お茶の水女子大学　共創工学部 文化情報工学科　　土山 玄 研究室

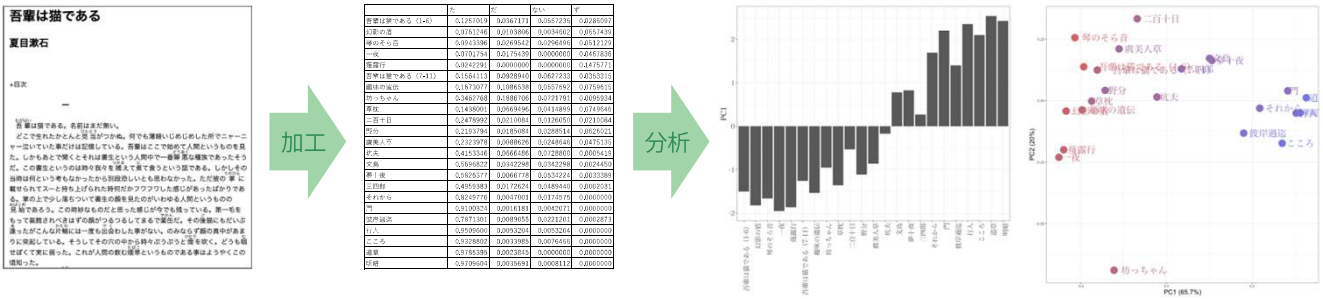
データサイエンスによる文学作品を対象とした計量分析：文体を分析することで書き手を推定する

研究室の目標

文化情報　統計学　文学

文学作品のテキストデータの作成から分析までのプロセスを学ぶ

テキストデータの取得　→（加工）→　単語の出現率データ　→（分析）→　分析結果の可視化



　文体には個性があらわれると考えられています。日本語の文章では助詞や助動詞の出現傾向、句読点などの用法に、英語の文章では接続詞や前置詞の出現傾向に個性、すなわち書き手の習慣的かつ形式的な特徴があらわれます。そこで、文章の計量分析では、まずテキストデータに対して形態素解析などの処理を行い、単語の出現率データを作成します。次いで、このようなデータを対象にデータサイエンスの手法を用い、計量的な分析を行います。
　文学的な文章を対象とした計量的な研究における目的は、①著者の識別、②文献の成立過程の推定の2つに大別されます。いずれの研究目的においても、書き手の文体的特徴を分析することで、その出現傾向の相違、あるいは変化を捉えることで、研究課題の解明を試みます。

研究事例

著者の識別　古典文学

源氏物語の作者は1人か?複数か?

　ここでは著者の識別に関連した研究事例を紹介します。
　平安時代に紫式部によって著された長編物語である『源氏物語』には、鎌倉時代から他作者説が提起されています。これは「宇治十帖」と称される『源氏物語』の最後の10巻が紫式部の手によるものではないという見解です。
　この他作者説について計量的な検討を行うために、まず現代文と同様に古典文においても助詞や助動詞の出現傾向が著者の識別に有効であるか確かめる必要があります。そのため、『源氏物語』とおよそ同時代に成立した和文体の長編物語である『うつほ物語』を採り上げます。主成分分析というデータサイエンスの手法を使用して、助詞の出現率を分析し、『源氏物語』と『うつほ物語』の作者の相違を識別できるのか検証します。その結果が図1の散布図です。『源氏物語』の諸巻は図中の右に、『うつほ物語』の諸巻は左にまとまって位置していることが分かります。これは古典文においても助詞の出現率を分析することで、著者の識別ができるということです。
　そこで、『源氏物語』における他作者説を検討するために、『源氏物語』の諸巻を対象に主成分分析を行います。その結果が図2になります。図1とは異なり、図2では「宇治十帖」の10巻が他の諸巻から独立して付置されません。したがって、計量的な判断に基づくと「宇治十帖」における他作者説を支持する積極的な根拠は認められません。よって、『源氏物語』は単独の作者によって執筆された可能性が高いと言えます。

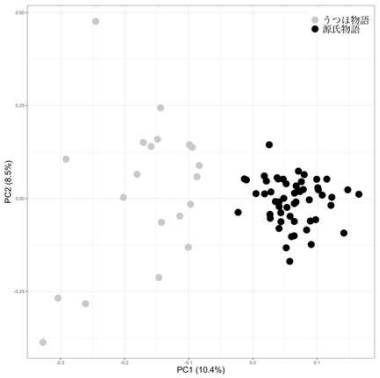


図1　『源氏物語』と『うつほ物語』

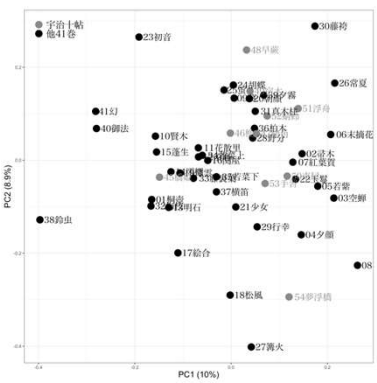


図2　宇治十帖他作者説の検討