The Hanover Tagger による品詞分解を用いた「ゲーテ時代」文学研究序論

片山 耕二郎（かたやま こうじろう）

本論文は、ドイツ語文章を品詞分けするプログラミング言語 Python のライブラリ The Hanover Tagger（通称 HanTa）を用いて、ドイツ語作品ライブラリである Projekt Gutenberg-DE（https://www.projekt-gutenberg.org/）の作品を分析し、それらの作品における品詞の割合から考察しうることを述べたものである。前半においては、HanTa を用いてテクストを分析するための方法について説明し、後半においてはそうして得られた数値を比較して、作品のこれまでの評価を元にした考察を述べる。
品詞分解は、あくまで研究の方法、手段であり、研究対象や目的ではない。研究は手段に振り回されず、ある目的のために必要なときにのみ手段が求められるというのが前提である。本論文の元となる研究も、メルヒェンと伝説の文体的特徴の違いや、共著とされる作品についての部分ごとの著者特定を目指して始められたものである。
しかし、前提となる研究手法とその有用性について、日本のドイツ文学研究では、まだ浸透しているとは言いがたい。このため、手法自体をまず解説することが必要である。本論文は例外的に品詞分解という手法自体に焦点を当てた論文となる。論文の狙いは、比較対象として用意した 50 冊の書籍を遊戯的に比べ、人間の判断と機械の判断が一致しあるいは相違する点を例示することによって、手法自体への関心を引き起こすことにある。

0. 背景

2013 年に刊行されたフランコ・モレッティの論文集『遠読――〈世界文学システム〉への挑戦』に見られるように、文学作品の計量的分析は理論として可能なだけでなく、すでに実践的な研究が行われている領域である。『遠読』に収められている論文「スタイル株式会社 ―― 七千タイトルの省察（一七四〇年から一八五〇年のイギリス小説）」(2009) では、小説の題名についての 7000 件のデータ（この規模は、やがては少ないと批判されうるが）を統計処理し、通時的な変化の特徴を抽出し、その理由を分析している。なかで

も題名が徐々に短くなっているという発見と、題名が内容を説明する必要がなくなったという時代背景による説明は、分かりやすく、本論文も模範とすべきものである。

計量的分析についてのよりくだけた内容の一般書として、2017 年に刊行されたベン・ブラット『数字が明かす小説の秘密　スティーヴン・キング、J・K・ローリングからナボコフまで』もある。モレッティがスタンフォード大学に〈リテラリー・ラボ〉を構えて、そのタイミングごとの先進的な研究を行うのに対し、ブラットの本は、個人の研究者でも Python やそのライブラリ NLTK を用いることで興味深い文学研究ができることを示している。

こうした研究は英語で率先して行われる。もちろん、たとえば日本でも杉浦清人が論文「文学研究におけるデジタル・ヒューマニティーズの可能性 ── 文章心理学・計量文献学・マクロ分析 ──」において、これまでの日本での（高速なコンピュータがなかった時代の）研究過程を紹介し、また発表「品詞分析から見る夏目漱石の前期作品の文体の特異性」（情報処理学会第 117 回人文科学とコンピュータ研究発表会、2018 年）あるいは「統計的な語彙の多寡は文学テクストにおいて何を意味するか？　夏目漱石の作品を中心に」（第 4 回　現代文芸論研究室報告会、2019 年）で日本語分析ツール KH Coder やプログラミング言語 R を用いて作品分析したように、研究例が見られないわけではない。吉澤弥生と奥彩子も論文「デジタル人文学の研究と教育に関する基礎的研究」において KH Coder を用いたテクスト分析を行っている。しかし、KH Coder の研究事例リストには、他に文学研究では井上明芳の「テキストマイニングによる森敦文学の基礎的研究」（2021）が掲載されている程度で、研究が活況にあるとは言いがたい。モレッティやブラットの本が訳されて、やや注目が増しているといった現況である。

英語で率先して行われる最も大きい理由は、言うまでもなく世界語としてもアカデミズムで用いる言語としても英語が一番有力だからである。もう一つ、これはドイツ語のテクストとの比較でより実感されることであるが、屈折語でありながら、活用が少ないなどその屈折性が弱く、また語順も固定されやすいなど、言語の機械処理を複雑化する要素が少ないことも理由として挙げられる。

本論文で扱うドイツ語に関して述べれば、活用がそれなりに豊富で、語順も動詞を除けば自由であるなど、西洋語として機械処理が容易とは言いがたい。分離動詞のように同じ単語の二要素がときには遠く離れて置かれるという構造は、単語を発見するのも困難であり、これを発見するために各文が分離動詞を持つか点検すれば、分析速度の低下にもつながる。また、分離動詞であるかは意味によって判明することもあり、機械が正確に判定する確率は下がる。

このような事情もあり、英語における NLTK（National Language Toolkit）のように広く用いられている言語処理の汎用ライブラリは未だドイツ語には存在しない。それでも土台は徐々に築かれてきている。まず、NLTK は英語については高い精度の品詞分解なども行える優れたライブラリであるが、ドイツ語についても文章を単語に切り分ける機能は実装されている。先述の理由から、単語の切り分け性能も英語ほど正確ではないが、機械による大量の文章を処理しての研究は、個々の項の分析において完全な正確性を持たない分を量で補う面があり、こうした精度の低さを踏まえてなお行う価値がある。むしろ、こうしたツールを積極的に用いることは、その精度を上げていくための貢献となる。たとえその結果、後年のより正確な研究によって自らの研究の価値が失われることになるとしても。

ドイツ語を扱えるライブラリとして、単語を切り分ける NLTK 以外に、汎用ライブラリ spaCy がある。こちらも NLTK 同様ドイツ語に特化しているわけではないが、品詞分解も行える。スラバヤ州立大学の Muhammad Kharis らは spaCy を用いてドイツ語教材を分析し、論文“Tokenization and Lemmatization on German Learning Textbook Level A1 of CEFR Standard”にまとめており、そこで spaCy は品詞分解が可能であると述べたうえで、なおエラーがあるため改善の余地があるとしている。これに対し、ドイツ語専用の品詞分解ライブラリとして登場したのが HanTa である。

1. The Hanover Tagger について

HanTa は正式名称を“The Hanover Tagger”という。地名のハノーファはドイツ語では Hannover と n を重ねてつづるが、こちらは定冠詞として英語の the が用いられているように英語名であるため、Hanover という表記である。ドイツ語圏に限らない利用が期待されていると思われる。なお、執筆時点でのバージョンは 0.2.0 である。

HanTa は単語を Lemmatisation、すなわち見出し語の形に変換する機能とともに、POS を Tag 付けする、すなわちそれぞれの単語に品詞情報を付加する機能を有している。この見出し語への変換と品詞情報付加のアルゴリズムを論文で 2019 年に公開したのがハノーファにある工科大学 Hochschule Hannover の教授 Christian Wartena である。彼が HanTa について解説した論文、“A Probabilistic Morphology Model for German Lemmatization”（ドイツ語の見出し語化のための有効な形態論モデル）で示されたデータを信じるならば、品詞情報付加の正確性は spaCy より高い。現在の英語向け Tagger には及ばないと思われるが、それらもより正確性の低い段階で研究に用いられてきたことを考えれば、HanTa を用いてドイツ語文章の品詞分解をする価値は十分にある。

本研究ではこの HanTa の Python 用ライブラリを用いて、ドイツ語の著作権切れ作品を公開している Project Gutenberg-DE（https://www.projekt-gutenberg.org/）のテクストを分析する。Project Gutenberg-DE は Zeno.org（http://www.zeno.org/）とならぶドイツ語作品のライブラリである。私がこれまで使用してきた限りでは、両サイトともテクストの正確性は統計的分析に用いるに十分なレベルにある。英語の分析に Project Gutenberg（https://www.gutenberg.org/）、日本語の分析に青空文庫（https://www.aozora.gr.jp/）が用いられることを考えれば、Project Gutenberg-DE を明示的に（すなわち、エラーの原因になりうるとすればその原因を探れるように）用いることに問題はないと考える。

HanTa の品詞情報付加について、その特徴を述べておく。そもそも品詞の分類は恣意的なものである。名詞、動詞、形容詞、副詞などに分けることについては共通了解があるとしても、たとえばある単語をどちらに属させるかの境界は曖昧であり、それぞれについてどこまで細かく区別するかについてはより判断が分かれる。

したがって、HanTa での品詞タグも NLTK が採用している Penn Treebank Project（https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html）のものとは数も対象も異なる。

HanTa の全ての品詞を拙訳を添えてリスト化すると次のようになる。

1．名詞群　NN（普通名詞）、NE（固有名詞）
2．代名詞群　PDAT（付加指示代名詞）、PDS（指示代名詞）、PIAT（付加不定代名詞）、PIS（不定代名詞）、PPER（人称代名詞）、PPOSAT（付加所有代名詞）、PPOSS（所有代名詞）、PRELS（関係代名詞）、PRELAT（付加関係代名詞）、PRF（再帰代名詞）、PWS（疑問代名詞）、PWAT（付加疑問代名詞）
3．動詞群（動詞・コピュラ動詞・助動詞の定形・不定形・命令形・完了形など）VVFIN（動詞定形）、VAFIN（コピュラ動詞定形）、VMFIN（助動詞定形）、VVINF（動詞不定形）、VAINF（コピュラ動詞不定形）、VMINF（助動詞不定形）、VVIMP（動詞命令形）、VAIMP（コピュラ動詞命令形）、VVPP（動詞完了形）、VAPP（コピュラ動詞完了形）、VMPP（助動詞完了形）、VVIZU（zu 不定詞の動詞部分）
4．形容詞群　ADJA（付加形容詞）、ADJD（叙述形容詞）
5．副詞群　ADV（副詞）、PWAV（疑問副詞）、PROAV（代名詞付き副詞）、PTKANT（応答副詞）

6．前置詞／後置詞群　APPR（前置詞）、APPRART（冠詞付き前置詞）、APPO（後置詞）、APZR（右側置詞、um … herum のような組合わせの右に置く要素）
7．接続詞群　KOUI（zu 不定詞を取る接続詞）、KOUS（従属接続詞）、KON（並列接続詞）、KOKOM（比較接続詞）
8．冠詞　ART（冠詞）、
9．数詞　CARD（数詞）
10. 感嘆詞　ITJ（感嘆詞）
11. 否定辞　PTKNEG（否定辞）
12. 外来語　FM（外来語）
13. その他　PTKZU（不定詞と用いられる zu）、PTKVZ（分離動詞の分離部）、PTKA（最上級の am と形容を強調する zu）、TRUNC（複合語分離部）、XY（不明）、$,（読点）、$.（句点）、$(（その他文記号）

合計 53 の品詞タグに分類されている。外来語を独立させることについては賛否あるだろうが、ドイツ語は外来語辞典も豊富にあるように、これを別まとめにする意欲が言語的に存在するようである。Penn Treebank Project では 36 に分けているから、これよりもかなり細かいことになる。とりわけ動詞が細かく分類されていることが特徴的で、Penn Treebank Project では 6 個なのに対し、12 個に区別されている。コピュラ動詞を区別することによって、文章中で用いられる動詞が形容的であるか動作的であるか、ある程度の推測がつくのは HanTa の長所であろう。

2. The Hanover Tagger の Python での実装

以下、私が Python において HanTa を用いた分析コードをどのように実装したかをやや詳しく述べる。日本人研究者にも HanTa およびドイツ語の品詞分解をしやすくし、また私の分析結果を検証したい人が同じ手順を辿れるようにするためである。
後者について少し補足する。文章の統計的分析は、対象の量を十分に確保することで価値ある研究を行えるが、疑いのない事実を積み重ねる研究とは異なり、不正確さが入り込む余地は常にある（もちろん、人間が機械を介さずにテクスト分析すれば正確に行えるというのも幻想であるが）。すでに、HanTa そのものの精度および Project Gutenberg-DE の不正確性については述べた。ここに私がコーディングミスをした可能性も加わるため、どこに不正確さが起因するか、手順を透明にしておく必要がある。なお、実装には Hochschule Hannover の Wartena 研究室による手引き（https://textmining.wp.hs-hannover.

de/Preprocessing.html）を参照した。

まずNLTKを用いて、与えられた文章を文および単語に分解する。なお、品詞の割合を調べるだけなら単語に分解するのみで良いが、あとから単語が出てきた文を検索しなおす場合には文への分解データも有益である。文への切り離しにはnltk.sent_tokenize関数を、単語への切り離しにはnltk.word_tokenize関数を用いれば良い。引数には言語としてgermanを指定する。これによって、NLTKは与えられた文章をドイツ語の規則に従って、文および単語に切り離す。言語ごとに設定があるのは細かい例外があるからだが、イメージしやすく言えば、句点や疑問符や感嘆符があればそこで文として区切り、空白があればそこで単語として区切るということである。こうして全ての文を収めた配列と全ての単語を収めた配列が生み出される。

ついで、HanTaのHanoverTaggerクラスのインスタンスを作る。ヴァルテナの手引き通り、HanoverTaggerをhtとしてインポートし、tagger = ht.HanoverTagger('morphmodel_ger.pgz')とすれば良い。これでtaggerはanalyze関数によって1単語の品詞を返したり、tag_sent関数によって多数の単語の配列から、それぞれの単語について見出し語と品詞を付した配列を返したりしてくれる。長い文章の場合、NLTKのword_tokenize関数によって単語に切り離し、それをHanTaのHanoverTaggerインスタンスのtag-sent関数に渡すことによって、全ての単語の見出し語と品詞のデータが得られるわけである。なお、私の環境ではtag-sent関数に渡すデータ量が多すぎるとエラーを吐いたため、50000語ごとに切り離して呼び出し、結果を結合している。

こうして得られた品詞ごとの数量について、文章の総単語数で割ることによって、その割合を求められる。全ての品詞の割合の合計が1になれば（今回の論文では個々の品詞をパーセンテージ表記とし、小数点以下2桁で偶数丸めしているため、合計が100の近似値になれば）、割合計算においてミスをしていないことが判断できる。今回取り上げた作品について、この点でミスは含まれていない。

3. Project Gutenberg-DEからのテクストのスクレイピング

本節の最後に、Project Gutenberg-DEからのテクストのダウンロードについて簡単に説明する。効率化のために必要であるが、ここでもエラーが生じうるからである。サイトからのページ内容のスクレイピングにはrequestsとBeautifulSoupの2つのライブラリを用いた。PythonのHTTPライブラリには標準のurllib2などがあるがやや使いづらく、またHTMLの解析にはBeautifulSoup以外にも標準のHTMLParserやより高速のいくつ

かのライブラリがあるが、大量の作品を一括でダウンロードするのでなければ使用しやすい BeautifulSoup で十分であり、広く用いられているためエラーが生じた際にも対処しやすい。

まず requests のインスタンスに該当作品の URL を与えて、サーバから HTML タグ付きテクストを受け取る。この手順は一般的なスクレイピングと全く変わらない。このテクストを BeautifulSoup のインスタンスに渡し、HTML ソースから作品のテクストを取り出す。requests のインスタンスで得たテクストをそのまま text として渡すとエンコードの判断を間違えるので、requests.apparent_encoding で判定しなおすか、reuests.text ではなく reuests.content を BeautifulSoup に渡すと良い（後者の方がより優れた解決である）。BeautifulSoup の HTML パーサ（HTML タグ付きのソースをタグに沿って解釈するプログラム）には何を用いても良いと思われるが、本研究では lxml を用いている。

Project Gutenberg-DE の作品ごとの HTML ソースは、短い作品であれば 1 ページにまとめられているが、そうでなければ章ごとなどの区切りで複数のページに分割されており、Inhalt（目次）から個々のページのリンクが得られる。したがって、後者の場合は全てのページへのリンクをリスト化し、頭から順にページの本文を取り出して結合すれば良いことになる。本文のテクストは html の <p> タグおよび各サイズの <h> タグで囲まれているため、これらを抜き出せば良い。ただし、<h> タグは同じ見出しに複数のサイズが重複してついていることがあるため、より正確な抽出のためには重複を確認する必要がある。一般にテクストの本文に対する見出しの分量は極めて小さいため、この Project Gutenberg-DE の仕様への対応は判断が分かれると思われるが、本研究では重複を避けたコーディングをした。この点で、同じテクストを用いても小さな数値の差が生まれる可能性があることをあらかじめ述べておく。

4. サンプルに用いた 50 作品

本論文はドイツ・ロマン主義期の作品を対象とした研究の一部であるため、サンプルとした 50 作品は、ロマン主義の全盛期と成立時期の近い、同時代または後世において高く評価されている作品とした。文学研究という点から小説を中心に、戯曲・学術書・哲学書等を加えた次の 50 作品を比較対象に用いた。選出については、各種の教科書等を参考にし、また Gutenberg-DE に重要著作があるかどうかも踏まえたうえで（たとえばヤコービやヴィルヘルム・フォン・フンボルトや A・W・シュレーゲルはこの点で候補から外した）、私の主観によった。

アルニム『エジプトのイザベラ』

ビューヒナー『ダントンの死』
シャミッソー『ペーター・シュレミールの奇妙な物語』
クラウゼヴィッツ『戦争論１巻』『戦争論２巻』『戦争論３巻』
フィヒテ『ドイツ国民に告ぐ』
フーケ『ウンディーネ』
ゲーテ『若きウェルテルの悩み』『ヴィルヘルム・マイスターの修業時代』『イタリア紀行』『ファウスト１部』『ファウスト２部』
グリルパルツァー『サッポー』『哀れな辻音楽師』
グリム兄弟『グリム童話』
カール・グロッセ『守護霊』
ハーマン『ソクラテス追想録』
ヘーゲル『精神現象学』
ハイネ『ドイツ古典哲学の本質』
ハインゼ『アルディンゲッロと幸福な島々』
ホフマン『悪魔の霊酒』
ヘルダーリン『ヒュペーリオン』
アレクサンダー・フォン・フンボルト『コスモス』
ジャン・パウル『巨人』『美学入門』
カント『判断力批判』
クライスト『こわれがめ』『ミヒャエル・コールハース』
コッツェブー『小都会のドイツ人』
レッシング『ラオコーン』『エミーリア・ガロッティ』
リヒテンベルク『雑記帳』
メーリケ『画家ノルテン』
モーリッツ『アントン・ライザー』
フリードリヒ・ニコライ『ゼバルドゥス・ノートアンカー氏の生涯と意見』
ノヴァーリス『青い花』
シェリング『世界霊について』
シラー『オランダ独立戦争史』『見霊者』『人間の美的教育について』『ヴァレンシュタイン』
フリードリヒ・シュレーゲル『ルチンデ』
シュライエルマッハー『宗教について』
ショーペンハウアー『意志と表象としての世界１部』『意志と表象としての世界２部』

ティーク『フランツ・シュテルンバルトの遍歴』『セヴェンヌの反乱』
ヴァッケンローダー『芸術を愛する一修道僧の真情の披瀝』
ヴィーラント『アガトン』

なお、これらの比較対象のテクストについて、どの版に基づいたテクストであるかは基本的に顧慮していない（敢えて言えば Project Gutenberg-DE 版である。同名のテクストが 2 種ある場合はサイトで上に表示されているものを用いたが、『グリム童話』については Vollständige Ausgabe（完全版）とされているものを選んだ）。これは、本論文は特定の版に絞った分析を目指したものではないからである。もちろん、特定の版に興味を絞らないにせよ、たとえばある作家の時代ごとの文体を分析するような試みであれば、テクストの版の年代を特定したうえで分析するのが望ましいのは言うまでもない。

5. 分析の手法 ── 中央値からの距離を用いた検討

本研究では、50 作品について品詞それぞれの割合を求め、品詞ごとの平均から個々の作品がどれほど離れているかを元にした検討を行う。

まず、品詞ごとの平均値を計算する。平均の取り方にはいくつか方法がある。有力なものとしては、全作品の単語数・品詞数を合計してから、品詞の割合を調べる方法（実質的には加重平均値）、そして個々の作品の品詞の割合を調べてから、品詞ごとに全作品の値を合計して作品数で割り、その平均を調べる方法（平均値）、さらにそれぞれの品詞割合において真ん中の順番に来る作品の数値を用いる方法（中央値）がある。

いずれにもメリットはあるが、ここでは中央値を採用する。前者 2 つの欠点としては、まず全作品の単語数・品詞数を合計して割合を調べる方法では、作品の影響力が文字数に比例してしまい、巨大な作品と短編との影響の差が大きくなってしまう。次に、平均値はこれに比べてはるかに有力な選択肢だが、外れ値の影響を受けやすい。たとえばリヒテンベルクの『雑記帳』は不明な品詞の割合が圧倒的に多く（各アフォリズムの区切り記号として用いられているアスタリスクの影響と思われる）、全作品の中央値が 0.02% なのに対して 2.13% に上るため、平均値を大きく押し上げてしまう（他のやや多い作品の影響もあり 0.09%）。こうした外れ値の影響を除くには、ひとつひとつ取り除くのも一つの方法だが、そうした作業にはきりがなく、また人による判断を持ち込むと再現性にも影響を与えてしまう。この点、中央値であればこのような外れ値を避けられる。また文章の品詞割合はそれほど幅広い分布ではないため、中央値が外れ値を除いた平均値と乖離することはほぼない。このため、本研究では中央値を採用する。すなわち、50 作品のうち中央に最も

近い 2 作品、25 番目と 26 番目の作品の平均値である。

次にこの中央値と各作品の値を比較する。一般的な統計の手続きとして、複数の項の差をまとめて比較したい場合、それぞれの項の差を二乗して合計したものを（そして元の次数に整えたければその平方根を）比較する。この数値が大きいほど品詞の割合が平均とは異なる作品と言え、小さいほど品詞の割合が平均に近いことになる。二乗の値が各項で大きく異なる場合には、一つの項の影響が大きくなりすぎるため、絶対値の差の合計を比較する方が良いこともあるが、両者の相関係数は下記の 50 作品を用いたデータの場合、約 0.98 あるため、結果に大きな違いはない。ここでは二乗の合計値の平方根での比較を中心とし、絶対値の差がこれと大きく異なる場合にはそれも参考にすることとする。

この結果を単語数・語彙数とともにまとめたのが次の表である。

	単語数	語彙数	二乗の和の根	絶対値の差の和
中央値	79268.5	9658	0	0
平均	112628.8	12334.48	1.3	4.51
『イタリア紀行』	198573	23201	1.82	7.8
『画家ノルテン』	153873	19193	2.17	8.66
『ドイツ古典哲学の本質』	93054	9011	2.6	10.61
『芸術を愛する一修道僧の真情の披瀝』	36734	6775	2.64	9.68
『ゼバルドゥス・ノートアンカー氏の生涯と意見』	148081	16902	2.99	10.45
『アントン・ライザー』	148397	13373	3.26	12.71
『ヴィルヘルム・マイスターの修業時代』	227623	18979	3.26	12.87
『悪魔の霊液』	126946	14796	3.28	13.19
『アガトン』	189653	15458	3.35	15.51
『セヴェンヌの反乱』	95550	13032	3.4	12.96
『アルディンゲッロと幸福な島々』	116994	14609	3.57	13.72
『守護霊』	199065	17518	3.6	14.1
『エジプトのイザベラ』	46394	7691	3.63	14.27
『見霊者』	55563	8345	3.67	15.08
『青い花』	55414	9130	3.97	16.03
『ルチンデ』	34224	5815	4.01	15.91
『巨人』	376565	40133	4.02	18.25
『雑記帳』	79320	10366	4.1	18.48
『哀れな辻音楽師』	17947	3934	4.11	15.66

『フランツ・シュテルンバルトの遍歴』	120666	13228	4.16	15.02
『ドイツ国民に告ぐ』	85957	9650	4.39	19.51
『宗教について』	86572	9317	4.45	20.61
『ウンディーネ』	30934	5763	4.46	17.44
『ヒュペーリオン』	57198	7888	4.84	19.6
『意志と表象としての世界１部』	288732	27163	4.96	19.42
『オランダ独立戦争史』	129557	15604	5.1	20.29
『意志と表象としての世界２部』	254035	20696	5.16	20.32
『ペーター・シュレミールの奇妙な物語』	23704	4742	5.18	19.04
『人間の美的教育について』	40776	5695	5.22	18.97
『若きヴェルテルの悩み』	48005	7226	5.25	18.95
『戦争論３巻』	54712	6675	5.38	21.06
『美学入門』	179080	24885	5.53	20.88
『戦争論１巻』	103229	9666	6.06	23.83
『精神現象学』	202826	10111	6.19	26.19
『世界霊について』	79217	7961	6.31	23.6
『戦争論２巻』	123829	10062	6.51	24.95
『ファウスト２部』	56477	11339	6.58	27.82
『判断力批判』	126197	9292	6.81	26.81
『ラオコーン』	73768	11515	7.06	23.96
『ミヒャエル・コールハウス』	41961	6190	7.36	24.2
『ソクラテス追想録』	9062	3043	7.45	26.91
『グリム童話』	314934	19758	7.71	27.6
『ファウスト１部』	38428	6932	7.97	28.92
『ダントンの死』	25569	4554	8.91	29.46
『ヴァレンシュタイン』	78605	10005	9.33	33.08
『サッポー』	21370	4168	9.57	36.1
『こわれがめ』	21022	3463	10.56	37.5
『コスモス』	728386	64420	11.91	49.25
『エミーリア・ガロッティ』	27948	3633	12.11	41.81
『小都会のドイツ人』	23124	3834	12.68	39.26

6. 表に基づく考察・予想

この 50 冊の各品詞の差の合計を比較してみるだけでも様々な発見がある。数値の理由が自明のものから、より深い考察を要するものまで、関連する順に列挙する。

一. 戯曲は HanTa による品詞分解において、他のジャンルに比べて平均から離れた数値を示しやすい。

散文と戯曲は文体が異なる。とりわけ戯曲は各話者の人名がセリフごとに示されるため、固有名詞の割合が多くなることが大きな影響を及ぼしていると考えられる。実際、戯曲の固有名詞の平均割合は約 4.8%、その他の平均割合は約 1.9% で、違いが生じている。ただし、下の二つのグラフ（戯曲のみマーキングしている）から分かるように、固有名詞を除いて比較してもほぼ順位は変わらなかった。複数の品詞の差が関わっていると考えられる。

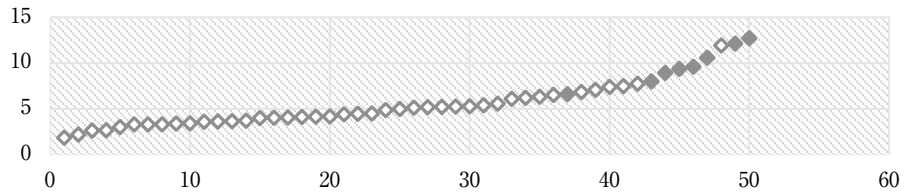

図 1　二乗の和の根のうち戯曲を示すグラフ

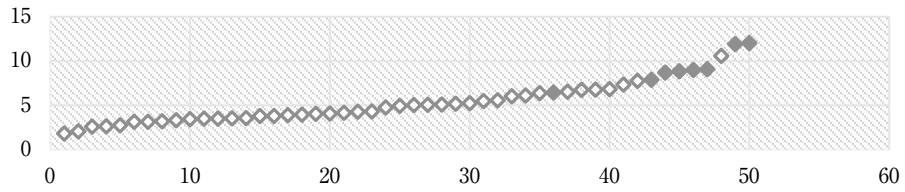

図 2　固有名詞を除いた同上のグラフ

固有名詞以外には、戯曲は付加形容詞が少なく、前置詞・後置詞が少なく、人称代名詞がやや多く、接続詞がやや少なく、関係代名詞が少なく、命令形が多く、疑問代名詞が顕著に多く（戯曲平均では約 0.6125、その他の平均では約 0.2143 であるから、3 倍に近い。ただし、疑問代名詞自体が全体に占める割合は小さいので、影響も小さい）、句点およびその他の文記号が非常に多かった。つまり、一文が短く、そこでの形容は少な

く、また文ごとの関係は弱く、命令や疑問が多い。口語的なジャンルとしての特徴をどの作品も帯びており、それが品詞の割合に大きな影響を与えているということになる。

二．ただし、その戯曲と散文との乖離は、確実に両者を区別できるほどの差ではない。

地誌という特殊なジャンルである『コスモス』が戯曲並みに平均から離れているのは例外としても、『ファウスト 2 部』は『戦争論 2 巻』『判断力批判』『ラオコーン』など極めて散文性の高い論文に囲まれている。小説を多く含むリストであるから論文が高い数値を示すにしても、『ミヒャエル・コールハース』は小説ながら『ファウスト 2 部』を上回っている(1)。絶対値の差であれば『コスモス』以外の散文と戯曲はきれいに分かれるが、『グリム童話』と『ファウスト 2 部』の差はごくわずかである。

三．戯曲と散文の違いと作家の違いでは前者の方が大きな影響を及ぼす。

ゲーテ、シラーなど複数のジャンルの作品を収めた作家について、それらの作品が固まった場所に位置することはなかった。それに比べ、戯曲の方がまとまりやすい傾向を示している。具体的に二乗の和の根について分散を求めると、戯曲 8 作品が約 4.26 なのに対し、ゲーテ 5 作品が 6.16、シラー 4 作品が約 6.05 である。もっとも、これはあくまで平均的な文体との比較であって、個々の作家の文体の幅を知りたければ、その作家のみの品詞ごとの中央値を基準として調べるべきである。このことは次の第四考察にも当てはまる。

四．作家は同じ傾向の作品を近い品詞の割合で書くことができるかもしれない。

『意志と表象としての世界』『戦争論』『ファウスト』の連作 3 作品は、いずれも執筆時期が長期に及び、あるいは巻ごとに執筆時期が離れているが、近い数値に位置している。この点については、他の連作についても調べるほか、それぞれの作家の同時期の他作品との比較を行うことで、より「作家が意図して同じトーンを保てるか」に関心を絞った研究が行える。ゲーテについては、『ファウスト』と他の作品で大きく差が出ているが、『ファウスト』が戯曲であることを考慮に入れるなら、彼による他の戯曲との比較がより有益である。

7. 上記の考察に基づくさらなる考察

ここまで注目してきたのは、品詞の違いを無視した全体的な差である。50 作品の全品詞の中央値からどの程度離れているかを総合的に見て、作品がどれくらい特徴的な文体で

あるかを判断してきた(2)。しかし、HanTa は品詞別に細かく割合を調べられるのであるから、この機能により、さらに一歩踏み込んだ考察ができる。

たとえば『コスモス』が戯曲とならんで（一般的な小説と比べて）特徴的な文体の作品であるとプログラムが判定することは前記した。では、それは品詞の割合においても似た特徴を持つのだろうか。それとも戯曲とは品詞の割合の特徴は異なるのだろうか。あるいは、ゲーテは『イタリア紀行』では最も中央値に近い文体を採用することもできているが、『若きヴェルテルの悩み』では大きく中央値から逸れた文体を用いている。執筆時期が異なるとはいえ、癖のない文章を書くことの可能なゲーテが『ヴェルテル』で中央値を逸れたとすれば、それを意図的と考えることもできる。では、どの品詞において特徴的な文体を用いているだろうか、そしてそれは作品を特徴付けているだろうか。

まず前者について見てみる。『コスモス』と中央値を品詞ごとに比較すると、次の表のように固有名詞、数詞、外来語、その他文記号が非常に多い。逆に代名詞、副詞、否定辞、動詞が非常に少ない（なお、表では代名詞や動詞は代表的なものを抜粋した）。

	固有名詞	副詞	数詞	外来語	人称代名詞	付加所有代名詞	否定辞	動詞定形	動詞不定形	コピュラ動詞定形	助動詞定形	その他文記号
中央値	1.815	6.675	0.62	1.245	5.87	1.775	0.82	5.34	2.095	2.855	1.135	0.235
『コスモス』	7.54	3.73	4.03	6.55	1.22	0.62	0.33	2.34	0.81	1.89	0.29	2.54

地誌というジャンルからは、固有名詞と外来語が多くなるのは当然と言える。数詞とその他文記号については、『コスモス』のテクストには括弧に挟まれた文献紹介および注釈 Fußnote が頻繁に挿入されており、頁数が数字の量を、括弧がその他文記号の量を増やしている。動詞が少ないのもこうした注釈部分で省かれることが一因と考えられ、副詞が少ないのは淡々とした記述だから、否定辞が少ないのは事実を記述していくから、代名詞が少ないのは固有名詞が多く、話題が次々に変わっていくために既知の名詞を指すことが少ないからであると「推測」できる。こうした推測を確信に変えるには、これまでの文学研究的な意味で作品に親しむことがなお必要であるが、本論文ではこれを行わない。作品研究においてデジタル的な分析と伝統的な精読が両立するテーマがあることを示唆するに留める。

なお、機械による品詞分析に基づいたこのような推測があまりに当たり前な内容である（つまり、プログラムでわざわざ分析する必要がない）という指摘については、研究者の経験による推測とプログラムの分析結果が一致するなら、より推測内容が補強されて望ましいということになる。そのうえで、もし推測と異なる分析結果が出た場合には、それを

説明しようと作品を読み返すことで新たな発見をもたらしうるのである。

先述した戯曲の特徴との違いから分かるように、品詞全体の平均からの距離で見れば似た数値になる『コスモス』と戯曲類であるが、品詞の割合にまで目を向ければ、はっきりと違った傾向を示している。コンピュータが近年、めざましい発展を遂げているジャンルの一つに機械学習がある。たくさんのデータを読み込ませることで、データを分類・評価する基準を機械自体が作り上げるもので、たとえばGutenberg-DEにある文章のうちどれが戯曲であるかを教え込めば、ある文章が戯曲であるかどうかを高い確率で判定してくれると思われる。もちろん、戯曲と分かっているものを戯曲と判定させることには機械の進化を示す以上の意味はないが、戯曲でありながらそうでないと判定される作品や、戯曲ではないのに戯曲と判定される作品については面白い考察が可能であろう。

次に、ゲーテ『ヴェルテル』について品詞レベルでの考察を行う。なおGutenberg-DEに掲載されているテクストは、1774年の初版ではなく、より広く普及した1787年の第2版である。この作品の品詞割合は『コスモス』や『ミヒャエル・コールハース』ほど極端な数値を示しているわけではないが、表1から分かるように、小説のなかでは中央値からかなり離れた数字を示している。『ヴェルテル』がドイツの文豪ゲーテの小説としての代表作（の一つ）であることを考えればやや意外な結果であり、はたして私たちが何も考えずに同作品を読んだときに、この小説は独特の文体だと感じる（したがって内容より文体に目が行く）かと言えば、そうではないかもしれない。では、具体的にどの品詞の割合が特徴的なのか、それを踏まえたうえでどう作品を解釈できるか。

『ヴェルテル』の品詞割合の特徴として、表の通り人称代名詞、付加所有代名詞、数詞、句点が多く、その分、固有名詞、形容詞、副詞、前置詞、冠詞の割合が低い。

	固有名詞	付加形容詞	叙述形容詞	副詞	前置詞	冠詞	数詞	人称代名詞	付加所有代名詞	句点
中央値	1.815	5.2	2.65	6.675	6.295	8.335	0.62	5.87	1.775	4.32
『若きヴェルテルの悩み』	1.52	3.75	2.2	5.7	5.53	6.61	1.24	9.2	2.62	5.34
『イタリア紀行』	2.01	5.75	3.05	7.47	6.56	8.53	0.44	5.14	1.22	4.37
『ヴィルヘルム・マイスター』	1.56	4.24	2.64	6.48	6.04	7.15	0.75	7.47	2.29	4.31
『ファウスト二部』	3.08	4.2	4.12	6.57	3.98	5.8	0.46	7.78	1.05	7.97
『ファウスト一部』	3.05	3.13	2.99	6.72	4.35	6.26	0.34	8.45	1.43	10.1

人称代名詞と付加所有代名詞が多く固有名詞が少ないということは、一つには登場人物

が少ないために、同じ人物が登場し続ける（固有名詞で誰の言動かを再度明示する頻度が低い）ということが考えられる。確かに『ヴェルテル』では、主要人物が限られており、それ以外の固有名詞によるセットに切り替わる頻度が低いと思われる。冠詞が少ないことも、付加所有代名詞が多いことを理由に説明できる。また、書簡体小説の性質上、場面が切り替わっても固有名詞から始まらず、一人称の ich やその所有形容詞の mein から語られ始めることになり、このことも頻度を高めていると考えられる。なお、書簡ごとに日付が付されているため、数詞も増える。

さらに句点が多いということは、一文の長さが短いと言うことである。文は短くとも、多くの場合、主語と動詞を含む。その結果、少なくなるのは形容表現である。『ヴェルテル』ではしばしば風景描写が行われ、主人公のヴェルテルはゲーテ同様に絵を描くのが得意なだけに、それらも見事ではあるが、作品の主題となっているのは主人公の心理であり、これは風景描写等に比べれば形容の少ないものになると思われる。上記の特徴も含めて、書簡で一人称で述べられていることの影響が大きいとすれば、書簡および書簡体小説についてのデータベースを作成し、それらの作品に共通する特徴と比較することでより確実なことが言えるはずである。

このように推測的に述べている内容については、分析にかける作品数を増やすとともに、テクストを精読することによってより確実にしうる。本論文はあくまで計量的分析の導入を目指すものであって、『ヴェルテル』一作品を取り上げて分析するものではないため、そうした研究は別の機会（あるいは別の方）に譲るが、これまで人文系で培われてきた精読の能力はデジタル技術を導入してもなお役立ちうることは述べておきたい。人文系の伝統的研究に根拠を与えたり、あるいは新たな視点をもたらしたりしてくれるツールとして、個々の研究者に有用な範囲で用いることは、研究全体の水準を上げることにつながると考えられる。何より、私が研究対象とする時代の作家達は、新しい知の在り方に意欲的であった。そうした態度を追体験しておくのは、作家の心情を知るうえで無駄にはならないと思うのである。

おわりに

本論文は、The Hannover Tagger を用いたドイツ語作品の計量的分析を小規模に行い、その有用性について示したものである。個々の作品の品詞分解からその特徴に気づくこともできるし、時代や作家によって選択した複数の作品によるデータベースを用意することで、他の作品と比較しての特徴把握も可能であることは示せたと考えている。本論文が個人研究者にできうることを示したことで、自らが研究対象とする作家について、同様の

データを用いたより深い考察を行う他の研究者が現れれば幸いである。

今後の計画としては、『グリム童話』と『伝説集』など、メルヒェンと伝説での固有名詞の比率の違いを検討したいと考えている。また、HanTa にこだわらず NLTK の単語切り分け機能を有効利用すれば、用いる語彙の特徴からティークとヴァッケンローダーの共著についてどちらの文章か判定することもできるだろうし、芸術家小説に頻出する語彙の割合を調べることで、これまでジャンルとしての境界が曖昧であった芸術家小説について、当てはまるか否かの線引きを試みることもできるだろう。さらに、英語だけでなく複数の言語での品詞分解の精度が高まった状況を利用するならば、翻訳による品詞割合の違いから、翻訳という行為や翻訳家ごとの特徴、また言語の特徴について考えることもできるはずである。

注

（1） なお、『ミヒャエル・コールハース』については、形容詞と副詞が少なく、前置詞が多く、従属接続詞と関係代名詞がそれぞれ全作品中最も多く、句点はやや少ないが読点が他を圧倒して多い。とりわけ読点はもともと文における割合が多いため、中央値が約 8.31％のところ約 14% となっていることが全体としての高さにつながっている。とはいえ、50 作品あるなかで 3 つの品詞類で一番高い割合を示す作品であるから、かなり特徴的な作品である。これが作家性か、作品単体の特徴であるかを知るためには、彼の作品をより多く分析することになろう。

（2） もちろん、「特徴的な文体」は品詞の割合だけで測れるものではない。たとえば動詞の頻度が同じでも、それが基礎的な語彙を中心とするか、使用頻度の低い語彙を中心とするかで読者に与える印象は全く異なる。また、特定の文字を使わないいわゆる「リポグラム」の手法を用いた小説も、それが品詞バランスに影響を与えない限り、本研究の手法では目立たない。ただし、品詞の割合が極端に平均から外れるのに平凡な印象を与える文体は稀であるから、特徴的な文体であることを証明する一つの方法にはなる。

参考文献

杉浦清人「文学研究におけるデジタル・ヒューマニティーズの可能性 ── 文章心理学・計量文献学・マクロ分析 ──」『れにくさ : 現代文芸論研究室論集』第 7 号、2017 年、80-96 頁。

吉澤弥生、奥彩子、堀新「デジタル人文学の研究と教育に関する基礎的研究」『共立女子大学・共立女子短期大学総合文化研究所紀要』第 27 号、2021 年、73-86 頁。

ブラット、ベン（坪野圭介訳）『数字が明かす小説の秘密　スティーヴン・キング、J・K・ローリングからナボコフまで』、DU BOOKS、2018 年。

モレッティ、フランコ（秋草俊一郎、今井亮一、落合一樹、高橋知之訳）『遠読 ── 〈世界文学システム〉への挑戦』、みすず書房、2016 年。

Kharis, Muhammad et al., “Tokenization and Lemmatization on German Learning Textbook Level A1 of CEFR Standard.” *Journal of Higher Education Theory and Practice* 22, no. 1 (2022): 141-152.

Wartena, Christian. “A Probabilistic Morphology Model for German Lemmatization.” In *Proceedings of the 15th Conference on Natural Language Processing* (*KONVENS 2019*), edited by the Chair of Computational Corpus Linguistics, 40–49.

Project Gutenberg-DE. Accessed March 30, 2022. https://www.projekt-gutenberg.org/.