中古仮名文学作品のコーパスに対する話者情報の付与とその活用例

竹内 綾乃・中村 壮範・小木曽 智信（人間文化研究機構　国立国語研究所）

概要：国立国語研究所が開発および公開を行っている『日本語歴史コーパス』（CHJ）に収録されている『CHJ 平安時代編 I 仮名文学』に新たな話者情報データが追加された．従来のデータのように一部の作品の一部の発話に対してのみ付与されていた話者情報とは異なり，今回追加された話者情報データは，『CHJ 平安時代編 I 仮名文学』に収録されている 16 作品中，古今和歌集を除く 15 作品の登場人物のことばに網羅的に付与されている．新たな話者情報の構築では，登場人物ごとに固有の識別子などの情報を付与することにより，源氏物語のような長編作品においても特定の登場人物のことばを網羅的に集計・分析を行うことを可能にした．また，本研究では，この新たな話者情報を使用した研究の一例として，源氏物語データを使用しコレスポンデンス分析を行った．その結果，出家・在家そして男・女のことばにおける感情形容詞の使用に一定の傾向があることがわかった．

キーワード：コーパス，アノテーション，平安文学資料，話者情報

New speaker information added to Heian literary works in the Corpus of Historical Japanese and its case study

Ayano Takeuchi / Takenori Nakamura/ Toshinobu Ogiso
(National Institute for Japanese Language and Linguistics)

Abstract: The National Institute for Japanese Language and Linguistics has released new speaker information for 15 out of 16 kana literary works of the Heian Period Series within the Corpus of Historical Japanese. The previous version of the corpus also contained the speaker information, yet it was only provided to a limited portion of characters' words in a few of the kana literary works. This new speaker information, however, has been added to characters' words including utterances, inner speech and short poems throughout each work, which makes it possible to thoroughly investigate characters' words. In this research, utilizing the new speaker information, we investigate characters' words in *The Tale of Genji* by using correspondence analysis, which shows relative tendencies in the use of emotive adjectives among characters in the tale.
Keywords: corpus, annotation, Heian literary works, speaker information

１．はじめに

『日本語歴史コーパス』（CHJ）は，国立国語研究所が開発・公開を行っている日本語史研究のためのコーパスである．現在，本コーパスには奈良時代編から明治・大正時代編まで様々なテキストが収録されており，ウェブアプリケーション「中納言」を通して公開されている．『CHJ 平安時代編 I 仮名文学』[1]は，その中でも早い時期に公開されたコーパスであり，現在，16 作品，短単位 101.3 万語が収録されている[2].

従来公開されている『CHJ 平安時代編 I 仮名文学』における全てのテキストには形態論情報が付与されており，さらに，一部の作品には小学館「新編　日本古典文学全集」（以下，新編全集）に基づき形態論情報の他に話者情報（発話に対する登場人物名）も付与されている．しかし，この話者情報は一部の作品における一部の発話に限られており，網羅的な研究を行うために用いるには不十分な状態であった．そこで，古今和歌集を除いた平安仮名文学 15 作品に対し，研究に耐え得る十分な話者情報を新たに追加することになった．本発表では，この新たに追加された話者情報についての報告を行う．また，新たに追加された話者情報を利用した研究の可能性として，『源氏物語』における話者情報を使用しコレスポンデンス分析を行った結果についても報告する．

２．従来の話者情報と問題点

これまで公開されていた『CHJ 平安時代編 I 仮名文学』データにおいても話者名は付与されていた．しかし，これは底本である新編全集の本文中において小書きで表示されている話者名であり，源氏物語や竹取物語などの一部の作品に限られていた．さらに，話者名が付与されている作品においても，網羅的な研究に用いるには 2 つの問題があった．1 つ目は「全ての発話に対して話者情報が付与されているわけではない」という問題，そして 2 つ目は「『源氏物語』や『落窪物語』などの長編作品では，付与されている話者名が官職の変更などに伴って変遷している」という問題で

ある．その一例として，源氏物語において光源氏のライバルとして登場し，一般的に「頭中将」として知られている登場人物の呼称の変遷の一部を下記に示す[a].

表 1 ：頭中将の呼称の変遷

帖	巻名	作中呼称
2	帚木	宮腹の中将、中将、君、頭の君
4	夕顔	君、中将、頭の君、中将殿
9	葵	三位中将、中将、中将の君
17	絵合	権中納言、中納言
21	少女	右大将、大将、殿、父大臣、大臣
34	若菜上	太政大臣、父大臣、大臣

表 1 からわかる通り，物語の進行とともに官職が変化し，それに合わせて作中における「頭中将」の呼称も変化していることがわかる．

そこで，このような問題に対応し，新たな話者情報の付与を行った．対象となった作品は『CHJ 平安時代編 I 仮名文学』に収録されている 16 作品中の古今和歌集を除く 15 作品であり，新編全集において括弧で括られている本文および歌を中心に新たな話者情報の付与を行った[b]. 新たな話者情報付与の対象となったデータは延べ語数 344841 語，異なり語数 7715 語である[c]．作品ごとの内訳は下記の通りである．

表 2 ：作品ごとのデータの内訳

作品名	延べ語数	異なり語数
土佐日記	1640	465
竹取物語	4819	783
伊勢物語	4584	908
落窪物語	33390	1887
大和物語	7662	1141
枕草子	15593	1676
源氏物語	151199	4487
紫式部日記	1300	469
和泉式部日記	6745	831
平中物語	4713	837
堤中納言物語	6665	1012
更級日記	5118	925
讃岐典侍日記	4366	794
蜻蛉日記	17471	1753
大鏡	79576	4115
合計	344841	

表 2 からわかるように，新たな話者情報が追加されたデータにおいて一番データ数が大きい作品は源氏物語（約 43%）であり，次に大鏡（約 23%），そして落窪物語（約 9%）が続き，この 3 作品が平安仮名文学における話者情報の約 75%を占めている．このことから，残り 12 作品の占める割合は非常に小さいことがわかる．

今回のアップデートでは，上記データを対象に登場人物のことばの研究を行う上で十分な情報を与えるために，「登場人物のことばを表出の形式で分類すること」そして「官職の変更などに伴う登場人物の名称の変化に左右されず，作品を通して一人の登場人物のことばを全て追うことができるようにすること」の 2 点を中心に情報の追加を行った．

３． 新たな話者情報の付与

今回のアップデートで追加された話者情報は，大きく分けて 2 種類ある．1 つ目は，登場人物のことばとして分類されたテキストをその表出方法に基づいて分類した情報である．従来の『CHJ 平安時代編 I 仮名文学』データには，テキストの役割に基づいて分類された「本文種別」というタグがあり，テキストを「和歌」「会話」「手紙」「地の文」の 4 種類に分類していた．今回のアップデートではテキストの役割ではなく，「登場人物によってどのように表出されたことばなのか」という点に焦点をあて，登場人物のことばに分類されたテキストを「会話」「心内文」「手紙」の 3 種類に分類した．「会話」は声に出して表出したことば，「心内文」は心で思ったことば，「手紙」は書きしたためたことばである[d].

2 つ目は，各作品における登場人物のことばを網羅するための情報で，「話者名」「作中呼称」「性別」の 3 つタグを付与した．「話者名」は，それぞれの作品における話者を同定するための ID のような役割を果たす情報であり[e]，源氏物語や落窪物語のような長編作品において物語の進行とともに登場人物の官職が変更しても一貫して特定の登場人物のことばを追うことができるように付与した情報である．この「話者名」を見ることにより，前述した「頭中将」のように物語の進行とともに呼び名が変化する登場人物や，「大君」や「女三の宮」など同じ呼び名を持つ複数の登場人物を区別し[f]，各登場人物のことばを

a) ここで示した呼称は，新編全集における巻末付録「各巻の系図」において巻毎に示されている呼称を表記している．
b) 本文において括弧で括られている場合でも，引用など発話以外のことばであると認定された場合には話者情報の付与は行われていない．
c) 補助記号を除き，語彙素で集計を行った．
d) 「手紙」に分類されたことばには，扇や火鉢の灰などに書かれたことばなども含まれており，必ずしも手紙として贈られたことばだけではない．
e) 「話者名」は，基本的に新編全集の各作品の巻末に収録されている系図に使用されている名前を採用している．しかし，源氏物語は各巻末に収録されている「各巻の系図」ではなく，新編全集『源氏物語 6』に収録されている付録「源氏物語作中人物索引」に使用されている名前を採用している[3].
f) 新編全集『源氏物語 6』の巻末付録「源氏物語作中人物索引」には，本文テキストにおいて「大君」とい

集計・分析することが可能になった．これに対して「作中呼称」とは，作中においておのおのの登場人物に使用される様々な呼び方である．「話者名」では作品のある時点においてある登場人物がどのようなに呼ばれているかを知ることはできない．しかし，「作中呼称」を参照することで，物語において登場人物がどのような名前で呼ばれているかを知ることができる．

このような話者情報を付与するにあたり，主要登場人物以外の登場人物のことばに対しても本文や訳文，頭注を参照し，可能な限り話者情報を付与した．個別の名前のない「女房」「女童」「使い」のような登場人物に対しては「話者名」には一般的な名称を，「作中呼称」には本文においてその登場人物を示すことばなどが使用されている場合にはそのことばを付与した．また，1 つの発話に対して複数の登場人物が話者候補となり本文，訳文，頭注からも判断がつかない場合には，「話者名」に話者として可能性のある複数の登場人物の名前を列挙し，できる限り「話者名」が不明とならないよう努めた[g].

今回付与した新たな話者情報は，『『日本語歴史コーパス』「平安時代編 I 」拡張話者情報データ ver.1.0』[4]としてコーパスに対するアノテーションとして表形式のデータでリポジトリ公開されている[h]. 拡張話者情報データにおける「開始サンプル ID」もしくは「作品名」で作品または作品中の巻を特定し，「出現書字開始位置」および「出現形終了位置」をみることで『CHJ 平安時代編 I 仮名文学』における話者情報該当位置を特定することができる．また本データは，『日本語歴史コーパス』の検索アプリケーション「中納言」（https://chunagon.ninjal.ac.jp/）上の検索結果にも反映されている．

語彙素読み	語彙素	語形	品詞	活用型	活用形	原文文字列	本文種別	話者	巻名等
イデ	いで	イデ	感動詞-一般			いで	会話/手紙	末摘花-女	玉鬘
イデ	いで	イデ	感動詞-一般			いで	会話	紫の上-女	胡蝶

図 1：新たな話者情報の「中納言」上での表示例

図 1 は，中納言上での検索結果における話者情報の表示例を示したものである．四角で囲われている部分に新たな話者情報が表示される．「本文種別」の列には「どのように表出されたことばなのか」が表示される．図 1 の上の例を見ると，「本文種別」に会話と手紙が表示されているが，これは従来のデータにおける本文種別と今回新たに付与された表出方法が同じ列に表示されているためである．従来のデータにおいて「本文種別」の情報がある場合にはその情報が先に表示され，今回新たに追加された表出方法はスラッシュの後に表示される．次に，「話者」の列には「話者名」と「性別」がハイフンでつながれて表示される．「作中呼称」については，話者が一人に特定できず複数人の話者候補の登場人物名が「話者名」に登録されている場合などには非常に長くなるため，「中納言」上の検索結果には表示されない．そのため，「作中呼称」を確認するためには，『『日本語歴史コーパス』「平安時代編 I 」拡張話者情報データ ver.1.0』を用いて確認する必要がある．

このように，各作品においてそれぞれの登場人物のことばに対して異なる観点から話者情報を付与することで，従来の話者情報とは違い，おのおのの登場人物のことばを網羅的に分析や集計を行うことができるようになったと同時に，巻や場面において当該登場人物がどのように呼ばれているかの把握も可能になった．

4．研究への応用可能性

新たな話者情報データを活用した研究の可能性を探るため，一例として源氏物語における話者

う名前で呼ばれている登場人物が 3 人，「女三の宮」という名前で呼ばれる登場人物が 3 人掲載されている．これは同じ名前で呼ばれる登場人物の 1 例に過ぎず，ひとつの呼び名が 4，5 人の登場人物に使用されている場合もある．源氏物語においては，このようにひとつの呼び名を共有している登場人物は多く存在する．

g）新たな話者情報の付与において本文，訳文，頭注を参照しても話者を特定できず「話者名」が不明となった語は，異なり語数で 1109 語存在する．

h）本データには，「作品章段の ID」「発話開始位置（先頭からの文字数）」「発話終了位置（同）」「作品名」「底本」「開始ページ」「話者名」「作中呼称」「性別」「表出型」「開始発話」「終了発話」「本文種別」「開始 URL」が付与されている．

情報データを利用し，コレスポンデンス分析を行う．物語作品における登場人物は，様々な場面において会話，心内文，和歌などの表現方法を通して各々の感情を表出している．このような感情の表出において，源氏物語の登場人物がどのような感情形容詞を使用しているのかを，男・女そして出家・在家の観点から分析する．コレスポンデンス分析とは，『データ表の行や列に含まれる情報を少数の成分（コレスポンデンス分析では次元[dimension]と呼ぶ）に圧縮し，それらの関係を散布図上に布置することで，視覚的なデータの俯瞰を可能にする』[5]．コレスポンデンス分析はデータの質の違いを問題としないため，質的データの定量的分析に使用することができる．　そこで，本研究では，出家した男性登場人物（出家男），出家した女性登場人物（出家女），出家していない男性登場人物（在家男），出家していない女性登場人物（在家女）の 4 つのカテゴリーにおける感情形容詞の出現頻度をもとに感情形容詞とカテゴリー間の関係を可視化することにより分析する．

使用したデータは，『CHJ 平安時代編 I 仮名文学』に収録されている源氏物語に付与された新たに話者情報データである．源氏物語における話者情報データは表 2 で示した通り，延べ語数 151199 語，異なり語数 4487 語である．このデータにおける各カテゴリーの延べ語数の内訳は下記の通りである[i].

表 3：男女および出家・在家別の発話語数

	出家	在家	合計
男	7238	88660	95898
女	10033	37157	47190
合計	17271	125817	143088

男性登場人物の延べ語数は，女性登場人物の延べ語数と比較して約 2 倍である．また，出家をしていない登場人物の延べ語数は出家をした登場人物の延べ語数の約 7 倍である．この集計では，話者が特定できず話者候補として複数の登場人物が並記されている場合には，在家の登場人物と出家した登場人物が混在していたり，男性登場人物と女性登場人物が混在していたりするため，集計の対象外とした．また，話者の性別が特定できていない場合にも集計の対象外とした．さらに，藤壺中宮や朱雀院など物語の途中で出家をした登場人物のことばは，出家をする前のことばは在家に，出家をした時点より後のことばを「出家」として分類した．また，紫の上は第三十五帖若菜下において受戒をしているが，正式な出家ではないため，若菜下以降のことばも在家に分類した．同じように，作品中において「俗聖」として知られている八の宮も正式な出家はしていないため，八の宮のことばは在家に分類した．

コレスポンデンス分析の対象となる感情形容詞は，源氏物語の話者情報データにおいて出現頻度が高く，その出現頻度が 15 以上のもの上位 34 語である．出家男，出家女，在家男，在家女の 4 つのカテゴリーにおける各感情形容詞の出現頻度は下記の通りである[j].

表 4：感情形容詞の話者分類別使用頻度

	出家男	出家女	在家男	在家女	合計
賢い	11	4	60	13	88
悲しい	7	16	61	29	113
心苦しい	3	5	62	39	109
心安い	3	1	61	13	78
口惜しい	2	10	62	23	97
憂い	2	9	35	41	87
嬉しい	2	13	39	21	75
心細い	2	10	39	19	70
傍ら痛い	2	5	9	22	38
懐かしい	2	0	19	4	25
いぶせい	2	3	18	2	25
いとおしい	1	7	49	45	102
覚束ない	1	3	43	24	71
忝い	1	10	21	32	64
後ろめたい	1	6	33	21	61
頼もしい	1	5	34	16	56
恥ずかしい	1	4	31	13	49
煩わしい	1	4	28	13	46
恨めしい	1	7	26	8	42
後ろ安い	1	3	27	10	41
憎い	1	2	23	9	35
味気無い	1	3	19	4	27
悔しい	1	3	16	5	25
妬い	1	1	17	1	20
苦しい	0	6	60	38	104
辛い	0	10	53	20	83
心憂い	0	11	51	30	92
恋しい	0	1	27	7	35
床しい	0	2	20	7	29
浅ましい	0	6	16	22	44
心許ない	0	1	14	6	21
惜しい	0	1	10	5	16
慎ましい	0	4	8	18	30
うたてい	0	0	4	11	15
合計	51	176	1059	591	1913

出家男，出家女に比べて，在家男，在家女における感情形容詞の出現頻度が高いことがわかる．これは，表 3 で示したようにそれぞれのカテゴリーにおける延べ語数の差に比例している．しかし，

i) 集計の際には，語彙素を使用した．
j) 表 4 の語の表記には，コーパスの「語彙素」を用いているため，現代語形となっている．

各カテゴリーにおける感情形容詞の出現頻度を 100 万語あたりの調整頻度（PMW：per million words）に換算すると，出家男が 7046 語，出家女が 17542 語，在家男が 11944 語，在家女が 15905 語となり，男性登場人物と比べて女性登場人物の方が上記感情形容詞を多用していることがわかる．

次に，表 4 に対するコレスポンデンス分析を行った結果は以下の通りである．コレスポンデンス分析には Python を使用し，mca ライブラリーを用いた．

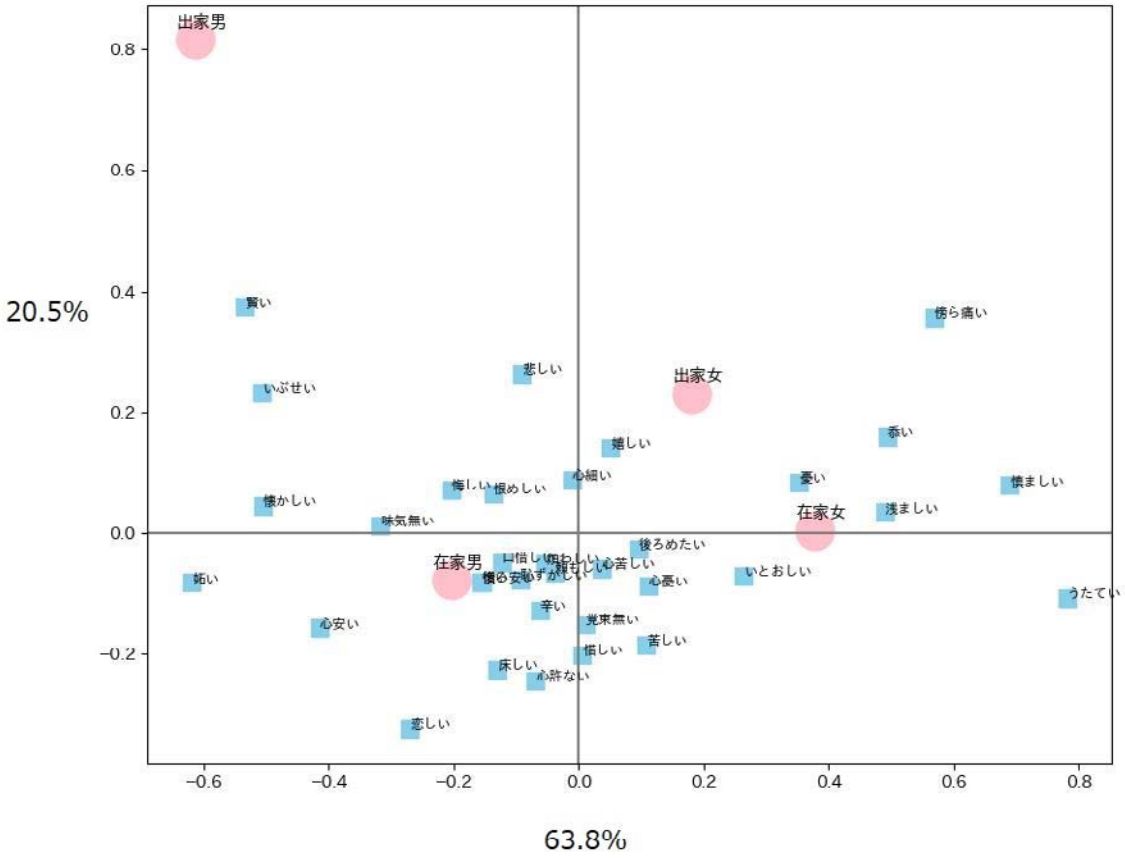

図 2：コレスポンデンス分析結果

第一次元の寄与率は 63.8%で，第二次元の寄与率は 20.5%である．図 2 においてピンク色の丸で示されているのはカテゴリーであり，青色の四角で示されているのが感情形容詞である．

カテゴリー間の相互関係を見てみると，横軸の正の方向に女性登場人物が位置し，負の方向に男性登場人物のカテゴリーが分布している．このことから，第一次元は男女の区別に対応していると考えられる．また，出家女と在家女の距離に比べて，出家男と在家男の距離が遠いことから，女性登場人物と比べて男性登場人物の出家・在家の違いにおける感情形容詞の使用に大きな違いがあることがわかる．

次に，感情形容詞間の相互関係を見てみると，横軸において大きな正の値を持つ感情形容詞は，「うたてい」「慎ましい」「傍ら痛い」などがあげられ，大きな負の値を持つ感情形容詞は，「妬い」「賢い」「いぶせい」などがあげられ，これらの感情形容詞の使用には男女差があることがわかる．また，原点近くにより多くの感情形容詞が集まっていることから，男性登場人物，女性登場人物の間でこれらの感情形容詞の使用において大きな差がないことがわかる．

カテゴリーと感情形容詞の相互関係を見てみる．これらの相互関係を見る際は，カテゴリーの位置から原点の距離，そして変数の位置から原点の距離が遠く，さらに二つの線を結んだ角度が小さいものほど相互関係が強いとされている．このことから，出家男と強い関係にある感情形容詞は「賢い」「いぶせい」などがあげられるが，あま

り多くない．在家男については，「妬い」「心安い」「恋しい」などが強い関係にあることがわかる．出家女と特に強い関係にある感情形容詞は「傍ら痛い」であり，これは在家女とも強い関係にあることがわかる．在家女と関係の強い感情形容詞はこの他にも「慎ましい」や「うたてい」などがあげられる．このことから，女性登場人物は出家・在家に関わらず共通して相対的によく使用されると考えられる感情形容詞があると考えられる．

５．おわりに

本発表では，新たな話者情報の追加による『日本語歴史コーパス』「平安時代編 I 仮名文学」データの拡張および新たなデータを用いた研究の可能性の一つとして，この話者情報データを使用したコレスポンデンス分析の報告を行った．この話者情報はリポジトリでの公開だけではなく，ウェブアプリケーション「中納言」上での検索結果にも表示されることから，様々な研究者に利用してもらえると考えられる．

また，この新たに付与された話者情報により，従来のコーパスでは不可能であった各作品における登場人物のことばを網羅し登場人物間で比較したり，物語の時間軸に沿ってどのようにそのことばが変化するかなどの研究を行ったりすることができるようになった．このことは，限りある平安時代のデータをさらに綿密に分析することを可能にし，言語学的観点からの研究だけではなく，文学的観点からの研究においてもデータ駆動型の研究によるさらなる発展の可能性をもたらすものであると期待される．

謝辞

本研究は，国立国語研究所共同研究プロジェクト「開かれた共同開発環境による通時コーパスの拡張」の成果の一部です．

参考文献

[1] 国立国語研究所：日本語歴史コーパス平安時代編 I 仮名文学 (短単位データ 1.1/ 長単位データ 1.1,中納言バージョン 2.2.0) https://clrd.ninjal.ac.jp/chj/heian.html#kanabungaku, （参照 2022-10-23）．
[2] 国立国語研究所：日本語歴史コーパス語彙統計量 https://clrd.ninjal.ac.jp/chj/chj-wc.html, (参照 2022-10-23).
[3] 阿部秋生，秋山虔，今井源衛，鈴木日出男（校注・訳）：新編日本古典文学全集 25 源氏物語 6，小学館，東京（1998）．
[4] 竹内綾乃，中村壮範，小木曽智信：『日本語歴史コーパス』「平安時代編 I 」拡張話者情報データ ver.1.0,（2022）．http://doi.org/10.15084/00003661.
[5]石川慎一郎・前田忠彦・山崎誠（編）：言語研究のための統計入門，くろしお出版，東京（2010）．