論　文　ニダバ (*Nidaba*) 30(2001), 24-33

語彙統計による文体分析

— Shakespeare と Marlowe における語彙の比較研究 —

富　原　裕　二

作家の個性はどの程度語彙のなかに反映するのか，また語彙の統計的な比較によって何が明らかになるのか．以下は，16〜7 世紀イギリスの劇作家についての語彙比較の試みである．対象とする作家は主に William Shakespeare(1564-1616)と Christopher Marlowe(1564-93)とするが，この二人以外に加えて，ほぼ同世代に活躍した Ben Jonson(1572-1637), Thomas Middleton(1580-1627), Cyril Tourneur(1575-1626), Francis Beaumont(1584-1616), Thomas Dekker(1572-1632)などの作家も扱う．Shakespeare と Marlowe を対象に選んだのは Shakespeare = Marlowe 説の検証という意味もあってのことである。今日では論じられることは少ないが，かつて代筆説は Shakespeare 学に重要な議論の一つだった[1]．現在でも A.D.Wraight(1994)のような熱心な遵奉者がいる．Shakespeare 代筆説が成立し得るという事実は，歴史的な考証は別において表現スタイルの問題として考えると，二人の作家の同一視を否定できるような決定的な違いを見つけるのが，現実には極めて困難であることを示している．では，語彙分析という方法は，個人に固有の表現スタイルの違いについて，何かを明らかにできるだろうか？ Shakespeare と Marlowe の語彙比較はこのような問題を考えるために有効であるように思える．

1．目的

対象とする作家，作品における語彙出現傾向を知るために，一定の語彙の 1000 語あたりの出現率を調べ，Shakespeare に対する Marlowe その他の劇作家の相関関係を求める．相関係数に個人間に有意の差が見られるか．有意の差があるとしたら，それを生み出した原因は何かを分析する．

2．テキストの選択

(a) 作家・作品

1500-1600 年代の演劇作品を主な対象とするが，同一のジャンル内での頻度差の相対的な価値を把握できるように，ほぼ同時期に書かれた他のジャンル（詩，散文）の

作品も参考のため加える．なお，すべてインターネット上で公開されたテキストを利用した[2]．

演劇作品

William Shakespeare: 全作品，ただし *Pericles*, *The Passionate Pilgrim* など真筆の保証がないものは除外する．

Christopher Marlowe: 全演劇作品と詩作品(*Hero and Leander*, *The Passionate Shepherd to his Love*)．*Ovid's Elegy* の翻訳．

Thomas Middleton: *A Chaste Maid in Cheapside, No Wit, No Help like a Woman's, The Phoenix, Revenger's Tragedy, The Second Maiden's Tragedy.*

Middleton と他の作家との共作: *The Old Law, The Changeling, The Family of Love, A Yorkshire Tragedy, Anything for a Quiet Life, The Blood Banquet, A Fair Quarrel, The Nice Valour, The Honest Whore, The Roaring Girl.*

Ben Jonson: *Epicoene, Volpone, Sejanus, New Inn, Every man in His Humour, Alchemist, Catiline, Bartholomew Fair, Cynthia's Revel's.*

Thomas Dekker: *The Witch of Edmonton, Match mee in London.*

Cyril Tourneur: *The Revenger's Tragadie.*

Francis Beaumont: *The Knight of the Burning Pestle.*

詩作品

Edmund Spenser: *The Faerie Queene.*

John Milton: *Paradise Lost, Paradise Regained, Samson Agonistes.*

同時代のソネット: Henry Constable: *Diana.* Edmund Spenser: *Amoretti.* William Percy: *Coelia.* Phillip Sidney: *Astrophil and Stella.* Michael Drayton: *Idea.* Samuel Daniel: *To Delia.*

その他の詩:

Michael Drayton: *Endimion.* Samuel Daniel: *The Complaint of Rosamond.* Edmund Spenser: *Epithalamion.* Lady Mary Wortley Montagu: *Selected Prose and Poetry.* Andrew Marvel: *Miscellaneous Poems.* 詩の部分のみを対象とする．

散文

Francis Bacon: *Essays*

Holy Bible: King James Version (Authorized in 1611)

なお，検索の対象はそれぞれの作品の本文のみとし，タイトル，献辞，人物名，ト書き等は含めない．また，Bacon や Mary Wortley Montagu の作品中のラテン語部分は検索対象領域から除外するなど，全ての作家について検索条件の平等を心がけた[3]．

3．検索語彙リストの条件設定

各作家，作品における頻度を調べる語彙は，次の基準でリストを作成する．

(1) 語彙レベル

原則として基本語彙をリストに登録する．ただし，Shakespeare 作品の頻度表をもとに，そこでの高頻度 50 位までの基本語を除く．語彙の出現頻度順に並べると，Shakespeare の場合，上位 10 語で全使用語彙数の 2 割を，上位 50 語が 5 割近くを占

めている．したがって，高頻度の基本語を数える Burrows & Hassal(1988)の方法は，マッチする件数を大量に利用できるという統計的な利点がある．しかし，今回の分析においては，次に述べる二つの理由で基本語 50 は除外し，リストに登録するのは 50 位以下の準基本語としたい．一つには，the and of などの基本語であればあるほど，たとえ個人的な頻度の差を認めることができたとしても，その違いをもたらした原因を分析することが難しくなるということ．もう一つの理由は，最上位の語彙と 100 位の語彙では出現頻度に一桁以上の差があり，同一の尺度で扱うのは適当ではないからである．

(2)機能語の重視

語彙リストに加える語彙は，ジャンルや主題の影響を受けにくい，いわゆる機能語（function word）を中心とする．名詞，動詞など内容語（content word）の頻度は主題に支配されやすいのでリストから除外する．喜劇では love が，悲劇では death や die が多用されて当然であり，リストに加えるには不適格である．また，人称代名詞についても登場人物の構成に大きく依存するので入れない．したがって，リストに入れるものは，まず，前置詞，接続詞，([w]h-疑問詞(関係詞)，代名詞である．形容詞，副詞類については，実体の属性についてではなく，もの・ことの有無，頻度，程度，全体と部分等について言及するもの(any no often all only half など約 50)のみを登録する．

(1)の条件から捨てられた基本語のなかには，50 位以降の複合語のなかで復活するものがある．たとえば，no は基本上位 50 語に含まれるが，代わりに，nothing nor nowhere などを数えることで，否定語の出現頻度にかんするデータに代用できる．同様に，what の代わりに whatever を数えるなどして，基本 50 語に含まれた機能語についてもなるべく網羅するようにした．

(3)品詞の問題

語が出現する文脈のなかでの品詞は問わない．前置詞は out や about などのように文法的には副詞として機能することかあるが，基本的にはもの・ことの関係の概念から派生した副詞化であり，一括して一語として数える．一般に語彙の文中での機能は一意的でなく，話者の語用に応じて柔軟である．*Twelfth Night* のなかで Toby が 'if thou *thou'st* him some thrice' と言うとき，代名詞の thou は動詞化(thou 呼ばわりする)されているが，その代名詞性が消滅するわけではない．上の(2)で前置詞，形容詞として登録した語彙が，現実の文のなかでは副詞や接続詞などの機能を持つことがあっても，品詞による区別を考慮しない．

(4)綴字異同の扱い

綴字に異同がある時代の文献であるので，統一的な綴字法を仮定して語数を数える．綴字法に時代的・個人的揺れがある場合は，なるべく統一的な基準で処理するよう努

める．たとえば，ever euer e'er e're 等は同一語とする．また，複合語を形成しがちな語彙についても作家間での異同を公平にする．したがって，else としては，else els ells だけではなく，elsewhere elswhere も登録する．反対に，whatever と what ever の表記法があるが，どちらも同一の語彙(whatever)として処理する．また，ハイフォン付きの語彙は切り離して２語として数えることで公平に努める．

以上(1)(2)(3)(4)の条件から作成された語彙は以下の通りである．

about above after after_* afterward against almost alone along although always amid among another any any_* anything at because before behind below beneath beside besides between betwixt beyond both despite down during each either else enough ere even every everybody everyone everything everywhere except far few fro from further half hence here here_* hither how into least less little many more more_* most most_* much near neither never next nobody nor nothing nowhere off oft often only or other otherwise out over rather around same since sith some some_* somebody something sometimes somewhat still such thence there there_* therefore these thither those though through throughout thus till too toward towards under unless until unto up upon upward very what when whence where where_* whereas wherefore wherever whether which while whiles whilst whither who whoever whole whom whose why with withal within without yonder

after_*, any_* とあるのは，たとえば，hereafter, thereafter を，あるいは anytime, anyway などの複合語を一語としてリストに一括登録したことを意味する．

4．結果と分析

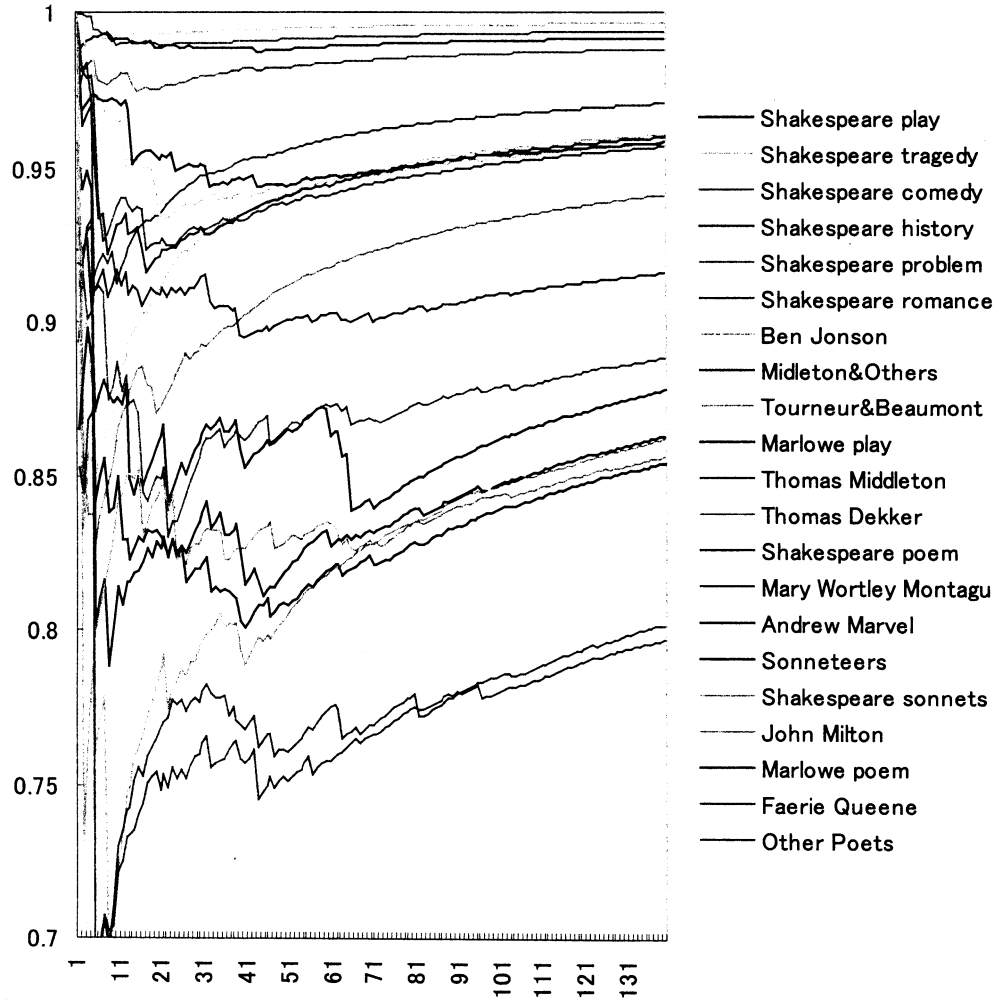
(1)相関係数の結果

綴字法が確立していない時代の文献であり，公開されたデジタル・テキスト資料のなかに誤字も散見される．出現語彙数の計上には綴字の揺れも考慮にいれてすべて網羅するように努めたが，不注意な漏れがあることも予想される．したがって，数値にはそれなりの誤差を含むものとして受け取るべきであるが，リストされた語彙の Shakespeare 演劇作品における出現頻度と，他の作家中での出現頻度とのあいだの相関係数は右の通りである．

Shakespeare play	1.000000
Shakespeare tragedy	0.996291
Shakespeare comedy	0.993646
Shakespeare history	0.991561
Shakespeare problem	0.987870
Shakespeare romance	0.970624
Ben Jonson	0.960686
Middleton&Others	0.959802
Tourneur&Beaumont	0.958188
Marlowe play	0.958075
Thomas Middleton	0.956769
Thomas Dekker	0.941156
Shakespeare poem	0.915956
Mary Wortley Montagu	0.888196
Andrew Marvel	0.877978
Sonneteers	0.862773
Shakespeare sonnets	0.861758
John Milton	0.856029
Faerie Queene	0.801699
Other Poets	0.796793
Francis Bacon	0.703154
Bible(King James')	0.539786

また，この最終的な値にいたる過程を示したのが次ページのグラフである．リスト中の語彙を Shakespeare 作品での頻度の高い順に

一語ずつ，他の作家・作品に関して検索をおこない，検索データの追加による相関係数が漸次変化する軌跡をグラフにしたものである．なお，Bacon と Bible はグラフの下限(0.7)以下であるため表示されていない．



グラフの縦軸は相関係数を，横軸は検索した語彙の延べ個数を表している．このグラフを見ると，相関係数自体には僅差しかないが，これ以上データを追加しても大した変化を予想できないほどの安定に達しているのが解る．したがって，作家間の相関係数のわずかな差は，決して偶然や誤差ではなく，特定すべき原因から生まれた結果であると見なさざるを得ない．実際に，データのなかに認めることができる次の(i)(ii)の事実は，その差が有意であることを示している．

(i)ジャンル間の相違

語彙選択においては作家間の個人差よりはジャンルの違いがはるかに優勢であ

る．Shakespeare の演劇作品との相関関係を見ると，最も近くには，同世代の劇作家が書いた演劇のグループがあり，次に相当の段差を挟んで，詩のグループ（Marlowe poem, Montagu, Marvel, Sonneteers, Milton）が，そしてはるかに遠いところに散文のグループ（Bacon, Bible）が分布している．

(ii)同一ジャンル内での個人間の差

Shakespeare 演劇は全ての劇が 0.97 以上であり，その中には Shakespeare 以外の誰も入らない．さらに，Shakespeare と他の作家との間には無視できない段差がある．他の劇作家は 0.956～0.96 の間に集中しており，0.97 は Shakespeare 演劇の識別値と見なすことができる．

以上の結果に基づいて，Shakespeare を中心にした他の作家の遠近図を描くことができる．Shakespeare 劇的な世界の中心に一番近いところに，まず悲劇（0.996291）があり，少し離れるにつれて喜劇（0.993646），史劇（0.991561），問題劇（0.987870）が広がっている．ロマンス劇（0.970624）はやや異質であり Shakespeare 演劇の辺境に接している．Shakespeare 演劇の領域の外に出ると，しばらく空白の領域が続き，当時のほとんどの劇作家の世界が広がっているのはその向こうである．Ben Jonson（0.960686），Middleton & Others（0.959802），Tourneur & Beaumont（0.958214），Marlowe (0.958075)，Middleton (0.957653) らが僅差でそこに集中している[4]．Dekker は同じ劇作家でありながら，この一団からやや離れているが，Dekker の後にはまた空白の大きな段差が存在する．次に来るのはもはや演劇ではなく，Shakespeare の詩（0.915956），Montagu（0.888196），Marvel（0.877978）など詩作品だけである．したがって，0.9 をもって演劇というジャンルの識別値と見ることができるのである．

(2)作家の区別の要因に関する分析

注目すべきことは，各作家の語彙の選択においては，主題の支配を受け難い基本的な語彙のレベルにおいてすら，個性の区別が認められるということである．では，Shakespeare と他の作家との間に，このような区別をもたらしているのは何だろうか？他の劇作家全てについて考察することは難しいので，ここでは Shakespeare と Marlowe の比較にだけ限って考察したい．

Shakespeare と Marlowe の頻度のデータを比較してみると，両者の違いを生み出すことに貢献している語彙には，一つの傾向的な性質があるように思える．Marlowe に比べて Shakespeare が好んで使った語彙を見ると，たとえば very nothing almost between などがあり，反対に Shakespeare が比較的使わなかった語彙には，whereas amid fro unto throughout などがある．勿論，二人の語彙傾向に区別をもたらした要因は複雑であろうが，要因の一つとして口語性の程度を挙げることができるように思える．

口語性の程度を認定する方法として，演劇作品と詩作品での語彙の頻度比を，口語性を示す便宜的な基準として使いたい．便宜的というのは，詩的であることと口語的であることとは必ずしも対立しないし，逆に，特に韻文が使われがちだった16〜7世紀英国演劇においては，演劇的であることと口語的であることとは単純な対立関係ではないからである．しかし，概して演劇作品は詩作品よりは口語的であるという前提に大きな間違いはないと思われる．したがって，ここではある語彙(n)の演劇作品における頻度をDn（1000語あたり），詩作品における頻度Pnを計算し，その語彙(n)のDn/Pn比をもって口語性を測る便宜的な尺度とする．ただし，ShakespeareとMarloweの分析が目的であるので，基準とするDn/Pn値の算出には，二人を除外してそれ以外の作家の演劇，詩作品だけを使うものとする．以下の表は，Dn/Pn比が3/2以上（演劇的）であるものと，逆に2/3以下（非演劇的）である語彙について，Sn/Mnが顕著なもの（1.5以上または0.667以下）のみを示したものである．なお，Snはある語彙(n)のShakespeare演劇での1000語あたり頻度であり，MnはMarloweのそれである。

	Sn/Mn	Sn	Mn	Dn/Pn
sith	0.536855	0.029694	0.055311	0.000000
fro	0.268428	0.004949	0.018437	0.040064
there_prep	0.640749	0.183110	0.285775	0.044942
thence	0.600421	0.105165	0.175152	0.089374
wherever	0.402614	0.007423	0.018437	0.107889
throughout	0.469762	0.008661	0.018437	0.143851
through	0.530835	0.327866	0.617642	0.181631
around	0.662664	0.097741	0.147497	0.183626
whereas	0.067104	0.006186	0.092185	0.191424
amid	0.205788	0.028456	0.138278	0.191985
nowhere	0.201334	0.003712	0.018437	0.208217
where_prep	0.622428	0.395914	0.636080	0.208629
everywhere	0.603949	0.022270	0.036874	0.247254
unto	0.393626	0.540670	1.373563	0.330990
whose	0.453516	0.760897	1.677775	0.335729
whilst	0.645491	0.124960	0.193589	0.344007
while	0.515541	0.299410	0.580768	0.358164
thus	0.587461	0.947718	1.613245	0.367199
far	0.658854	0.334052	0.507020	0.371653
whether	0.421145	0.112588	0.267338	0.546873
among	0.666581	0.184347	0.276556	0.563143
least	0.515369	0.118774	0.230464	0.686304
about	0.581581	0.482520	0.829669	0.855476
near	0.508085	0.262293	0.516238	1.175649
these	0.562305	1.612111	2.866967	1.243338
**somewhat	0.402621	0.018558	0.046093	3.113852
*oft	1.599342	0.176924	0.110623	0.133482
*each	1.784220	0.279614	0.156715	0.155792
*during	2.013017	0.018558	0.009219	0.290103
*afterward	3.086669	0.028456	0.009219	0.331795
*same	1.831586	0.287037	0.156715	0.405215
*behind	2.303972	0.127435	0.055311	0.478027

there	1.694607	2.718195	1.604027	1.510766
between	5.413209	0.299410	0.055311	1.545339
whole	1.878952	0.155891	0.082967	1.635176
little	1.627309	0.600057	0.368742	1.689438
out	1.590342	1.656651	1.041695	1.825175
hither	2.303952	0.382304	0.165934	1.900973
somebody	---	0.019796	0.000000	1.942084
another	2.423253	0.402100	0.165934	2.113848
off	1.582275	0.554279	0.350305	2.150929
almost	2.650675	0.195482	0.073748	2.398370
something	1.868634	0.223939	0.119841	2.511256
nothing	2.100401	0.774506	0.368742	3.051831
too	1.706997	1.510658	0.884980	3.439182
after_plus	1.677659	0.061862	0.036874	3.758253
very	4.929204	0.999682	0.202808	7.630517
any_plus	---	0.028456	0.000000	17.408416
anything	1.650800	0.152179	0.092185	22.628552

Shakespeare が Marlowe に比して避けた語彙（S/M <0.5）は，**somewhat という例外を除いて全てD/Pが低い．つまり，演劇的であるよりは詩的であるような語彙を使うことを Shakespeare は好まなかった．反対に，Shakespeare が好んで使った語彙（S/M > 1.5）は，D/Pが全て高い訳ではないが比較的高い．語彙の前に* がついているのは，D/P 値が高いにも拘わらずS/M が低い例外である．* または** を付けた「例外」は 49 語中わずか 7 語にすぎない．つまり，Shakespeare が好んで使った語彙は，詩作品でよりは演劇でよく使われた語彙を使う傾向が，Marlowe に比べると著しく認められるのである．ちなみに，全ての語彙(n)に関する各劇作家における出現頻度と，演劇作品での頻度Dn との相関，ならびに詩作品での頻度Pn との相関係数を下に示す．なお，ここでもDn とPn の中には Shakespeare と Marlowe を含めていない．

演劇作品との相関		詩作品との相関	
Middleton & Others	0.989132	Marlowe play	0.890963
Ben Jonson	0.983509	Shakespeare play	0.864832
Thomas Middleton	0.982375	Ben Jonson	0.804044
Thomas Dekker	0.975109	Tournner&Beaumont	0.797478
Shakespeare play	0.974850	Thomas Middleton	0.775501
Tourner & Beaumont	0.966219	Middleton&Others	0.764238
Marlowe play	0.914761	Thomas Dekker	0.756088

詩作品との相関について見ると，Marlowe のほうが Shakespeare よりやや高い．しかし，他の劇作家と比較すると，この二人には詩作品との相関が高く，詩的語彙の頻度の高さは二人に共通した特徴と見なすべきである．ところが，Marlowe には Shakespeare からはっきりと区別する要素がある．演劇作品との相関について見ると，Marlowe は演劇的な語彙の頻度において Shakespeare との間に大きな較差がある．実は，演劇的語彙の不足は Marlowe 一人に際だった特徴であると言うことができる．なぜなら，演劇作品との相関係数は，Marlowe（0.9114761）よりも一様に高く，しかも

Marlowe 以外の全ての劇作家で大きな差が見らない（0.966219～0.989132）からである．

要するに，語彙選択においてそれぞれの劇作家は次のような特徴を持っていると考えることができる．同時代の演劇のなかで，Marlowe 一人だけが詩的で，非口語的なスタイルである．Shakespeare は Marlowe とほとんど同じ程度に詩的であるが，同時に口語的である点で Marlowe とははっきりと区別される．エリザベス朝演劇において最高峰とされる二人以外の劇作家について見ると，だれもほぼ均一に口語的（演劇的）であるが，詩的な語彙において不足していることが解る．

５．結論

Shakespeare と Marlowe，さらに同時代の劇作家における語彙の使用頻度を分析することで明らかになったのは次の２点である．

(1)機能語のレベルにおいてでさえも個人を識別する有意の差が認められる．

(2)Shakespeare と Marlowe さらに他の劇作家のスタイルの違いを決定する一つの要因として，演劇的な語彙と詩的な語彙の比率の違いを挙げることができる．

注

[1] H.N.Gibson(1962)は *The Shakespeare Claimants* のなかで Shakespeare 代筆説のどの候補者も論駁している．Christopher Marlowe 説の根拠も恣意的で薄弱とされている．

[2] テキスト入手先サイトについては，一括して以下に示すだけにとどめる．

http://promonet/pg/ (Project Gutenberg)
http://darkwing.uoregon.edu/~rbear/ren.htm (Renascence Editions)
http://www.hti.umich.edu/p/pd-moderng/bibl.html (HTI Modern English Collection)
http://ota.ahds.ac.uk/ (The Oxford Text Archive)
http://etext.virginia.ed/modeng/ (Electronic Text Center)

[3] 検索，データ処理のためにはプログラミング言語 Perl を使った．

[4] もちろん，Shakespeare 以外の作家の相関係数に差が極めて小さいことは，それらの作家同士の相関が高いという意味ではなく，単に Shakespeare からの隔たりが一様であるという意味にすぎない．

参考文献

斉藤俊雄他(編) (1998) 「英語コーパス言語学」研究社.

田中穂積(監修) (1999) 「自然言語処理 - 基礎と応用 - 」電子情報通信学会.

長尾真他 (1998) 「岩波講座 言語の科学 9 言語情報処理」pp.149-199. 岩波書店.

Burrows, J.F. (1987) *Computation Into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Gibson, H.N. (1962) *The Shakespeare Claimants,* pp.124-150. Methuen & Co Ltd.

Larry Wall, L. and Schwartz R.L. (1990) *Programming Perl.* O'Reilly and Associates, Inc. 「Perl プログラミング」近藤嘉雪訳. ソフトバンク.

Milic, L.T. (1967) *A Quantitative Approach to the Style of Jonathan Swift.* The Hague: Mouton.

Wraight, A.D. (1994) *The Story that the Sonnets Tell.* Adam Hart Ltd. London.