の接続助詞中テが11個, 14個の接続助詞中テが13個, 24個の接続助詞中テが17個出現)で、共に出現する接続助詞は「ながら」「ば」「が」が見られる。他に「洪水」(21個の接続助詞), 「あいびき」「玉を懐いて」(14個の接続助詞)が挙げられる。これらは、前近代的な文体である戯作派的な特徴というより、写生を重視した原文の影響による翻訳ならではの欧文派的な特徴とすることができる。

以下に、「洪水」の例を挙げる。

- (2) 此デットロウの澤は乾いて居る時だに、この通りに面白くないが、扱丁度あの満潮が力一杯に差掛つて来る時、丁度あの湿つた風が冷く無作法にちらつく水の面を擦つて通り、脇を向いて見れば、次の潮が正面に吹付けて来る時、丁度あの沼の涯のない深みが、鋼鐵の様な青色に光つて来る時、丁度あの隙間もなく蠅殻に喰付かれて倒れて居る大木の幹が、

又た起き直つて、目當もない陰氣な道中を始めて、神に見離された猶太の老人の様に、絶ず彼處此處と迷ひ歩く時、丁度あの鴨が光る羽を擴げて音もせさせず、又た平な水の面にさゝ波も立てずに、潮の上近く過ぐる時、丁度あの狭霧が潮と一所に寄せて来て、水が柳の緑を包むと俱に空の青さを隠す時、丁度あの漁師がほぐすにほぐされない霧の網にからまれて、行先も知らず、櫓をびちや〜と動かし乍ら、舳が鳴る度には、水に住む怪しい女の悪い手が觸れたかと思ひ、情なくも泥を離れた一叢の艸が水の面に浮んだに出逢ふ度には、溺れて死んだ人の髪に比べて喫驚し、さてはデットロウの澤に迷込んだかと心付いて、しん氣な一夜を明さうと思ひ諦むる時は、この澤の景色は、まあ何程悲しからう。  
(下線部「テ」15個 ※自動解析で複合辞「ている」「てくる」を一部認定)

表 18 1文中に含まれる接続助詞の数と「接続助詞テ」の数の相関

データの個数 / 接続助詞タイプ数	1文中に含まれる接続助詞の数																	総計
1文中のテの数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	21	24	0	総計
0																		
1	1	28	76	25	2	1												133
2		1	31	73	15	3												123
3			1	27	33	13	3											77
4				1	27	19	5											52
5					1	13	9	5		1								29
6						1	10	5	2		1							19
7							1	4	5	3	2	1						16
8								1	2		1	1						5
9									1									1
10										1				2				3
11													1					1
13														1				1
15															1			1
17																1		1
総計	1	29	108	126	78	50	28	15	10	5	4	2	1	3	1	1		462

この調査は、あくまで予備的な位置づけとなる。それというのも、接続助詞「て」には複合辞の問題や、いわゆる南(1974)のA・B・C分類で、接続助詞「て」がA類からC類すべてに含まれる問題があり、また接続助詞「て」だけでなく連用中止ともあわせて観察すべきものである。また、言語単位・短単位的设计上の問題として「ので」は「準体助詞の+助動詞だ(連用形)」の2短単位で認定されており、接続助詞とはなっていないため、上記の調査には含まれていない。これらの問題は、節情報に特化したアノテーションを付与することにより解決できるのではないと思われる。

## 5. おわりに

本稿では、「近代口語文翻訳小説コーパス」構築の概要を述べ（3節）、構築されたデータの計量的分析によりそれらの文体特徴を探った（4節）。用いた情報・手法は、基礎統計量から算出された文の長さや文節の長さ、および品詞比率（3.4節）、また品詞比率を用いたMVR（4.1節）、文書間類似度（4.2節）、そして接続助詞の配列（4.3節）である。比較にBCCWJ図書館サブコーパス（LB）に収録されたサンプルの一部の内「小説」のデータを取り出して用いた。

基礎統計量による文の長さや文節の長さ、品詞比率、そして品詞比率を用いたMVRについては、有意な差がないという結果を得た。これは、翻訳小説6作品と近代口語文体創作小説2作品の文書が、計量可能な観点では類似していることを意味している。収録された翻訳小説6作品は、筆者が純粋な読者として通読しても少々読みにくいとを感じる程度の文章で、本稿の調査結果によると、当時活躍していた他の作家たちのそれとは比べものにならないほど“現代的”であったと言える。文書間類似度についても、品詞・語彙素・書字形、それぞれにユニグラム・バイグラムと粒度をこまかくしていったが、いずれも相互に文書間類似度が高く（一番、粒度のこまかい書字形分布においても低くはない）、多くの品詞もしくは語を共有していることになる。こちらも基礎統計同様、あまり差のない結果となった。

しかし、接続助詞の配列においては、作品ごとに差異を見出すことができた。1文中に含まれる接続助詞の数について、その配列タイプで頻度を出し、分布を調べた。すると、四迷訳「あひびき」「めぐりあひ」や鷗外訳「玉を懐いて罪あり」「洪水」については、1文中で接続助詞が8個や9個以上出現する文があり、1文中の接続助詞数にばらつきが大きいという結果になった。また、文の語数（接続助詞の数）と接続助詞「て」の数には正の相関があることも明らかとなった。一方で、本稿でのデータ抽出では不十分であることも承知しており、節情報に特化したアノテーションの必要性を強く感じた次第である。

今後は、節情報を足がかりに、文を構成する各要素を、意味レベルで抽象化したアノテーションを行い（例文（3））、言文一致体と近代口語文体の特徴的な差異を明らかにしていきたい。

(3)	《 本文 》	《 アノテーション 》
	斯うなると	<u>Conj: 順接: こうなると</u>
	ラスコーリニコフの勇氣は忽ち十倍し	N ハ + <u>V,</u>
	鉦を取直して	N ヲ V <u>テ,</u>
	眞向微塵に二度まで打下すと	<u>V ト :</u>
	鮮血泉の如く迸って	N φ + <u>V テ,</u>
	死躰はどたり轉がた。	N ハ + <u>V タ。</u>
	眼球は飛出さんとし	N ハ + (V <u>ント</u> ) + <u>V,</u>
	顔は形に變じて	N ハ + <u>V テ,</u>
	全く死相を現じた。	N ヲ V <u>タ。</u>

## 参考文献

- 服部隆 (2008) 「二葉亭四迷『浮雲』における文意識：節 (Clause) を用いた文体分析の試み (一)」『上智大学国文学科紀要』 25: 35-64.
- 服部隆 (2011) 「二葉亭四迷『あひまき』初訳・改訳における文章展開一節 (Clause) を用いた文体分析の試み (三)」『上智大学国文学科紀要』 28: 1-33.
- 服部隆 (2012) 「嵯峨の屋おむろ『薄命のすま子』の「である」体一節 (Clause) を用いた文体分析の試み (四)」『上智大学国文学科紀要』 29: 113-142.
- 服部隆 (2013) 「山田美妙『武蔵野』における文末時制一節 (Clause) を用いた文体分析の試み (五)」『上智大学国文学科紀要』 30: 17-54.
- 樺島忠夫・寿岳章子 (1965) 『文体の科学』 京都：綜芸舎.
- 加藤百合 (2012) 『明治期露西亜文学翻訳論攷』 東京：東洋書店.
- 川戸道昭 (2014) 『欧米文学の翻訳と近代文章語の形成—資料集成 近代日本語(形成と翻訳)別巻』 東京：大空社.
- 木村毅 (1972) 『明治文學全集 7 明治翻譯文學集』 (解題) 東京：筑摩書房.
- 木坂基 (1976) 『近代文章の成立に関する基礎的研究』 東京：風間書房.
- 国立国語研究所 (編) (2005) 『太陽コーパス 雑誌『太陽』日本語データベース』 東京：博文館新社.
- 近藤明日子 (2014) 「『国民之友コーパス』解説書 第 1.1 版」国立国語研究所. ([http://pj.ninjal.ac.jp/corpus\\_center/cmj/doc/kokumin\\_manual\\_v1\\_1.pdf](http://pj.ninjal.ac.jp/corpus_center/cmj/doc/kokumin_manual_v1_1.pdf) 2015 年 5 月確認)
- 小西光 (2015) 「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」国立国語研究所 (編) 『第 7 回コーパス日本語学ワークショップ予稿集』 257-264.
- 南不二男 (1974) 『現代日本語の構造』 東京：大修館書店.
- 宮内佐夜香 (2012) 「接続助詞とジャンル別文体特徴の関連について—『現代日本語書き言葉均衡コーパス』を資料として—」『国立国語研究所論集』 3: 39-52.
- 水野的 (2007) 「近代日本の文学的多元システムと翻訳の位相—直訳の系譜」『翻訳研究への招待 (日本通訳訳学会翻訳研究分科会)』 1: 3-43.
- 森岡健二 (1991) 『近代語の成立 文体編』 東京：明治書院.
- 小椋秀樹・小磯花絵・富士池優美・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』形態論情報規程集 第 4 版 (下)』 (国立国語研究所内部報告書 LR-CCG-10-05-02) 東京：国立国語研究所.
- 齊藤美野 (2012) 『近代日本の翻訳文化と日本語—翻訳王・森田思軒の功績』 東京：ミネルヴァ書房.
- 佐藤武義・前田富祺 (編) (2014) 『日本語大事典』 (上) 東京：朝倉書店.
- 田中牧郎 (2012) 「近代語コーパスにおける資料選定の考え方」『近代語コーパス設計のための文献言語研究成果報告書』 (国立国語研究所共同研究報告 12-03) 13-26. 東京：国立国語研究所.
- 柳父章・水野的・長沼美香子 (編) (2010) 『日本の翻訳論 アンソロジーと解題』 東京：法政大学出版局.
- 山本正秀 (1965) 『近代文体発生の史的的研究』 東京：岩波書店.
- 山本正秀 (1981) 『言文一致の歴史論考 続篇』 東京：桜楓社.

## 関連 Web サイト (すべて 2016 年 3 月確認)

- 近代語のコーパス (国立国語研究所) [http://pj.ninjal.ac.jp/corpus\\_center/cmj/](http://pj.ninjal.ac.jp/corpus_center/cmj/)
- 近代デジタルライブラリー (国立国会図書館) <http://kindai.ndl.go.jp/>
- 『現代日本語書き言葉均衡コーパス』 (国立国語研究所) [http://pj.ninjal.ac.jp/corpus\\_center/bccwj/](http://pj.ninjal.ac.jp/corpus_center/bccwj/)
- 『現代日本語書き言葉均衡コーパス』品詞構成表 (同上)
- 中・長単位解析器「Comainu」 <https://osdn.jp/projects/comainu/>
- 日本語係り受け解析器「CaboCha」 <https://taku910.github.io/cabocha/>
- コーパス管理アプリケーション「ChaKi.NET」 <https://osdn.jp/projects/chaki/>

## 参照資料 (作品本文冒頭より一部引用)

### (1) 「あいびき」

秋九月中旬といふころ、一日自分がさる樺の林の中に座してみたことが有つた。今朝から小雨が降りそゞぎ、その晴れ間にはおり〜生まやかな日かげも射して、まことに氣まぐれな空ら合ひ。

あわ／＼しい白ら雲が空ら一面に棚引くかと思ふと、フトまたあちこち瞬く間雲切れがして、無理に押し分けたやうな雲間から澄みて伶俐し氣に見える人の眼の如くに朗かに晴れた蒼空がのぞかれた。自分は座して、四顧して、そして耳を傾けてゐた。木の葉が頭上で幽かに戦いだだが、その音を聞たばかりでも季節は知られた。それは春先する、面白さうな、笑ふやうなさゞめきでもなく、夏のゆるやかなそよぎでもなく、永たらしい話し聲でもなく、また末の秋のおど／＼した、うそさぶさうなお饒舌りでもなかつたが、只漸く聞取れるか聞取れぬ程のしめやかな私語の聲で有つた。

#### (2) 「めぐりあひ」

何處へと云つて夏の中は「グリンノエ」村へほどよく遊獵に往つた所はなかつた、自分の莊園から二十露里計の所に有る（「グリンノエ」村へほど）。此村の近傍には野鳥撃ちには郡中第一とも謂ふ可き場所が澤山有つた。最寄の灌木または野田などを残らず獵り盡して、自分はいつも日没ごろになると此邊には滅多に無いといふ隣の沼へ立寄つて、其處からはもう旅宿へ歸る事としてゐた、此旅宿といふは何時も取り付けで、主人は村の差配人信切な男で有つたので。沼から「グリンノエ」までは道程凡そ二露里計り總て中凹の道で、只中ごろで大きくも無いが丘を一ツ躡さなければならなかつた。此丘の頂に一構の莊館、と云つた所が住み棄てた母屋一棟と園一區の外何も無かつたが、それが有つた。

#### (3) 「洪水」

水の引いた迹で見れば、デットロウの澤の憫れ氣な姿が、丸で打出して目の前に見え、低い、海綿の様な表面はゆら／＼として、爹兒かとおもはるゝ程に黒い沼になつて居て、そこから泥雜りの水が、蛇のやうに委它りながら、靜かにねば／＼した滑りさうな溝をこしらへて、太平洋の廣い入江に流れゆく。彼處此處に青黄いろい艸の取残されたやうな斑が見えて、此艸の瘦せた、ぬる／＼した莖には厭な、濕つた、土臭い香がするから、日の目を見ずに育つたものといふことが直ぐに知れう。この沼の平たい、變りのない淋しさは、人の妄想を呼出すことはない筈だが、それでも妄想を起せば、あの波のうねりのやうな、打寄せられた浮木は、直ぐに又た、今一旦引きは引いたが、また厭でも高潮になつて戻つて来る水の、悲しい知らせの使になるだらう。

#### (4) 「緑葉歎」

重傷を負うた亞非利加獵兵聯隊の軍曹カドユウル、ベン、ゲリツファをザウエルバツハの百姓リツベルトの家へ擔ぎ込んだは、丁度五週間程前であつた。彼は始終人事不省で、熱度が高く、久しく生死の海に漂うて居つた。そして譚語を云ふを聞けば、いつもワイセンプルクの麻畑の上で戦つて居る話でなくば、故郷のアルジェリヤで村長をして居る「マトマタ」種の所に歸つて居る話であつた。

やつと今日始めて目を見開いて驚いた。まだ見た事のない大きな室の窓には、白い布が懸つて居つて、雲の間から折々漏れて来る日の光を最一度遮つて居つた。



## (5) 「玉を懐いて罪あり」

路易第十四世の寵愛が、メントノン公爵夫人の一身に萃まつて世人の目を驚かした頃、宮中に入入をする年寄つた女學士にマグダレン、ド、スキュデリイと云ふ人があつた。丁度千六百八十年の秋の事で、或る夜の十二時過ぎに、其女學士が住つて居るセントノレイ町の家の戸を劇しく敲くものがあつた。其夜下男のバプチストは妹の婚禮に招ばれて行つてまだ歸らず、家の内で目を醒まして編物をして居たは仲働きのマルチエ、ルといふものであつたが、今此響を聞くと、何となく怖氣立ち、急に自分と主人と女二人で家に居ることに氣が付き、昔から巴里であつた人殺しや、押込の怖い話が皆一時に胸に浮んで、只慄々と震ひ乍ら室の片隅に蹲んで居た。

## (6) 「小説 罪と罰」

七月上旬或る蒸暑き晩方の事。S……「ペレウーロク」(横町)の五階造りの家の、道具付きの小坐敷から一少年が突進して、狐疑逡巡の体でK、の橋の方へのッそり出掛けた。首尾よく階子の下口で主婦に出會はなかつたが、此家の主婦は下坐敷に住つて、臺所が常々戸の開いたまゝ階子に對て居るので、いつでも少年が出掛ける時は、餘義なく敵の竈前を通り過ぎ、骨身に染みるほどの恐怖を生じ意久地なく眉に皺を寄せるが常であるといふは宿料の停滞があるから、それで顔を合はせるが怖ろしいのだ。

## (7) 「破戒」

蓮華寺では下宿を兼ねた。瀬川丑松が急に転宿を思ひ立つて、借りることにした部屋といふのは、其蔵裏つゞきにある二階の角のところ。寺は信州下水内郡飯山町二十何ヶ寺の一つ、真宗に附属する古刹で、丁度其二階の窓に倚凭つて眺めると、銀杏の太木を経て、飯山の町の一部も見える。さすが信州第一の仏教の地、古代を眼前に見るやうな小都会、奇異な北国風の屋造、板葺の屋根、または冬期の雪除として使用する特別な軒庇から、ところへ高く聳れた寺院と樹木の梢まで——すべて旧めかしい町の光景が香の烟の中に包まれて見える。

## (8) 「高瀬舟」

高瀬舟は京都の高瀬川を上下する小舟である。徳川時代に京都の罪人が遠島を申し渡されると、本人の親類が牢屋敷へ呼び出されて、そこで暇乞いをするを許された。それから罪人は高瀬舟に載せられて、大阪へ回されることであつた。それを護送するのは、京都町奉行の配下にいる同心で、この同心は罪人の親類の中で、おも立つた一人を大阪まで同船させることを許す慣例であつた。これは上へ通つた事ではないが、いわゆる大目に見るのであつた、黙許であつた。

## Compilation of a Corpus of Translated Works of the Mid-Meiji Era and Its Quantitative Analysis

KONISHI Hikari

Adjunct Researcher, Center for Corpus Development, NINJAL

### Abstract

The modern colloquial style of Japanese had begun to be constructed since the genbun itchi movement, which involved the unification of spoken and written old Japanese. This style had been established and popularized as the standard form of Japanese through the national textbook system (*kokutei kyōkasho*) in the Meiji era. From then on, it became the basis for contemporary written Japanese. Recent discussions have indicated that the translation of European and American literature had affected the establishment of the colloquial style. To quantitatively investigate the effect of the translation work on the style, we compiled a corpus of six translated works of the mid-Meiji era and one each from the late Meiji and Taisho eras.

This paper provides an overview of the corpus and the findings of a statistical survey, such as the length of sentence, the similarity between the works, and so on. The survey showed that there are no significant differences between the two sets of works in the number of bunsetsus in a sentence, the rate of POS tagging, Modifier Verb Ratio (MVR), or in their cosine similarities. The quantitative results indicated similarities between the translated works of the mid-Meiji era and those of the late Meiji and Taisho eras. It also supported the assumption that the colloquial style had already become familiar through translated works and read by people of the mid-Meiji era. However, as an exception, the statistics showed that the variance of the conjunctive postposition in a sentence tends to decrease with the passage of time.

**Key words:** Mid-Meiji era, genbun-itchi, stylistics, novels in translation, conjunction particle