# 母語学習者コーパスの基礎調査

# 母語学習者コーパスの基礎調査

小野　望・田中省作・持尾弘司

# Basic Study on the Corpus of Native Language Learner

Nozomi ONO・Shosaku TANAKA・Hiroshi MOCHIO

本稿は、国語表現学習者（母語話者）に対して表現熟練者の文章情報を提示することにより表現学習を支援するツール開発の可能性を探るための基礎的調査報告である。学習者・熟練者それぞれの文章を計量的に分析し、有意の差があることを明らかにした。

なお、本研究は、コーパスの構築と言語情報の統計処理・分析を田中（立命館大学）、文章素材の提供と表現教育の考察を小野、e-learning 環境の構築と統計分析を持尾が分担する共同研究である。

## 1．コーパス言語学の教育への応用

コーパス言語学の教育への応用については、小野2006[1]に述べたとおり、母語話者コーパスと第二言語学習者コーパスの対比分析とその提供という形で実用されている。これらの多くは、第二言語学習者と母語話者との差異を析出し、その部分を教育／指摘のポイントとするというものである。この方法は、母語話者の習熟のために応用することも可能であるはずだが、そのためには「(母語) 学習者」と「(母語) 熟練者」との間に、有意の差があることが前提となる。

文章表現に関して、学習者を大学生レベルに設定した場合でも、両者に差がある（全ての学生ではないにしても、その文章には未熟さ／学習すべき点がある）ことは経験的に分かっている。本研究の目的は、教師の経験によって指摘されてきた未熟さのうち、学習者・熟練者双方のコーパスの計量的な対比・分析によって抽出しうる要素を特定すること、そして、それらを利用した教育支援／学習支援のシステム実装を計画することにある。本稿はその基礎調査として、両コーパスの比較を行って有意の差があることを確認するとともに、さらなる観測ポイントを発見しようとするものである。

## 2．データと分析方針

学習者の文章データとしたのは文系女子大学生のエッセイ[2] 440本。熟練者データは、新聞の女性投稿欄エッセイ[3] 463本である。投稿エッセイが500字以内（平均約440字）という制限を持つのに対し、学生エッセイは2,000字を超えるものもある（平均約830字）[4]。文書量の違いが言語状況に影響を及ぼすことが考えられるが、その点については分析の際に考慮することとした。

分析に当たっては、計量的な文体研究に使用されてきた指標の中から、基本的なアプローチとして以下の５項目を選び、検討することとした[5]。

（１）　語彙のバラエティ
（２）　品詞構成比率
（３）　MVR
（４）　接続詞を含む文の割合
（５）　文末表現の種類

また、形態素解析には「茶筅」を使用し、品詞体系は「研究開発用知的資源タグ付きテキストコーパス報告書（テキストサブワーキンググループ、技術研究組合新情報処理開発機構、1998）によった[6]。

## 3．データ分析

### 3－1．語彙のバラエティ

学習者・熟練者の間には使用語彙量の差があると考えられる。その差が、文章作成に際して語彙のバラエティの多寡となって現れ、学習者の未熟さの一因となっているはずだ。このことを、計量的に比較、確認してみる。

観測する語の単位を次のように設定した[7]。

文節に区切った上で付属語を除き、自立語（内容語）を基本単位とする。

活用語については基本形（終止形）を代表形とし、これを見出し語とする。

語彙のバラエティを測る指標としては、「TTR」(Type-Token Ratio) と「D」(SimpsonのD) の両方を適用した。TTR は語彙力を測る尺度として知られるが、文書量の影響を大きく受ける。それに対し、D は理論的には文書量に無依存の統計量である。

#### 3－1－1．TTR (Type-Token Ratio)

TTRは、異なり語数 (Type) を延べ語数 (Token) で除して比率 (Ratio) を表すもので、値が高いほど語彙のバラエティが多いことを示す。

表1

| | 学習者 | 熟練者 |
|---|---|---|
| TTR の平均 | 0.638 | 0.749 |
| TTR の不偏分散 | 0.015 | 0.004 |
| TTR／文書量の Pearson 相関係数 | -0.708 | 0.547 |

両データのTTRの値・TTRと文書量のPearson相関係数を表１に、TTR値の分布を図１に示す。「Chikujo ✚」が学習者データ、「Benizara ✖」が熟練者データである。

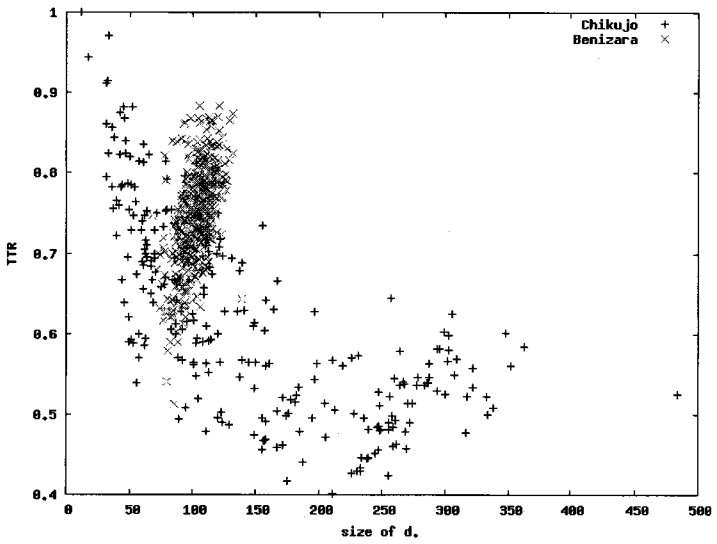


図１

学習者のTTRはばらつきが大きいが、その平均は熟練者に比して低く、語彙のバラエティに乏しいことを示している。

文書量との相関を見ると、特に学習者データは、従来指摘されている傾向が顕著で、「文書量が大きくなればTTRが下がる」という意味で高い負の相関が現れている。一方、熟練者データは、逆に比較的高い正の相関となっている。

今回使用した熟練者データは、500字という（短めの）制限付き投稿エッセイであり、もともと文書量は揃っている方だが、400～500字の間でばらつきはある[8]。そのばらつきとTTRとの間に正の相関があるわけだから、限られた字数の中に多数の情報量を盛り込むために極力用語の重複を避け、あるいは同語反復による文章の稚拙さを嫌った結果であるように見える。このような態度自体が、熟練度の要件になると考えられるが、それについては別途定性的な考察が必要となる。また、熟練者の文章量が500字を超えて大きくなっていけばTTRは下がるはずで、この切り替え――つまり、言葉を切りつめねばならない文章の短さによる制約からの解放はどのように起こるのかということも興味深い。同様に、学習者データを文書量の違いによって区分し、区分同士のTTRを対比することも必要だが、これらの考察は別稿に譲る。

今回の両データ全体を比較する段階では、TTRはやはり文書量の影響を大きく受けており、語彙のバラエティーの程度を表す指標としては不足であるということになる。

### ３－１－２．D（SimpsonのD）

標本量（ここでは文書量）に無依存の統計量として、SimpsonのDがある。D（diversity）は集中度を表す指標で、次の式によって与えられる。

$$D = \sum_{m=1}^{N} V(m,N)\frac{m}{N}\,\frac{m-1}{N-1}$$

（$V(m,N)$は大きさNの標本中にm回現れる異なり語数を表す。）

Dは、ある語が集中的に用いられる（語彙のバラエティが少ない）場合に、高い値となる。

両データのDの値・Dと文書量のPearson相関係数を表２に、D値の分布を図２に示す。「Chikujo✚」が学習者データ、「Benizara ✖」が熟練者データである。

表２

| | 学習者 | 熟練者 |
|---|---|---|
| Dの平均 | 0.00996 | 0.00694 |
| Dの不偏分散 | 1.58 $e^{-5}$ | 0.71 $e^{-5}$ |
| D／文書量の Pearson 相関係数 | -0.268 | 0.033 |

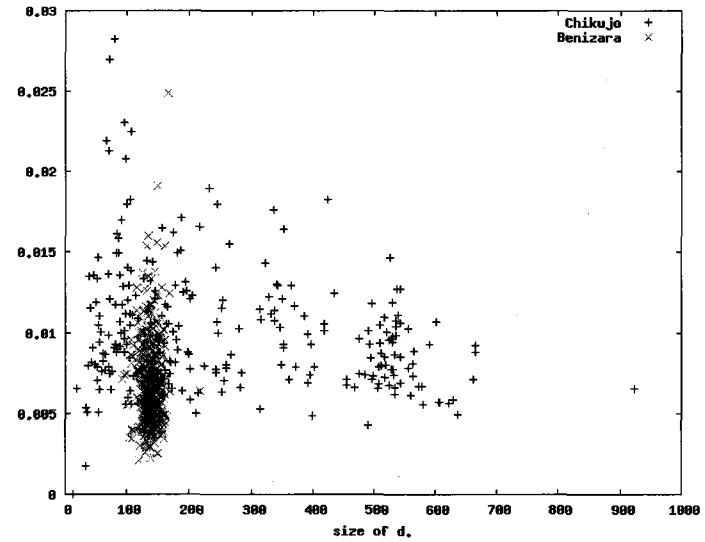



図２

Dと文書量との相関を見ると、熟練者データにはほとんど相関関係がない。学習者データはある程度の負の相関を示すが、TTRの場合に比べて相当低いことが確認できる。Dが予想のとおり、文書量に依存しにくく振る舞っていることが分かる。

両者のDの平均値の差が有意のものであるかどうかSPSS でｔ検定を行うと、Leveneの検定から等分散は仮定できず、その下でのｔ検定のｐ値は .000となる。別途 Excel で非等分散を仮定して片側ｔ検定を行うと、ｐ値は$9.18\ e^{-16}$となり、熟練者データの方が有意に集中度が低いという結論が導かれる。

学習者／熟練者の語彙のバラエティについて、文書量の大小に関わらず、学習者データは語彙のバラエティに乏しいことが確認された。

### ３－２．品詞構成比率

次項MVRとも関連するが、文章の品詞構成比率は、その文体的特徴を調べる指標の一つとして注目される。今回観測する品詞は、「名詞・動詞・副詞・形容詞・連体詞・接続詞・接頭詞・感動詞とフィラー・未知語・その他」の10項目とした。学習者／熟練者データの延べ語数による品詞別構成比率と、それに母比率の検定を施した結果を表３に示す。

表3

| 品　詞 | | 学習者 | | 熟練者 | |
|---|---|---|---|---|---|
| | 検定結果 | 比　率 | 延べ語数 | 比　率 | 延べ語数 |
| 名詞 | ＊＊＊ | 0.5420 | 57,218 | 0.5714 | 35,640 |
| 動詞 | ＊＊＊ | 0.3083 | 32,553 | 0.3009 | 18,768 |
| 副詞 | ＊＊＊ | 0.3083 | 5,412 | 0.0438 | 2,734 |
| 形容詞 | | 0.0423 | 4,469 | 0.0411 | 2,562 |
| 連体詞 | ＊＊＊ | 0.0225 | 2,375 | 0.0173 | 1,077 |
| 接続詞 | ＊＊＊ | 0.0180 | 1,900 | 0.0077 | 479 |
| 接頭詞 | ＊＊＊ | 0.0057 | 603 | 0.0071 | 445 |
| 感動詞＋ﾌｨﾗｰ | ＊＊＊ | 0.0023 | 296 | 0.0049 | 334 |
| 未知語 | ＊＊＊ | 0.0071 | 750 | 0.0053 | 332 |
| その他 | | 0.0000 | 3 | 0.0000 | 2 |

＊＊＊　$p<.001$

ほとんどの項目で$p<.001$であり、有意の差があることを示している。

名詞の比率については、「サマリー的な文章ほど名詞の比率が大きい」(樺島1961) [9] ことが指摘されている。両データとも名詞の比率は比較的高い [10] が、特に熟練者はその傾向が強く、より「サマリー的な文章」であるということになる。

樺島1961に次のような指摘があるが、

> 一般に言語表現において、事件の筋道を綜合して述べようとする場合には、事柄の関係に叙述の重点がおかれ、何が、何を、何になどを明らかにする骨格的表現になる。そしてこれによって名詞の比率が大きくなり、他の品詞の比率が減少することが見られる。

特に熟練者データが文字数制限のある新聞投稿であり、コンパクトな文章の中になるべく多くの情報量を盛り込み、しかも自分の思いを説明的にまとめたいという表現意図が働く結果、「サマリー的な様相」を表すことになると考えられる。これらのことについては、別途定性的な検討を行う必要がある [11]。

### 3－3．MVR

MVRは、「形容詞・（形容動詞）・副詞・連体詞」(Modifier) の合計数を「動詞」(Verb) で除して比率 (Ratio) を表すもので、値が高いほど「ありさま描写的」、低いほど「動き描写的」であることになる [12]。学習者／熟練者データの MVR の平均を表４に示す。

両データにおける MVR の等分散性が仮定できなかったため、Welch の検定を行ったところ、$p<.001$で有意差のあることが確認された。相対的に、学習者が「動き描写的」で、熟練者が「ありさま描写的」であるという結果である。

表4

| | 学習者 | 熟練者 |
|---|---|---|
| MVRの平均 | 0.3500 | 0.4008 |
| MVRの不偏分散 | $1.588766\times10^{-2}$ | $2.113290\times10^{-2}$ |

上掲の品詞構成比率と合わせ考えると、熟練者は「サマリー的な様相」を示すが、同時に相対的に「ありさま描写的」であることになる。一方、Modifier の品詞比率は学習者＞熟練者だから、特に副詞について、その振る舞いを構文的に確認する必要がある。

### 3－4．接続詞を含む文の割合

接続詞は、多くの場合一つの文に一回使用されるので、その比率は文長（文書中の文の総数）に大きく影響される。それゆえ、接続詞については、これを含む文の比率を計測することが行われる[13]。本稿では、特に文頭に接続詞を持つ文の割合に注目した。接続詞全体の内、文頭に現れる比率は学習者（96.8%）熟練者（97.7%）で、ともに高い。文頭に接続詞を持つ文の割合を表５に示す。

表５

| | 学習者 | 熟練者 |
|---|---|---|
| 文頭接続詞文の割合 | 0.1700 | 0.0685 |
| 文頭接続詞文の数 | 1,840 | 468 |
| 文の総数 | 10,821 | 6,836 |

この結果に母比率の検定を適用したところ、ｐ＜.001で有意差のあることが確認された。

樺島・寿岳1965に、近代小説100作品の分析に基づく、接続詞を持つ文の割合についての５段階評価尺度が示されている[14]。すなわち、「極めて小＜３小＜７普通＜21大＜27＜極めて大」である。この尺度を適用すると、学習者は「普通（のやや〈大〉寄り）」、熟練者は「（〈普通〉に近い）小」と評価される。

文頭の接続詞は、ほとんどの場合前文との論理関係を示すために使用されるから、同様の機能を果たすことのある指示語（文脈指示）の用法と合わせて観察する必要がある。さらに、特に学習者について、接続詞（や文脈指示の語）の用法が論理的に適切であるか、無駄なあるいは過多の使用がないか等について、評価した上で検討しなければならない[15]。

### 3－5．文末表現の種類

文末表現の構成要素について、両データから150ずつ、計300の文末サンプルを無作為に抽出して観察した。文末表現は「テンス・アスペクト・ヴォイス・モダリティ・肯否」の５種類について可視的なもの[16]を計量し、これら可視的な要素を持たないものを「構成要素なし」で計上した。また、同一文末に同じ構成要素が複数回出現した場合でも１として計上した[17]。学習者／熟練者サンプルにおける各

表６

| 構成要素 | | 出現数 | |
|---|---|---|---|
| | 検定結果 | 学習者 | 熟練者 |
| テンス | ＊＊＊ | 37 | 65 |
| アスペクト | | 34 | 33 |
| ヴォイス | ＊ | 6 | 14 |
| モダリティ | ＊＊＊ | 77 | 36 |
| 肯否 | | 5 | 7 |
| 構成要素なし | | 38 | 44 |

＊＊＊ p<.001
＊＊ p<.01
＊ p<.05

構成要素の出現数と、それに独立性の検定（Fisher の検定）を適用した結果を表６に示す。

３－５－１．テンス

「テンス」は、日本語で可視的に表現される「過去表現」を計数しているから、熟練者データに過去表現が多いことを示している。これは、「名詞の比率」で見た熟練者の「サマリー的な様相」と矛盾しない結果である[18]。

一方、学習者も名詞比率は低い方ではないが（相対的に熟練者より低いが)、テンス表現において異なった様相を示す。この結果を解釈するためには、過去表現をしない例について、アスペクトやモダリティとクロスさせ、さらには可視的でないテンス（基本形は何を表しているかなど）に踏み込んで考察する必要がある。

３－５－２．モダリティ

「モダリティ」の結果は、「学習者は相対的に自分の心情・態度を表現する」ことを表している。この結果から、丁寧表現／デス・マス／を除いた結果を表７に示す。

同様に Fisher 検定を行うと、ｐ＜.01で有意差が認められ、より内容に近いレベルで学習者の心情・態度表明傾向を確認することができる。このことについては、前項に述べたように、名詞比率・MVR やテンス・アスペクトなどと合わせて検討すること、さらに定性的にも考察する必要があろう。

**表７**

| デス・マスを除くモダリティの出現数 | | |
|---|---|---|
| 検定結果 | 学習者 | 熟練者 |
| ＊＊ | 44 | 20 |

＊＊　ｐ<.01

## ４．今後の課題

本稿において、学習者／熟練者の間で有意の差があることを確認した項目は次の通りである。

（１）　語彙のバラエティ（文書量に無依存で）

（２）　品詞構成比率（10項目中８項目について）

（３）　MVR

（４）　接続詞を含む文の割合

（５）　文末表現の種類（テンス・モダリティ）

これらのうち、（５）の中の「丁寧表現を除くモダリティ（有意水準１％)」以外の有意水準は１‰であって、信頼性は高い。

文体的な特徴を表すとされる指標を選んで検討したわけだが、その全てに亘って有意差を確認できたことの意味は大きい。本研究の目的とするコーパス言語学の国語表現教育への応用を考える場合、「差がある」ことが前提だからである。この前提が確認できたので、次はどのようなポイントを指導／学習すればよいか、あるいは第二言語教育において採られてきた手法で

母語学習者に有効なものはどれか、など、実用段階の検討に移ることができる。

とはいえ、今回の指標の中で国語の習熟度をストレートに表すのは、（1）語彙のバラエティである。（この度合いを示すのに、標本量に無依存のDが有効であることを確認できたのは大きな成果である。このDの適用については、別項で詳細を報告する予定である。）これ以外の指標については、今回のような計量データがそのまま学習者の未熟度を表すわけではない[(19)]。ここからさらに、表現教育への応用に進むためには、未熟さについての定性的検討が必要となる。今後は、教師による添削／指導項目の情報を付加するなどの方法を試みる予定である。

本稿は、平成18年度筑紫女学園大学特別研究助成を受けた研究成果の一部である。

**注**

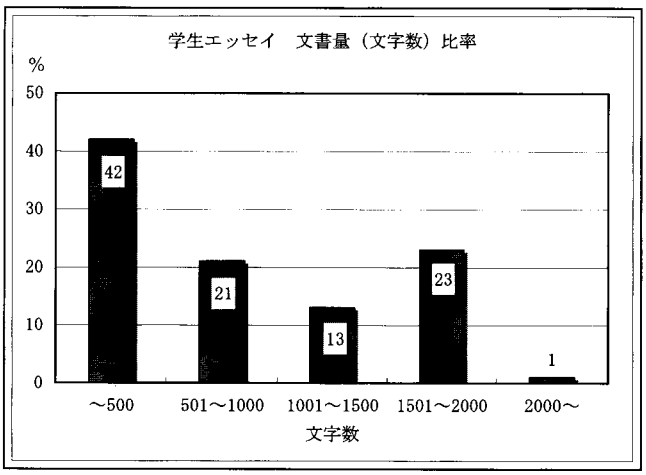
本稿の作成に当たっては、田中省作の指導による田野多龍一（立命館大学）の卒業論文「*低年次大学生の文章における稚拙さに関する計量的研究*」（2006年度）を参照した。

（1） 小野2006：「*国語表現教育におけるコーパス利用の可能性*」小野望、『論叢』第17号、筑紫女学園大学・筑紫女学園大学短期大学部国際文化研究所、2006.8

（2） 主として2年次生。2004～2005年度「日本語表現演習」科目の課題として提出されたもの。

（3） 西日本新聞「紅皿」欄。2005年1月～12月分。投稿者の年齢が大学生以下のものは除いた。なお、本研究に際し、西日本新聞社の許可を得て、同社記事データベースを使用している。

（4） 学生エッセイの文書量（文字数）分布は下図の通り。授業時の課題として、800字・1600字といった指定をしない場合は、500字未満となるものが多かった。



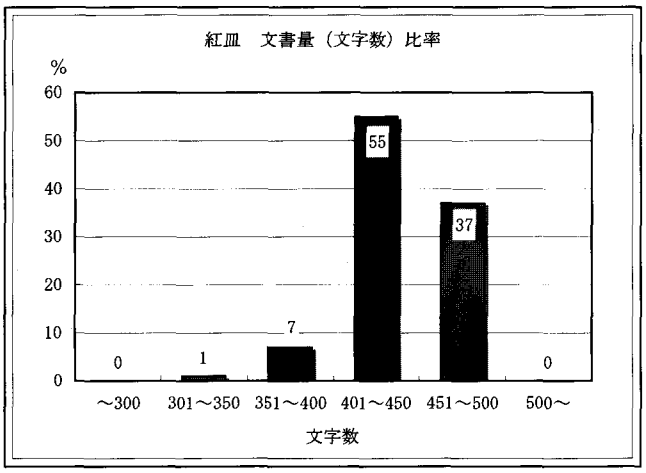

（5） 樺島・寿岳1965：「*文体の統計的観察*」樺島忠夫・寿岳章子、『*文体の科学*』第6章、綜芸舎、1965.6（『*論集日本語研究8　文章・文体*』山口仲美編、有精堂出版、1979.4に所

収）などを参照した。

（6） http://chasen.naist.jp/hiki/ChaSen/
http://chasen.aist-nara.ac.jp/chasen/doc/ipadic-2.6.3-j.pdf

（7） 『*計量言語学入門*』伊藤雅光、大修館書店、2002.4を参照した。

（8） 「紅皿」の文書量（文字数）分布は下図の通り。300字未満１本、500字を超えるもの１本（いずれも比率としては０％という表示になる）がある。




（9） 樺島1961：「*文体の変異について*」樺島忠夫、『*国語国文*』30巻11号、1961.11（『*論集日本語研究８　文章・文体*』山口仲美編、有精堂出版、1979.４に所収）。
「過去の事件をまとめて綜合的に述べたてようとする場合に、名詞の比率が大きくなり」「品詞比率の差は、以上のような叙述態度の異なりによって生じたと考えられる、」とし、『サマリー的な文章ほど、名詞の比率が大きい』ことを「第一の仮説」として立てている。（『*論集日本語研究８　文章・文体*』P.153）

(10) 樺島・寿岳1965に、近代小説100作品の分析に基づく、名詞比率の出現率評価尺度が次のように示されている。
極めて小＜45小＜48普通＜54大＜56＜極めて大
各評価段階の出現率は、「極めて小（10％以下）、小（30％以下）、大（30％以下）、極めて大（10％以下）」である。（『*論集日本語研究８　文章・文体*』P.183）
今回の調査とは品詞の設定が異なるため、この評価値をそのまま当てはめることはできない。ことに本調査では同書で Modifier の一とする「形容動詞」を「名詞」に区分しているため、これを差し引けば、名詞比率は数％低くなるはずである。その場合、両データともこの評価の「大」に近い「普通」の辺りに位置することになると推測される。

(11) 樺島1961で、本文直上の引用に続けて、
またこのような場合に言葉の量の制限（少しの言葉数で表現するなどの）が加わると文はくりこみの文となって、長さを増すことになる。たとえば新聞記事の文章は他の

種類の文章よりも名詞の比率が大きく、文が長い。これは以上に述べた理由によるものである。

とする。「くりこみの文」とは「『私が本を買った。その本を弟が破ってしまった』を『私が買った本を弟が破ってしまった』のように一つの文とした場合」と注記されている。名詞比率の意味を考えるためには、このような文構造・文長に関する観察を合わせて行う必要がある。

(12) 注10に述べたように、本調査では「形容動詞」を設定していないため、他の調査のMVR 値と対比することはできない。しかし、今回の両データ同士の差異は、従来の調査で指摘された MVR のありさまと同様の傾向を示すと考えてよい。

(13) 樺島・寿岳1965（『*論集日本語研究8 文章・文体*』P.181）

(14) 注９に同じ。

(15) このような検討を行うため、学習者データに教師の指摘・添削の情報を付加し、分析の準備を行っている。

(16) 例

テンス：／た／（過去）

アスペクト：／ている／（継続）／しはじめる／（動作の開始）／しおわる／（動作の終了）

／したことがある／（過去の経験）など

ヴォイス：／られ／（受身）／させ／（使役）

モダリティ：／だろう／（推量）／したい／（意思）／ね・よ等の終助詞／／です・ます／（丁寧）など

肯否：／ない／（否定）

(17) 例

「食べたいな」の場合、[たい] [な] がいずれもモダリティの要素であるが、「この文末にはモダリティあり」（１）として計測する。

(18) 特に、注９に掲げた樺島1961の記述が注目される。

(19) 例えば、学習者データの MVR「53」は、熟練者データ「40」より高いが、だから未熟なのだということは、すぐには言えない。試みに、樺島・寿岳1965に示された近代小説100編を見ると、MVR53 を超えるものが26編ある。（ただし、MVR の前提となる数値が本稿と異なることについては、注10に記したとおり。）

（おの　のぞみ：日本語・日本文学科　教授
たなか　しょうさく：立命館大学文学部　准教授
もちお　ひろし：発達臨床心理学科　准教授）