英語コーパス学会10周年記念シンポジウム

『日本における英語コーパス言語学の現状と展望』

コーパスとテキスト

田畑　智司
大阪大学

本発表の視点

電子化された言語資料を活用したテキスト・文体・レトリックの研究.

ここでは「狭義のコーパス」だけでなく，電子テキスト一般を活用したものをも視野に入れる.

国内外でこれまでに行われてきた研究の潮流を俯瞰的に分類・整理し，今後の動向・課題を考える.

電子化された言語資料をもとにしたテキストの研究

その目的をもとに分類すると：

- Authorship attribution 著者推定論 (& 文体模写)
- Stylistics 文体論
- Text typology & variation studies 類型論・言語変異

  register variation, regional variation, social variation, authorial variation, chronological variation, etc.

内外の業績を振り返る

～時系列に沿って～

三つの時代区分：

I. ～1979年（電子コーパス黎明期）
II. 1980年代
III. 1990～現在

I. 電子コーパス黎明期（～1970年代まで）

1.1 著者推定(Disputed authorship)

- A. Ellegård's study of 'Junius Letters' (1962a, b)

  1769年から1772年にかけて*The Public Advatiser* 紙にJuniusの名の下に投稿された政治評論.
  著者として40名あまりの候補者があがっていたが，Sir Philip Francis 説が有力であった。

- F. Mosteller & D. L. Wallace, 'Federalist Papers' (1964)

  1787年から1788年にかけてニューヨーク市民に新憲法を承認させようという意図をもって書かれた88篇の論文で，最初新聞に掲載され，その後本の形にまとめられたもの.
  一連の論文はJohn Jay, Alexander Hamilton, James Madisonの三人が執筆したものであることが判っており，著者推定の問題となっていたのは，その内12篇であった．しかも，問題の論文の著者は HamiltonかMadisonのどちらか一方であるということがあらかじめ判明していたため，著者推定の問題としてはJunius Lettersよりはstraightforwardな課題であった.

Ellegård, Mosteller & Wallaceの方法論

- ‘Marker words’（識別語）の特定
- 候補者のテキストにおける識別語生起率の比較

Ellegård

458語句(plus words : minus words)

51組の同義語間の選択傾向

(*on* vs *upon*, *sort* vs *kind*, etc.)

Mosteller & Wallace

HamiltonとMadisonのテキストで有意差のある28語

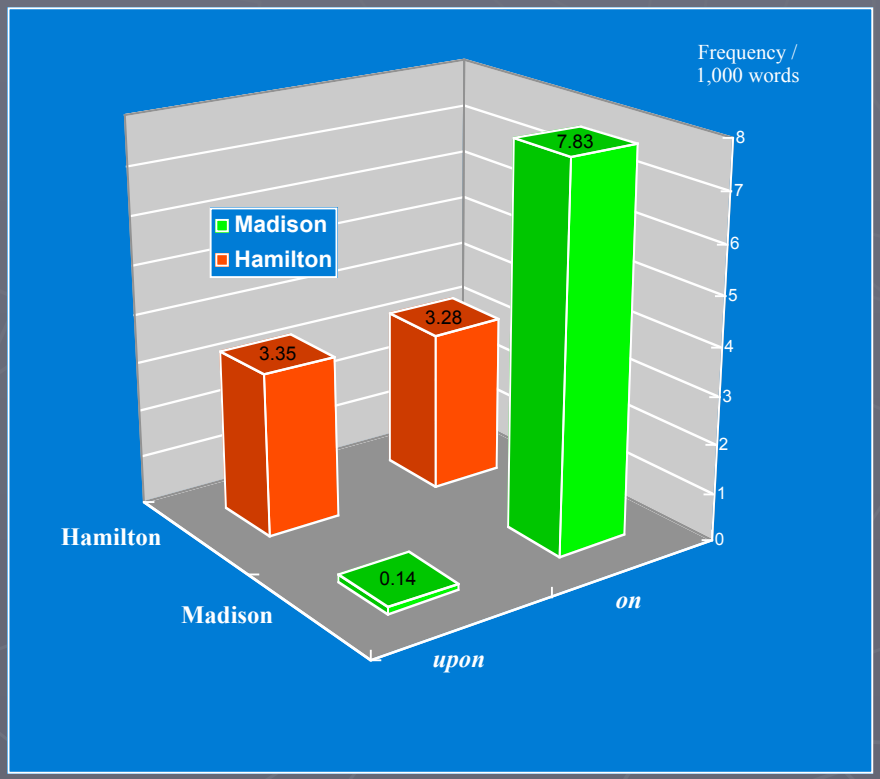


Fig. 1 Marker words in the ‘Federalist’ (Mosteller & Wallace, 1964)

1.2 Stylistic study

L. T. Milic (1969) によるSwiftの文体研究

R. Cluett, *Prose Style and Critical Reading* (1976)

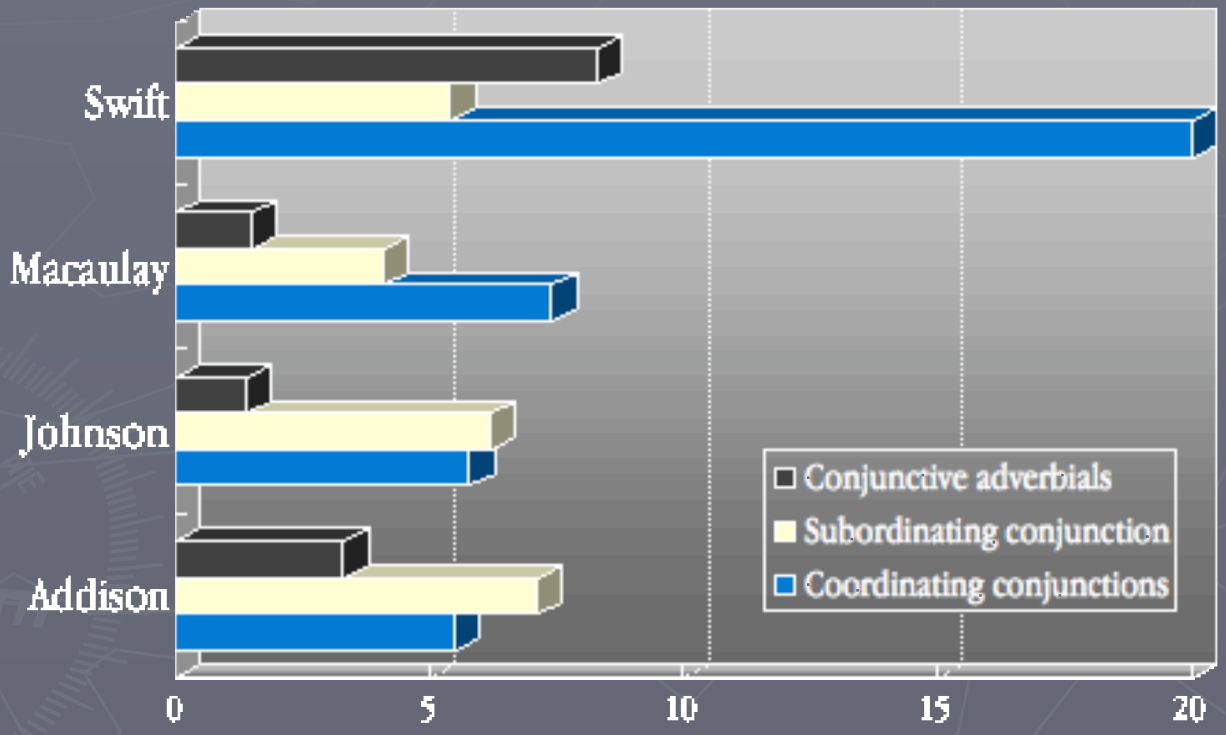


Fig. 2 Sentence initial connectives in the four authors: Percentage in 2000-sentence samples (Milic, 1969: 125).

1.3 Chronological studies and text typology

R. Cluett (1976)

テキストの各単語を2桁および3桁のWord-classに翻訳して入力した約30万語のコーパス．Philip Sydneyから Anthony Burgessにいたる80人の著者をサンプルとして，英語散文の史的動向をとらえようとする試み.

B. Brainerd (1979)

人称代名詞の分布をもとにしたShakespeare劇のジャンル分類の試み．Discrimination Analysisなどの統計解析法の援用．人称代名詞という内的基準による作品の分類が，伝統的なジャンル区分と合致していることを示している.

II. 1980年代
～計算機環境の向上，統計学の進歩とともに～

ICAME Corpus Collection, Part-of-speech tagging, Electronic text archive

J. F. Burrows

Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method (1987)

D. Biber

Variation across Speech and Writing (1988)

テキストから引き出した語彙頻度データ解析ツールとしての多変量解析法の有効性

Multivariate Statistics (Principle Components Analysis [PCA], Factor Analysis, etc.)

2.1 Stylistic study

Burrows (1983～)

investigates Jane Austen's narrative style, character differentiation through idiolects and free indirect discourse by the multivariate analysis of 12-60 most common words.

Implications of his results:

computer-assisted literary criticism, literary and linguistic stylistics, identification of stylistic 'fingerprint,' authorship attribution, stylistic imitation, register variation, etc.

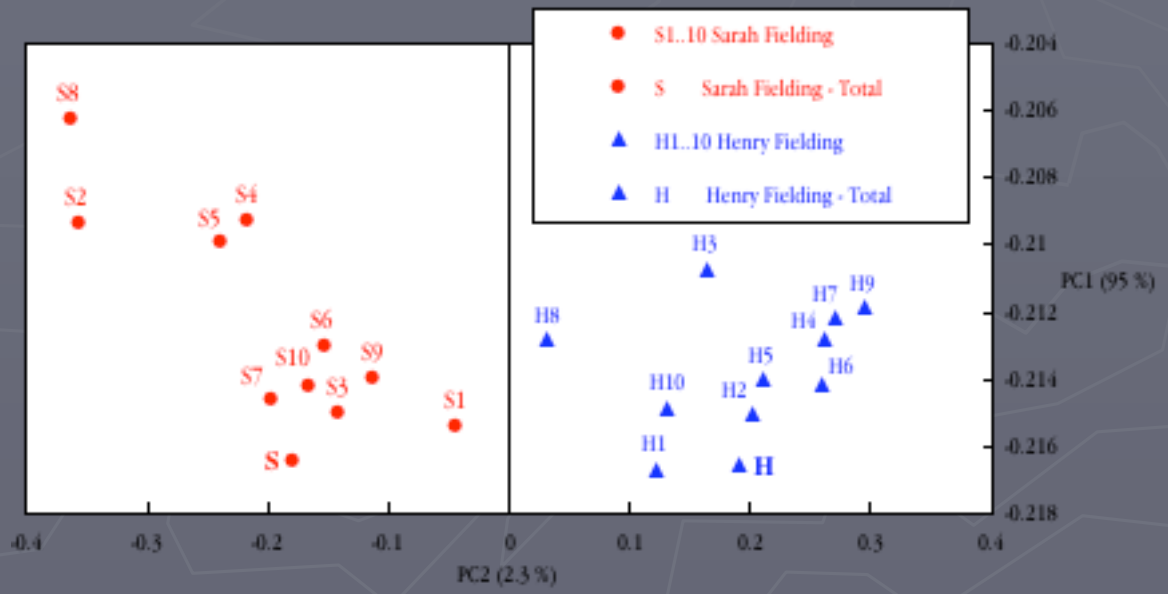
2.2 Authorship attribution

Burrows & Hassal (1988)

最も頻度の高い50タイプの高頻度語を変数としてPCAで解析．高頻度語の共起関係とテキスト間の関係．Fielding 兄妹の文体の違いを散布図を用いて視覚的に表示．*Anna Boleyn*, 'A Vision'など，disputed authorshipの問題を扱った．

Fig. 3

Narratives by Sarah & Henry (Burrows & Hassal, 1988).



総語数6760語からなるAnna Boleynは，Henryのテキストと共通する特徴を示す最初の1100語ほどの部分，Sarahが書いたであろう推定される次の2800語程の部分，そして，最後の2860語に分割することが可能であると述べ，最後の部分に関しては，Sarahが下書きをしたものをHenryが書き改めたか，もしくはHenryがSarahを模倣して書いた可能性が高いことを示唆した．

2.3 Register variation

竹蓋 (1981)

Conversation, letters, classroom speech, Literary texts, 中学教科書, 入試問題, News, Business letters, Science abstract などからなるコーパス

使用域間での語彙の類似度比較
Cluster Analysisも利用

Biber (1988)

67タイプの言語項目（ハンドアウト参照）を変数としてFactor Analysisで解析.

言語項目の共起関係を主要因子を抽出.
使用域による言語変異を6つの次元('Dimensions')で捉える.

23におよぶ Spokenレジスターとwrittenレジスターを多次元的に比較分析.

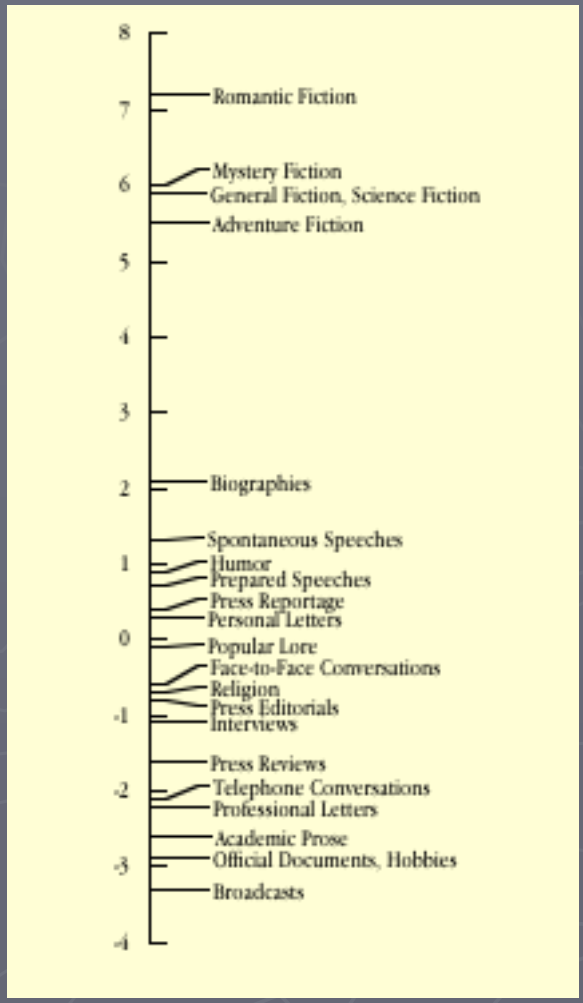


Fig. 4 Relationship of 23 registers in Dimension 2 (Biber, 1988).

Biber (1988)の分析モデルで用いられる67の言語項目：その 1

- **A. Tense and aspect markers**

1. past tense
2. perfect aspect
3. present tense

- **B. Place and time adverbials**

4. place adverbials (e.g., *above*, *beside*, *outdoors*)
5. time adverbials (e.g., *early*, *instantly*, *soon*)

- **C. Pronouns and pro-verbs**

6. first person pronouns
7. second person pronouns
8. third person personal pronouns (excluding *it*)
9. pronoun *it*
10. demonstrative pronouns (*that*, *this*, *these*, *those* as pronouns)
11. indefinite pronouns (e.g., *anybody*, *nothing*, *someone*)
12. pro-verb *do* (e.g., the cat *did* it)

- **D. Questions**

13. direct *WH*-questions

- **E. Nominal forms**

14. nominalizations (words ending in *-tion*, *-ment*, *-ness*, *-ity*)
15. gerunds (participial forms functioning as nouns)
16. total other nouns

- **F. Passives**

17. agentless passives
18. *by*-passives

- **G. Stative forms**

19. *be* as main verb (e.g., His house *is* big.)
20. existential *there*

- **H. Subordination features**

21. *that* verb complements (e.g., I said *that* he went.)
22. *that* adjective complements (e.g., I'm glad *that* you like it.)
23. *WH* clauses (e.g., I believed *what he told me*.)
24. infinitives
25. present participial clauses (e.g., *Stuffing* his mouth with cookies, Joe ran out the door.)
26. past participial clauses (e.g., *Built* in a single week, the house would stand for fifty years:)
27. past participial WHIZ deletion relatives (e.g., *the solution produced by this process*)
28. present participial WHIZ deletion relatives (e.g., *the event causing this decline is...)*
29. *that* relative clauses on subject position (e.g., the dog *that bit be*)
30. *that* relative clauses on object position (e.g., the dog *that I saw*)
31. *WH* relatives on subject position (e.g., the man *who likes popcorn*)
32. *WH* relatives on object position (e.g., the man *who Sally likes*)

Biber (1988)の分析モデルで用いられる67の言語項目：その2

33. pied-piping relative clauses (e.g., the manner *in which he was told*)
34. sentence relatives (e.g., Bob likes fried mangoes, *which* is the most disgusting thing I've ever heard of)
35. causative adverbial subordinators (*because*)
36. concessive adverbial subordinators (*although*, *though*)
37. conditional adverbial subordinators (*if*, *unless*)
38. other adverbial subordinators (e.g., *since*, *while*, *whereas*)

- **I. Prepositional phrases, adjectives, and adverbs**

39. total prepositional phrases
40. attributive adjectives (e.g., the *big* horse)
41. predicative adjectives (e.g., the horse is *big*)
42. total adverbs

- **J. Lexical specificity**

43. type/token ratio
44. mean word length

- **K. Lexical classes**

45. conjuncts (e.g., *consequently*, *furthermore*, *however*)
46. downtoners (e.g., *barely*, *nearly*, *slightly*)
47. hedges (e.g., *at about*, *something like*, *almost*, *sort of*, *kind of*)
48. amplifiers (e.g., *absolutely*, *extremely*, *perfectly*)
49. emphatics (e.g., *a lot*, *for sure*, *really*)
50. discourse particles (e.g., sentence initial *well*, *now*, *anyway*)
51. demonstratives

- **L. Modals**

52. possibility modals (*can*, *may*, *might*, *could*)
53. necessity modals (*ought*, *should*, *must*)
54. predictive modals (*will*, *would*, *shall*)

- **M. Specialized verb classes**

55. public verbs (e.g., *assert*, *declare*, *mention*, *say*)
56. private verbs (e.g., *assume*, *believe*, *doubt*, *know*)
57. suasive verbs (e.g., *command*, *insist*, *propose*)
58. *seem* and *appear*

- **N. Reduced forms and dispreferred structures**

59. contractions
60. subordinator that deletion (e.g., *I think* [*that*] *he went*)
61. stranded prepositions (e.g., *the candidate that I was thinking of*)
62. split infinitives (e.g., He wants *to convincingly prove* that ...)
63. split auxiliaries (e.g., they *are objectively shown* to ...)

- **O. Coordination**

64. phrasal coordination (NOUN *and* NOUN; ADJ *and* ADJ; VERB *and* VERB; ADV *and* ADV)
65. independent clause coordination (clause initial *and*)

- **P. Negation**

66. synthetic negation (e.g., *No answer is good enough for Jones*)
67. analytic negation (e.g., *that's not likely*)

2.4 Chronological study

D. Biber & E.Finegan,

‘Drift and the Evolution of English Prose Style: A History of Three Genres.’ (1989)

17世紀から20世紀にかけての3ジャンルの英語散文（Essay, Fiction, Letter)

Biber (1988)の分析モデルに基づき，67タイプの言語項目の正規化頻度をもとに，テキストの比較.

18世紀以降，次第に口語的要素が強まって行く相を呈示．(*cf.* Milic, Cluett, Burrows)

III. 1990年代～21世紀へ

Bank of English, British National Corpus, OCR, Internet …
大規模コーパス時代
テキストの電子化が進む
統計的手法（特に多変量解析ツール）の普及

3.1 Authorship attribution

H. Baayen, J. F. Burrows & H. Love, D. H. Craig, W. E. Elliott & R. Valenza, R. S. Forsyth & D. I. Holmes, J. Hope, G. Ledger, D. L. Mealand, T. V. N. Merriam, M. W. A.Smith

中でも，Craig (1992)による‘Additions’ to *The Spanish Tragedy* の著者推定研究，および *Edward III* が ‘Shakespeare canon’ に属するものであることを客観的な証拠で示したHope (1994)特筆に値する.

PCA, Correspondence analysis, Discriminant analysisなどの手法が盛んに用いられ始める

3.2 Variation studies

3.2.1 Register & Text typology

D. Biber, Biber *et al*., A. Jucker, 久屋， J. Nakamura, 齊藤， M. Short *et al*., K. Takahashi, 高橋・古橋, S. Tsukamoto, A. F. Umesaki

* この分野では我が国の研究者の貢献も多い.

Biber model (Multi-Feature/Multi-Dimensional approoach)

数量化III類, Quantification of contingency table, ANACOR (Correspondence Analysis)

Cluster analysis などの手法を駆使している

Nakamura model—QCT—

語彙生起頻度の分割表を数量化．語彙項目とテキストサンプル間の相互関係を三次元空間に視覚化

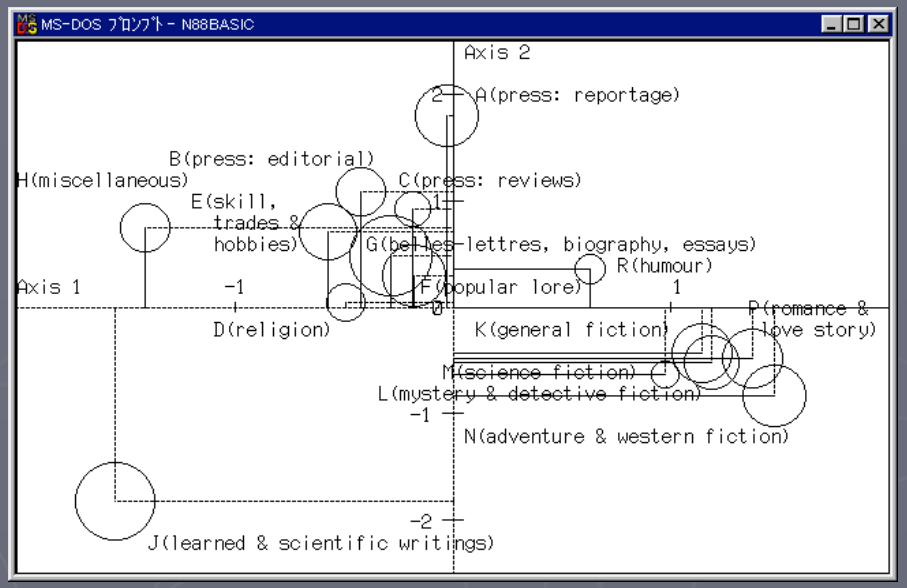


Fig. 5
動詞の分布に基づく三次元空間でのテキストの位置づけ(LOB Corpus)

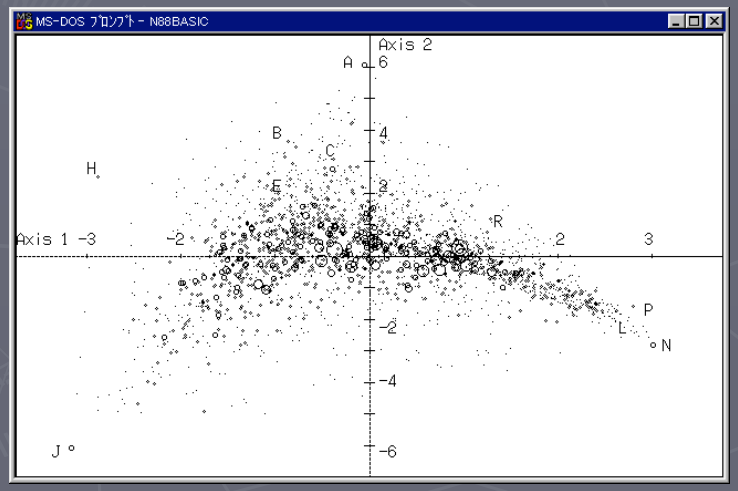


Fig. 6　LOB Corpus の動詞の宇宙

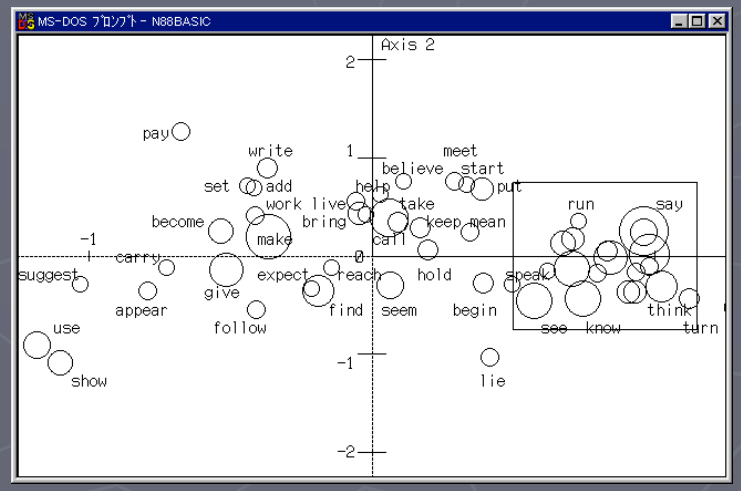


Fig. 7　LOB Corpus の高頻度動詞の分布

3.2.2 Chronological variation

J. F. Burrows, J. N. G. Binongo, D. H. Craig, Tabata

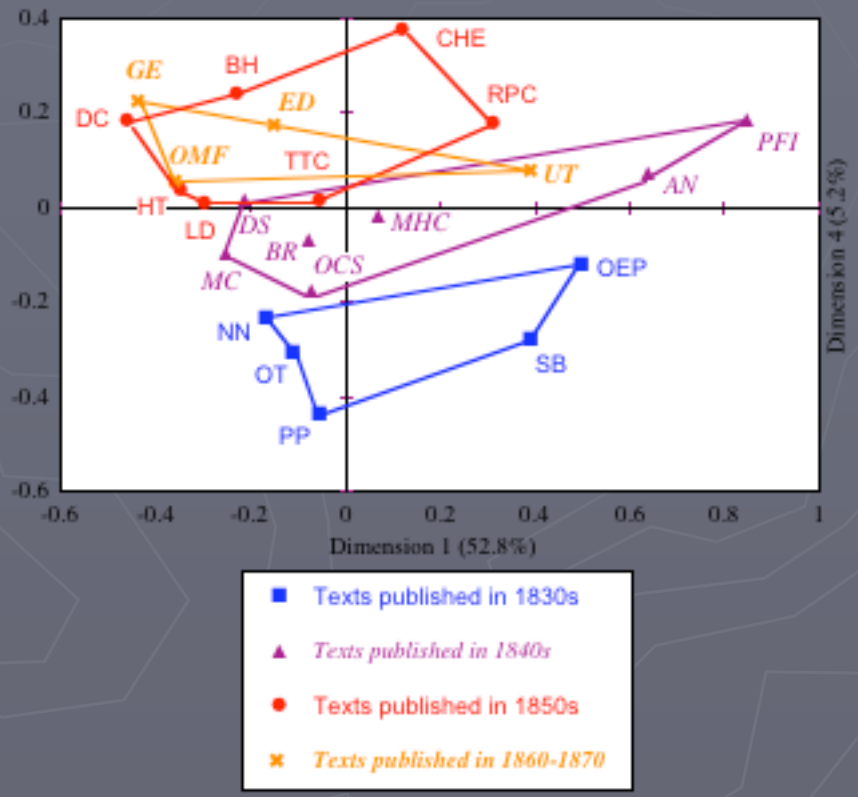
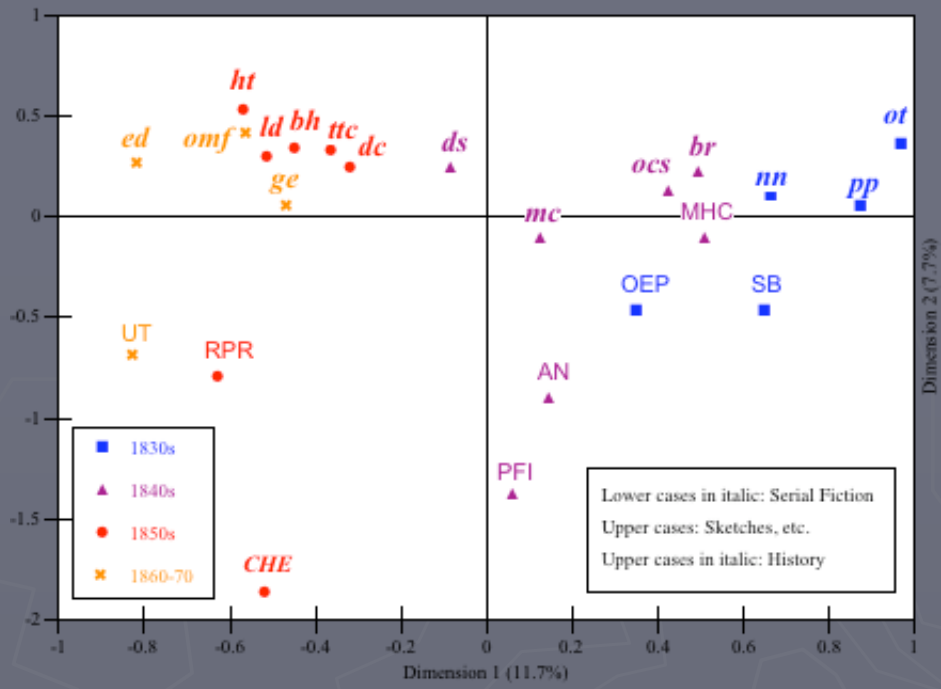


Fig. 8 Dickens in Four Decades: Correspondence analysis of 34 word-classes across his oeuvre (Tabata, 2002)

‘—ly’ 副詞の生起頻度分布とテキストの出版年代，ジャンルとの相互関係

Dickensが複数の作品で用いている1246タイプの ‘—ly’ 副詞



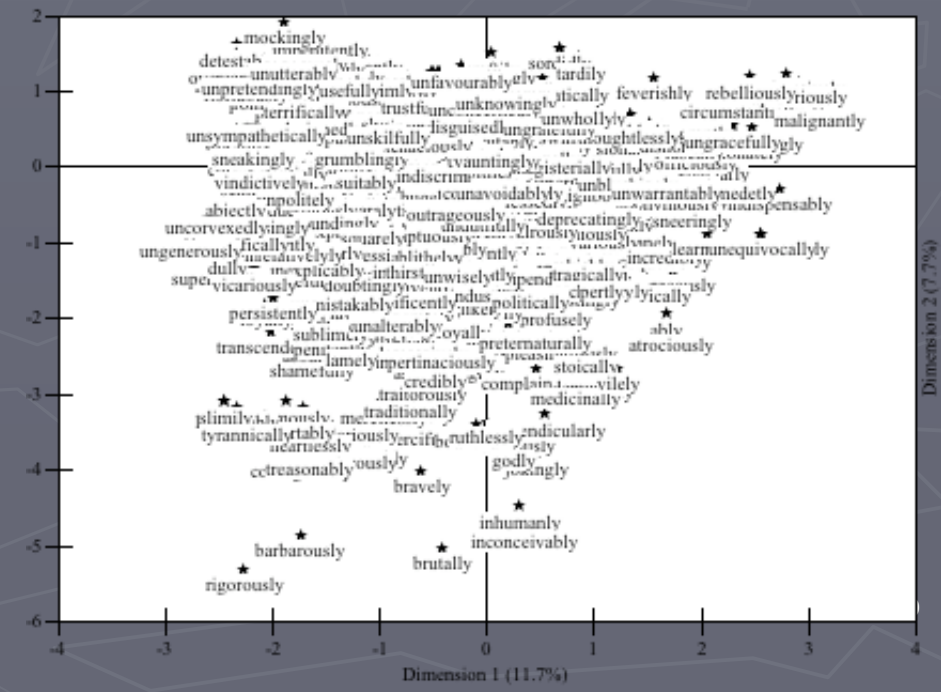


Fig. 9 （上）
Dickensian Galaxy of Adverbs: Text-plot
1246タイプの‘—ly’ Adverbsの分布をもとにしたDickensのテキスト23点の位置付け
（経年的推移・テキストジャンルによる違いが表れている）

Fig. 10 （下）
Dickensian Galaxy of Adverbs: Word-plot
1246タイプの‘—ly’ Adverbsの分布

3.2.3 Literary idiolects and stylistic imitations

文学作品中のイディオレクト，文体模写

J. F. Burrows, D. H. Craig, E. Irizarry, E. Johnson, C. W. F. McKenna & A. Antonia, McKenna *et al.*, L. L. Opas, etc.

3.2.4 National differences in narrative style

J. F. Burrows (British, American, Australian, New Zealand)

IV. —文学テキストの言語・文体研究—
～わが国における計算機，コンコーダンスを活用した研究の系譜～

4.1 中世英語英文学・フィロロジー，古英語(1978年～)

西村，西出・川端，山縣 (OE)，中尾，地村

早い時期から電子的なツールの利用に積極的であったフィロロジストの研究・教育活動が，わが国のコーパス言語学の普及の源になっている.

4.2 コンコーダンス編纂 (1980年代半ば～)

齊藤・今井

コンコーダンスは語彙・文法・文体など様々なレヴェルで言語研究のための強力なツールとなっている.

4.3 計算機によって引き出されるデータと作品の読み

中川(Wordsworth)，西村 (Lawrence)，Hori (Dickens)，稲木・沖田(アリス小説)

'Through' in *The Prelude* (中川, 1990 & 1997)

V.　最後に～今後の展望・課題

- コーパスや電子テキスト資源の相互利用
- 共通の規格，または互換性を持ったフォーマット
- 共同研究・チームプロジェクト
- データ分析とテキストの読みをいかに結びつけるか