

ドイツ語話者の話し言葉コーパスの開発

はじめに

近年、科学研究補助金などの公的助成を受け、数々のコーパスが作成・公開されてきており、日本語学・日本語教育においても、コーパスを利用した研究が推奨されている。言語習得研究もその例外ではなく、多くの分析で母語話者コーパスや学習者コーパスなどの言語資源が用いられるようになってきた。このようにコーパス基盤研究が進む中、ドイツ語圏、あるいはドイツ語母語話者に関する日本語学習者コーパスは存在せず、データに基づく定量的研究もきわめて少ないという現状がある。筆者らは、この課題を解決すべく、ドイツ語話者の話し言葉コーパスの開発に取り掛かった。本稿では、「ドイツ語話者日本語学習者話し言葉コーパス (Spoken Corpus of German Learners of Japanese)」のコーパスデザインを示すと同時に、どのような調査研究ができるか、ケーススタディを示す。

先行研究と学習者コーパス

ここでは、学習者コーパスについて、歴史的な流れに沿って、簡単に述べる。なお、本稿で紹介したコーパスについてのリンク情報は、本論文の最後にまとめて記載する。

まず、英語教育に関しては、1990年代、PCなどの技術的な発展により、コーパス作成が個人でも容易に行える環境が整ったことから、Granger (1998) が誤用例を含む学習者コーパスに注目し始めた。その後、1990年代後半から2000年代にかけて Granger が中心となって作った英語の話し言葉コーパス「Louvain International Database of Spoken English Interlanguage: LINDSEI」が公開された。さらに、2000年代後半から英語学習者の書き言葉コーパス「International Corpus of Learner English: ICLE」が公開されるようになる。また、コーパス開発と同時にコーパスを用いた研究が第二言語習得や外国語教育のフィールドで広まり、現在まで英語教育の研究や現場への還元に貢献している (Granger 2002)。

一方、日本語の学習者コーパスについて言えば、1990年代後半から次第に注目され始め、「KY コーパス」などの日本語会話データベースが構築されるようになる。2000年代に入り、データの規模拡張化が進み、対訳データベースなどが公開されるようになる。その後、2010年代からは

多様化の時代へと進み、BTS (Basic Transcription System) 準拠の会話コーパスや「Corpus of Japanese As a Second language: C-JAS」などの様々なコーパスが開発されるようになった。

表 1 日本語教育分野で利用されている主な学習者コーパス

| 種別 | コーパス | 規模 | 特徴 |
|----------|-----------------------------|---------------------------|------------------|
| 話し言葉コーパス | 日本語学習者会話データベース | 330 人分 | OPI コーパス |
| | KY コーパス | 90 人分 | OPI コーパス |
| | タグ付き KY コーパス | 90 人分 | OPI コーパス |
| | C-JAS | 6 人×3 年間 | 自然習得者 |
| 書き言葉コーパス | 日本語学習者作文コーパス | 304 人分 | 母語訳付き |
| | 作文対訳 DB | 1754 作品 | 母語訳付き |
| | 日本・韓国・台湾の大学生による日本語意見文データベース | 母語話者 134 人分 学習者 112 人分 | 母語訳付き |
| | YNU タスク別書き言葉コーパス | 1080 作品 | 12 種類のタスクによる作文収集 |

表 1 は、日本語教育分野で利用されている主要な学習者コーパスの事例である。話し言葉コーパスは、対話を文字化することで作成されているが、書き言葉コーパスは、作文として収集する形式で構築されている。それぞれの特徴として、話し言葉コーパスは、OPI (ACTFL-Oral Proficiency Interview)¹ (牧野ほか (2001)) の方法で構築されたものが多く、書き言葉コーパスは、母語訳付きの作文として構築されたものが多い。

¹ OPI とは、ACTFL (American Council on the Teaching of Foreign Languages) によって開発された汎用的会話能力テストであり、学習者の会話のタスク達成能力を、対面のインタビュー方式で判定するテストのことである。OPI のレベル判定は、超級、上級、中級、初級という主要レベルが 4 つ、超級以外のレベルはさらにそれぞれ、上中下の下位レベルが設けられている。インタビューの時間は、レベルによるが一人 20 分から長くても 30 分である。

問題意識

すでに述べた通り、日本語学習者コーパスにおいても、これまでに様々なコーパスが作成、公開されるようになった。しかし、現在のところ、ドイツ語母語話者の学習者コーパスは存在しない。そのため、コーパスを用いたドイツ語母語話者の日本語使用に関する研究も非常に少なく、また、彼らを対象とした誤用研究や習得研究も進んでいないのが現状である。こうした課題を踏まえ、本研究では、OPI コーパスの開発手順に従い、習得研究に利用可能なコーパスを開発した。このようなデータを利用することで、ドイツ語圏の日本語学習者の言語使用の実態を明らかにすることを目指す。また、このような調査の成果は教育現場レベルの指導法の改善につながる可以考虑している。

本研究が提案するコーパス

本研究が提案するコーパスは「ドイツ語話者日本語学習者話し言葉コーパス (Spoken Corpus of German Learners of Japanese、以下 GLJ コーパス)」である。データ収集は、2014年7月～2015年6月にかけて行った。調査協力者はミュンヘンに在住するドイツ語を母語とする日本語学習者である。「母語」の定義は複雑だが、ここでは学校教育をドイツ語で受けた者とする。コーパス名に「ドイツ語母語話者」と入れていないのは、今後「母語」を限定せず、ドイツ語圏で学ぶ日本語学習者のデータも加えたいと考えているからである。データ収集には、OPIを採用した。その理由は、既存のコーパスと収集方法を合わせることで、コーパスどうしの比較、分析を可能にするためである。収集したインタビューは OPI の基準に沿ってレベル判定をし、その後、音声データを全て文字化した。

表 2 では調査協力者の人数と文字化資料の延べ語数を示す。表 2 のとおり、レベルが上がっていくにつれ、延べ語数が増加傾向にあること、すなわち発話量が増えていくことが確認できる。

GLJ コーパスのもうひとつの特徴として、主観テストである OPI に加え、SPOT (Simple Performance-Oriented Test、以下 SPOT)²を実施し、客観テストによる調査も行っている点あげられる。本稿の後半で紹介するケーススタディ 2 では、言語テスト SPOT と GLJ コーパスの発話データの関連を分析した結果を報告する。

² SPOT:詳細は <http://ttbj-tsukuba.org/> (2017年11月3日)参照

表2 GLJ コーパスのサイズ

| 熟達度区分 | 調査協力者数 | 延べ語数* |
|-------|--------|---------|
| 初級学習者 | 15名 | 67,751 |
| 中級学習者 | 15名 | 83,107 |
| 上級学習者 | 15名 | 95,837 |
| 総計 | 45名 | 246,695 |

*延べ語数は、解析辞書の UniDic と形態素解析エンジンの MeCab の解析結果に基づいて計算

GLJ コーパスの初級学習者の発話サンプル

ここでは、初級学習者の発話のサンプルを一部紹介する。なお、S は学習者を、T はテストターを、〈 〉内はテストターの発話を表し、()内は学習者データ番号を示す（以下、中級・上級も同じ）。

- (1) 「どんなときにそのゲームをしますか」という質問に対して

S: んー、いえ、んー、あ、週末に、あー、んー、あん、時間、時間が、あるのとき、ゲームをします。〈あ、そうですか〉えーと、悪い、天気があります、あん、私も、えー、ゲームをしません。(初級-上 05)

- (2) 「月曜日から金曜日まで何をしますか」という質問に対して

S: あー、月曜日、私は、休みです。あん、私、は、あん、んー、Unterricht は、日本語で何ですか。(中略) あん、月曜日は休みです。あん、あーん、火曜日、から、金曜日、私は、日本語、と、フランス語、の、えー、Vorlesung です。³ (初級-上 07)

また、ロールプレイでのテストターとのやりとりとして次のサンプルを紹介する。これはバスで隣り合わせになった見知らぬ日本人が読んでいた本が気になって声をかけた場面である。

- (3) ロールプレイ (初級)

S: えー、すみません、この本は、どうですか。

T: あ、面白いですよ。

S: うん。えーと、なん、なんご、あなたは、読みますか。

³ Unterricht は日本語で「授業」、Vorlesung は「講義」の意味

T: なんご?

S: なんご, 英語, と, あん, 日本語, ですか。

T: あ, これは日本語の本です。

(後略) (初級-上 05)

初級レベルの大きな特徴は、文レベルの発話が維持できず、単語や句の羅列、暗記した表現を使って最低限のコミュニケーションができるのみである。下位レベルが「-上」になると、中級レベルの特徴も見られるようになるが、完全ではない。

GLJ コーパスの中級学習者の発話サンプル

次に、中級学習者の発話のサンプルを一部紹介する。

- (4) 「キムゼー⁴についてどんなところか教えて下さい」という質問に対して

S: あの一, 例えば, キムゼーの中に, あの一, 島が, さん, あの一, さん, あります。〈うん〉あの一, フラウエンキムゼーという, そして, ヘルンキムゼーという, 〈ええ, ええ〉それに, クロイタインゼルという島, です。〈ええ〉それに, あの一, ヘレンキムゼーの, 島に, あの一, 有名な城, が, あります。〈へー〉ちょっと, ベルサイユみたいな城です。(中級-上 08)

ロールプレイでは、仕事を始めたばかりなのに、急用で 4 時間早退しなければならぬ場面で、上司にその状況を説明するというタスクを出した。(5)はその発話の一部である。

- (5) ロールプレイ (中級)

S: あの一, ちょっと, 問題があります。

T: あ, はいはい。どうしましたか? ええ。

S: あの一, 問題があります。

T: はい。

S: え一, 今, ほんとに, 急な, え一, 電話, 電話。弟は, 電話, 電話をしました。〈はいはい〉ほんとに, 急に, 問題が, あります。

T: あ一, はいはい。

⁴ ドイツ語で Chiemsee (キムゼー、またはキームゼー) というキーム湖のことで、バイエルン州にある湖

S: あの一病院に、行った、えー、い、病院に行けなかった、あ、いえ、行かなければなりませんでした。

T: あら、大変ですね。ええ。

S: そう。あの、俺の質問は、あの、今日、よん、よん時、あー、あの、今日、早く、えー、仕事を、で、出たらいいですか。

(後略) (中級-上 08)

中級レベルの大きな特徴は、正確さや流暢さには問題があるものの、日常的な活動や身近な話題なら自ら文を作り、会話に参加できるとされている。また、目標言語の話されている環境で生活するための必要最低限のことができるのが中級レベルの特徴である。

GLJ コーパスの上級級学習者の発話サンプル

(6)と(7)は上級学習者の発話サンプルである。

(6) 心理学の本について、その中のあるセラピーについての説明をしているところ

S: 例えば、その、あの、精神分析とその似ている？あのセラピーは、例えば、過去に起こったことについてけっこう話しますが、〈はい、はい〉行動セラピーは、もっと今の現在、を基にして、今の問題について話すセラピーです。(上級 13)

ロールプレイは、部屋の窓ガラスが割れたため、管理人に連絡して状況を説明し、早急に対処してもらえるように依頼するタスクである。(7)はその発話の一部である。

(7) ロールプレイ (上級)

(前略)

S: あ、私は、あの、私のアパートの窓？が〈あ、はいはい〉壊れてしまって、あのー、外に遊んでる子供が、ボールを、あのー、そうですね、ボールを、んー、ボールを使って、えー、そのボールが、窓に、ぶつけて、ぶつかってしまって、今、窓が壊れているので、あの電話をします。

T: あ、はいはい。大丈夫ですか。

S: あ、そうですね。あのー、んー、えー、あのー、早くそれを、えーと、あー、それを、せい、せい、整理していただけませんか。

(後略) (上級 13)

上級レベルの大きな特徴は、詳しい説明や叙述ができること、段落で話せること、予期していなかった複雑な状況にも対応できることなどがあげられる。また、正確さに関しては、母語の影響が残っていても外国人の日本語に慣れていない人がわかる程度である。

以上、3 レベルの発話サンプルを紹介した。もちろん、紹介した一部のみではレベルを判定することはできない。次節では、これら発話データをコーパスとして統計的手法を用いて調査したケーススタディを二つ紹介する。

ケーススタディ 1：学習者の発話特徴量

調査概要

本節では、筆者らが開発した GLJ コーパスのケーススタディとして、ドイツ語母語話者の発話特徴について述べる。具体的には、初級・中級・上級の発話にはそれぞれどのような特徴が表れるのかを量と質の観点から明らかにするため、1) 平均文長と発話数、2) 発話に含まれる語種の分布を調査した。1) では、話者交替が起きるまで学習者がどのぐらい発話を維持できるのかをレベルごとに明らかにし、2) では、語種の異なりから学習者がコントロールする話題の変化を明らかにする。

データと方法

GLJ コーパスの全データに対して調査を行った。コーパスサイズは 246,695 語 (UniDic と MeCab の解析に基づいて集計) である。調査方法は、次のとおりである。

- ① プレーンテキストデータを形態素解析エンジンの MeCab と解析辞書の UniDic で解析し、形態素単位で区切りを入れた。
- ② 学習者単位で、総形態素数、平均文長、語種の比率を計算した。
- ③ GLJ コーパスについている OPI レベルを因子にして、2) の変数の平均値に統計的な有意差があるかを調べるため、SPSS Ver. 23 で一元配置分散分析を行った。

結果と考察

1) 平均文長と発話数

学習者の発話量をレベル別に比較してみると、初級から上級へとレベルが上がるにつれて、一文の長さが平均すると倍以上長くなることが明らかになった（図1）。一方、文の数自体は上級に上がるにつれて少なくなるが明らかになった（図2）。

図1 レベル間の平均文長

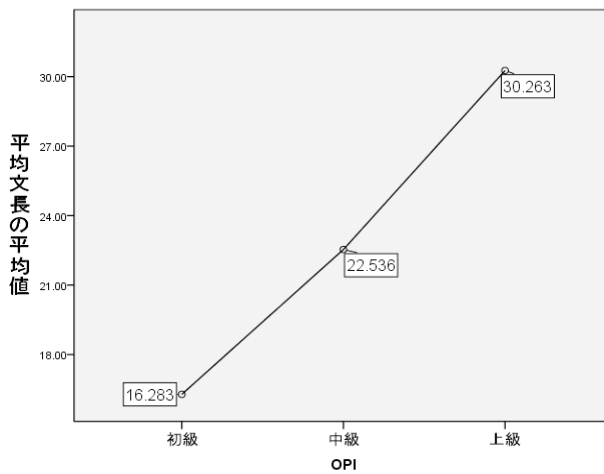


図2 レベル間の文数

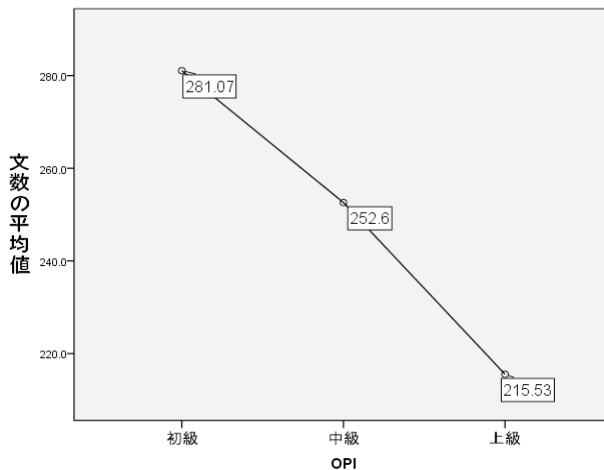


図 1 では、平均的な発話の長さを示すものとして、初級から上級までの平均文長の推移を示した。図 2 では、発話数を示すものとして、初級から上級までの文数の推移を示した。図 1 で読み取れることとして、初級の場合、1 発話の長さは平均 16.283 語で構成されているのに対して、中級では、22.536 語、上級では 30.263 語で構成されていることが明らかになり、OPI レベルに対する平均値に統計的に有意な差が見られた ($F(2,42) = 48.745, p < .001$)。

一方、図 2 では、レベルが上がるにつれ、全体の発話数（文数）が減っていく様子が確認できる。平均値として、初級では 281 発話、中級では 252 発話、上級では 215 発話が発せられたことが明らかになった。これに関しても OPI レベルに対する平均値に統計的に有意な差が見られた ($F(2,42) = 6.328, p < .005$)。図 1 と図 2 の結果を総合的に考えてみた場合、初級レベルの特徴としては、短い発話をたくさん発しているのに対して、上級レベルでは、発話の数としては少ないものの、1 発話の長さとしては、長いものを発しているという特徴が観察される。こうした分布の原因として、中級や上級においては、複文や連体修飾のような複雑な構文を使うことで、発話が長くなったと予想される。

2) 語種の分布

学習者の発話における語種（漢語・和語・外来語・混種語）の分布をレベル別で比較した。分析の結果、漢語と混種語の使用頻度に関しては OPI レベルに対して統計的な有意差が確認できた。

図 3 レベル間の漢語の使用頻度

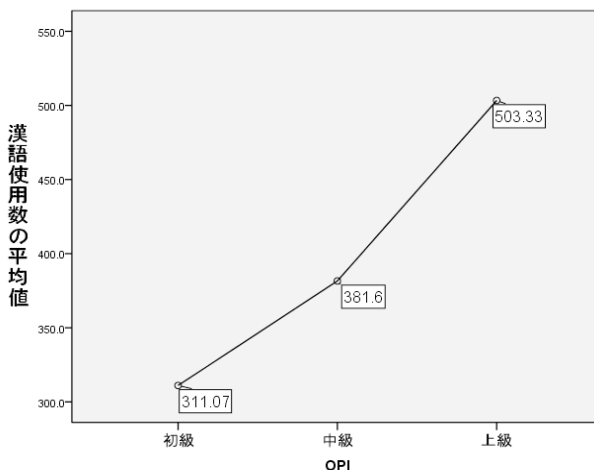


図 4 レベル間の混種語の使用頻度

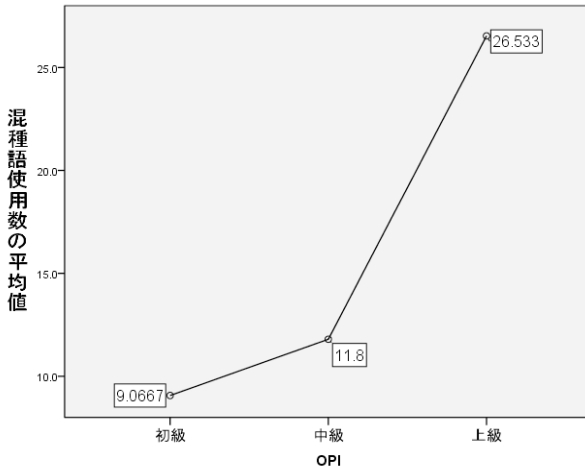


図 3 では、漢語の使用頻度の平均値の推移を、図 4 では、混種語の使用頻度の平均値の推移を示した。図 3 の漢語の使用数に関しては初級から上級へとほぼ一定の頻度で増えていることが明らかになり、一元配置分散分析の結果としても OPI レベルにおいて統計的に有意な差が見られた ($F(2,42) = 28.146, p < .001$)。次に、混種語の使用数に関しては初級・中級では使用頻度が低く、上級で使用頻度が高くなることが明らかになった ($F(2,42) = 54.463, p < .001$)。

語種の使用頻度に見られる差の原因として、話題の抽象性が関係している。牧野 (2001: 16-24) でも指摘されていることであるが、OPI の場合、上級という認定において抽象的な話題ができるかどうかは重要なファクターになる。日本語の語種として漢語の場合、抽象的な概念を表しやすく、和語の場合、具体的で身近な概念を表しやすくとされているので、こうした要因が関係していると考えられる。混種語に関しては、「ガス器具異」「駅前ビル」「ガラス窓」のように、形態的に異なる語種を組み合わせることで派生的に作られた語種であるため、初級や中級レベルでは、理解も産出も難しいということが考えられ、こうした制約が図 4 の分布においても反映されていると思われる。

ケーススタディ2：日本語学習者の発話量と言語テストの得点の関連性

GLJ コーパスのレベル判定の妥当性を確認するため、言語テスト SPOT との関連を分析した。具体的には重回帰分析と分散分析を行い、SPOT の得点とコーパスデータにおける発話量の関連を考察した。

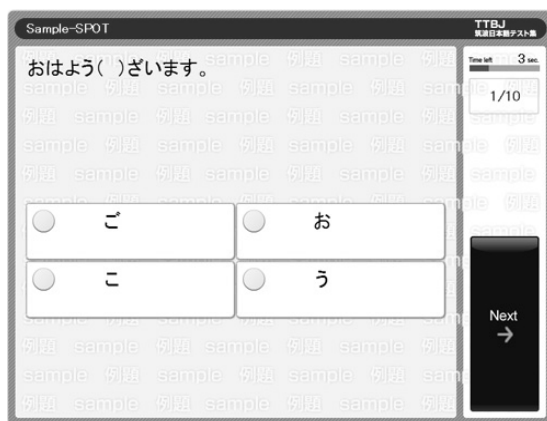
SPOT について

SPOT は、日本語の客観テストの一つであり、以下の特徴を持っている。

- ① 言語運用能力を間接的・客観的に測定するテストである。
- ② 短時間で実施できるテストである。
- ③ 能力差が比較的大きな集団を2～4段階程度の能力別グループに分けるテストである（小林 2015: 123）。

SPOT は自然な速度で読み上げられる文を1文ずつ聞きながら、1カ所の空欄に平仮名1文字を挿入するという形式の問題である。今回はWEB版 SPOT を使っており、図5のテスト項目が音声とともにディスプレイ上に表示される。

図5 「SPOT 90」の問題提示画面



資料：<http://ttbj-tsukuba.org/p2.html> (2017年11月3日)

SPOT にはいくつかのバージョンがあるが、本研究では「SPOT 90」を利用した。「SPOT 90」は30問ずつの「SPOT 90-1」「SPOT 90-2」「SPOT 90-3」で構成されており、それぞれのテストセットの難易度は次

のように設定されている。まず、「SPOT 90-1」は初級向けで、日本語能力試験の N4-N5 レベルの学習者を対象にしている。「SPOT 90-1」の音声は声優による明瞭なものを使用している。次に、「SPOT 90-2」は初級後半から上級前半向けで、日本語能力試験の N2-N4 レベルの学習者を対象にしている。「SPOT 90-2」の音声は日本語教師のやや不明瞭なものを使用している。最後に、「SPOT 90-3」は上級向けで、日本語能力試験の N1-N2 レベルの学習者を対象にしている。音声は日本語教師のやや早口の発話に加え、現実世界によく見られる雑音をノイズとして挿入している。

分析方法

言語テストの成績と学習者の発話量の関連を明らかにする目的で、以下の分析を行った。

分析①：GLJ コーパスの学習者の発話特徴量を独立変数、SPOT の得点を従属変数にして、重回帰分析を行った。

分析②：GLJ コーパスにおける熟達度による集団間で、SPOT の得点にどのような差があるかを分散分析で検討した。

分析①では SPOT の得点の違いが GLJ コーパスの発話量にどのように関連しているかを調査する。分析②では SPOT の得点と GLJ コーパスにおける OPI の熟達度がどのように関連しているかを調査する。なお、学習者の発話特徴量を抽出するため、全文字データを形態素解析エンジンの MeCab と解析辞書の UniDic で解析し、李・松岡・林（2013）で有効とされた変数を利用し、発話特徴量を抽出した。具体的には、語種別の使用頻度（和語、漢語、外来語、混種語）、品詞の比率（助詞率、名詞率、動詞率）、平均文長である。

結果と考察

分析① 重回帰分析の結果

重回帰分析では、「SPOT 90」の合計得点を従属変数、発話特徴量を独立変数にして、ステップワイズ法で分析してみた。分析の結果、助詞率と平均文長による回帰モデルが得られ、高い予測力を持つことが明らかになった ($R^2 = .807$)。

「SPOT 90」の得点 = $-164.791 + \text{助詞率} * 114.050 + \text{平均文長} * 63.498$ の回帰式が得られた。この結果を受け、平均文長と助詞率の散布図を作成してみた（図 6）。

図 6 の数値は SPOT の得点であるが、左下に 30~40 点前後の学習者が分布しており、右上に進むにつれ、50 点、60 点、さらには 70 点前後の学習者が分布している。これにより、助詞率と平均文長が言語テストの得点に対して強い予測力を持つことが明らかになった。このことを踏まえ、2 つの変量は、言語能力の差を反映する変数である可能性があると考え、より詳細に分析してみた。

図 6 平均文長と助詞率の散布図

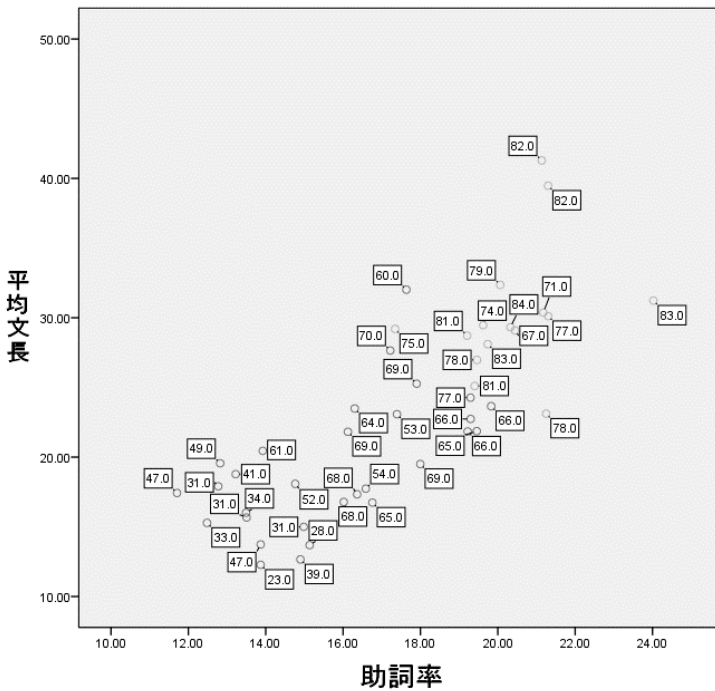


図 7 の横軸は、GLJ コーパスに付与されている学習者の熟達度であり、縦軸は変量を表す。図 7a は平均文長であり、図 7b は助詞率である。平均文長の場合、初級では、16 語程度の長さで構成されているが、中級では、23 語、上級では 29 語で構成されている。次に助詞率で言えば、初級では 14%、中級では 17%、上級では 20%程度が助詞である。

図 7a 平均文長の箱ひげ図

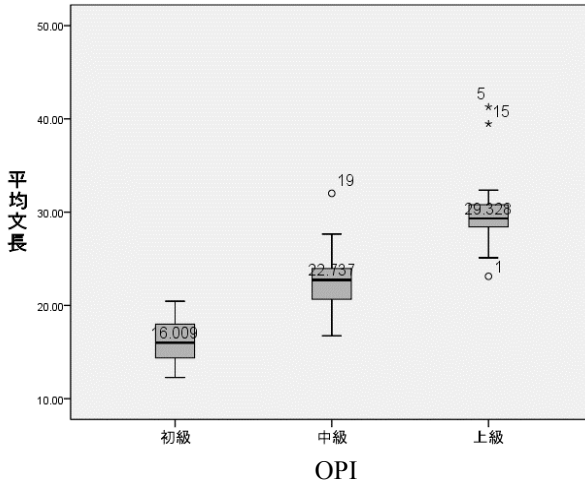
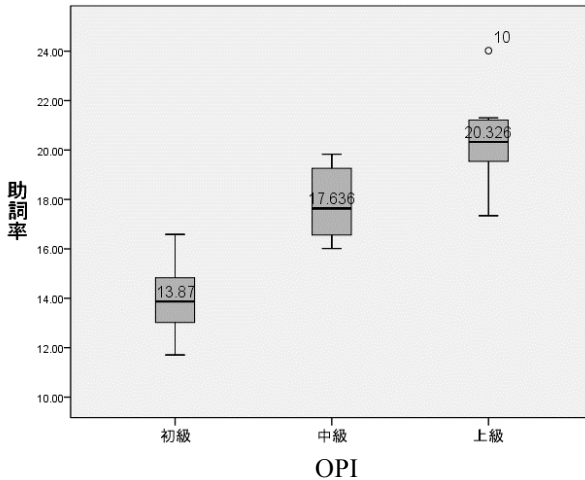


図 7b 助詞率の箱ひげ図



分析② 分散分析の結果

「SPOT 90」の得点を独立変数、GLJ コーパスの熟達度を従属変数にして平均値に差があるか調査した。調査結果として、「SPOT 90-1」、「SPOT 90-2」、「SPOT 90-3」のいずれにおいても、有意差があった（「SPOT 90-1」は $F(2,42) = 51.422, p < .001$ 、「SPOT 90-2」は $F(2,42) = 74.004, p < .001$ 、「SPOT 90-3」は $F(2,42) = 68.505, p < .001$ 、「SPOT 90」全体は $F(2,42) = 99.080, p < .001$ ）（得点分布は図 8 参照）。

図 8a 「SPOT 90-1」 (初級向け) の箱ひげ図

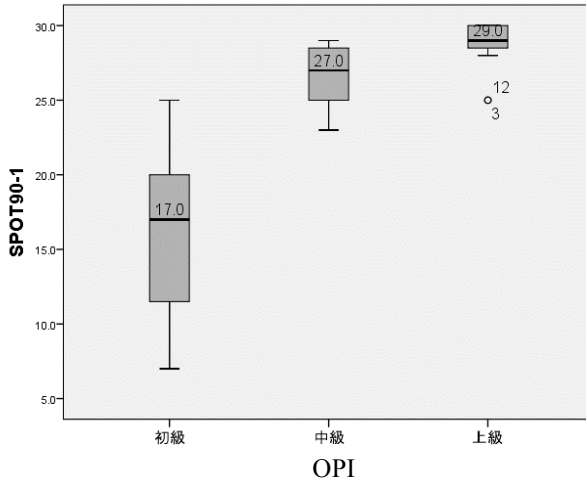


図 8b 「SPOT 90-2」 (初級後半から上級前半向け) の箱ひげ図

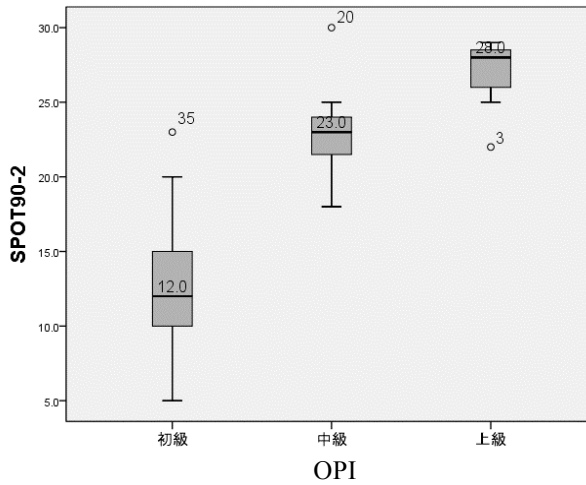


図 8c 「SPOT 90-3」 (上級向け) の箱ひげ図

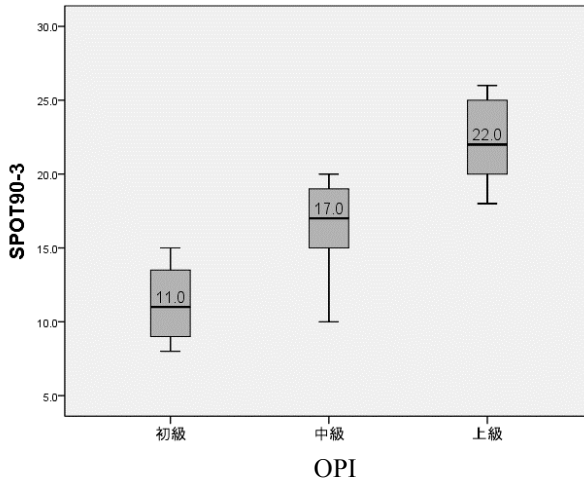


図 8d 「SPOT 90」 全体の箱ひげ図

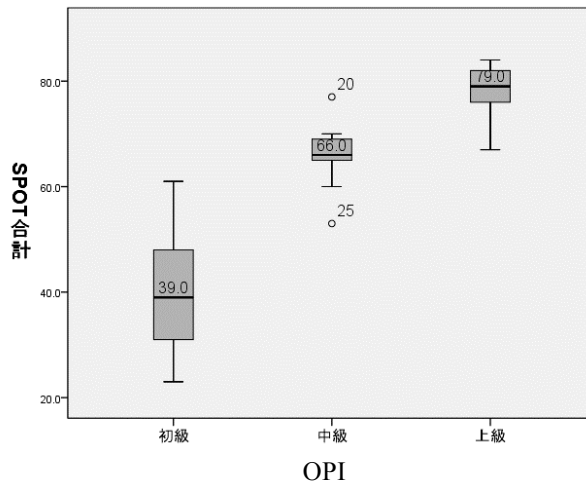


図 8a で示されている初級向けの「SPOT 90-1」の場合、初級においてはかなり幅があるが、中央値は 17 点であった。中級は 27 点であるが、上級は 29 点となり、中上級に関しては初級に比べ、差が小さい。図 8b で示されている中級向けの「SPOT 90-2」では、初級の幅が大きいことは、「SPOT 90-1」と同じであったが、中級と上級の間で中央値で言えば、5 点ほど差がついた。図 8c で示されている上級向けの「SPOT 90-3」では上級と中級と初級でほぼ等間隔の差がついており、能力の差を適確に捉え

られていると考えられる。図 8d で示されている「SPOT 90 合計」では初級の幅が非常に大きいものの、3つの熟達度で十分な得点差が見られた。

まとめと今後の課題

本稿では、まず筆者らが開発した「ドイツ語話者日本語学習者話し言葉コーパス (GLJ コーパス)」について紹介した。本コーパスは、「OPI コーパス」として開発しており、日本語学習者 45 名分 (初級 15 名、中級 15 名、上級 15 名) のデータが収録されている (コーパスサイズは 246,695 語)。ドイツ語を母語とする学習者の話し言葉コーパスとしては今のところ唯一である。

後半は GLJ コーパスに関して、2つのケーススタディを行った。

1つ目は、学習者の発話特徴をレベル別に明らかにした。その結果、次の2点が明らかになった。

1. 初級レベルの特徴としては、短い発話をたくさん発しているのに対して、中級、上級にレベルが上がるにつれ、発話の数としては少なくなるものの、1発話の長さとしては、長いものを発しているという特徴が明らかになった。
2. 語種に関しては、漢語と混種語の使用頻度においてレベル間で有意差が見られた。漢語については、初級から上級へとほぼ一定の頻度で増えていることが明らかになった。混種語については、上級レベルになってから産出が確認でき、初級と中級では難しいことが明らかになった。

2つ目のケーススタディでは、SPOT の得点と発話量について統計的な分析を行った。その結果、次の2点が明らかになった。

1. GLJ コーパスの学習者の発話量で SPOT の得点を予測させると 80% 以上、妥当な結果が得られるため、発話量と客観テストの結果に強い関連があると言える。
2. OPI レベルと SPOT の得点の関連を分析した結果、OPI レベルによって、すべてのテストセットの得点に関して、OPI レベル間で統計的に有意な差が認められた。これにより、GLJ コーパスのレベル判定の妥当性が確認できた。

今後の課題としては、本コーパスを公開し、さらに研究を進めることである。これまでドイツ語母語話者の学習者コーパスがなかったため、彼らの日本語使用に関する研究は少なく、誤用研究や習得研究も進んでいない。こうした現状を踏まえ、研究を進めることがドイツ語圏の日本語学習

者の言語使用の実態を明らかにし、そして、その成果を教育現場へ還元できると考えている。

参考文献

- 金澤裕之編 (2014) 『日本語教育のためのタスク別書き言葉コーパス』 ひつじ書房。
 小林典子 (2015) 「SPOT」 李在鎬編『日本語教育のための言語テストガイドブック』 くらしお出版 pp. 110-126.
 牧野成一 (2001) 「理論編 OPIの理論と日本語教育」 牧野成一ほか『ACTFL - OPI入門—日本語学習者の「話す力」を客観的に測る』 アルク pp. 8-49.
 牧野成一ほか (2001) 『ACTFL - OPI入門—日本語学習者の「話す力」を客観的に測る』 アルク。
 李在鎬・宮岡弥生・林炫情 (2013) 「学習者コーパスと言語テスト—言語テストの得点と作文のテキスト情報量の関連性」 『言語教育評価研究 (AELE) 』 3号 pp. 22-31.
 Granger, Sylviane (1998): *Learner English on Computer*. Addison Wesley Longman: London & New York.
 Granger, Sylviane (ed.) (2002): *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: Benjamins.

リンク・コーパス情報源

◆SPOT:

- <http://ttbj-tsukuba.org/> (2017年11月3日閲覧)
<http://ttbj-tsukuba.org/p2.html> (2017年11月3日閲覧)

◆コーパス:

C-JAS

- <http://c-jas.ninjal.ac.jp> (2017年11月3日閲覧)

ICLE

- <https://www.uclouvain.be/en-277586.html> (2016年11月15日閲覧)

KYコーパス

- http://www.opi.jp/shiryo/ky_corp.html (2016年11月15日閲覧)

LINDSEI

- <https://www.uclouvain.be/en-352660.html> (2016年11月15日閲覧)

YNUタスク別書き言葉コーパス

金澤裕之編 (2014) 『日本語教育のためのタスク別書き言葉コーパス』 ひつじ書房。
 作文対訳DB

- http://contr-db.ninjal.ac.jp/essay_01.html (2016年11月15日閲覧)

タグ付きKYコーパス

- <http://jhlee.sakura.ne.jp/kyc/corpus/> (2016年11月15日閲覧)

日本・韓国・台湾の大学生による日本語意見文データベース

- <http://www.tufs.ac.jp/ts/personal/ijuin/terms.html> (2016年11月15日閲覧)

日本語学習者会話データベース (横断調査編)

- <https://nknet.ninjal.ac.jp/nknet/ndata/opi/> (2016年11月15日閲覧)

日本語学習者会話データベース（縦断調査編）

http://nknet.ninjal.ac.jp/judan_db/（2016年11月15日閲覧）

日本語学習者作文コーパス

<http://sakubun.jpn.org>（2016年11月15日閲覧）