



ゲルマニスティーク分野間の学術テキストに関する 計量的考察 : 語彙・品詞・N グラムに注目して

今道, 晴彦

(Citation)

統計数理研究所共同研究リポート, 465:55-65

(Issue Date)

2023-03-10

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD0I)

<https://doi.org/10.24546/0100479385>

(URL)

<https://hdl.handle.net/20.500.14094/0100479385>



ゲルマニスティック分野間の学術テキストに関する計量的考察

—語彙・品詞・N グラムに注目して—

今道 晴彦(広島大学)

imamichi@hiroshima-u.ac.jp

A Quantitative Study on Texts among the Sub-disciplines of German Studies —Focusing on Frequency of Word, Part of Speech, and N-gram— IMAMICHI, Haruhiko (Hiroshima University)

概要

本研究では、ゲルマニスティック分野間の学術テキストにおける言語特性を概観するため、ドイツ文学、ドイツ語学、ドイツ語教育に関連する学術論文ないし書籍と、比較の際の参考資料として新聞記事をコーパスとして使用し、計量的観点から分析を試みた。その結果、高頻度語(RQ1)、品詞構成(RQ2)、N グラム(RQ3)のいずれにおいても、学術テキストと新聞記事との差が最も顕著に見られ、次に文学テキストと語学テキスト(とくに語学論文)との対立が確認された。文学テキストではとりわけ所有代名詞、人称代名詞、時を表す表現などが、語学テキストでは関係代名詞、話法助動詞、助動詞、依存や根拠を表す表現、問題点を指摘する際に使用される表現などが特徴的に見られ、隣接分野であるとは言え、多用される表現が両テキストでは異なることが示唆された。また、語学テキストよりも文学テキストのほうがテキスト間の類似性が高いことなどが確認された。

キーワード

学術テキスト, ゲルマニスティック, 高頻度語, 品詞構成, N グラム

1. はじめに

国内のドイツ語学習者の大半は初修外国語としてドイツ語を履修する大学生が占める(Auswärtiges Amt, 2020)。そのため、大学ドイツ語教育では、ドイツ文学、ドイツ語学、ドイツ語教育などのいわゆるゲルマニスティック(Germanistik, 英 German studies)分野のためのドイツ語よりも初修ドイツ語の教育に重きが置かれがちである。また、ドイツでは留学生にとってゲルマニスティックは必ずしも人気の高い分野であるとは言えず(Statistisches Bundesamt, 2022), Fachsprache(専門分野に関するドイツ語)や Wissenschaftssprache(学術分野に関するドイツ語)としてこれまで注目されてきたのは、もっぱら留学生の多い工学系や社会科学系の分野のドイツ語であった(Hoffmann, 1976; Roelcke, 2020)。

一方で、近年研究者は研究成果を外国語で公表することが推奨されており、学生も例外ではなくなっている。ゲルマニスティックを専攻する学生のなかには、修士論文や卒業論文であってもドイツ語で執筆する学生も一定数おり、ドイツ語による論文概要の提出を課している大学も少なくない。

また、当該分野で必要とされるドイツ語は、初修ドイツ語の延長線上に位置付けられるものでもない。つまり、ゲルマニスティック分野を学ぶには、高度なドイツ語運用能力のみならず、当然ながら当該分野の言語特性についても合わせて心得ておかなければならない。しかし、筆者の知る限り、現在までのところ、ドイツ語の論文要旨の作成や、ゲルマニスティック分野のドイツ語を学ぶための教育プログラムや授業を提供する大学はなく、仮にあったとしてもごく少数の大学であると思われる。その限りにおいて、大学ドイツ語教育では、当該分野を学ぶためのドイツ語にスポットを当て、その言語特性に関する知見を蓄積していくことで、より良い学習方法とは何かを探っていくことが今後の重要な課題になる。

言語特性に関する知見を蓄積する際には、客観的基準に従うことが重要であり、計量的手法に基づく分析は、有効な手法のひとつになると考えられる。ドイツ語の学術テキストに関しても、これまで質的研究に加えて、いくつかの量的研究がなされてきた。たとえば、Köhler(1968)は、技術分野関連のテキストを対象に、動詞の使用傾向を調査し、頻繁に使用される動詞が、特定の動詞や、意味の希薄な動詞であったと報告している。

また、Kempter(1969)は、化学と物理学の教科書を対象に、文構造を複雑にする要因のひとつである付加語(Attribut)の有無を調査した結果、およそ5割が後置修飾の付加語をひとつ取る事例で、約3割が付加語を取らず、基本的にこれらのテキストでは、文構造を複雑にしない表現が好まれるとしている。

また、Scheffe(1975)は、統語レベルにおいて医学テキスト、経営学テキスト、文学テキストの比較を試み、医学テキストでは名詞表現が、文学テキストでは動詞表現や従属文が多用され、経営学テキストではこれらのテキストの中間的な傾向を示したと報告している。また、どの言語手段を優先するかは各テキストの動機と関連しており、経営学テキストでは抽象的な関係を叙述するために従属文が選択されるのに対して、文学テキストでは言語手段のバリエーションのひとつとして従属文が多用される傾向にあるとしている。

しかしながら、計量的観点に基づくこうした研究は、もっぱら1960年代から1970年代にかけて行われてきた。さらに、これらの研究は小規模のサンプル調査であるため、テキスト処理技術やデータ分析手法が進展した現在においては、より包括的な調査が期待される。

2. リサーチデザインと手法

2.1 研究目的とRQ

そこで、本研究では、ゲルマニスティック分野の学術テキストをコーパスとして、どのような言語特性が見られるのかを計量的観点から探索的に調査する。ただし、国内の大学のゲルマニスティック分野に関しては、通例、専攻の学生を対象にドイツ語の授業が行われ、他分野の学生が参加する

ことは稀である。そのため、当該分野のドイツ語が他分野で使用されるドイツ語と比較してどのようにユニークなのかという問題よりも、ドイツ文学、ドイツ語学、ドイツ語教育の主要 3 分野でどのような違いや共通点が見られるのかという問題のほうがドイツ語教員にとっては指導上より重要な意味を持つ。したがって、本研究ではゲルマニスティック分野間の比較を通して当該分野の言語特性を探る。

本研究では、分析項目として、機械的に収集可能で、コーパス言語学で取り上げられることが多い高頻度語、品詞構成、N グラム(語彙連鎖)の 3 項目に限定する。したがって、以下 3 点のリーチクエスチョン(RQ)を設定し、3 分野の計量的傾向を抽出することで、ゲルマニスティック分野のためのドイツ語に関しての教育的示唆を得ることを試みる。

RQ1: 高頻度語に差が見られるか?

RQ2: 品詞構成に差が見られるか?

RQ3: N グラム頻度に差が見られるか?

2.2 データ

本研究では、ドイツ文学、ドイツ語学、ドイツ語教育に関連する学術テキストと、比較用の参考資料として新聞記事のコーパスを構築し、分析に利用する。学術テキストに関しては、ドイツの学術出版社大手の De Gruyter 社 (<https://www.degruyter.com/>) が Open Access として公開する 2000 年以降の電子書籍および電子ジャーナルを、新聞記事については、ライプツィヒ大学情報学研究所の Deutscher Wortschatz (<https://wortschatz.uni-leipzig.de/de/download/German>) が公開する新聞記事データのひとつ(2021 年の新聞記事より無作為に抽出された 1 万文のデータ)を利用する。

具体的には、以下の表 1 に示す(1)ドイツ文学書籍コーパス(以下、文学書籍)、(2)ドイツ文学論文コーパス(以下、文学論文)、(3)ドイツ語学書籍コーパス(以下、語学書籍)、(4)ドイツ語学論文コーパス(以下、語学論文)、(5)ドイツ語学書評論文コーパス(以下、語学書評)、(6)ドイツ語教育論文コーパス(以下、教育論文)と(7)新聞記事コーパス(以下、新聞)を分析データとする。

なお、専門分野の分類については、De Gruyter 社のサイトの分類に準じることとし、Literaturwissenschaft(文学)で抽出されるテキストを「ドイツ文学」、Linguistik und Semiotik(言語学と記号学)で抽出されるテキストを「ドイツ語学」とする。ただし、「ドイツ語教育」は、Linguistik und Semiotik に含まれる上、収録文献が少ないことから、電子ジャーナルとして公開される Info DaF のみを扱うこととする。

また、電子書籍については、特定の著者の影響を回避するため、分析対象を単著の書籍ではなく、各章で著者が異なる共著の書籍に限定し、1 冊あたり 3~5 章程度を目安に無作為に収集する。

表 1

コーパスサイズ

コーパス	検索条件・出典等	総語数
(1) 文学書籍	Fachgebiet(専門分野)→Literaturwissenschaft(文学), Dokumenttyp(文書タイプ)→Kapitel(章), Sprache(言語)→Deutsch(ドイツ語)で抽出	805,429 語
(2) 文学論文	Fachgebiet(専門分野)→Literaturwissenschaft(文学), Dokumenttyp(文書タイプ)→Artikel(論文), Sprache(言語)→Deutsch(ドイツ語)で抽出 ※大半は Zeitschrift für interkulturelle Germanistik に収録の論文	570,117 語
(3) 語学書籍	Fachgebiet(専門分野)→Linguistik und Semiotik(言語学と記号学), Dokumenttyp(文書タイプ)→Kapitel(章), Sprache(言語)→Deutsch(ドイツ語)で抽出	247,821 語
(4) 語学論文	Fachgebiet(専門分野)→Linguistik und Semiotik(言語学と記号学), Dokumenttyp(文書タイプ)→Artikel(論文), Sprache(言語)→Deutsch(ドイツ語)で抽出 ※大半は Zeitschrift für Sprachwissenschaft に収録の論文	761,157 語
(5) 語学書評	Zeitschrift für Rezensionen zur germanistischen Sprachwissenschaft ※書評論文のみで構成されるジャーナル	771,634 語
(6) 教育論文	Info DaF	636,265 語
(7) 新聞	Universität Leipzig Deutscher Wortschatz News 2021	148,668 語

2.3 手法

2.3.1 データの事前処理

本研究で使用する上記 7 つのコーパスには、品詞タグ付けソフトの **TreeTagger** を用いて品詞情報を付与し、語彙頻度表と品詞頻度表(ともに 10 万語あたりの調整頻度表)を作成する。その際、数字、記号、句読点は、分野間の比較を行うことを目的とする本研究には大きく影響しないことから分析対象から除外する。作成する語彙頻度表に関しては、総頻度の上位 100 位までの語彙とする。

品詞については、ベルリン・ブランデンブルク科学アカデミー(BBAW)の DWDS コーパス(<https://www.dwds.de/>)の品詞タグ分類(20 分類)に従い、「その他」、「文記号」、「名詞」を除く 17 種の品詞(形容詞、副詞、前置詞、決定詞、接続詞、指示代名詞、不定代名詞、人称代名詞、所有代名詞、関係代名詞、再帰代名詞、疑問代名詞、代名詞副詞、不変化詞、動詞、助動詞、話法助動詞)を分析対象とする。

また、抽出する N グラムについては、なるべく構文レベルの特徴を拾い上げるため、コーパス言語学で一般的な 4 グラムに限定し、同じく 4 グラム頻度表(10 万語あたりの調整頻度表)を作成する。語彙頻度表同様に、総頻度の上位 100 位のデータを以下の分析に使用する。

2.3.2 分析の手順

RQ1(高頻度語)では、上記の 10 万語あたりの語彙頻度表を用いて 7 コーパスを第 1 アイテム、

上位 100 語を第 2 アイテムとしてコレスポンデンス分析を行い、各コーパスと語彙との間にどのような関連が見られるかを検討する。

RQ2(品詞構成)では、上記の 10 万語あたりの品詞頻度表を用いて 7 コーパスを第 1 アイテム、17 品詞を第 2 アイテムとして同じくコレスポンデンス分析を行い、各コーパスと品詞との間にどのような関連が見られるかを検討する。

RQ3(N グラム)では、上記の 10 万語あたりの 4 グラム頻度表を用いて 7 コーパスを第 1 アイテム、上位 100 位の 4 グラムを第 2 アイテムとしてコレスポンデンス分析を行い、各コーパスと N グラムとの間にどのような関連が見られるかを検討する。

3. 結果と考察

3.1 RQ1 高頻度語に差が見られるか？

10 万語あたりの語彙頻度表を用いて 7 コーパスを第 1 アイテム、上位 100 語を第 2 アイテムとしてコレスポンデンス分析を行なった結果、図 1 が得られた。第 1 次元の寄与率は 48.44%、第 2 次元の寄与率は 21.32%で、合わせると 69%の説明力を有する次元が取り出せたことになる。

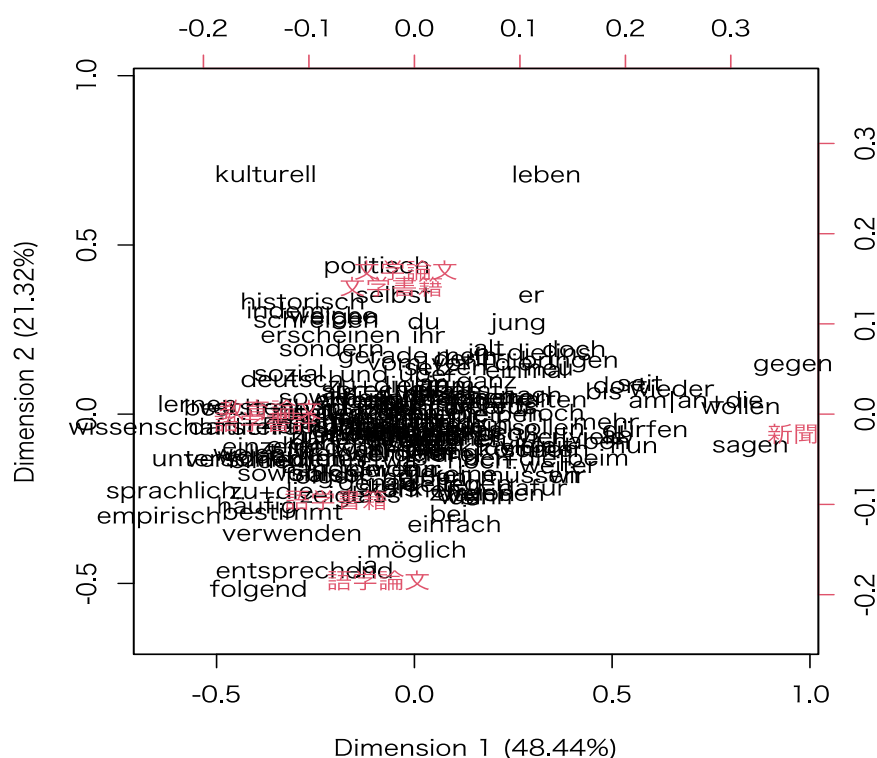
第 1 次元を示す横軸に注目して 7 コーパスの位置関係を確認すると、プラス方向に新聞が、マイナス方向にその他の学術テキストがポジショニングされているのがわかる。語彙頻度に関しては、分野間の違い以上に、新聞記事と学術テキスト間のジャンル差のほうが大きいと言える。

次に、2 番目に説明力の強い縦軸を見ると、プラス方向に文学論文、文学書籍が、マイナス方向に語学論文、語学書籍が、その中間の原点付近に教育論文、語学書評、新聞が布置されているのが見て取れる。したがって、第 2 次元の縦軸が分野差を示す軸になっており、文学テキストと語学テキストの対立になっていると考えられる。

さらに、文学テキストの 2 つのコーパス(文学学論文、文学書籍)はともに重なる形で布置されるのに対して、語学テキストはそれぞれやや距離を取る形でポジショニングされており、語学テキストよりも文学テキストのほうが等質性は高いことがわかる。

図 1

第 1 アイテムと第 2 アイテムのスコア散布図(名詞, 略語を除く頻度上位 100 語)



また, 上位 100 語の分布を見ると, 多くは図の中央にある原点付近に集中するが, 左右上下に広がる形で分布しているのがわかる。以下の表 2 は, 新聞記事と学術テキストの差を示す第 1 次元において, それぞれに特徴的な各語のスコアを示したものである。便宜的に訳語を与えているが, 必ずしも各語を代表したものであるとは限らない。

表 2

新聞と学術テキストの識別に関わる語彙

新聞	得点	学術テキスト	得点
gegen (～に対する)	5.75	empirisch (経験的な)	-3.88
sagen (言う)	5.10	sprachlich (言語の)	-3.69
wollen (～したい)	4.96	wissenschaftlich (学問の)	-3.66
am an+die (～で)	4.28	lernen (学ぶ)	-3.29
wieder (再び)	3.92	unterschiedlich (様々な)	-2.51
dürfen (～してもよい)	3.50	darstellen (叙述する)	-2.48
seit (～以来)	3.42	wobei (その際に)	-2.42
nun (今や)	3.33	häufig (たびたび)	-2.39
dort (そこに)	3.12	folgend (以下の)	-2.37
ab (～から)	3.08	verstehen (理解する)	-2.31

新聞に特徴的な語を見ると, まず 1 番目に「～に対する」という意味の前置詞 gegen がランクさ

れる。これは、新型コロナウイルスの影響(例:**gegen** das Coronavirus)のほか、時刻(例:**gegen** 12.50)についての言及や、サッカーをはじめとするスポーツの試合(例:**gegen** Frankfurt)が話題にされたためである。次の発話動詞 **sagen** は、第 3 者の発言が取り上げられることが多い新聞記事の特徴を示していると言える(例:..., **sagte** Bundeskanzlerin Angela Merkel...)。話法助動詞(**wollen**, **dürfen**)は、意志や、許可・推量などの意味を担っており、新聞記事では動詞が表す事態を直接表現するよりも、これらのニュアンスが付け足される場面が多いことを意味している。また、前置詞や副詞(**am**|**an**+**die**, **seit**, **ab**, **nun**, **dort**)は、「いつ」、「どこで」といった基本情報を記す新聞記事の特徴を反映したものと解釈できる(例:**am** Freitag, **seit** Beginn der Corona-Pandemie)。

一方、学術テキストに関しては、**empirisch**, **sprachlich**, **wissenschaftlich**, **unterschiedlich** といった形容詞や、**lernen**, **darstellen**, **verstehen** といった動詞など、研究活動に関わる具体的な内容を伴う内容語が特徴的に使用されていることがわかる。

次の表 3 は文学テキストと、語学論文や語学書籍の差を示す第 2 次元における各語のスコアを示したものである。文学テキストに特徴的な語として、**kulturell**, **politisch**, **historisch** は、文化、政治、歴史との関わりが強いことを示している。また、**leben**, **selbst**, **eigen** は、文学が生や自己を話題にすることを示している。また、人称代名詞 **er** は男性名詞を受けることも可能だが、ここでは概ね男性を受けている。

表 3

文学テキストと語学論文・語学書籍の識別に関わる語彙

文学テキスト	得点	語学論文・語学書籍	得点
kulturell (文化的な)	6.43	folgend (以下の)	-4.72
leben (生きる)	6.42	entsprechend (～に応じて)	-4.23
politisch (政治的な)	3.94	ja (はい、確かに)	-4.08
selbst (自分で)	3.20	möglich (可能な)	-3.67
er (彼は)	3.15	verwenden (使う)	-3.17
historisch (歴史的な)	3.02	einfach (簡単な)	-2.93
indem (～によって)	2.79	empirisch (経験的な)	-2.76
welche (どの)	2.65	bei (～のそばに)	-2.65
eigen (自分の)	2.56	bestimmt (特定の)	-2.60
schreiben (書く)	2.54	häufig (たびたび)	-2.49

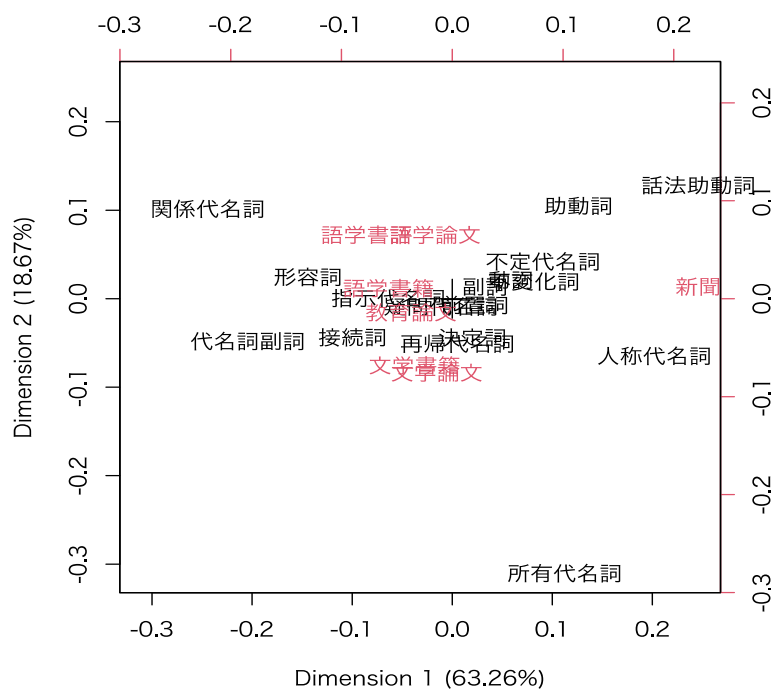
次に、語学論文や語学書籍に特徴的な語を見ると、**folgend** が最も高く、読者の意識を「次の」項目に導く場面が多いことが窺える(例:**im Folgenden**)。2 番目の **entsprechend** は前置詞としての用例が多く、語学論文や語学書籍では依拠を示す場面が多いことを示唆している(例:**entsprechend der Reihenfolge ihres Vorkommens**)。また、その次の **ja** は通常、肯定表現として使用されるが、書き手の主観的心情を表す心態詞(**Modalpartikel**)として用いられることが多いと言える。

3.2 RQ2 品詞構成に差が見られるか？

次に 10 万語あたりの品詞頻度表を元に、7 コーパスを第 1 アイテム、17 品詞を第 2 アイテムとしてコレスポンデンス分析を行なったところ、次の図 2 が得られた。

図 2

第 1 アイテムと第 2 アイテムのスコア散布図(名詞を除く 17 品詞)



第 1 次元の寄与率は 63.26%, 第 2 次元の寄与率は 18.67%で、8 割を超える説明力を有する次元が取り出せたことになる。第 1 次元の横軸に注目して、コーパスの分布に目を向けると、プラス方向に新聞が、原点付近に学術テキストが布置されており、ここでも新聞と学術テキストの対立関係が確認できる。新聞ではとりわけ話法助動詞、人称代名詞、助動詞などが、学術テキストでは関係代名詞、代名詞副詞、形容詞、接続詞などが特徴的に使用されているのがわかる。

また、第 2 次元の縦軸に目を向けると、プラス方向に語学書評と語学論文が、マイナス方向に文学論文と文学書籍が布置され、その中間に語学書籍、教育論文、新聞が位置しており、RQ1 同様に、文学テキストと語学テキストの対立が窺える。ただし、品詞頻度を対象とする RQ2 では、語学書籍と語学書評の位置関係が入れ替わっている点で、RQ1 とは異なる結果が得られたと言える。品詞の分布に着目すると、原文からの引用も多い文学テキストではとりわけ所有代名詞、人称代名詞などが、プラス方向の語学テキストではとりわけ話法助動詞、助動詞、関係代名詞などが顕著に使用されていることが観察される。このうち助動詞は、受動文(動作受動)での使用例が多いことが関係しており、語学テキストにおいて関係代名詞が特徴的に使用されるのは、文学テキストよりも具

にする点で共通していると言える。一方、語学論文では、依存や根拠に関する表現に加えて(in Abhängigkeit von der, auf der Basis der), 問題を言及する際に使用される表現が多い(es sich dabei um, handelt es sich bei, sich die Frage ob, dass es sich um)。とくに、後者の表現は、句表現ではなく、動詞を伴っており、構文レベルに及ぶ特徴と言える。

表 4

新聞・文学テキストと語学論文の識別に関わる 4 グラム

新聞・文学テキスト	得点	語学論文	得点
mit Blick auf das (～に関して)	3.70	in Abhängigkeit von der (～に依存して)	-2.15
in der Geschichte der (～の歴史において)	3.02	auf der Basis der (～に基づき)	-2.14
in dem sich die (.....)	2.82	es sich dabei um (～が問題である)	-1.84
Anfang der er Jahre (～年代初頭に)	2.43	in dem Sinne dass (～という意味で)	-1.83
nicht in der Lage (～できない)	1.87	handelt es sich bei (～が問題である)	-1.77
im Zusammenhang mit der (～に関して)	1.83	an dieser Stelle nicht (この点で...でない)	-1.71
Mitte der er Jahre (～年代の中頃に)	1.73	im Gegensatz zu den (～と比べて)	-1.60
in die er Jahre (～年代まで)	1.65	sich die Frage ob (～かどうか問題)	-1.56
in den er Jahren (～年代に)	1.64	z B in der (例えば～において)	-1.56
die sich in der (.....)	1.46	dass es sich um (～が問題である)	-1.53

4. まとめ

本研究では、ゲルマニスティック分野間の言語特性の共通点や違いを捉えるため、文学書籍、文学論文、語学書籍、語学論文、語学書評、教育論文、新聞の 7 つのコーパスを用いて、高頻度語、品詞構成、N グラムに基づき計量的観点から分析を試みた。

RQ1(高頻度語)については、新聞記事と学術テキスト間の差が最も強く見られ、次に文学テキストと語学テキストの対立が見られた。その際、文学テキストでは、文化、政治、歴史に関わる語彙が多く用いられることが確認された。

RQ2(品詞構成)においても、まず、新聞と学術テキストの対立関係が見られ、次に文学テキストと語学テキストの対立が見られた。文学テキストでは、所有代名詞、人称代名詞などが、語学テキストでは話法助動詞、助動詞、関係代名詞などが特徴的に使用されることが観察された。

RQ3(N グラム)においては、文学テキストと語学論文が対立し、文学テキストのさらに先に新聞が位置づけられる関係にあることがわかった。とくに文学テキストでは時間表現が多いのに対して、語学論文では依存や根拠、問題について言及する表現が顕著に用いられることが判明した。

以上のことから、隣接する分野であっても、多用される表現には違いが見られることが見て取れる。ドイツ語学習への示唆という観点で上記の結果を踏まえると、ドイツ文学専攻の学生は、とりわ

け時間表現を、ドイツ語学専攻の学生は、話法助動詞、受動文、関係文に慣れるとともに、依存や根拠、問題について言及する表現などを優先的に学習する必要があると考えられる。

ただし、本研究で使用したデータは、一般化するにあたって数量面で十分とは言えない。今後コーパスのサイズ、種類、分析項目等の拡充が必要である。

謝辞

本研究は JSPS 科研費(JP21K00735)の助成を受けた。

引用文献

- Auswärtiges Amt (2020). *Deutsch als Fremdsprache weltweit: Datenerhebung 2020*.
<https://www.auswaertiges-amt.de/blob/2344738/b2a4e47fdb9e8e2739bab2565f8fe7c2/deutsch-als-fremdsprache-data.pdf>
- Hoffmann, L. (1976). *Kommunikationsmittel Fachsprache: Eine Einführung*, Akademie Verlag.
- Kempter, F. (1969). Die präpositionale Wortgruppe in der Sprache der Chemie und Physik: Beziehungen zwischen Struktur und Mitteilungsgehalt. *Deutsch als Fremdsprache*, 6(6), 439-445.
- Köhler, C. (1968). Zur Verwendung des Verbs in technischer Literatur — insbesondere bei der sprachlichen Realisierung von Zuordnungen: Ein Beitrag zur Beschreibung der Fachsprache der Technik und zu ihrer Behandlung im Deutschunterricht für Ausländer (1). *Deutsch als Fremdsprache*, 5(2), 89-95.
- Roelcke, T. (2020). *Fachsprachen*. Erich Schmidt.
- Scheffe, P. (1975). *Statistische syntaktische Analyse von Fachsprachen mit Hilfe elektronischer Rechenanlagen am Beispiel der medizinischen, betriebswirtschaftlichen und literaturwissenschaftlichen Fachsprache im Deutschen*, Göttingen.
- Statistisches Bundesamt (2022). *Bildung und Kultur: Studierende an Hochschulen Sommersemester 2022*. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/studierende-hochschulen-ss-2110410227314.pdf?__blob=publicationFile