様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　5 月　13 日現在

機関番号：14401
研究種目：基盤研究(C)
研究期間：2008−2010
課題番号：20520439
研究課題名（和文）多変量文体分析モデルによる近代英語散文の通時的コロケーション研究

研究課題名（英文）Multivariate Approaches to Diachronic Study of English Collocations

研究代表者
田畑 智司（TABATA TOMOJI）
大阪大学・大学院言語文化研究科・准教授
研究者番号：10249873

研究成果の概要（和文）：
本研究では，多変量文体分析モデルを近代英語散文のコロケーション分析に応用することにより，通時的視点から近代英語散文における *gentleman* のコロケーションを調査した。研究基盤となるコーパスとして, 19 世紀の作家 Dickens を核に, 18 世紀, 19 世紀の通時コーパス ORCHIDS (Osaka Reference Corpus for HIstorical and Diachronic Stylistics: 収録語数 1400 万語超)を編纂した。多変量文体分析モデルを応用することによって，(1) キーワードと共起する多数の語彙項目間の相互関係，(2) キーワードが生起するテクスト間，ジャンル間，時代区分間の相互関係，さらに，(3) 共起語とテクスト，ジャンル，時代区分等との相互関係を視覚化し，キーワードのコロケーションを通して近代英語散文の文体変異を俯瞰した。
研究成果の概要（英文）：
This study attempts a multivariate approach to the collocation of *gentleman* in English prose texts in 18th and 19th Centuries. By applying a stylo-statistical analysis model based on correspondence analysis to the investigation of the collocation of the word gentleman, the present study visualizes the complex interrelationships among *gentleman*'s collocates, interrelationships among texts, and the association patterns between the *gentleman*'s collocates and texts in multi-dimensional spaces. By so doing, the author illustrates how the collocational patterns of *gentleman* reflects a stylistic variation over time as well as the stylistic fingerprint of the author.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：人文学
科研費の分科・細目：言語学・英語学
キーワード：文体，コロケーション，共起語，Dickens，多変量アプローチ，共起強度，近代英語，視覚化

１．研究開始当初の背景

コロケーション研究は, 電子コーパスの登場によって著しい発展を遂げた領域の一つである。コロケーションの概念 “mutual expectancies of words”は，先述の Firthが半世紀以上前に披瀝したものだが, 特定の文学形式やジャンル，あるいは特定の作家のコロケーションを調査することが記述言語学研究に寄与すると主張し，

コロケーションの重要性を指摘したことは慧眼であったと言える。しかし，特定の言語使用域や個人語を特徴付けるコロケーションを実証的に記述するには大規模な言語資料と分析ツールが不可欠であり，コロケーション研究に本格的な光が当てられるようになったのは，J. M. Sinclairに代表される英国 Birmingham学派による現代英語大規模コーパスの構築，分析が始まって以降のことである。特に，近年はコンピュータの処理性能の飛躍的な向上により，数千万語のコーパスでもラップトップで処理することが可能となったのに加え，*t*-Score，MI-Score（相互情報量），Log-likelihood ratio（対数尤度比）などコロケーションの強度を測定するさまざまな統計ツールを実装したコンコーダンス作成ソフトウェアも普及し，プログラミングとは縁が遠い言語研究者でも簡単にコロケーションを抽出できる環境が整ってきた。その結果，コロケーション（および‘phraseology’）の研究は言語研究・教育の重要な一角を占めつつある。

とはいえ，Sinclair (1991), Kjellmer (1994), Stubbs (1995, 2001), Hunston and Francis (1999)などに代表されるコロケーション研究は語彙，語法・文法，パターン文法，辞書編纂，言語教育などに関するものがほとんどであった。Firthはコロケーションを通して個人の文体的特徴 (idiosyncrasies)から，さらに巨視的には文体の史的推移を観察する可能性を示唆していた。同じくGreenbaum (1970)もコロケーションが文体研究の重要な視点になりうることを指摘しているが，これは，近年 Louw (1993), Adolphs and Carter (2003), Partington (2003, 2006), Hori (2004)などが現れて，ようやく実践が始まったところである。

現在のコロケーション研究では，コンコーダンスと統計尺度を用いて，ノード（node,中核語）と共起語 (collocates)との結び付きを記述するアプローチが主流である。しかし，この方法ではノードと共起語のネットワークは記述できても，共起語間の複雑な関係や，コロケーションと著者，ジャンル，時代区分などの重層的な関係を記述することは困難である。とりわけ，多数のテクストにおける共起語の分布データのように，数十行×数百列からなる高次元のデータに存在する潜在的パターンを目視で捕らえることはほとんど不可能であろう。他方，筆者がこれまで進めてきた多変量アプローチは，多数の語彙項目間の関係，テクスト間の関係，さらに語彙項目とテクスト（著者，ジャンル，時代区分など）との関係を視覚化することにより，マクロ的な分析の視点を提示するのに適した方法論である。これまで筆者はこの分析モデルを基に， Dickens の文体的特徴，文体の経年変化，作品執筆年代の識別，小説中の個人語の類型分析，米国大統領就任演説の文体の類型分析などの諸問題を考察してきた。この方法を *gentleman* 共起パターンの解析に応用することで，コロケーション分析に新たな着眼点を導入するのが本研究の狙いである。

２．研究の目的

本研究では，多変量文体分析モデルを近代英語散文のコロケーション分析に応用することにより，通時的視点から近代英語散文における *gentleman* のコロケーションを究明する。*gentleman* という語は18，19世紀の小説やDickensの作品において文体的に重要な語の一つであるが，この語を取り巻く共起語が，18世紀から19世紀にかけて，（修飾語や動詞との）統語関係，意味範疇，‘semantic prosody’などの点でどのように変遷しているか，また，共起語間の関係や，共起語とそれらが生起するテクスト，ジャンル，年代間の関係を，多変量分析アプローチによって俯瞰的に捉えることで，近代英語散文における*gentleman*という語の文体的「ふるまい」を明らかにするのが本研究の目的である

３．研究の方法

本研究が拠り所とする一次資料は，通時的文体論研究のために筆者が編纂したコーパス，Osaka Reference Corpus for HIstorical/Diachronic Stylistics (ORCHIDS)である。ORCHIDSは，18世紀の代表的作家の作品 23点，4,163,353語，Dickensの作品 24点，4,835,158語，そして他の19世紀の代表的作家の作品31点，5,118,346語の総計 14,116,857語より構成されている。コーパスデザインという観点からは，ORCHIDSの 19世紀サブコーパスは男女のバランスが大きく女性に傾いているように見える。9人のうち 6人が女性作家である。しかし，収録されている総語数での男女比は 45:55であり，見かけ程バランスを欠いてはいない。少なくとも， 以下に示す分析結果からは作家の性別に由来すると考えられるクラスターは認められないことから，コーパスの男女比は *gentleman*の共起語の分布にほとんど影響を及ぼしていないと言えよう。

ORCHIDS における *gentleman* の生起度は8,432回，サブコーパスごとの内訳は Table 1に挙げるとおりである。一方，*gentleman*の共起語は，スパン（共起範囲）をノードの左右 4語に設定した場合，7,344タイプにのぼる。Table 3に頻度上位100位までの共起語を挙げている。頻度上位は冠詞，等位接続詞，前置詞，

代名詞・指示詞など機能語が占めているが，*young, old*が10位以内に，50～100位以には*single, good, little, poor, fine, great*などの形容詞が動詞（主に過去形）とともにランクインしている。

Table 1: *Gentleman* in ORCHIDS

	tokens	gentleman	Freq./million
Dickens	4,835,158	4,601	951.57
18th C	4,163,353	2,267	544.49
19th C	5,118,346	1,564	305.57
TOTAL	14,116,857	8,432	

方法論に関する諸問題—分析の前処理—

実際の分析に入る前に *gentleman* のコロケーション分析に多変量アプローチを適用する上で考慮すべきいくつかの方法論上の問題について述べておきたい。

(1) コロケーションのスパン

本稿で言うコロケーションは，Sinclair (1991：170)の定義 “the occurrence of two or more words within a short space of each other in a text”に基づく。ここで問題になるのは，“within a short space”をどの程度に定めるかである。つまり，ノード *gentleman*からどれくらいのスパンに生起する語との結びつきをコロケーションと捉えるかという問題である。スパンをどこまで広げるか, あるいは絞り込むかという問題に関しては, コーパス言語学者の間でまだ完全には意見が一致していないが, おおむねノードの左右 4語の範囲に有意なコロケーションが見いだせることについては一定のコンセンサスがある (Jones and Sinclair, 1974)。スパンを 4語とすることの論拠としては，Fig. 2 (Sinclair *et al*., 2004)参照。グラフでは，ノードからのスパンが 4以上になると漸近線となっている。つまり，ノードからの位置が離れるほど, ノードが語の生起に及ぼす影響力が小さくなっていることがわかる。

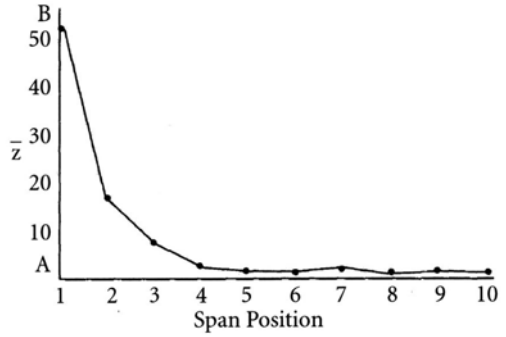

Fig.1: Graph showing average node predictions over span positions 1–10*

* From Sinclair (1969) as reprinted in Sinclair *et al*. (2004)

もちろん, 品詞によってはノードから 5語以上の位置に有意なコロケーションが表れることもあり得る。たとえば， Fig. 3の KWICコンコーダンスラインにおいて，副詞 *originally*が生起する文には *but*も共起しており，両者の間に談話的結束性が認められるが， *but*の生起位置はいずれもノードから 6語以上離れている。確かにこのようなケースは存在するものの, スパンを 5語以上に広げた場合, そのトレードオフとしてコロケーションの検出力は弱まってしまう。情報検索の概念で言えば, スパンを広げると recall（再現率）は高まるが， precision（適合率）は低くなるわけである。従って，本研究では operational definitionとして，ノードの左右 4語をスパンに設定し, その範囲に共起する語を分析対象とする。名詞 *gentleman*のコロケーション検出のためのスパンの最適化については稿を改めて論じることにしたい。

(2) コロケーションの指標，閾値の設定

コロケーションの結びつきの強さを計る指標には，粗頻度 (raw frequency), *t*-score,対数尤度比 (log-likelihood ratio), 相互情報量 (MI-score) などがある。このうち，粗頻度は最も基本的で単純な指標であるが，Table 3 からも明らかなように，高頻度の機能語がリストの上位を占める傾向があり，（少なくとも文体論的見地からは）コロケーションの検出力という点で難がある。対数尤度比， *t*-score は粗頻度よりはバランスのよい指標だと考えられており, WordSmith Tool, CasualConc その他のコンコーダンサーや BNC*web* にも実装されている。コロケーションの各指標の特徴を分析した石川 (2006: 11)は，*t*-score，対数尤度比は頻度の高い一般的なコロケーションの検出に強い指標であり, 形容詞 *large* の共起語の上位 10 語について比較した場合，粗頻度と似たような結果を返すことを報告している。他方，相互情報量は，二つの語の実測共起度とそれぞれの語の生起度から導かれる期待値との比の底を 2 とする対数で表す指標で, 次の式から得られる。

$$MI = \log_2 \frac{F_{x,y} \times N}{F_x \times F_y}$$

相互情報量は，頻度情報を対数圧縮するため, 低頻度の項目を高く評価する傾向があり, 意味論的な特徴が強調される指標である。ただし, あまり頻度の低い語の場合, コロケーションの結びつきが過大評価されるので, 適用の際注意を要する。

上記の各指標はそれぞれの原理, 特性をよく把握した上で, 用途に応じて使い分けるのが賢明であろう。結局のところ，高頻度の項目にウェイトを置く指標を採用するか, 低頻度ではあ

るが特徴的なコロケーションの検出力が高い指標を採用するかは，分析対象と目的によって決まる。辞書編纂や文法書の記述など，高頻度の現象や一般性を重視した記述を指向する場合は *t*-scoreや対数尤度比を使う方がよいと考えられるが，対して，文体論研究は書き手やテクストの idiosyncrasyをあぶり出すことに主眼を置くため，相互情報量の方が適していると考えられる。したがって，本研究では相互情報量を指標として用いる。ただし，共起度 10未満 ($F_{x,y} < 10$)の項目はデータに含めない。ORCHIDSには *gentleman*の共起語のうち生起度 10回以上の項目が 644タイプある。そこで，さらに Church and Hunks (1990)にならって分析対象の閾値を3.0に設定すると，390タイプに絞られる。相互情報量で降順にソートしたリストのうち上位 100タイプを挙げたのがTable 4である。Table 3とは内容，順位ともに大きく異なる。冠詞を始めとする機能語はほぼリストから消え，内容語がリストを占めている。ただし，よく観察すると*Brass's*, *Oliver*など特定のテクストにのみ生起する固有名詞が見つかる。そこで，そのような固有名詞 12項目を除外した残りの 378項目を変数に設定する。なお，変数とした 378項目の共起語が一度も生起しないテクスト一点 (Smollett, *The History and Adventures of an Atom* (1769) [59,503 words])もデータから外れるため，最終的に 77点のテクストにおける 378項目の共起語のコロケーション結合度をクロス分割表にしたデータを解析する。この分割表は77行× 378列，計 29,106個のセルからなる巨大なデータ行列である。目視で全体像を掴むことは到底不可能なため，データを縮約するテクニックの一つである対応分析を用いて解析する。

4．研究成果

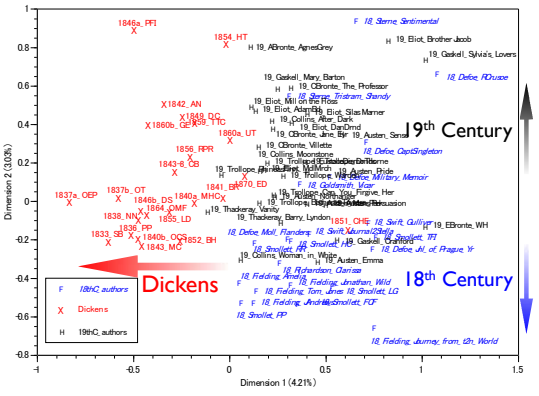

Fig. 2: Correspondence Analysis of the collocates of *gentleman* (378 collocates across 77 texts)

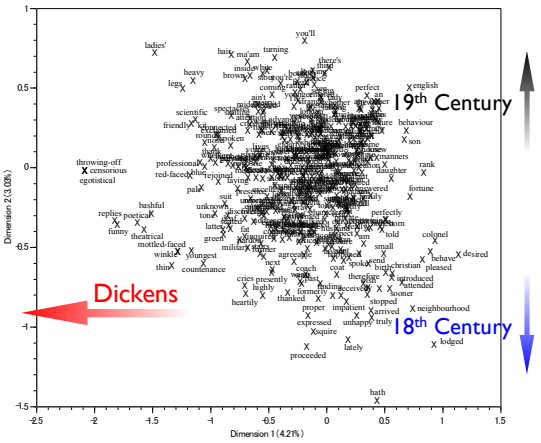

Fig. 3: A galaxy of gentleman's collocates: Word-map of 378 collocates

Figs 2–3は解析結果を要約したものである。対応分析は変数（共起語）間の複雑な関係，テクスト間の関係，さらには共起語の分布とテクストとの関係を少数の次元 (dimensions)に縮約して表現する手法である。

分析結果の特徴として，まず，各次元の寄与率（百分率表示）が非常に小さいことが挙げられる。これは 378 タイプの共起語間の関係，77 点のテクストの関係が極めて複雑であることを反映している数値だと言える。 Row weight は各次元におけるテクスト間の関係を表しており， Column weight は，各次元における共起語間の関係を示している。解析結果として得られる 377 次元のうち，最も寄与率の高い Dim 1 (4.2%)と Dim 2 (3.03%)の数値を二次元散布図に布置したものが Fig. 2 と Fig. 3 である。テクスト，共起語の散布図のいずれにおいても，図中で近接する項目は似通った性質があることが示されている。換言すると，Fig. 2 において近隣に位置するテクストどうしは *gentleman* のコロケーションのパターンが似通っているということである。逆に，図中の項目間を隔てる距離が大きいほど異質性が高いことが示されている。さらに，テクストと共起語の散布図とを照らし合わせることに より，テクストの配置にどのような語彙項目が関与しているかを読み取ることができる。ただし，軸 (Dimension)の解釈を行う際，各軸は直交している，すなわち無相関であることを念頭に置いておく必要がある。

(1)：テクスト間の関係

Fig. 2は対応分析の解の Dim1, Dim2をプロットした散布図である。Fig. 2の特徴は一目瞭然である。図の横軸 (Dim 1)に沿って，Dickens のテクストが図の左側，18世紀および 19世紀サブコーパスのテクストは， Trollope の *Phineas Finn* (1869)と Thackeray の *Vanity Fair* (1848)を除いて全て正 (+)の帯域に位置してい

る。つまり，Dickensの作品における *gentleman* のコロケーションが他の 19世紀作家や 18世紀作家の作品と一貫して異なっていることがこの図に表れている。しかし，Dickensの作品で一つだけ他の作品群から大きく外れたところに位置しているものがある。 1851–3年に書かれた *A Child's History of England* (*CHE*)である。*CHE*は子供向けに書かれた英国の歴史物語であり，Dickens作品の系譜では極めて異質なテクストである。この作品は *–ly*副詞や最上級表現を変数にした分析 (Tabata, 2009a/b)においても他の作品と異なる特徴が特定されており，これはほぼ予測された結果である。*CHE* での生起箇所を確認すると，*gentleman*は歴史上の人物の由来や史実について語る文脈で *Buckinghamshire, Catholic, England, English, Leicestershire, Welsh, Worcestershire, Yorkshire* など地名，国名，宗派を表す固有名詞と共起していることがわかる。歴史物語を特徴付けるコロケーションと言えよう。他の Dickens作品では 1830年代の作品はお互いに近い位置に分布している。 Dickensの初期の作品群は前掲の *–ly*副詞や最上級表現を変数にした分析の他，高頻度語や品詞範疇を変数にした分析でも中期～円熟期の作品とは示差的な特徴が認められている (Tabata, 1995; 2002;田畑, 1998他)。データとする言語項目に関わらず一貫した傾向が認められるのは興味深い事実である。

一方，図の右半分に目を転じると，19 世紀作家の作品はほとんどが上半分に，18 世紀作家の作品はその多くが下半分に位置している。このパターンから大きく外れたテクストは Sterne, *A Sentimental Journey* (1768), 奇書 *Tristram Shandy* (1759–67), および Defoe の *Robinson Crusoe* (1719)で，19 世紀のテクスト群の領域に深く入り込んでいる。多少の例外はあるものの，全体として見ると，Dickens の作品 ,他の 19 世紀作品，18 世紀作品との間にはある程度一貫したコロケーションの違いがあることを示唆する結果が出ていることは注目に値する。

(2)：共起語間の関係

前節で述べたテクストグループの布置に対応する共起語の布置を観察してみよう。図の中央付近に密集している共起語群はほとんど識別不可能であるが，作品の分布にあまり肯定的に寄与していない語であるため本稿では特に光を当てない。そこで，まず， Fig. 3の左方向に分布している共起語に目を向けると，主に形容詞，中でも多音節の形容詞が Dickensの作品群を特徴付けるコロケーションであることがわかる (*e.g. throwing-off, censorious, egotistical, poetical, theatrical, bashful, mottled-faced, scientific, friendly, professional, red-faced, pale, unknown, military, funny, thin, pale, youngest, green, round, etc.*)。KWICコンコーダンスを提示して後述するが，これらの語は特に初期の作品で固定したコロケーションで用いられており，特定の人物の呼称として機能していることを指摘しておく。

次に，18世紀のテクストを特徴付ける共起語に焦点を当ててみよう。Fig. 3の右下には主に動詞の過去形が分布している (*lodged, hath, desired, behave, pleased, introduced, attended, stopped, arrived, received, expressed, proceeded, etc.*)。このことから 18世紀のテクストは *gentleman*の所作，動作の描写に重点があると解釈できるが，詳細な分析に関しては稿を改めて論じたい。

図の右上，19 世紀のテクストが分布する領域に位置している共起語にはいかなるパターンが通底するのだろうか？この領域には *English, behaviour, son, manners, rank, daughter, fortune, father, lady, etc.*などが分布していることは確認できるものの，コンコーダンスで文脈を読み込んでみても，これらの語に共通する特徴を読み取ることは困難であった。その結果筆者が到達した解釈は次の通りである。19 世紀の作品は，Dickens や 18 世紀作品に対して「ネガティブ」に特徴付けられていると考えることが出来るのではないか。つまり，19 世紀サブコーパスは，Dickens を特徴付けるやや特殊な形容詞とのコロケーションや， 18 世紀のテクストにおいて顕著な動詞とのコロケーションのように際だったパターンが欠如している，という意味でネガティブに特徴付けられるという解釈である。その主な理由の一つとして考えられるのは，*gentleman* の生起度は Dickens において著しく高いが，18 世紀から 20 世紀にかけて大きく減少していっていることである。19 世紀のサブコーパスは全体で見た場合，ノード *gentleman* の生起度が低いため，Dickens や 18 世紀のテクストに特徴的に表れるコロケーションも生じにくいと考えるのが妥当であろう。さらに付け加えるならば，小説の勃興期であった 18 世紀のテクストは（*Tristram Shandy* のような奇書を除けば）いわば 'Augustan prose'の規範に沿ったものが多い。それに対し，19 世紀に入ると小説の形式の多様性が高まったため，世紀全体をセットとして見た場合，個々の作家あるいはテクストの特徴が相殺されて数値上平坦化されてしまった，ということも考えられる。実際，Dickens を合わせると，Fig. 2 における 19 世紀のテクストの分布は広範囲に及んでいる。ORCHIDS の 19 世紀サブコーパスには，同時代の Dickens にやや近い位置取りをしている

Thackeray や Trollope, あるいは年代的にも 18 世紀のテクストに近い特徴をもった Austen, あるいは Collins の *The Woman in White* (1859), Mrs Gaskell の *Cranford* などの作品から, 目立った特徴のないテクストまで，ある程度の variation が存在していることは確かである。本研究では多数のテクストにおける *gentleman* の共起パターンを俯瞰的に捉えるために，各テクストにおける相互情報量によって重み付けをしたコロケーション強度指標値行列を作成し，対応分析を用いてそのデータを縮約した。その結果，Dickens, 18 世紀サブコーパス，19 世紀サブコーパスを特徴付けるコロケーションのパターンを可視化することができた。これにより，Dickens の作品においていかにコロケーションが文体意匠として機能しているか，その一端を提示することができたと思う。従来，多変量解析を適用したテクストの計量研究は，高頻度語や機能語など，おおむね無標の語彙項目を変数として分析を行うものが多かったが，本稿で提示した方法論によって，内容語のコロケーション・パターンの量的な分析法が有効であることが明らかとなった。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）
(1). 田畑 智司「コンピュータを利用した英語研究の方法—多機能型テクスト分析ポータル TAPoR への招待—」『大阪大谷大学英語英文学研究』38 号(2011), 145–63. （査読無）
(2). 田畑 智司「TF-IDF 値を通してみるテクストの特徴—文体論研究への応用可能性を探る (1)—」『電子化言語資料分析研究 2009−2010』（大阪大学大学院言語文化研究科, 2010 年）65–80.（査読無）
(3). 田畑 智司「歴代米国大統領就任演説の言語変異―多変量アプローチによるテクストマイニング―」『英語コーパス研究』17 号 (2010), 143-60. （査読有）
(4). Tomoji Tabata, ‘More about *gentleman* in Dickens’, *Digital Humanities 2009 Conference Proceedings* (2009), 270–275. （査読有）
(5). 田畑 智司「*Gentleman* in Dickens―多変量アプローチで見る文体意匠としてのコロケーション―」統計数理研究所共同研究リポート『多変量アプローチによるテキストの計量研究』231号(2009), 1–22. （査読無）
(6). Tomoji Tabata, ‘*Gentleman* in Dickens: A Multivariate Stylometric Approach to its Collocation’, *Digital Humanities 2008 Book of Abstracts* (2008), 199–202. （査読有）

〔学会発表〕（計 5 件）
(1). Tomoji Tabata, ‘A Multivariate Approach to Linguistic Variations in the Century of Prose Corpus Part B: An experiment in corpus stylistics’, PALA 2010 (Poetics And Linguistics Association), 2010 年 7 月 24 日 ジェノア大学（イタリア共和国）
(2). Masahiro Hori, Tomoji Tabata, *et al*., ‘Dickens Lexicon and its Practical Use for Linguistic Research’, Digital Humanities 2010, 2010 年 7 月 9 日 King’s College London（連合王国）
(3). Tomoji Tabata, ‘A corpus stylistic study of collocation in Dickens’, PALA 2009, 2009 年 7 月 29 日 Roosvelt Academy（オランダ王国）
(4). Tomoji Tabata, ‘More about *gentleman* in Dickens’, Digital Humanities 2009, 2009 年 6 月 24 日 University of Maryland, College Park, MD（アメリカ合衆国）
(5). Tomoji Tabata, ‘*Gentleman* in Dickens: A Multivariate Stylometric Approach to its Collocation’ Digital Humanities 2008. 2008 年 6 月 29 日 オウル大学（フィンランド共和国）

〔図書〕（計 4 件）
(1). 田畑 智司他 4 名『電子化言語資料分析研究 2009-2010』大阪大学大学院言語文化研究科，2010 年，65–80.
(2). 田畑 智司他 4 名『電子化言語資料分析研究 2008-2009』大阪大学大学院言語文化研究科，2009 年，3–24.
(3). Masahiro Hori, Tomoji Tabata, et al. (eds.) *Stylistic Studies of Literature*: *In Honour of Professor Hiroyuki Ito*. Peter Lang AG. 2009. 113-134.
(4). 田畑 智司他 4 名『電子化言語資料分析研究 2007-2008』大阪大学大学院言語文化研究科，2008 年，65–81.

６．研究組織
(1)研究代表者
田畑 智司（TABATA TOMOJI）
大阪大学・大学院言語文化研究科・准教授
研究者番号：10249873