文学研究のためのデータサイエンス：女流作家の文体的特徴について

お茶の水女子大学文理融合AI・データサイエンスセンター　土山 玄

数理・データサイエンス・AI教育強化拠点コンソーシアム2023年度関東ブロック第3回ワークショップ
「女子大学におけるデータサイエンス教育事例」（2023年9月1日）

本日の構成

- **お茶の水女子大学について**

- **データサイエンス科目**
  - 文理融合データサイエンス
  - 共創工学部の必修科目群
  - 人文学領域のデータを採り上げるメリット

- **男性作家と女性作家**
  - 欧米圏の研究
  - 平安時代の女流作家の特徴

- **結びにかえて**

お茶の水女子大学について

- **学部**
  - 文教育学部
  - 理学部
  - 生活科学部
  - **共創工学部**（2024年度設置）
    - ✓ 人間環境工学科
    - ✓ 文化情報工学科
    - データサイエンスが**必修**
  - 3学部合わせて、1学年500人弱
    ➡4学部体制になっても定員増はない

▸ データサイエンス科目

- **文理融合データサイエンスⅠおよびⅡ**
  - 2019年後期から開講
  - **人文学領域のデータ**を題材に、データサイエンスの手法を学ぶ
    ➡文学作品のテキストデータなど
  - **全学データサイエンス学際カリキュラム** の必修科目
    - ✓ データサイエンスに関する開講科目を体系化した科目群
    - ✓ データサイエンスに関心を持つ学生が学部・学科を超えて学際的・系統的に履修することを目的とする
    - ✓ 2021年に **MDASHリテラシーレベル** の認定

  ▸ 受講生数
    - ✓ 文理融合データサイエンスⅠ：70人程度
    - ✓ 文理融合データサイエンスⅡ：35人程度

▸ データサイエンス科目

- **共創工学部の必修科目**
  - ✓ データサイエンス（基礎）
  - ✓ データサイエンス（中級）
  - ✓ データサイエンス（上級）
  - ✓ 機械学習

  - 2022年後期から先行開講

  - ▸ 受講生数
    - ✓ データサイエンス（基礎）：30人程度
    - ✓ データサイエンス（中級）：10人程度

データサイエンス科目

- **人文学領域のデータを採り上げるメリット**
  - データサイエンスガイダンス
    - **1年生全員**に対して実施
    - 文理融合データサイエンスの告知
    - 1時間程度の**模擬授業**
      - ✓ 夏目漱石の小説の分析
      - ✓ ルネサンス期のフィレンツェのネットワーク
  - ガイダンス受講生の反応
    - ガイダンス後に受講生へアンケートを実施
    - **文系の学生への訴求力が高い**
    - 理系の学生の反応が悪い訳ではない

文理融合データサイエンスとは

- **文章の計量分析**
  - ジョルジュ＝ルイ・ルクレール・ド・ビュッフォン（1707-1788）
    - 文は人なり

  - オーガスタス・ド・モルガン（1806-1871）
    - **2人の人間が同じ主題**について記述した２つの文章よりも、
      1人の人間が異なる主題について記述した２つの文章の方が
      より**類似度が高い**、ということが明らかになることが期待される

  - ピエール・ギロー（1912-1983）
    - **文体とは標準からの逸脱**
    - 一般的規範からの個人的な文体的**偏差**

文理融合データサイエンスとは

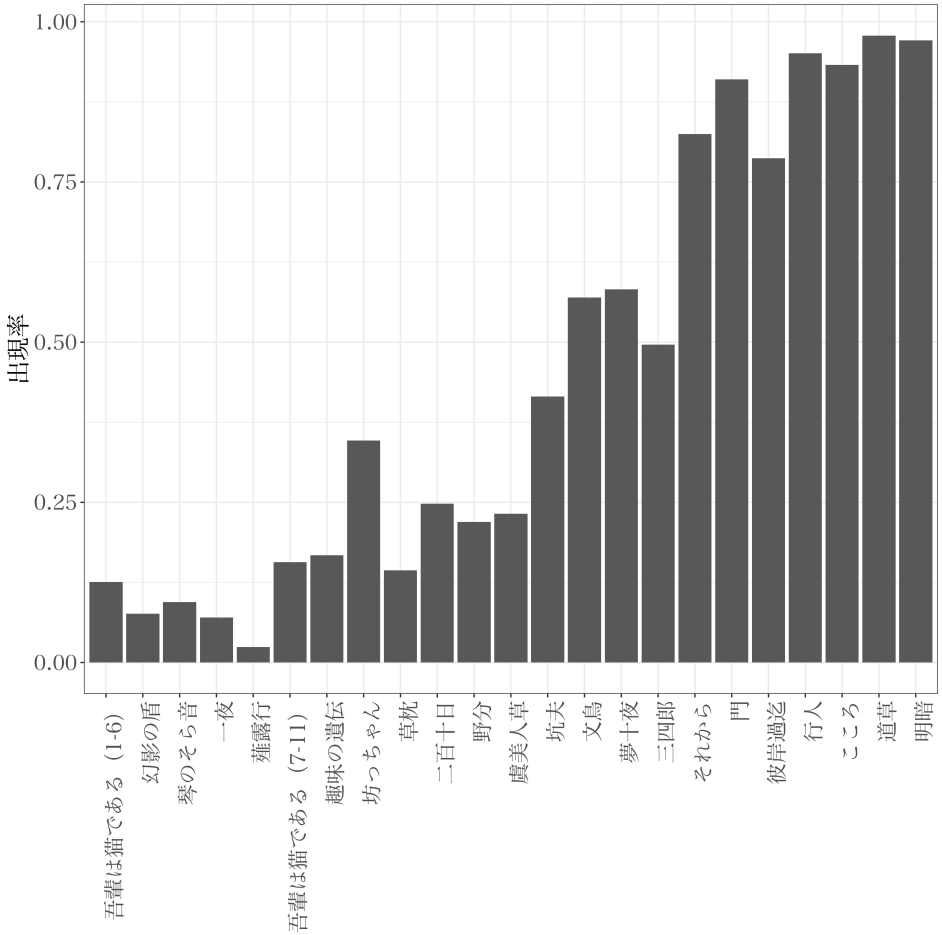
- **夏目漱石の小説における文体的特徴の変化**
  - 対象
    - 夏目漱石の小説22作品
  - データ
    - ✓ 青空文庫
      https://www.aozora.gr.jp/
  - 形態素解析
    - 品詞分解と品詞のタグ付け
    - ✓ Web茶まめ（国立国語研究所）
      https://chamame.ninjal.ac.jp/

タイトル	発表時期
吾輩は猫である（一〜六）	1905年1月
幻影の盾	1905年4月
琴のそら音	1905年5月
一夜	1905年9月
薤露行	1905年11月
吾輩は猫である（七〜十一）	1906年1月
趣味の遺伝	1906年1月
坊っちゃん	1906年4月
草枕	1906年9月
二百十日	1906年10月
野分	1907年1月
虞美人草	1907年6月23日〜1907年10月29日
坑夫	1908年1月1日〜1908年4月6日
文鳥	1908年6月
夢十夜	1908年7月
三四郎	1908年9月1日〜1908年12月29日
それから	1909年5月31日〜1909年8月14日
門	1910年3月1日〜1910年6月12日
彼岸過迄	1912年1月1日〜1912年4月29日
行人	1912年12月6日〜1913年11月15日
こころ	1914年4月20日〜1914年8月11日
道草	1915年6月3日〜1915年9月14日
明暗	1916年5月26日〜1916年12月14日

文理融合データサイエンスとは

- **夏目漱石の小説における文体的特徴の変化**
  - 考察
    - 文末表現は1908年頃に計量的な傾向の変化が認められる
    - 文末に助動詞の使用が増加
      ➡特に「た」の使用が増加
    - 文末の「た」が増加傾向ならば
      ➡**発表年を予測するモデル**を構築できるのではないか？
    - 回帰分析

      出版年 $= 9.1960 \times$ **た** $+1904.1334$



男性作家と女性作家

- **欧米圏での研究**
  - 対象
    - ✓ 古典文学作品や大衆小説など200作品

  - **男性**作家の特徴
    - she より he を多く使用する
    - **she kissed** という表現が**女性作家により多い**

  - **女性**作家の特徴
    - he より she を多く使用する
    - **he kissed** という表現が**男性作家により多い**

  - 出典

    ベン・ブラット『数字が明かす小説の秘密』, DU BOOKS, 2018.

男性作家と女性作家

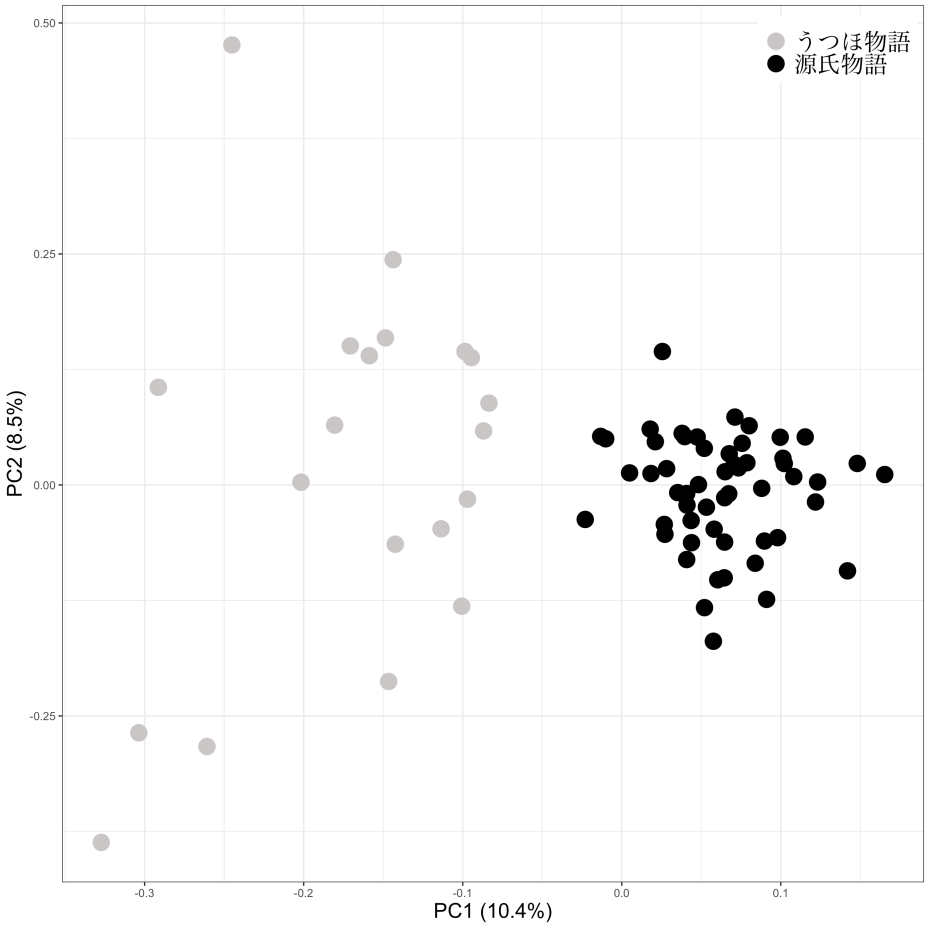
- **平安時代の女流作家の特徴**
  - 対象
    - ✓ 源氏物語
    - ✓ うつほ物語

  - 源氏物語
    - 平安時代に成立した**和文体**の長編物語
    - 紫式部によって執筆とされる
    - 全54巻（延べ語数はおよそ38万語）

  - うつほ物語
    - 平安時代に成立した現存最古に類する長編物語
    - **和文体**の物語
    - 作者は後撰集の撰者、源順（みなもとのしたごう）か？
    - 全20巻（延べ語数はおよそ25万語）

男性作家と女性作家

- **平安時代の女流作家の特徴**
  - 考え方
    - 物語の内容の影響を受けない要素を用いる
      - ✓ 助詞
      - ✓ 助動詞
  - 著者の識別
    - 助詞の出現率
    - 主成分分析



男性作家と女性作家

- **平安時代の女流作家の特徴**
  - 女流作家の特徴の抽出
    - カイ二乗値を用いる

	源氏物語	うつほ物語	行和
単語Aの頻度	13	7	20
それ以外の頻度	987	993	1980
列和	1000	1000	2000

    - カイ二乗値が大きければ、物語間で出現傾向が異なる

  - 結果
    - 漢文訓読文体の表現が少ない
      ➡助動詞の直喩表現「ごとし」など
    - 様態をあらわす単語の出現率が高い

結びにかえて

- **期待される効果**
  - 興味の喚起
    - 少し検索しただけでは分析事例が見つからない
      ➡データサイエンスへの興味が知的好奇心に変わることを期待
    - 複数の分析手法を使うことで、知識を連結させる
      ➡研究へのモチベーション向上を期待
  - 実践力の涵養
    - 例示と演習を繰り返す
    - 演習ではRを使用
    - 期末課題で実践
      - ✓ データを配布
      - ✓ 受講生自らが課題を設定し、分析する
      - ✓ 分析結果をレポートにまとめて提出