# 源氏物語のテキストマイニングと特徴抽

細　井　尚　子[a)]

# Text Mining and Extraction of Special Features in the Tale of Genji

**by Hisako HOSOI**

本論文は，全４章で構成されている。第１章は序論であり，本研究の背景と目的および本論文の構成についてまとめた。第２章は源氏物語の概要や特長について文学的な観点からの考察を行い，過去の研究や現在の課題についてまとめた。第3章は源氏物語の助動詞を対象として，テキストマイニングによる助動詞の抽出および数量化理論解析とクラスタリングによる特徴抽出を行なった結果である。最後に第４章は本論文の結論を述べている。

第１章は，源氏物語を始めとする古典文学作品に対する文学的あるいは科学的なアプローチによる研究の背景や目的について詳しく説明した。特に源氏物語は日本の古典文学の最高傑作であるといわれており，過去1000年に渡って研究が行なわれていることから，それらの研究成果や現在の課題についてまとめた。また，現在でも明らかになっていない源氏物語の謎について述べるとともに，近年行なわれている科学的なアプローチを用いた研究についてもまとめ，それらの研究の課題と本研究での目的について記述している。

第２章では，源氏物語の概要や文学作品としての特徴をまとめ，文学上の謎として考えられている作者や物語の成立の過程に関する文学的な研究について調査し，その考察を行なっている。その結果，源氏物語と紫式部日記の内容から両作品の著者が同一人物であり，源氏物語の作者は紫式部だと考えられるが，54帖全てを紫式部が執筆したかどうかは未解明であることを述べた。また，源氏物語の54帖は大きく３つのグループに分類でき，その中の第１部（１帖から33帖）は紫の上系と玉鬘系に分類される。それぞれの系の登場人物や各帖の巻末と巻頭の接続を考察すると，初めに紫の上系が執筆され，後に玉鬘系が挿入されたものと考えられることを説明した。しかしながら，文学上では作者や成立の謎に関する結論を出すことが困難であることから，作者による助動詞の使用方法の違いに注目し，特に助動詞の「き」や「けり」を例として，源氏物語での使用例から作者の特徴抽出を行なう方法について考察した。

第３章では，源氏物語の助動詞の出現数を対象として，プログラムを用いたテキストマイニングによる助動詞の検出，数量化理論３類を用いた統計解析，クラスタリングを用いた各帖のグループ化による特徴抽出を行なった。その結果，プログラムによる助動詞の自動検出では，アルファベットで４文字以上の単語については通常の検索で，２文字以下の単語については助動詞の活用形や前後の単語を用いた判別を行なうことで，ほとんどの単語はRMS誤差15%以下での検出が可能であった．数量化理論３類およびクラスタリングによる解析では，「紫の上系」と「玉鬘系」の助動詞の出現頻度の違いを定量的に示すことができ，２つのグループでは助動詞の使われ方に明確な差があることが確認され，クラスタリングによる任意性を排除した判定でも「紫の上系」と「玉鬘系」が明確に分けられることが明らかとなった。源氏物語を基にした評価により，これらの特徴

[a)] 新潟大学大学院自然科学研究科

〔新潟大学博士（学術）　平成27年９月24日授与〕

抽出では，助動詞の出現率の誤差はRMS誤差で15%以下に抑える必要があることが示された。また，助動詞の「けり」の出現数と文の長さに注目すると，「紫の上系」と「玉鬘系」とでは使用頻度や文の長さの点で使われ方に差があることが明らかとなった。

第4章は本論文の結論であり，本研究で得られた知見をまとめている。