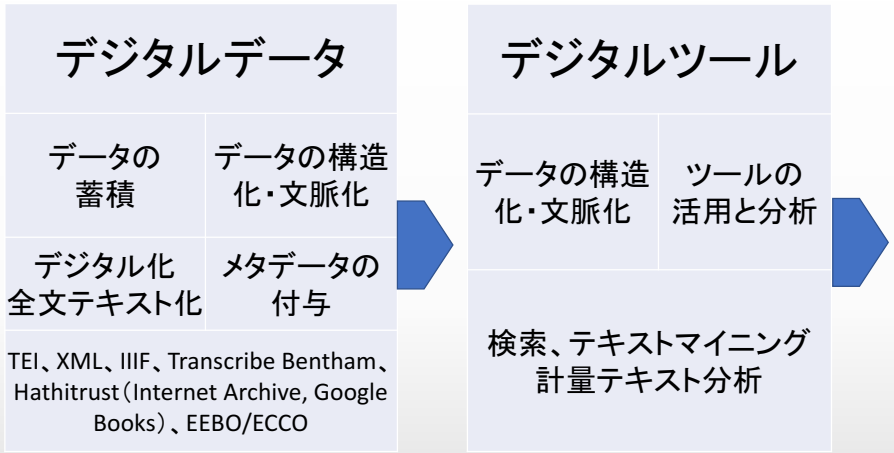
データフィケーションの時代における思想・哲学研究

デジタルデータ、デジタルツール（検索、計量分析）をどう利活用できるか

日本イギリス哲学会　2017年度研究大会（2018.3.28）
シンポジウム1「イギリス哲学研究とデジタル・ヒューマニティーズ――思想史の事例を手がかりに」におけるコメント報告

<http://inuzukah.ws.hosei.ac.jp/inuzuka_20180328_jsbp.pdf>

犬塚　元（法政大学法学部）



データアクセス
デジタル格差
知的所有権

成果の公開
協働、発表
コミュニケーション

3つの結論

1データフィケーションにより生じたのは質的変化
- 対象と手法（材料と道具）における量的・質的変化
（1）大量のデジタルデータ（フラット化・非カノン化）
（2）デジタルツールを用いる半自動化された高速処理
（2a）大量データの検索
（2b）定量的データを集計して統計解析
- とくに（2b）のインパクトが大きい

2素朴な反発・敵視・誤解にとどまる段階でない
- 「浅薄で、乱暴な、非人間的分析」「研究者のアイデアや独創性を否定」「テキストに虚心坦懐に向かって精読あるのみ。方法・方法論の吟味は不要」「問題意識の欠如」
- 量的か質的か、デジタル化は万能か無用か、機械か人間か、はいずれも極端な二元論（擬似区分）

- 万能ではないが有用。自動化できる部分は多いが、研究者の（非定量的な）判断に依存する部分も多い
  - 技術蔑視（方法蔑視）も技術崇拝（方法崇拝）も採用しがたい
- デジタルデータ／ツールをいかに有効に利活用するか、いかに従来の方法と併用するか、検討する段階
  - 思想・哲学研究における「混合研究法」（収斂型・順次型）
  - 日本におけるイギリス哲学・思想研究は、デジタルデータにおける相対的優位性にもかかわらず利活用が遅れている？

3他人事ではない。「見ないふり」はできない
- 自分がデジタルデータ／ツールを用いなくても、それを駆使した研究（その方法と内容）の妥当性を、たとえば査読者・審査員として評価・判定する必要
  - 十分な調査・理解もなくすぐに「成果らしきもの」をつくれる
  - よい研究、悪い研究を慎重に区別する必要→以下の2例

1　キーワード検索

- 検索を中心とする研究はどこまで妥当か？
  - EEBO/ECCOの検索から「初期近代における〇〇」
  - 書籍の索引にたよる「拾い読み」となにが違うか
- 例1）Hume, *History of England* (128万語)を‘prudence’という語の検索から分析する研究
  - 語の出現頻度や関係（単語、*n*-gram、共起語の頻度分析）は、軽快なソフトウエア（コンコーダンサー、コーパス分析ツール）で簡単に検索可能
  - CasualCon 2.0.6　Mac、フリーウエア、大阪大今尾康裕氏
  - AntConc 3.5.2 Windowsほか、ドネーションウェア、早稲田大 Laurence Anthony氏

KWIC（Key Word In Context）コンコーダンス

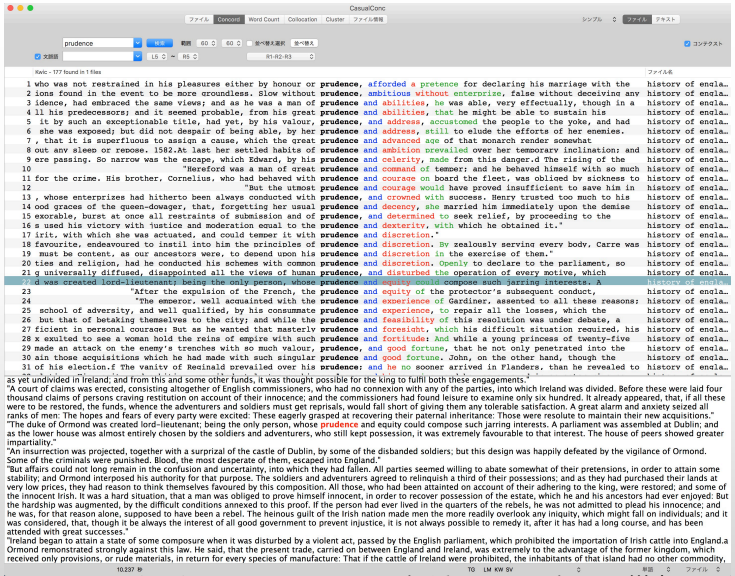


Fig. 1　CasualCon（2.0.6）/Hume, *History of England*における‘prudence’のキーワード検索

検索語の位置（コンコーダンスプロット）

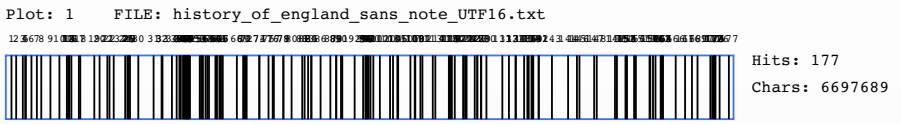


Fig. 2 AntConc（3.5.2）/Hume, *History of England*における‘prudence’の位置

n-gram（単語クラスター）の検索・頻度集計

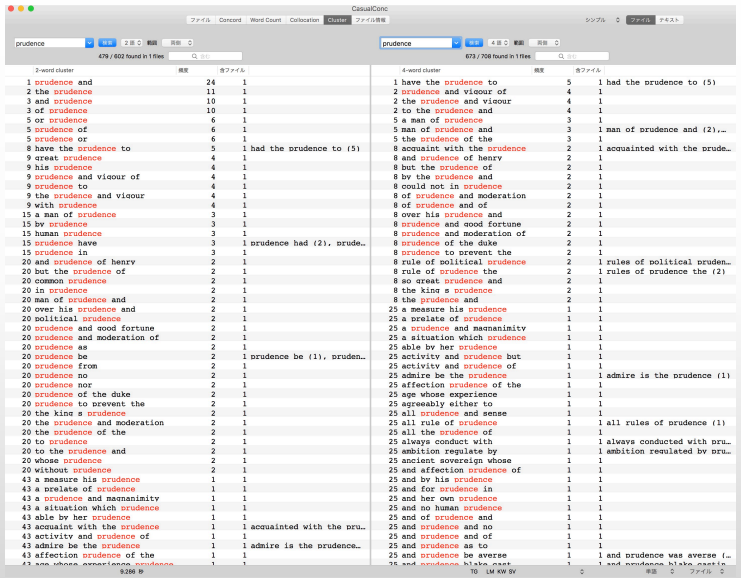


Fig. 6 CasualCon（2.0.6）/Hume, *History of England*における‘prudence’を含む単語バイグラム、フォーグラム

共起（コローケーション）の検索・頻度集計

CasualConc

ファイル　Concord　Word Count　Collocation　Cluster　ファイル情報

prudence　検索　範囲　L5 ~ R5　並べ替え

視覚化　単語

	共起語	左右合計	左合計	右合計	L5	L4	L3	L2	L1	キーワ…	R1	R2	R3	R4	R5	
1	and	130	49	81	9	6	5	4	25	0	65	5	5	3	3	
2	the	127	75	52	4	20	15	5	31	0	2	16	13	12	9	
3	of	112	51	61	9	9	5	8	20	0	19	3	18	11	10	
4	to	44	15	29	3	3	2	5	2	0	13	2	5	7	2	
5	his	42	26	16	2	5	3	6	10	0	2	4	0	8	2	
6	be	38	23	15	11	7	4	1	0	0	4	1	2	7	1	was-b
7	with	36	27	9	5	1	6	8	7	0	1	1	5	1	1	
8	by	35	27	8	4	6	6	6	5	0	1	2	4	1	0	
9	in	28	11	17	3	3	1	1	3	0	6	1	4	4	2	
10	he	26	9	17	3	4	2	0	0	0	1	1	4	3	8	
10	a	26	17	9	8	2	6	0	1	0	0	2	2	3	2	a-pru
12	have	25	15	10	3	0	5	7	0	0	3	1	4	2	0	had-b
12	that	25	13	12	4	2	2	3	2	0	1	4	2	2	3	that-
14	which	18	6	12	0	3	1	1	1	0	1	2	3	3	3	
14	or	18	8	10	1	0	0	0	7	0	8	1	0	1	0	
16	great	16	14	2	1	1	1	0	11	0	0	0	1	0	1	
17	as	15	7	8	1	3	1	1	1	0	6	0	1	0	1	
17	from	15	7	8	2	2	3	0	0	0	2	0	3	1	2	
17	her	15	10	5	2	1	1	3	3	0	0	2	1	0	2	
20	him	14	3	11	1	2	0	0	0	0	0	5	1	1	4	
21	so	12	8	4	1	2	1	4	0	0	0	1	0	1	2	
21	conduct	12	6	6	3	1	0	2	0	0	1	0	2	0	3	condu
23	could	11	4	7	0	2	2	0	0	0	2	0	4	1	0	
23	not	11	7	4	0	2	2	2	1	0	1	1	0	2	0	
23	but	11	7	4	2	2	1	2	0	0	2	0	0	2	0	
26	for	10	6	4	0	0	2	0	4	0	0	1	1	2	0	
26	make	10	1	9	0	1	0	0	0	0	0	1	4	2	2	made-
26	all	10	7	3	1	3	1	1	1	0	0	0	1	1	1	
26	vigour	10	2	8	1	0	0	1	0	0	2	5	1	0	0	
30	moderation	9	3	6	0	1	0	2	0	0	0	6	0	0	0	
30	temper	9	4	5	0	1	1	2	0	0	0	1	1	2	1	tempe
30	man	9	8	1	1	2	1	4	0	0	0	0	1	0	0	man-b
30	this	9	2	7	2	0	0	0	0	0	0	0	2	1	4	this-
34	no	8	3	5	0	1	0	1	1	0	2	2	1	0	0	
34	rule	8	7	1	0	0	2	5	0	0	0	1	0	0	0	rules
34	such	8	4	4	0	1	0	2	1	0	0	0	0	1	3	
34	valour	8	3	5	0	0	0	1	2	0	1	3	0	1	0	
34	without	8	4	4	0	1	1	0	2	0	1	1	1	0	1	
39	king	7	3	4	0	0	1	2	0	0	0	1	2	0	1	
39	it	7	4	3	1	2	0	1	0	0	0	0	1	1	1	
41	human	6	5	1	0	0	1	0	4	0	0	1	0	0	0	
41	much	6	5	1	0	0	2	1	2	0	0	0	1	0	0	
41	want	6	5	1	0	2	1	2	0	0	0	0	1	0	0	wante
41	on	6	4	2	1	0	0	3	0	0	0	0	1	0	1	
41	who	6	3	3	1	1	1	0	0	0	1	1	0	0	1	
41	well	6	5	1	2	2	0	1	0	0	0	1	0	0	0	
47	courage	5	2	3	0	0	1	1	0	0	0	2	1	0	0	

10,237 秒　TG　LM　KW　SV

Fig. 3 CasualCon（2.0.6）/Hume, *History of England*における‘prudence’の共起語の頻度表（コロケーションテーブル）

関連語検索

Search Entry:

#直接入力

コーディングルール・ファイル：　参照　選択

直接入力：　and　prudence

AND検索　集計単位：

Result:

N	抽出語	品詞	全体	共起	Jaccard
1	vigour	Noun	225 (0.042)	27 (0.160)	0.0736
2	conduct	Noun	534 (0.101)	46 (0.272)	0.0700
3	moderation	Noun	104 (0.020)	17 (0.101)	0.0664
4	temper	Noun	247 (0.047)	25 (0.148)	0.0639
5	monarch	Noun	432 (0.081)	36 (0.213)	0.0637
6	situation	Noun	421 (0.079)	35 (0.207)	0.0631
7	fortune	Noun	287 (0.054)	27 (0.160)	0.0629
8	experience	Noun	156 (0.029)	19 (0.112)	0.0621
9	predecessor	Noun	122 (0.023)	17 (0.101)	0.0620
10	dangerous	Adj	499 (0.094)	39 (0.231)	0.0620
11	virtue	Noun	193 (0.036)	21 (0.124)	0.0616
12	ambitious	Adj	113 (0.021)	16 (0.095)	0.0602
13	able	Adj	415 (0.078)	32 (0.189)	0.0580
14	possess	Verb	512 (0.096)	37 (0.219)	0.0575
15	favourite	Noun	171 (0.032)	18 (0.107)	0.0559
16	valour	Noun	210 (0.040)	20 (0.118)	0.0557
17	render	Verb	557 (0.105)	38 (0.225)	0.0552
18	affection	Noun	290 (0.055)	24 (0.142)	0.0552
19	Europe	ProperNoun	292 (0.055)	24 (0.142)	0.0549

コピー　KWIC　ソート：　Jaccard　フィルタ設定　共起ネット　文

単語　ファイル

Fig. 4 KH Coder（3.Alpha.11）/同左
*Jaccard係数（a; b）＝|a∩b| / |a∪b|

共起ネットワーク

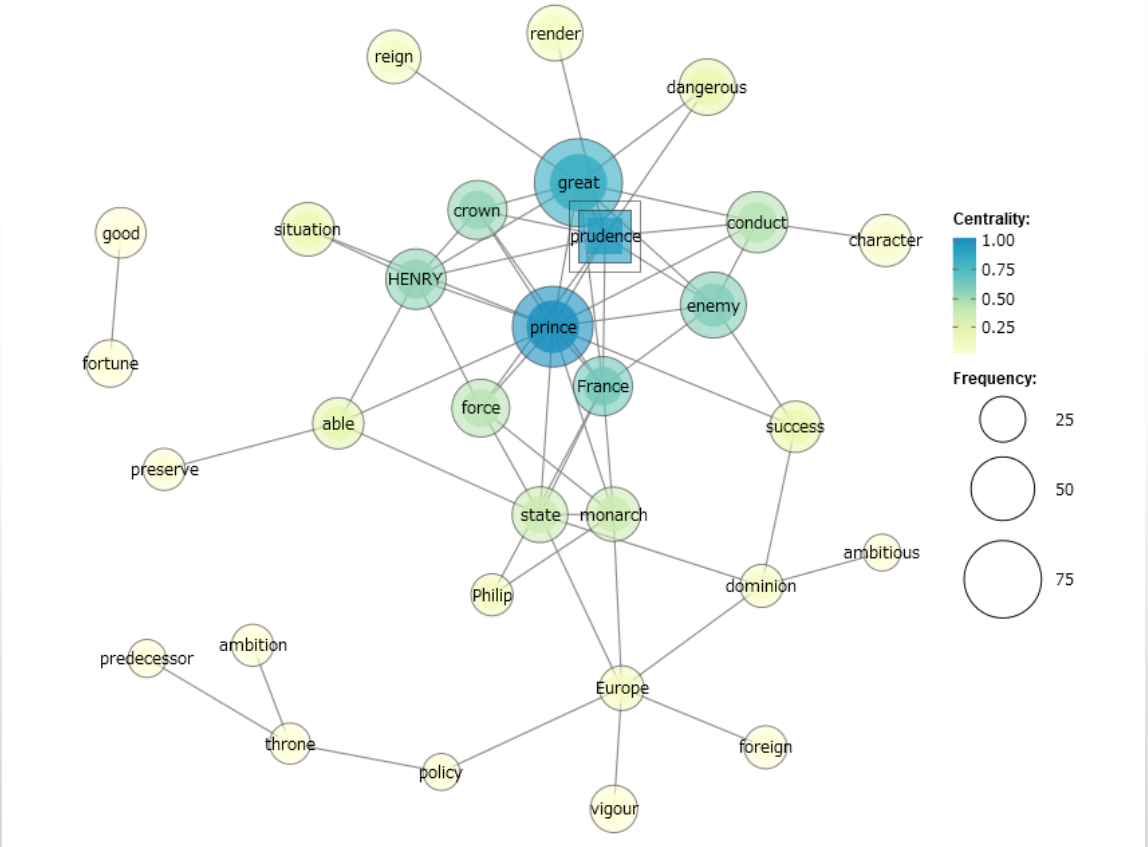


Fig. 5 KH Coder（3.Alpha.11）/Hume, *History of England*における‘prudence’と関連が強い語の共起ネットワーク

コンコーダンスでの気づきにもとづく調査と発見

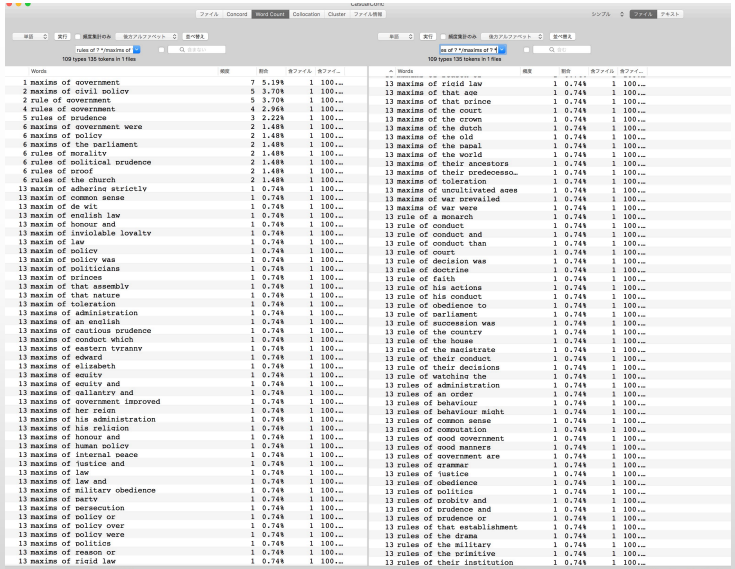


Fig. 7 CasualCon（2.0.6）/Hume, *History of England*における‘(rules|maxims) of ?*’の検索

キーワード検索の長所・短所

- キーワード検索、文脈調査（KWIC検索）、語の関係の調査（*n*-gram、共起語の検索）は、デジタルツールで瞬時に可能
  - 全文入手可能ならコーパス（データベース）の規模を問わない
- 検索は、書籍の索引検索と同じステータス（「非直線的な」読み）
  （1）情報アクセスの補助ツールとして有効。気づきや発見
  （2）しかし、キーワード検索（半自動化された作業）を中心手法とする研究は、学位審査やピアレビューで、望ましくないと判定される可能性が高い（という直感的判断）
- 望ましくないとされる厳密な理由はなにか？独自性の欠如？
  - 検索語の選択や、検索結果の解釈において、むしろオリジナリティが必要（これは説明責任とセット）
  - 網羅性・代表性や信頼性は、コーパスやメタデータに依存
- どういう条件のもとなら、キーワード検索という手法は妥当か？
  - 学位審査や査読での、緩やかな基準は設定できるか？
  - 補助的・副次的な手法？　問題の所在を知る第一ステップ？

２　テキストの計量的分析

- テキストマイニング、計量（統計的）テキスト分析
  - テキストのさまざまな量的データを集計して、それらを統計的に分析
  - 全体構造（言説構造、概念関係）の可視化
  - 人文社会科学でも、言語学・文体学、国文学・漢文学、社会学では珍しくない
- 本格的なソフトウエアの使用
  - 統計計算とグラフィック作成のソフトウエア環境としてR（オープンソース）を使用するのが一般的
  - しかし、たとえば、KH coder（Windowsほか、フリーウェア、立命館大学樋口耕一氏）では、コマンド、関数、統計の知識なしでもGUIで分析可能

計量テキスト分析はなにをしているのか

前処理（テキストの準備）	問われること
電子化（OCRやウェブスクレイピング）、テキスト整形（クリーニング）、メタデータ（タグや外部変数）の付加、文字コードの調整	データの正確性（同一性・再現性）・信頼性、分析対象の設定、メタデータの設定
自然言語処理（テキストの、基本単位への分割）	
形態素解析、補正（語形変化・単複・活用をふまえた名寄せ、複合語の設定）、メタデータ（品詞情報や属性などのタグ、外部変数）の付与、構文解析、除外語の設定	解析の信頼性、どのように分けるか（例：「日本イギリス哲学会」）、どう名寄せするか、辞書整備、メタデータや除外語の設定
頻度集計（出現頻度や関係（同時出現）の集計）	
単語・概念（任意のコーディングルールにもとづいた語彙グループ）・*n*-gram・共起の頻度集計、可視化（ワードクライド、ネットワークグラフなど）	なにに注目するか、どの単位・スパンでなにを分析するか、頻度や強度をどう判定するか、コーディングルールをどう設定するか
統計処理（多変量解析）	
複数データ（集計結果や外部変数）の関連性を一般的な統計的手法で解明：因果・相関の分析（重回帰分析、相関分析）、グループ化（対応分析、クラスター分析）、主成分分析など。可視化	どのような統計処理をしているか、各解析ではどの係数・アルゴリズムを使用しているか

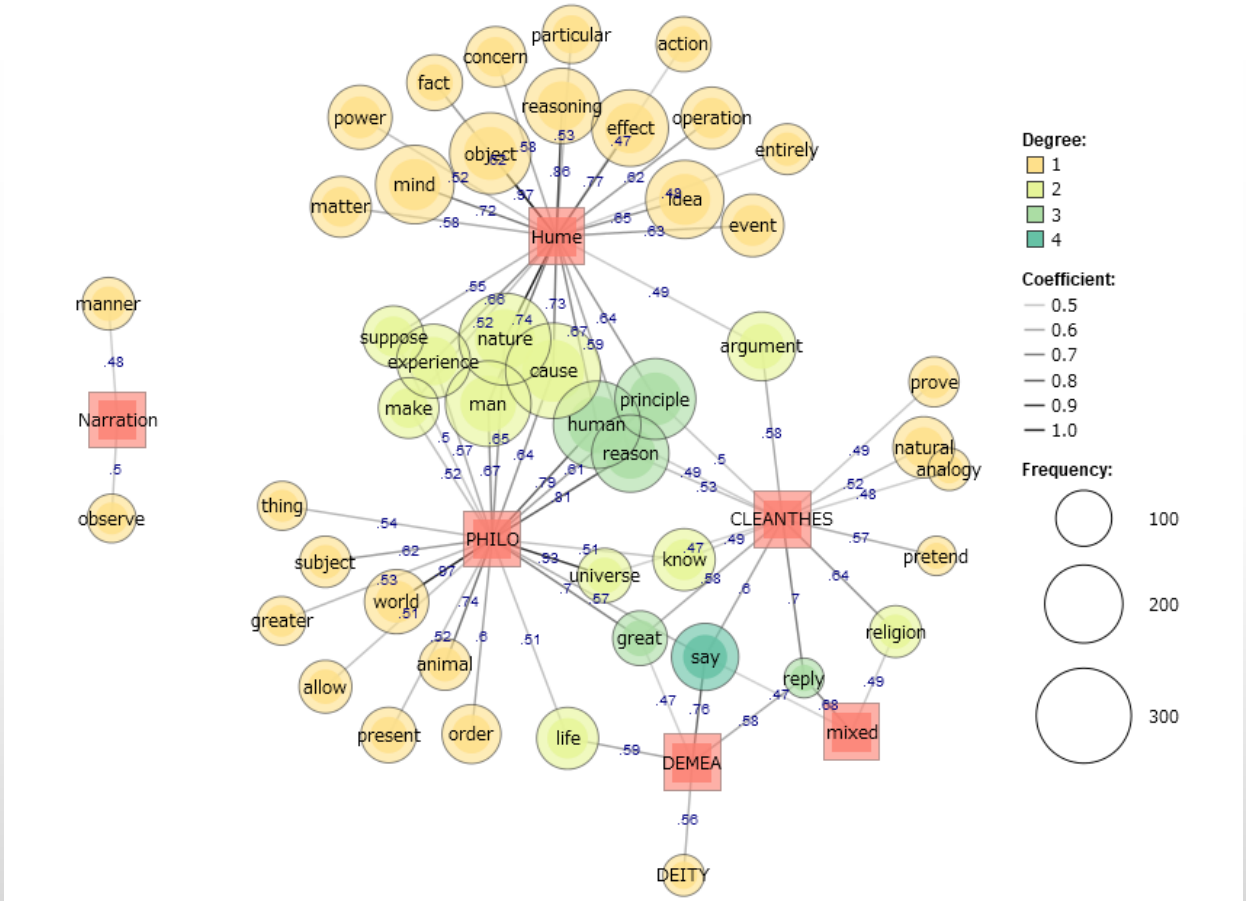


Fig. 8 KH Coder（3.Alpha.11）/Hume, *Dialogues concerning Natural Religion*と*First Enquiry*における、発話主体（外部変数、四角で表記）と特徴語の共起ネットワーク　＊話者の固有名は除外語

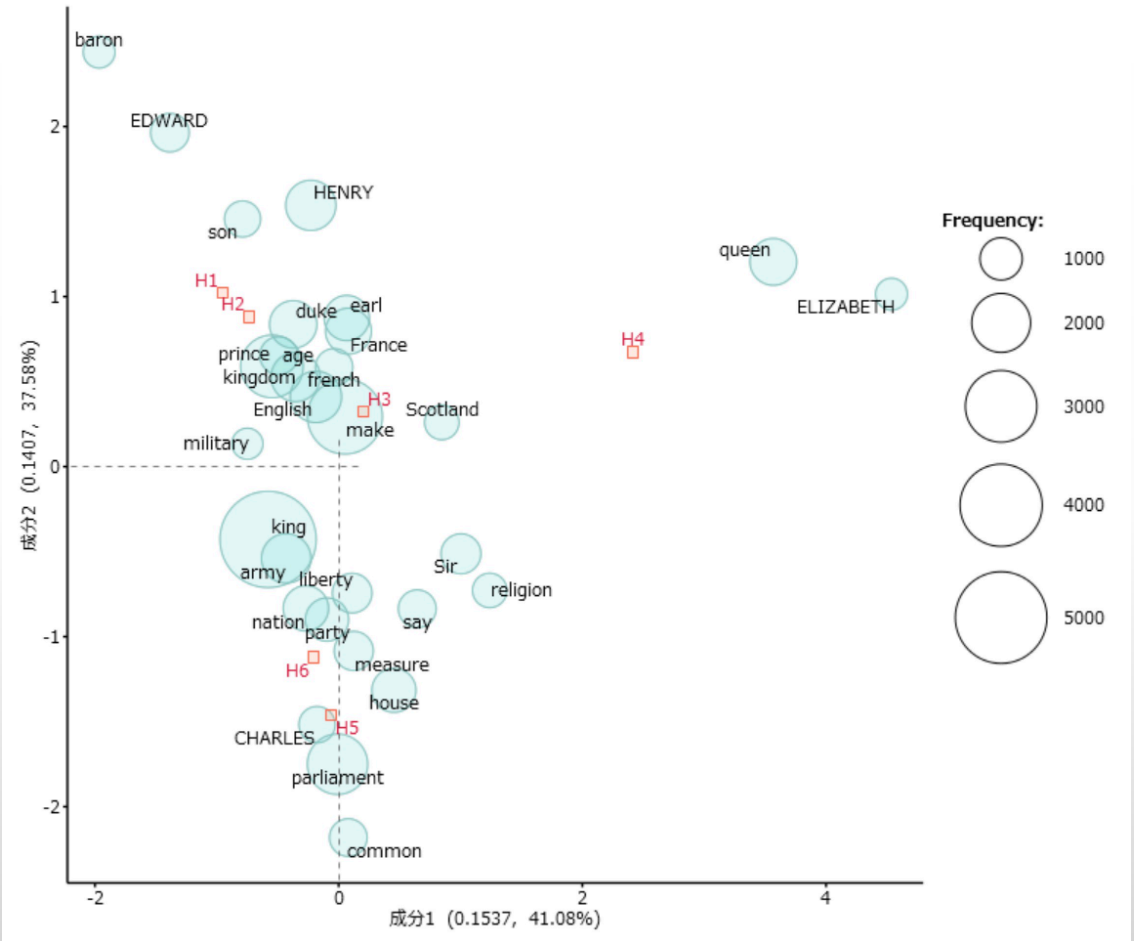


Fig. 9 KH Coder（3.Alpha.11）/Hume, *History of England*の対応分析

- 例2）Hume, *Dialogues concerning Natural Religion* の執筆意図を解明するための計量的分析
  - 「対話篇の3名のうち、だれがヒュームに近いか」
  - 3名の主張の特徴を定量的に測定し、ほかのテキスト（*Enquiry concerning Human Understanding*）と対比
  - 発話主体の情報（3名＋ヒューム）を外部変数として設定して、それぞれの特徴語を析出

  →かなりクリアに関係（H-P、H-P-C）を視覚化（*Fig.*8）
- 例3）Hume, *History of England* 各巻のグループ化
  - 各巻をグループ化するために「対応分析」
  - 2次元の散布図に可視的に表現。頻度パターンの近いテキストや単語は近くに布置される

  →Vols.1-2、5-6の近さ、Vol.4の特異性が明らか（*Fig.*9）
- いずれの例でも、テキストを実際に読解していないと、解析結果の分析・解釈は難しい

計量的分析をどう評価するか

- 簡単に「それらしい結果」を視覚的に提示できる
- 妥当性の評価には、一定の理解や知識が必要
  - 分析者が、理解・判断・選択すべき事項は多い
    - データ（材料）、ツール（道具）、解釈（アウトプット）
    - 具体的手法の詳細（データ、ツール）を明記する必要
  - 評価者はこれらの妥当性を吟味しなければならない
    - どのように追試（確認）できるか
- 言い換えれば、データやコーパス、アルゴリズム（どんな関数・係数か、どんな計算か）が、不明確なもの・理解していないものには注意が必要
  - たとえば、Google n-gramやGoogle Books Ngram Viewerは、どこまで厳密な研究で使ってよいか？
  - エンドユーザーは、テキストの統計情報にアクセスできるだけ。著作権問題のために、コーパスそのものにはアクセスできない（その信頼性は不明）

さいごに 「見ないふり」はできない

- 自分自身が手法を採用しなくても、計量テキスト分析を評価・吟味しなければならない可能性
  - それらは、従来の手法による成果と競合しうる
  - たとえば、*n*-gramや多変量解析によって、執筆時期・順序や書き手や異本系統を特定する研究（近藤2005、金2009a,2009bなど）
- 本学会は、特定の手法・アプローチを排除してない
  - 「本会はイギリス哲学の研究と普及をはかることを目的とする」（会則第2条）
  - 「狭義の哲学専門研究の立場からのみならず、人文・社会諸科学のあらゆる分野からアプローチしてこそ」（WSより）
  - たとえば、投稿論文を査読できる態勢かどうか
- 技術の進展に応じた継続教育プログラムの必要性

参考文献

石田基広［2017］『Rによるテキストマイニング入門（第2版）』森北出版.

今尾康裕［2017］「CasualConcでのアカデミック英語分析－単語検索からデータの視覚化まで－」水本篤 (編)『ICTを活用した英語アカデミック・ライティング指導－支援ツールの開発と実践－』金星堂.

金明哲［2000］「自然言語における統計手法を用いた情報処理」『統計数理』48(2).

金明哲［2009a］『テキストデータの統計科学入門』岩波書店.

金明哲［2009b］「文章の執筆時期の推定」『行動計量学』36(2).

クレスウェル, J.W.（抱井尚子訳）［2017］『早わかり混合研究法』ナカニシヤ出版.

小林雄一郎［2017］『Rによるやさしいテキストマイニング』オーム社.

小峯敦・下平裕之［2017］「ベヴァリッジ『自由社会における完全雇用』のケインズ的要素. ～テキストマイニングを加味した量的・質的分析～」Discussion Paper Series 17-01（龍谷大学）.

近藤みゆき［2005］『古代後期和歌文学の研究』風間書房.

近藤みゆき［2015］『王朝和歌研究の方法』笠間書院.

近藤泰弘・近藤 みゆき［2001］「N-gramの手法による言語テキストの分析方法--現代語対話表現の自動抽出に及ぶ」『漢字文献情報処理研究』2.

永崎研宣［2013］「人文学分野とサイバーインフラストラクチャ：デジタル・ヒューマニティーズにおける現状と課題」『情報の科学と技術』63(9).

永崎研宣［2017］「デジタル文化資料の国際化に向けて：IIIFとTEI」『情報の科学と技術』67(2).

仁平典宏・藤田真文［2017］「特集「テキストマイニングをめぐる方法論とメタ方法論」によせて」『社会学評論』68(3).

樋口耕一［2017］『社会調査のための計量テキスト分析：内容分析の継承と発展を目指して』ナカニシヤ出版

樋口耕一［2017］「計量テキスト分析およびKH coderの利用状況と展望」『社会学評論』68(3).

福田名津子［2016］「「デジタル・ヒューマニティーズ2.0」がもたらす人文・社会科学への影響」『一橋大学附属図書館研究開発室年報』4.

古谷豊［2014］「テキストマイニングを用いたスミス『国富論』普及の分析」Discussion Paper 325（東北大学経済学研究科）.

松村真宏・三浦麻子［2014］『人文・社会科学のためのテキストマイニング（改訂新版）』誠信書房.

美馬秀樹・丹治信・増田勝也・太田晋［2012］「近代文献のデジタルアーカイブ化とテキストマイニング：岩波書店「思想」を題材に」『情報処理学会研究報告』2012-95(4).

山田崇仁［2007］「N-Gram 方式を利用した漢字文献の分析」『立命館白川靜記念東洋文字文化研究所紀要』1.

吉見俊哉［2010］「コンピュータは思想史を書き換えられるか？MIMAサーチによる20 世紀日本の人文知への挑戦」『丸善ライブラリーニュース』10.