言語テストに基づくドイツ人学習者の対話型コーパス構築

村田 裕美子（ミュンヘン大学）
李 在鎬（筑波大学）

要旨

本発表の目的は,「ドイツ語話者日本語学習者話し言葉コーパス」におけるOPIとSPOT90の関連を統計的に分析することでOPIのレベル判定の妥当性を検証することである。利用したデータはドイツ語母語話者45名（初級, 中級, 上級の各15名）の発話データとSPOT90の得点データ，OPIのレベル情報である。分析方法は，OPIとSPOT90の関連を調べるため, 分散分析と回帰分析を行い，レベル分けの妥当性を検証するため, 判別分析を行った。

分散分析ではOPIのレベルを従属変数に，SPOT90の得点を独立変数にして分析した結果，有意効果が認められた（$F(2,42)=99.080$,$p<.001$）。また，回帰分析では全体の言語的要素の使用頻度からSPOT90の得点を高い精度で予測できた($R^2=.807$)。そして，判別分析では言語的要素の使用頻度からOPIのレベルを予測させたところ，93%正答できた。以上の調査により，本コーパスが客観性の高いデータであることが示された。

【キーワード】 OPI，学習者コーパス，ドイツ語母語話者，回帰分析，判別分析

1 背景と目的

村田・李（2015）では，OPIに準拠し，ドイツ語母語話者45名のデータを格納した「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）」の開発について報告した。GLJコーパスでは，OPIが持つ主観テストとしての欠点を補うべく，客観テストのSPOT（Simple Performance-Oriented Test; 以下SPOT，詳細は小林2015参照）を同時に実施している。本調査で使用したのは，90問で構成されたSPOT90である。

本研究は，GLJコーパスにおけるレベル判定の妥当性を検証し，研究リソースとして精度を評価することが目的である。データとしては次の3つを使用する。1) OPIのレベル判定情報（初級，中級，上級），2) SPOTの得点（0点～90点），3) GLJコーパスのテキストデータ（延べ語数246,695語）である。この3つのデータの相互関連性を明らかにするため，分散分析，回帰分析，判別分析を行う。調査の結果，3 つのデータには高いレベルの相互関連性が認められ，GLJコーパスの学習者コーパスとしての妥当性が認められたことを報告する。

2 先行研究

学習者コーパスとは, 言語学習者の産出データを格納したデータベースのことである（石川2008）。一般的には，学習言語の熟達度（proficiency）の差が言語使用にどのようなバイアスを与えるかを調査する目的で使用する。そのため，学習者コーパスの開発者は何らかの方法で学習者の（学習言語に対する）熟達度を判断し，アノテーション情報として公開

している。

熟達度を判断する方法としては，1) 産出データそのものに対して，直接的に熟達度を判断する方法，2) 言語テストを使用し，産出データとは独立して熟達度を判断する方法である。1) の方法については，ACTFL OPI（Oral Proficiency Interview；以下 OPI）の枠組みがよく利用される。そして，2) の方法については，SPOT がよく利用される。1) の方法を利用したコーパスとしては，鎌田修氏と山内博之氏による「KY コーパス」が広く知られている。2) の方法を利用したコーパスとしては，伊集院郁子氏が構築した「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」および金澤（2014）の「YNU 書き言葉コーパス」があげられる。なお，本研究が利用する GLJ コーパスは OPI と SPOT をともに利用しているコーパスである点で，ハイブリット的データベースと言える。

本研究では，「GLJ コーパス」のハイブリッド的特徴を活かし，SPOT90 と OPI の関連性を統計的な手法で分析する。とりわけ，学習者の発話データを形態素解析し，発話特徴量を抽出し，それらが SPOT90 の得点や OPI のレイティングとどのような関連を持つか考察する。この分析を通して，「GLJ コーパス」のレベル分けの妥当性を検証する。なお, SPOT90 と OPI の関連づけに関する先行研究として岩崎（2002）および鈴木（2014）があり，両者の関連性を具体的に示している。しかし，これらの研究は OPI のレイティングと SPOT90 の得点を分析したもので, 発話データそのものと SPOT90 の関連を分析したものではない。

3　データと方法

GLJ コーパスとは，村田・李（2015）によって開発されている学習者コーパスで，ドイツ語母語話者 45 名の発話データを収録した学習者コーパスである。コーパスの基本設計において，KY コーパスと同様に，OPI を用いて熟達度を判断している。

コーパスの中には，テスターと学習者による 2 者の対話データが文字化されているが，OPI の判定ルールに基づいて，初級学習者，中級学習者，上級学習者にカテゴリー化されている。各集団の学習者数およびコーパスサイズを表 1 に示す。

表 1. GLJ コーパスのサイズ

熟達度区分	学習者数	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，形態素解析エンジン「MeCab」の解析結果に基づいて計算

GLJ コーパスの特徴として，すべての学習者はテスターと対話を行ったあとに，インターネット日本語テストである「SPOT90」（http://ttbj.jp/）を受けており，コーパスデータ（話し言葉データ）と言語テストの得点が比較できるように構成されている。

本研究では，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:OPI のレベルを従属変数, SPOT90 の得点を独立変数にして分散分析を行った。
- 分析 2 : GLJ コーパスの（文字化資料における）発話特徴量から SPOT90 の得点を予測する統計タスクを回帰分析で行う。
- 分析 3 : GLJ コーパスの（文字化資料における）発話特徴量から OPI のレベルを予測する統計タスクを判別分析で行う。

分析 1 の仮説としては, OPI のレベル差は言語能力の差であり, 言語能力の差は, SPOT90 の得点群の差として表れるはずであると考えられ，このことを検証する。分析 2 の仮説としては, OPI のレベル差は言語的要素の使用頻度の差として確認できるはずであり, また, 分散分析で差があるとするなら，言語的要素の使用頻度の差と SPOT90 の得点差に因果モデルが仮定できると考え，このことを検証する。分析 3 の仮説としては，OPI のレベル差は言語的要素の使用頻度の差として確認できるはずであると考え，このことを検証する。なお，分析 2 と分析 3 で使用した発話特徴量の変数は「平均文長，漢語率，和語率，外来語率，名詞率，助詞率，動詞率，述語率」である。また，分析 2 と分析 3 で異なる統計分析を行ったのは，従属変数が異なるためである。分析 2 では間隔尺度（0 点～90 点の得点データ）であるのに対して，分析 3 は順序尺度（初級～上級のレベルデータ）であり，尺度水準を考慮した分析を行う必要があることから分析 2 では回帰分析，分析 3 では判別分析を行っている。紙幅の都合上，統計分析の詳細は述べないが，石川・前田・山崎（2010）を参照してほしい。

4 結果

4.1 分析 1 の結果

OPI のレベル差によって，SPOT90 の得点「0 点～90 点」の分布に統計的有意が認められるかどうかを調べるため，分散分析を行った。分析の結果，OPI のレベル差によって，SPOT90 の得点差に統計的な有意差が認められた（$F(2,42)=99.080$,$p<.001$）。

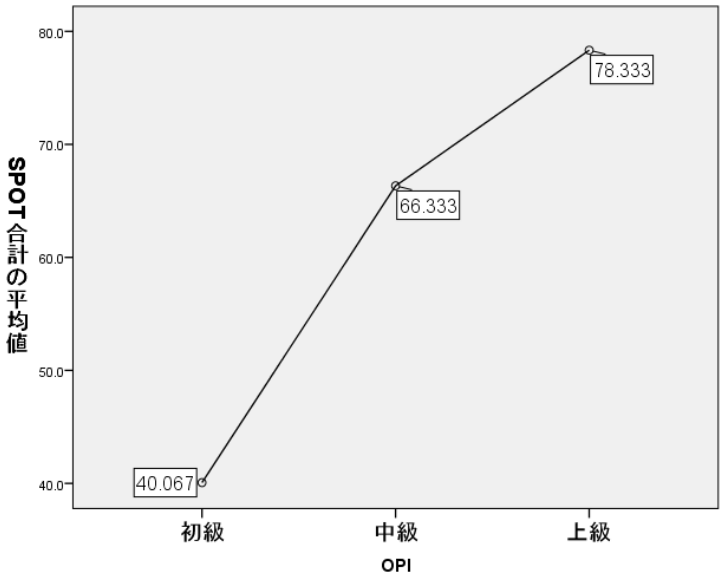

図 1. OPI レベル×SPOT90 の得点差の平均

図 1 で確認できるように，OPI の初級話者の場合，SPOT90 の得点が平均 40 点であり，全体の 50%未満の正答率である。OPI の中級話者の場合，平均 66 点であり，70%以上の正答率である。最後に，OPI 上級話者の場合，平均 78 点であり，80%程度の正答率である。レベルが上がるにつれ，得点が線形的に上昇している事実が確認できる。

4.2　分析２の結果

回帰分析では，SPOT90 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）。「SPOT90 の得点=-164.791+助詞率*114.050+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 2）。

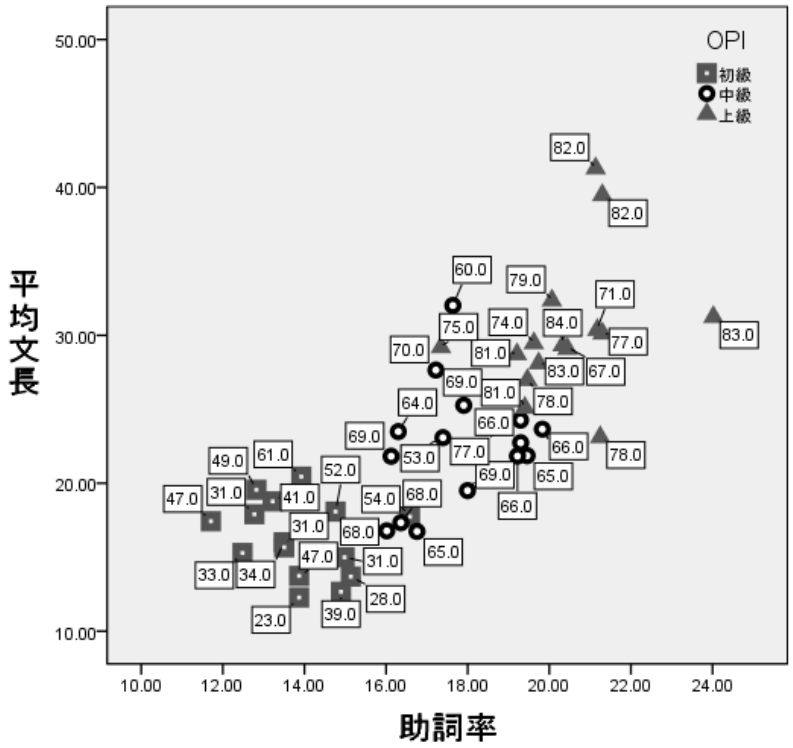

図2. 助詞率×平均文長の散布図

図2では，横に助詞の使用率，縦に平均文長が配置されており，図内のケースにはSPOT90の点数が表示されている。図2で確認できることとして，OPIの初級話者の中でもSPOT90の得点が高い学習者は平均文長が長いことが確認される。OPIの中級話者は，平均文長と助詞率のいずれにおいても初級と上級の真ん中に分布している。OPIの上級話者は，SPOT90の得点や平均文長，助詞率のいずれにおいてももっとも大きな値を示している。

4.3　分析３の結果

判別分析では，OPI のレベルを従属変数，発話特徴量を独立変数にして，分析を行った。分析の結果，判別関数による元のレベル判定の予測精度として 93%が正しく分類され，OPI

のレベルによって発話特徴量が異なっていることが示唆された。判別の結果を表 2 に示す。

表 2. 判別分析による分類の結果

OPI			予測レベル			合計
			初級	中級	上級	
元レベル	度数	初級	14	1	0	15
		中級	0	13	2	15
		上級	0	0	15	15
	%	初級	93.3	6.7	0	100
		中級	0	86.7	13.3	100
		上級	0	0	100	100

表 2 では，横の列に GLJ コーパスの元のレベルが，縦の列に判別関数による予測レベルが示されている。初級話者の場合, 14 名に関しては元レベルと予測レベルが一致しているが, 1 名に関しては，予測レベルは中級であるという結果になった。中級話者の場合，13 名は一致，2 名に関しては上級レベルである可能性が出てきた。上級話者では，15 名全員に関して元レベルと予測レベルが一致し，レベル判定において妥当性が高いことが示された。

5　考察

以上の分析結果を踏まえ，次の二つの論点において考察を行う。

1. OPI と SPOT の関係性について。
2. OPI のレベル差と言語的要素の使用頻度の関係性について。

1 の問題として，OPI も，SPOT も，運用力を図るテストである点で共通しているが，OPI は発話そのものに対するテスターの判定，SPOT は四肢選択による客観テストである点で異なっている。両者の関連に関しては，岩崎（2002）および鈴木（2014）によって OPI のレベルが高ければ，SPOT の得点も高いことが報告されているが，発話量に関する考察は行われてこなかった。しかし，本研究では GLJ コーパスを形態素解析し，発話特徴量を抽出，定量的な分析を行った結果，80%の精度で SPOT90 の得点が予測できることが明らかになった。2 の問題に関連づけて考えてみた場合，助詞率と平均文長の値によって，80%の受験者に関しては SPOT90 の得点が予測できる，ということになる。これは OPI による産出能力の測定と SPOT90 の得点の間には高い相関関係が認められることを示す証拠であり，同時に産出能力を示す指標の中でも助詞率と平均文長が SPOT90 の得点を予測する上で有効な指標であることを示すものでもある。

助詞率と平均文長の問題は, 2 の問題にも関連する。本研究が明らかにした点としては，発話特徴量として使用した「平均文長，漢語率，和語率，外来語率，名詞率，助詞率，動詞率，述語率」の中でも助詞率と平均文長が学習者の言語能力をとらえる上で，重要だということである。この 2 つの変量は，発話の長さに依存する要素である。特に平均文長の場合，1 つの発話を単位にしたものであり，長くなればなるほど，大きな値になる。長い

発話が持つ特徴としては，複文であったり，長い名詞句といった特徴が認められるが，こうした長い発話と言語能力の高さ，さらには，流暢さには関連があると言える。ただ，他の語種や品詞ではなく，なぜ助詞率が高くなることが言語能力の高さと関連を持つかについては今後さらなる検討が必要である。

6　まとめと今後の課題

本研究では，GLJ コーパスの妥当性検証を目的に 3 つの統計分析を行った。統計分析の結果は，以下のようにまとめることができる。

分析１：OPI のレベル差と SPOT の得点群の差について
　　→統計的な有意効果が確認できた（$F(2,42)=99.080$,$p<.001$）。
分析２： OPI の文字化資料における発話量と SPOT の得点の関係について
　　→80%の精度で予測できた。密接な関係が認められる。
分析３： OPI の文字化資料における発話量と OPI のレベル差について
　　→93%の精度で予測できた。密接な関係が認められる。

このことから，GLJ コーパスにおけるレベル分けの妥当性は証明されたと考えられる。
今後の課題としては，本コーパスを公開し，広く研究を進めることである。ドイツ語母語話者の学習者コーパスはこれまでなかったため，学習者の言語使用に関する研究は未開拓の部分が多い。こうした現状において GLJ コーパスが果たす役割は大きいと言えよう。したがって，まずは，ドイツ語を母語とする学習者対象の誤用研究や習得研究に GLJ コーパスを活用したい。そして，それらの研究を踏まえて，教育コンテンツの作成など教育現場への還元を目指す。

＊謝辞:本研究に協力してくれたミュンヘン大学, ミュンヘン工科大学の学生に感謝する。

＜参考文献＞

石川慎一郎（2008）『英語コーパスと言語教育：データとしてのテクスト』大修館書店.
石川慎一郎・前田忠彦・山崎誠（編）（2010）『言語研究のための統計入門』くろしお出版.
岩崎典子（2002）「日本語能力簡易試験（SPOT）の得点と ACTFL 口頭能力測定（OPI）のレベルの関係について」『日本語教育』114, pp.100-105.
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』くろしお出版.
鈴木祐一（2014）「第二言語の文法知識の自動化の簡易的な測定方法：WEB 版 SPOT と ACTFL 口頭能力測定（OPI）の比較」『電子情報通信学会技術研究報告 = IEICE technical report : 信学技報』114(100), pp.49-54.
村田裕美子・李在鎬（2015）「ドイツ語母語話者の話し言葉コーパスの開発」『Japanologentag 2015 - LMU München 予稿集』, pp.92-93.