The University of Osaka
Institutional Knowledge Archive

Title	Tess of the d' Urbervillesの会話部によるキャラクタライゼーション
Author(s)	曹, 芳慧
Citation	言語文化共同研究プロジェクト. 2023, 2022, p. 59-78
Version Type	VoR
URL	https://doi.org/10.18910/91702
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

Tess of the d'Urbervilles の会話部による キャラクタライゼーション

曹 芳慧

大阪大学人文学研究科
〒560-0043 豊中市待兼山町 1–8
Email: u327503a@ecs.osaka-u.ac.jp

概要　Thomas Hardy は 19 世紀の有名なイギリスの小説家および詩人であり，著作が多く，作風に個性があり，非常に研究の価値のある作家である。その代表作である *Tess of the d'Urbervilles*（日本語訳：『テス』）についてすでに様々な解釈や論評がなされてきたが，本研究はこれまでの研究と異なる角度で小説の内容の分析を試みる。『テス』の会話部（直接話法）のみを対象として，各登場人物固有の言語的特徴や人物像，および話者間の人間関係に注目し，Hardy がいかに巧みに会話部を用いてキャラクタライゼーション（登場人物の特徴付け）を行なっているかを論じる。方法論として，TEI（Text Encoding Initiative）ガイドラインに基づき，小説テクストの構造化および会話部のアノテーションを行い，XML 解析によって登場人物間の対話ネットワークを考察し，主要な登場人物の会話に用いられている語の頻度パターン解析を対応分析で行う。

キーワード　Text Encoding Initiative (TEI)，XML 解析，会話部，キャラクタライゼーション，対応分析

Characterization by Idiolects in *Tess of the d'Urbervilles*

Cao Fanghui

Graduate School of Humanities, Osaka University
1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　*Tess of the d'Urbervilles* is one of the representative works of Thomas Hardy, who was a famous British novelist of the 19th century. Most of the research on the novel *Tess of the d'Urbervilles* is qualitative. There are few studies conducted based on quantitative analysis on Hardy's literary works and quantitative research on the idiolects is also restricted due to technical limitations. This study focuses on the direct speech of the dialogue part and discusses how Hardy characterizes the characters of the work. In this paper, I will attempt to consider whether the findings shown by the quantitative research on the idiolects can generate new insights about Hardy's works, using methods that have not been implemented in the previous research. Firstly, I mark up and annotate the text of this novel based on the TEI (Text Encoding Initiative) guidelines to create a structured XML file. Secondly, I parse the file with

Python to extract the dialogue sections from the entire novel and represent the interaction network among the characters visually. Then, referring to the network, I analyze the frequency patterns of the words used in the idiolects of the main characters with CA (Correspondence Analysis) to clarify the linguistic features and personality traits of them, as well as the relationships within the speakers. In this way, this research overcomes the technical limitations in abstracting idiolects in novels and proposes a quantitative research approach for the dialogue part. By doing so, we are able to discuss how the author skillfully uses the idiolects to perform characterization in the novel.

Keywords Text Encoding Initiative(TEI), XML Analysis, Idiolects, Characterization, Correspondence Analysis

1. はじめに

Tess of the d'Urbervilles は，イギリスで創刊された絵入り週刊誌 *Graphic* に連載され，1891 年に James R.Osgood, Mc Ilvaine&Co.によって出版された小説であり，Thomas Hardy（トマス・ハーディ，1840 年 6 月 2 日～1928 年 1 月 11 日）の代表作の一つである。日本語訳ではよく『テス』の題名が使用されているため，以下は『テス』と称する。

1. 1. 『テス』の位置付け

本研究は，Hardy の代表作の一つ『テス』を研究対象に絞っている。名前からわかるように，『テス』の主人公は Tess である。Johnson T. (1968) で，「Tess は Hardy の最も巧妙に描き出した女性だ」[1]と述べられている。Laird J. T. (1975) で，「彼女（Tess）が Hardy の最も成功したヒロインの 1 人であるという判断に異議を唱える読者はほぼいないだろう」[2]と書かれている。また，1891 年に『テス』を，1896 年に『日陰者ジュード』を出版後，酷評されたこともあり，Hardy は執筆活動の中心を小説から詩作へと移し，以後は詩作に専念するようになった。ここからも窺えるように，『テス』は Hardy の数多くの著作の中で非常に重要な位置付けにある。

1. 2. 本研究の問題設定

これまでの文体や文学作品に関する研究は，作品全体をデータとして扱うものが多い。『テス』を会話部と地の文の二部分にはっきり分けて分析する質的研究もなかったようである。文学作品，特に発話順序が複雑な小説の会話部に関する量的分析もやや乏しい。

しかし，会話部が小説の成り立ちにおいては無視できないほど重要な役目がある。Page (1973) によれば，小説の対話（the dialogue in a novel）は多機能である。筋書きを立てたり、キャラクターを発展させたり、設定や環境を説明したり、道徳上の議論を提示したり、様々な話題についての議論を行なったりするのに対話が役立つ。これらの機能の任意の組み合わせも可能である。対話のさまざまな機能の中で、「キャラクターの提示と発展(the presentation and development of character)」が最も重要である。

そこで，本研究は，『テス』に関して，会話部のみに着目し，小説の会話部の語彙使用について

[1] 'Tess is Hardy's most masterly portrait of a woman'. Johnson T. (1968: 151).

[2] '…there are few readers of Hardy's fiction who would disagree with the judgement that she is one of Hardy's most successful heroines…'. Laird J. T. (1975: 118).

量的分析を行ってみる。会話部を対象とした量的研究が示した知見が Hardy の作品について新たな洞察を生み出しうるか，本研究では先行研究では（技術的な制約もあり）実施されていない手法も加えて考察を試みる。

2. 研究方法と目的

以下は，まず，『テス』本文の電子データをプレーンテクスト（.txt）として収集し，TEI（Text Encoding Initiative）に準拠したタグで，文学作品テクストの構造化，および文学作品の会話部と地の文を区別するマークアップを施し，XML（Extensible Markup Language）ファイルを作成する。次に，小説テクストの精読に基づく質的考察を行いながら，会話部に対してアノテーション（話し手と聞き手の情報付与）を行った上で，小説全体から会話部の直接話法を抽出する。そして，XML 解析により，ネットワークを描画し，主要な登場人物を特定する。それにより，主要な登場人物の個人語を詳細に分析する取り組みを行う。最後に，会話部の語彙について量的分析（対応分析）を行い，得られた結果に基づき，小説の各登場人物の発話の語彙的特徴および登場人物の間の関係性を明らかにしたい。

具体的なリサーチ・クエスチョン（RQ）は，（1）使用語彙を手掛かりとして登場人物の差が区別できるか，区別する場合，発話者をグループ分けする基準は何か，（2）会話で使われている語彙から，その登場人物の性格が見られるだろうか，見られる場合，どのような人物像が思い浮かぶか，（3）会話部の語彙から，登場人物の間の関係性は見られるだろうか，の 3 点である。

3. データ

3. 1. 小説本文の電子データの収集（TXT）

『テス』のテクストをデジタル化するに当たっては，CasualTextractor[3]というツールを使用し，電子書籍アーカイブ Project Gutenberg[4]から小説『テス』の本文をテキストファイル・プレーンテクスト（.txt）として収集した。小説テクストを抽出する際には，余分な記号（改行やタブ）を削除する，curly quotation marks（2 バイトの引用符）を ASCII quotation marks（1 バイトの引用符）に変換するなど，テクスト分析を行う前に，フォーマットを揃えておいた。抽出した小説テクストの総語数（Tokens）は 149,532 であり，異なり語数（Types）は 12,803 である。

3. 2. 小説のマークアップとアノテーション（TEI/XML）

3. 2. 1. XML と TEI

XML は，文章の見た目や構造を記述するためのマークアップ言語の一種である。XML はタグ名も属性も基本的に利用者が自由に設定できるため便利ではあるが，「各自が異なるルールでタグ付けをしてしまうと，共通のツールで利便性を高めたり，それぞれの成果を共有したりするこ

[3] https://sites.google.com/site/casualconcj/
[4] https://www.gutenberg.org/

とが難しくなってしまう」（石田他 2022: 128）。それを回避するために, TEI のガイドラインに沿ったテキストデータの構築が重要である。本研究は TEI/XML ファイルの形で小説テクストをデータ化することを試みる。

3. 2. 2.マークアップとアノテーション

『テス』は7つのフェーズと59の章に分かれている。『テス』の文書構造形式について, <text>, <body>, <div1>, <div2>, <head>タグを用いてマークアップを施した後, 会話部に対してアノテーション（話者の情報付与）を行う作業に入る。本研究は会話部の直接話法のみに着目するため, 小説の本文にある二重引用符を切り口とし, 二重引用符に囲まれた内容から分析対象となる会話データを選別する。

TEIガイドラインによると, 標準的なquotation marksには, 発話や思考（speech and thought）を表せる<q>と<said>, 引用（quotations）を表す<quote>, 書誌参照を伴う他の文書からの引用（citations）を表す（<quote>と<bibl>を伴う）<cit>, 言及された語句（words or phrases mentioned）を表す<mentioned>, 著者や語り手が責任を持ちたくない語句（words or phrases over which the author or narrator disclaims responsibility）を表す<soCalled>, 術語（terminology）を表す<term>, 注釈（glosses）を表す <gloss>がある[5]。本研究の会話部に関するアノテーションは, 引用内容が直接話法なのか間接話法なのかを示せる<said>タグを使用する。一方, 二重引用符には囲まれているが, 本研究の分析対象としない部分に対しては, より一般的な<q>タグで区別することにした。<said>タグと<q>タグの区別, および本研究で扱うタグに関する説明（図1）は以下の通りである。

<said>: (speech or thought). a more explicit element for the encoding of speech or thought, which allows the encoder to distinguish these from other quoted text

<q>: (quoted). the general element for quotation

```
1  <teiHeader>
2    <!-- ... -->
3    <encodingDesc>
4      <tagsDecl>
5        <namespace name="http://www.tei-c.org/ns/1.0">
6          <tagUsage gi="text">Contains a single unitary text.</tagUsage>
7          <tagUsage gi="body">Contains the whole body of a single unitary text, excluding
…  any front or back matter.</tagUsage>
8          <tagUsage gi="div1">Marks a phase. The <att>type</att> attribute gives the
…  division type. The <att>n</att> attribute gives the canonical phase number, where
…  appropriate.</tagUsage>
9          <tagUsage gi="div2">Marks a chapter. The <att>type</att> attribute gives the
…  division type. The <att>n</att> attribute gives the canonical chapter number, where
…  appropriate.</tagUsage>
10         <tagUsage gi="head">Provides the phase/chapter header.</tagUsage>
11         <tagUsage gi="said">Marks a speech within the text. The <att>who</att> and
…  <att>toWhom</att> attribute identifies the characters associated with that
…  speech.</tagUsage>
12         <tagUsage gi="q">Contains quoted sections of text and other speeches which
…  speaker and/or listener is/are unclear. The <att>who</att> attribute identifies the
…  characters associated with that speech. </tagUsage>
13       </namespace>
14     </tagsDecl>
15   </encodingDesc>
16   <!-- ... -->
17 </teiHeader>
```

図 1 『テス』TEI/XML ファイルに使用したタグに関する説明（<teiHeader>の一部）

[5] https://teibyexample.org/tutorials/TBED03v00.htm（2.5. Quotation を参照）
https://tei-c.org/release/doc/tei-p5-doc/ja/html/CO.html#COHQQ

『テス』の会話部に関するアノテーションの作業を終え，作られた XML 文書が well-formed（形式が整った，整形式）であるかどうか, valid（妥当的）であるかどうかを確認[6]してから次の作業に進む。作られたファイルの冒頭部分を図 2 で示す。

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <TEI xmlns="http://www.tei-c.org/ns/1.0">
3     <text>
4       <body>
5         <div1 type="phase" n="1">
6           <head>Phase the First: The Maiden</head>
7           <div2 type="chapter" n="1">
8             <head>Chapter I</head>
9             On an evening in the latter part of May a middle-aged man was walking homeward
…   from Shaston to the village of Marlott, in the adjoining Vale of Blakemore, or Blackmoor.
…   The pair of legs that carried him were rickety, and there was a bias in his gait which
…   inclined him somewhat to the left of a straight line. He occasionally gave a smart nod,
…   as if in confirmation of some opinion, though he was not thinking of anything in
…   particular. An empty egg-basket was slung upon his arm, the nap of his hat was ruffled, a
…   patch being quite worn away at its brim where his thumb came in taking it off. Presently
…   he was met by an elderly parson astride on a gray mare, who, as he rode, hummed a
…   wandering tune.
10            <said who="#John" toWhom="#Parson Tringham">"Good night t'ee,"</said>
11            said the man with the basket.
12            <said who="#Parson Tringham" toWhom="#John">"Good night, Sir John,"</said>
13            said the parson.
14  The pedestrian, after another pace or two, halted, and turned round.
15            <said who="#John" toWhom="#Parson Tringham">"Now, sir, begging your pardon; we
…   met last market-day on this road about this time, and I said 'Good night,' and you made
…   reply 'Good night, Sir John,' as now."</said>
16            <said who="#Parson Tringham" toWhom="#John">"I did,"</said>
17            said the parson.
18            <said who="#John" toWhom="#Parson Tringham">"And once before that-near a month
…   ago."</said>
19            <said who="#Parson Tringham" toWhom="#John">"I may have."</said>
20            <said who="#John" toWhom="#Parson Tringham">"Then what might your meaning be in
…   calling me 'Sir John' these different times, when I be plain Jack Durbeyfield, the
…   haggler?"</said>
21            The parson rode a step or two nearer.
```

図 2　TEI に準拠したタグでマークアップされた『テス』XML ファイル（一部）

紙幅のため，図 2 で示されているのは長い<teiHeader>の部分を省いたものである。実際に扱うファイルはヘッダ（<teiHeader>）を持っている。注意すべきなのは, TEI ではヘッダ（<teiHeader>）が必須である。全ての妥当な TEI ファイルは<TEI>エレメントの内側に，<titleStmt>と<publicationStmt><sourceDesc>を含む<fileDesc>を伴う<teiHeader>エレメントを追加しなければならない。石田他（2022: 49）によれば, <teiHeader>エレメントに沿って，データがいつ誰によってどのような資料に基づいて作られ，どのようなルールでデジタル化されているか，といったことを記述しておくことは，データの信頼性を担保する上できわめて重要である。

3. 3. 登場人物の interaction network 描画

登場人物の会話部に絞り込んで分析を行うためには, <said>要素を抽出し XML 解析を行う必要がある。XML 解析とこれから実行する登場人物の interaction network 描画に関しては, Karsdorp *et al.* (2021)を参照して Python で実行する。本研究は，『テス』のキャラクターインタラクションネットワークを次のように定義する。全ての<said>タグに対して, @who 属性にある人名 Name A を

[6] validation check: https://www.xmlvalidation.com/
syntax-check: https://www.w3schools.com/xml/xml_validator.asp

登場人物 i, @toWhom 属性にある人名 Name B を登場人物 j と見なし，登場人物 i と登場人物 j のペアが n 回現れると，A と B の間に interaction が n 回あると認識する。この定義に関する部分の Python コードは図 3 に示す。この定義に従って描いたネットワークの簡略化したものは図 4 である。

```
def character_network(tree):
    G = nx.Graph()
    # extract a list of speaker turns for the novel
    for said in tree.iterfind('.//said', NSMAP):
        # ... and extract pairs of adjacent speakers
        try:
            speaker_i = said.attrib['who'].replace('#', '')
            speaker_j = said.attrib['toWhom'].replace('#', '')
            # if the interaction between two speakers has already been attested, update their interaction count
            if G.has_edge(speaker_i, speaker_j):
                G[speaker_i][speaker_j]['weight'] += 1
            # else add an edge between speaker i and j to the graph
            else:
                G.add_edge(speaker_i, speaker_j, weight=1)
        except KeyError:
            continue
    return G
```

図 3　本研究で扱うキャラクターネットワークの定義（Python code block）

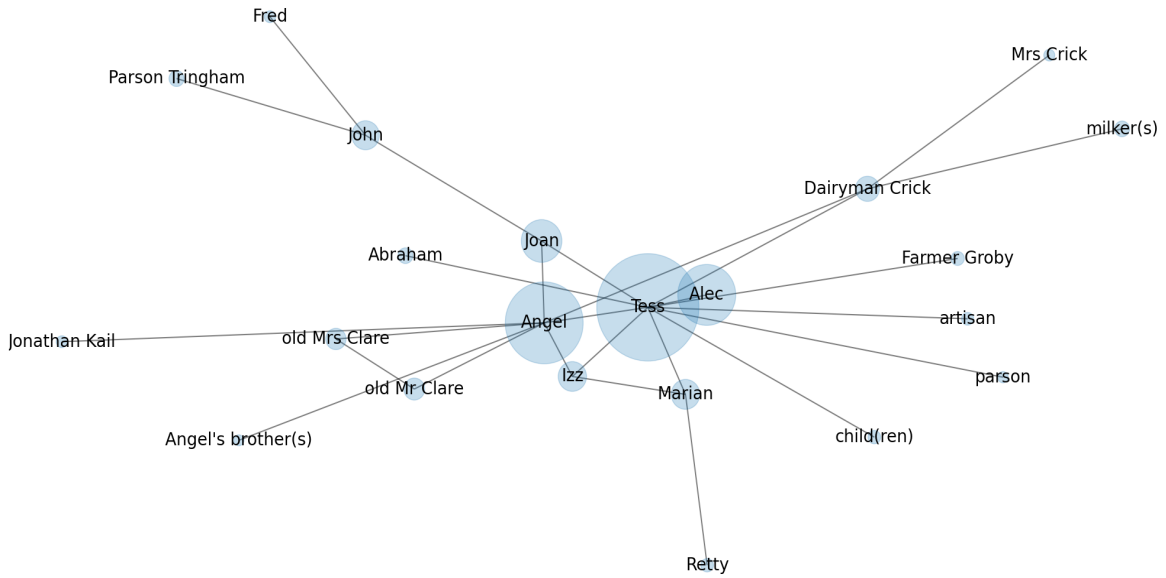

図 4　『テス』のキャラクターインタラクションネットワーク（簡略図）

図 4 で示した登場人物のネットワークはノード（node：節点・頂点，点）の集合とエッジ（edge：枝・辺，線）の集合から構成されている無向グラフである。nodes は小説の登場人物に対応しており，node labels は登場人物の名前を示している。登場人物 i と登場人物 j の間のインタラクション

（interaction：対話）の回数が多ければ node size が大きくなる。node と node を繋げている edge は登場人物 i と登場人物 j の間に interaction があることを表す。『テス』の全ての登場人物の対話ネットワークでは, 主要でない登場人物も数多く含まれており, 見づらいため, このように簡略化して見やすくした[7]。図 4 のネットワークでは, interaction の回数 が 10 回以上のキャラクターのみが含まれており, node 数は 22 で edge 数は 26 である。この図の中に, 幾つか三角形になっている部分もあるため, まずはそれらに注目して説明しておきたい。

Angel と Clare 夫妻は家族であり, この 3 人がネットワークで三角形になっているのは, Angel と old Mrs Clare, Angel と old Mr Clare, old Mrs Clare と old Mr Clare の間にそれぞれ対話があったことを意味している。しかし, 主人公の Tess を見ると, Tess と母 Joan, 父 John は一直線状に並んでいる。つまり, Tess と Joan, Joan と John の間に対話があったが, Tess と John は（同じ場面にいることはありながら）一度も一対一の会話をしたことがない。しかも, John は Joan に比べ, ネットワークの周辺辺りにある。このことから, 小説で Joan は John より重要な役割を果たしていることが推測される。同様に, Tess と Marian, Izz, Retty の四人は友人であるが, Retty だけが離れた位置にある。Marian と Izz は Retty に比べ, 物語の発展でより重要な役割を果たすことが言えよう。

他に, Tess-Angel-Dairyman Crick, Tess-Angel-Izz, Tess-Angel-Joan も三角形になっている。Tess と Angel は恋人で親密な関係にあり, 対話ネットワークでもこの二人は共通の知り合いが複数人いることが反映されている。Tess と Angel は共に酪農場で働いていたので, それぞれ酪農場主 Dairyman Crick と話したことがあるのは当然のことである。Angel と Izz の対話は『テス』の第 40 章にあり, Angel が Tess を見捨ててブラジルへ渡航する前に, Izz を誘って一緒にブラジルに行く部分にある。Angel と Joan の対話は『テス』の第 54 章で現れ, Angel がブラジルからイギリスへ戻り, Tess の居場所を探すために Joan のところに行った部分にある。

このように, 『テス』の会話部から登場人物間の人間関係をとらえることができ, インタラクションネットワークを描画することで, 登場人物の重要性が視覚化できたと考えられる。

3. 4. 主要な登場人物の会話部抽出

会話部の分析において対象となるキャラクターは, 簡略化したネットワーク（図 4）に含まれ（interaction の回数≧10）, 且つ発話語数が 500 語以上（Tokens≧500）の登場人物に絞り込んでいる。抽出された 10 人の登場人物の個人語に関する統計値は表 1 の通りである。10 人のうち, 男性（Angel, Alec, Dairyman Crick, John, Old Mr Clare）は 5 人で, 女性（Tess, Joan, Marian, Old Mrs Clare, Izz）は 5 人いた。男女比は正に 1 対 1 であった。分析対象となる会話テクストの総語数（Tokens）は 32,555 であり, 全小説の<said>タグに囲まれた会話テクストの総語数(36,186)の約 90%である。この 10 人の発話語数は『テス』の総語数（149,532）の約 5 分の 1 を占めている。なお, ここでいう「主要な登場人物」の概念は「発話語数が極めて少ない脇役」と対照的なものである。

[7] 簡略化せずにネットワークを描画すると, node 数は 48 で edge 数は 77 である。

表 1 主要な登場人物（10 人）の個人語に関する統計値[8]

	登場人物	性別	Tokens	Types	TTR	STDTTR	RTTR
1	Tess	女	9,125	1,280	0.14	0.43	13.40
2	Angel	男	7,547	1,566	0.21	0.48	18.03
3	Alec	男	6,634	1,352	0.20	0.48	16.60
4	Dairyman Crick	男	2,463	773	0.31	0.52	15.58
5	Joan	女	2,233	648	0.29	0.50	13.71
6	John	男	1,276	490	0.38	0.51	13.72
7	Marian	女	1,162	410	0.35	0.47	12.03
8	Old Mrs Clare	女	768	337	0.44	0.51	12.16
9	Old Mr Clare	男	675	334	0.49	0.53	12.86
10	Izz	女	672	280	0.42	0.43	10.80
	平均値	-	3,255.5	747	0.32	0.48	13.89

4. 結果と考察

このセクションでは，抽出された10人の主要な登場人物の個人語がいかに描き分けられているかに対して考察を行なっていく。まず，各キャラクターの発話の語彙に対して品詞タグを付与し，各登場人物の品詞の使用について探る（4.2）。次に，法助動詞に着目し，個人語の法助動詞の出現頻度で差が見られるかについて検討する（4.3）。最後に，『テス』で一番重要なTess, Angel, Alecの3人にさらに絞り込み，会話部の頻出語からこの3人の関係性が窺えるかを分析する（4.4）。分析に当たっては主に対応分析を使用するので以下はまず対応分析について説明する（4.1）。

4. 1. 対応分析とは

対応分析（Correspondence Analysis）は，データ表の行と列の対応（correspondence）関係に着目しており，行や列に含まれる多くのカテゴリの情報を少数の成分（次元［dimension］）に圧縮し，各次元を組み合わせて散布図を作成することで，分析結果を視覚化し，カテゴリ間の関係を直観的に解釈することを可能にする多変量解析手法の一つである。

石川他 （2010）によれば，対応分析はデータを分類するその他の統計手法[9]に比べ，以下のように多くのメリットを持っている。

[8] TTR, STDTTR と RTTR は語彙の多様性・豊富さを測定する指標である。TTR(Type/Token Radio)は延べ語数（Token 数）の中に占める異なり語数（Type 数）の割合である。TTR が基本的な考え方となるが，文章の長さに依存するため，長さの異なるテクストの比較には適切でない。STDTTR(Standardised Type/Token Ratio)は標準化 TTR で n 単語ごとに計算される。デフォルトでは, n=1,000 が，本研究では n=500 に設定した。RTTR（Guiraud Index ともいう）は Type 数を Token 数のルートで割ったものである。

[9] 対応分析以外に，データの分類手法として，判別分析，クラスター分析，主成分分析，因子分析等がある。

ア）変数とケースといったデータの性質の違いが問題にならない[10]。

イ）第 1 カテゴリ，第 2 カテゴリの同時分類が可能である。

ウ）分析の主目的が，データの合成や分解ではなく，カテゴリの整理と分類そのものにある。

エ）分析者が決定すべきオプションの幅が狭く，分析者の判断によって結果が左右されにくい。

（石川他 2010: 248）

また，多変量解析では，多くのカテゴリを少ない次元に圧縮するという分析の目的に照らして考えれば，およそ 5 個以上，通例 10 個以上のカテゴリを持つアイテムに対して分析を行うことが望ましいと言える（石川他 2010: 250）。今回は 10 人の登場人物の会話部を対象にして分析するので，対応分析を使用するのが適切であると考えられる。

4. 2. 品詞構成の比較

本節では, 10人の主要な登場人物の言語データを用いる。データ処理は，まず, CasualTagger[11]を起動し，抽出された10人の発話に対し, TreeTaggerのEnglish BNC tagset[12]で品詞タグ付けを行っておく。ここでは, TreeTaggerによる形態素解析の誤りについて特に修正を行わない。次に, CasualConc[13]を用いて，個々の登場人物の語彙リストを用意する。表1ですでに示したように，この10人の発話語数に尺度差がある。そのため，以下は個人語の品詞の出現頻度を500語あたり調整頻度に備えておく。得られた頻度表は表2に示す。計38種類の品詞タグがあったが，ここでは上位5位までの品詞のみを示す。

会話部全体で出現頻度の高い上位5位の品詞は, pnp（Personal pronoun：人称代名詞）, nn1（Singular common noun：名詞の単数形）, av0（General adverb：一般副詞）, vbb（The present tense forms of the verb BE：動詞の現在形）, vbi（The infinitive form of the verb BE：動詞の不定形）である。この点から会話部において人称代名詞が非常に重要な役割を果たしていることがわかる。

[10]他の多変量解析手法では，列方向（縦方向）に入るデータと行方向（横方向）に入るデータをそれぞれ異なる性質を持つものと考え，前者を変数，後者をケース（検体，個体，サンプル）として区別してきた（石川他 2010: 245）。しかし，対応分析では，行と列の入れ替え（転置）を行っても，分析結果は変わらない。

[11] https://sites.google.com/site/casualconcj/

[12] http://www.natcorp.ox.ac.uk/docs/c5spec.html

[13] https://sites.google.com/site/casualconcj/

表 2 主要な登場人物（10人）の個人語の品詞頻度表（相対頻度）[14]

	登場人物	Tokens	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	Tess	9,141	pnp (89.43)	nn1 (38.95)	vbi (35.66)	vbb (35.23)	av0 (34.73)
2	Angel	7,515	pnp (69.93)	nn1 (52.03)	av0 (35.40)	vbb (33.93)	vbi (30.87)
3	Alec	6,649	pnp (74.82)	nn1 (49.48)	av0 (36.55)	vbb (33.31)	prp (32.86)
4	Dairyman Crick	2,524	nn1 (67.55)	pnp (49.33)	av0 (34.67)	at0 (34.47)	prp (33.48)
5	Joan	2,298	pnp (59.62)	nn1 (54.61)	av0 (34.60)	vbi (30.46)	prp (28.29)
6	John	1,315	nn1 (65.78)	pnp (51.71)	prp (32.70)	av0 (32.32)	at0 (25.86)
7	Marian	1,197	pnp (84.38)	nn1 (41.77)	av0 (36.34)	vbi (31.75)	vbb (27.57)
8	Old Mrs Clare	769	pnp (66.97)	nn1 (64.37)	av0 (31.86)	prp (31.86)	aj0 (31.21)
9	Old Mr Clare	678	nn1 (64.90)	pnp (53.10)	prp (36.14)	aj0 (33.19)	at0 (29.50)
10	Izz	678	pnp (67.85)	nn1 (47.20)	av0 (44.99)	vbi (32.45)	prp (26.55)
	総語数	32,764	pnp (4,809)	nn1 (3,268)	av0 (2,300)	vbb (2,028)	vbi (2,021)

4. 2. 1. 品詞構成のクラスター分析

対応分析を行う前に，まずは，各登場人物の品詞頻度表に対してクラスター分析を用いて登場人物間の品詞使用の類似度を探る。クラスター分析（cluster analysis）とは，データが持つ情報を手掛かりにして，距離の近いデータ同士をまとめてクラスター（cluster：群，集落）を構成する統計手法である。言語コーパス研究では，語彙・品詞・ジャンルなどの分類の目的で幅広く使用される（石川他 2010: 163）。本研究では, 500語毎の相対頻度を基準に，全ての品詞種類の相対頻度

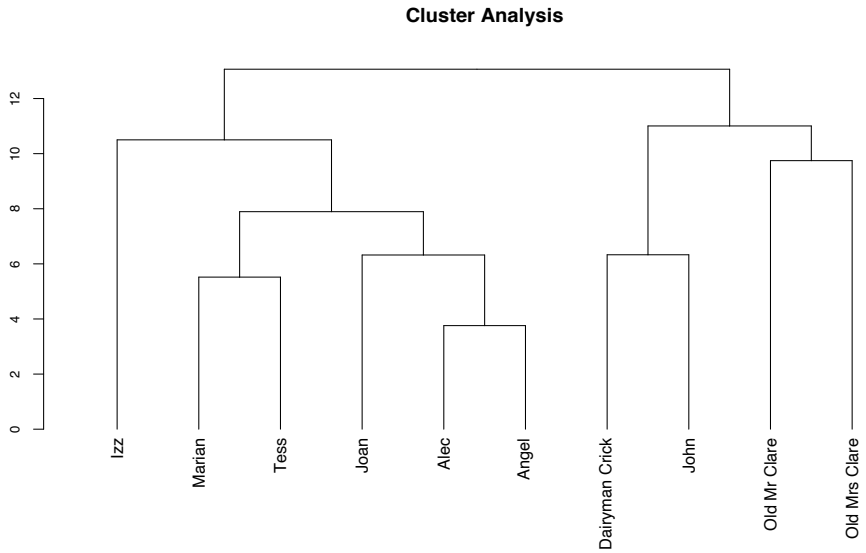

図 5 主要な登場人物（10 人）の個人語の品詞構成の樹形数

[14] こちらは品詞タグが付与されたデータなので，個人語の Token 数がタグのないデータに比較すると，数はやや増加か減少した。

をデータとして, Ward法, Euclidean距離を用いる（図5）。

クラスター分析の結果，この10人は大きく2つのグループに分けられた（図5）。左側の6人は正に図4のネットワークの中心に位置しているキャラクターで，右側の4人はより周辺的な場所に位置するキャラクターである。これは, Hardyが自分の小説で肝心な人物たちの発話とそれ以外の人物の発話に対してそれぞれ似たようなパターンで描いているかもしれない。また，樹形図の低い下位分類を見ると, MarianとTess, AlecとAngel, Dairyman CrickとJohn, Old Mr ClareとOld Mrs Clareの品詞使用が似ていることがわかる。MarianとTess, Old Mr ClareとOld Mrs Clareの品詞使用には類似性があるが，それは『テス』における設定を考えると，理解は容易かと思われる。一方，中央のAlecとAngelは共に物語の主要人物であるが，彼らは一度も会ったことがない。Dairyman CrickとJohnも知り合いではなく，同じ場面で登場したことがない。しかし，クラスター分析の結果を見ると，似たような品詞の使い方をしている。考えられるのは, AlecとAngel, Dairyman CrickとJohnが『テス』での重要性は同等レベルであることである。このように, Hardyが似たような設定にしたい人物の個人語では，使用する品詞も似通っている。

4. 2. 2. 品詞構成の対応分析

クラスター分析から読み取れたのは, Hardyが『テス』を作成する際に，関係が近い人物同士や似たような設定にしたい人物同士の会話部に用いる語彙に一定のパターンがあることである。しかし，クラスター分析で各登場人物間の類似度は見られたが，登場人物と品詞の間でどのような対応関係にあるのかは明らかにならなかった。そのため，次に，クラスター分析の結果に踏まえ，

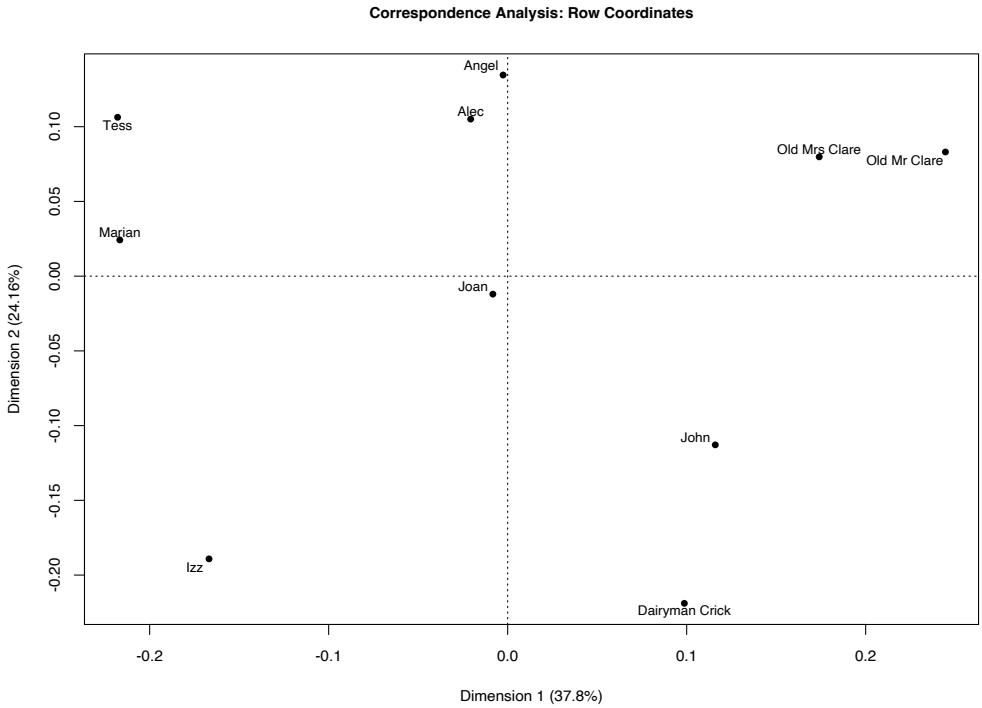

図 6　品詞の相対頻度による対応分析（第 1 アイテム）

品詞構成の対応分析を行う（図6，図7）。

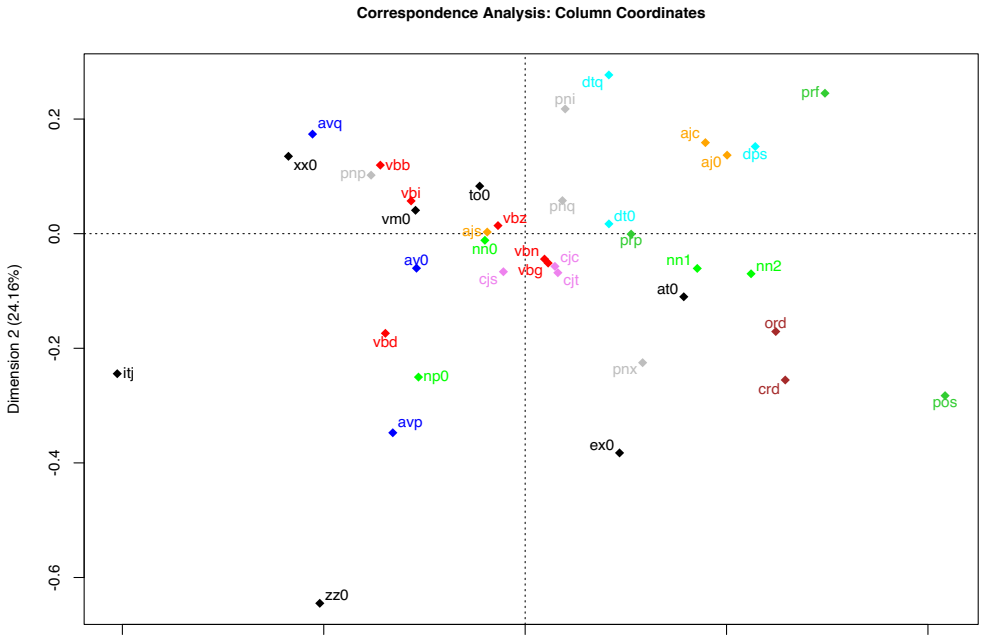

図 7　品詞の相対頻度による対応分析（第 2 アイテム）

第1アイテム（登場人物）がプロットされた散布図（図6）を見ると，クラスター分析の樹形図（図5）と同じ結果が得られているので，その説明を省略する。ここでは，第2アイテム（品詞）がプロットされた散布図（図7）に注目する。この図は第1アイテムの散布図（図6）と対応している。図6では，品詞の頻度パターンの近い人物同士が近く布置される。同様に，図7では，個人語における頻度パターンの近い品詞が近くに現れる。

図7の動詞の分布に着目すると，上から下へと現在形（vbb），不定形（vbi），過去形（vbd）が分布している。動詞の-*s*形（vbz），過去分詞（vbn）や-*ing*形（vbg）は原点あたりにある。ここから，登場人物の間で動詞の時制の違いがあったことがわかる。第1軸（横軸）の上半分に分布している人物は現在形を，下半分に分布している人物は過去形を頻繁に使用している。図6と併せて見ると，第2象限にあるTess, Angel, Alec, Marianは現在形を多用し，第3象限にあるIzzは過去形を多用していることがわかる。コンコーダンスを調べた結果，Izzの個人語で過去形の使用は主に第40章でAngelとの長いやり取りに現れているが，全体を見ると，Izzの発話で過去にあった出来事の説明・伝達や状況の確認のために過去形が使われることが多数である。同じく対面でコミュニケーションをしているにも関わらず，過去形を会話で多用する人物は主に既に発生した出来事を他のキャラクター（と読者）に伝える役割を果たしており，現在形を多用し現在の物事を語る人物に比較すると重要度がやや低いことが考えられる。このように，『テス』で会話部の動詞の時制からもキャラクターの役割が反映されている。

また，他にも散布図から読み取れる情報ある。例えば，図7で副詞（avq, av0, avp）が第2軸（縦

軸）の左側に分布しており，図6の左側の6人（ネットワークの中央にある人物）が相対的に副詞をより頻繁に使用することがわかる。一方，形容詞や名詞は主に軸の右側に寄っている。最上級の形容詞（ajs），固有名詞（np0）と普通名詞（nn0）は原点近くにあるが，比較級（ajc）または一般的な形容詞（aj0），単数形の名詞（nn1）と複数形の名詞（nn2）は右側に分布している。つまり，図6の右側の4人（より周辺的な人物）の品詞使用は異なる傾向にある。さらに，形容詞以外，疑問限定詞（dtq），所有限定詞（dps），一般限定詞（dt0）と前置詞（prf）が第1象限に分布しているのに対し，所有格または属格のマーカー's または'（pos）は第4象限に分布している。Old Mr ClareとOld Mrs Clareの発話の文構造の複雑さ，JohnとDairyman Crickの文構造の簡単さがそれぞれ反映されている。

4. 3. 法助動詞の比較

品詞構成について分析した後，法助動詞に着目して，法助動詞の使用よる登場人物の分類ができるか試みる。語彙リストから法助動詞（*_vm0）を全て抽出し，10人の登場人物と全18種類の法助動詞の頻度表を作成する。ここでも，TreeTaggerによる形態素解析の誤りは特に修正しない。

4. 3. 1. 法助動詞の対応分析（粗頻度）

以下ではまず，頻度補正を行わない頻度表（粗頻度）[15]に対して対応分析を実行してみる（図8，図9）。

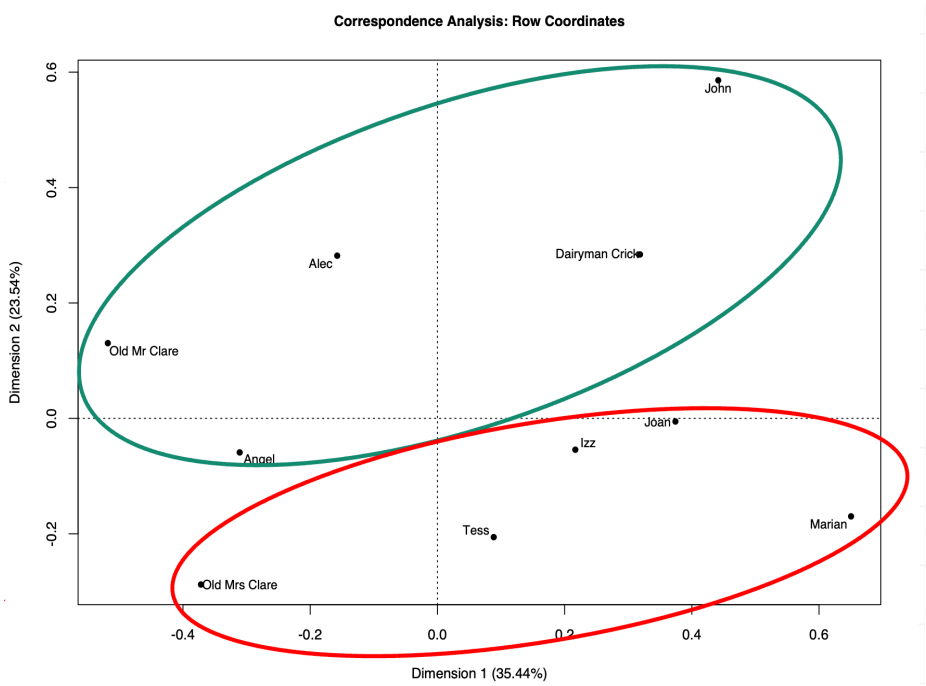

図 8　法助動詞の粗頻度による対応分析（第 1 アイテム）

[15] 一般に，ファイル間やコーパス間に総語数の差（尺度差）がある場合は○○語あたりの調整頻度によって尺度差を補正するが，対応分析では，補正前の生データを使うことも多い。

まず，第1アイテム（行：登場人物）がプロットされた散布図（図8）に注目すると，第1次元の正負によって，Angel一家（Old Mr Clare, Old Mrs Clare, Angel），Alecとそれ以外が区分されており，第1軸（横軸）はこれらを分ける軸になっていることがわかる。同様に，第2軸（縦軸）は，Angel（男性）以外，第2次元の正負によって男性群（Old Mr Clare, Alec, John, Dairyman Crick）と女性群（Old Mrs Clare, Tess, Izz, Joan, Marian）を区分する軸になっている。第1軸（横軸）に対してより詳しく分析してみると，図中の左から右に向かって登場人物の教養レベルが下がっていくという解釈が可能になる。

次に，第2アイテム（列：法助動詞）がプロットされた散布図（図9）を併せて考察する。この図は図8と対応しているため，教養レベルの高い人物に顕著な語が図中の左側にプロットされており，教養レベルの低い人物に特徴的な語は図中の右側にプロットされている。教養レベルの低い人物に特徴的な語は，*d*, *ca*, *ll*, *sha*, *wo*のような省略形の語である[16]。方言の影響があるともいえよう。それに対して，教養レベルの高い人物に特徴的な語として，*dare*, *may*, *will*, *shall*, *need*といった省略されていない形の法助動詞が現れている。

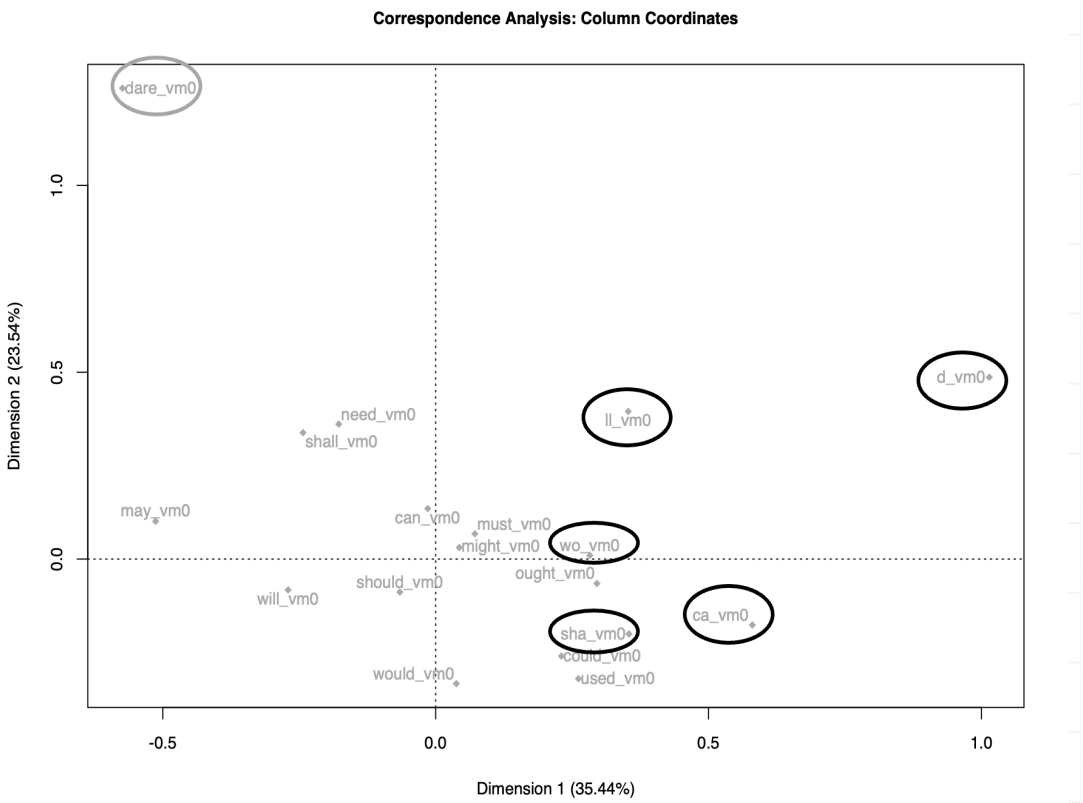

図 9　法助動詞の粗頻度による対応分析（第 2 アイテム）

[16] コンコーダンスを確認した結果，*d* は could, would, had に，*ca* は can の否定縮約形 can't の ca に，*ll* は will に，*sha* は shall の否定縮約形 shan't の sha に，*wo* は will の否定縮約形 won't の wo に，それぞれ対応している。なお，had は法助動詞ではないが，ここではタグの誤りに関して特に修正を行わない。

法助動詞の粗頻度についての考察では，『テス』の登場人物での男女別および教養レベル別から法助動詞の使用に差が見られたが，これは個人語で実際に使われる法助動詞の頻度による対応関係であり，異なるキャラクター間の発話語数の尺度差は考慮に入れなかった。そこで，相対頻度の対応分析を以下で実行してみる。

4. 3. 2. 法助動詞の対応分析（相対頻度）

粗頻度の図と比較すると，相対頻度の散布図（図10，図11）では，*need*, *sha*, *used*, *dare*の4つが（頻度が低いかつそれを用いる登場人物が少ないため）減少し，14種類の法助動詞のみが対応分析の対象になる。

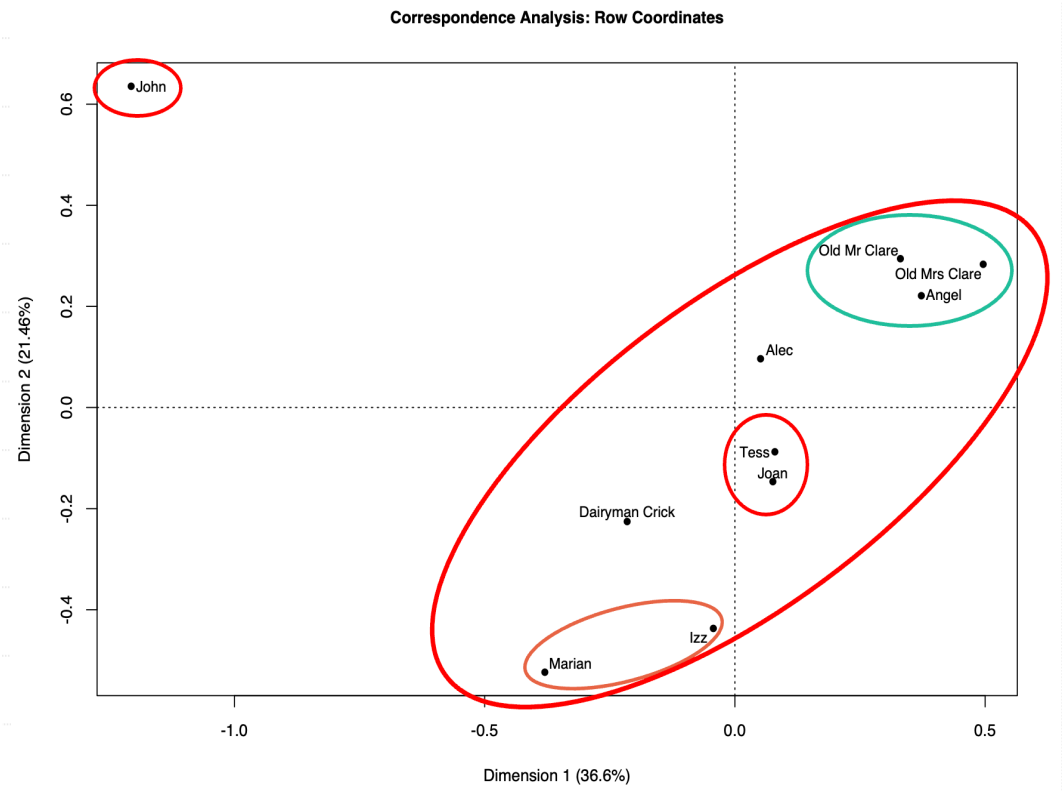

図 10 法助動詞の相対頻度による対応分析（第 1 アイテム）

図10で描かれた人物の分布から，人物同士の関係性を観察することができる。例えば，右上のClare夫妻（Old Mr Clare, Old Mrs Clare）とAngelの3人は家族なので，似たような場所に分布しており，使用する法助動詞も似ているという結論が得られる。原点あたりのTessとその母Joanも近いところに分布しており，底辺あたりのIzzとMarianは友人関係で，分布が似たところにあるので，法助動詞の使用も似ていることがわかった。しかし，この相対頻度に基づくデータでは，Tessの父Johnだけが遠く離れた場所（左上）にあることに注目すべきである。

図11と併せて見ると, *d*もそれ以外の法助動詞に比べると外れた場所（左上）にある。Johnは*d*や*wo*, *ll*, *ca*といった省略形の語を使用する傾向にあり, John以外の人物は省略されていない法助動詞を使う傾向にあることが読み取れる。これは, Johnの個人語における一つ顕著な特徴であると考えられる。

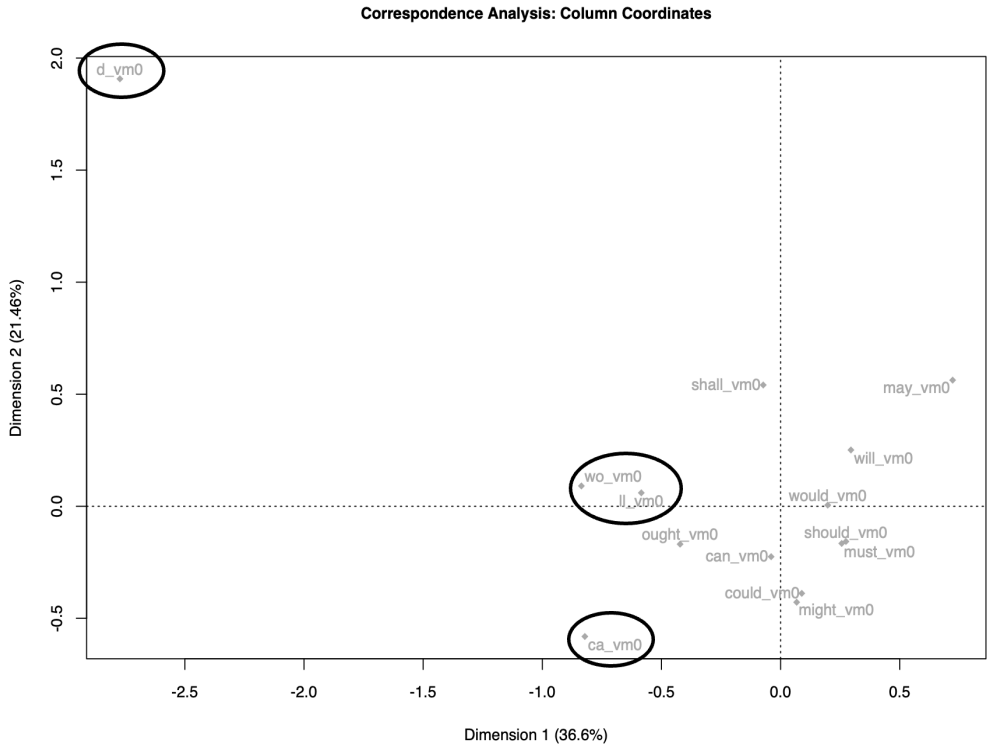

図 11 法助動詞の相対頻度による対応分析（第 2 アイテム）

以上をまとめると, 対応分析を用い, 個人語に出現した法助動詞（粗頻度）に注目して登場人物の会話部を分析した結果, 登場人物を男性と女性の2つのグループに分類することができた。また, 相対的に教養レベルの低い人物が法助動詞の省略形を使用する頻度が高いこともわかった。さらに, 個人語に出現した法助動詞（相対頻度）を考察した結果, 関係が近い人物同士の法助動詞の使用も似ていることが明らかになった。尤も, 個人語は常に身近な人に影響を及ぼされることから, 得られた結果はある意味では当然と言える可能性もあるが, 一方, 登場人物の中で, Johnの法助動詞の使用が特異な位置にあることが明らかになるなど, いくつかの興味深い知見も観察された。ただし, 言語データだけで言えることは限りがあり, これらの点についてさらに議論を進めるためには, 小説原文への考察を加えたより詳細な分析が必要となろう。

4. 4. Tess, Angel, Alec の発話における最頻出語の比較

次の作業では，既に述べたこと以外に, Tess, Angel, Alecの3人の個人語にある頻出語を中心として，登場人物間の関係性を調査・分析していきたい。ここではこの3人の頻出語（相対頻度）を降順で並び，上位55語を図12のワードクラウド（word cloud）で視覚化する。Word cloudでは，テクストに高い頻度で表れている語が大きいフォントで表示される。ワードクラウドから読み取れるのは, Tess, Angel, Alecの3人の特徴語である。

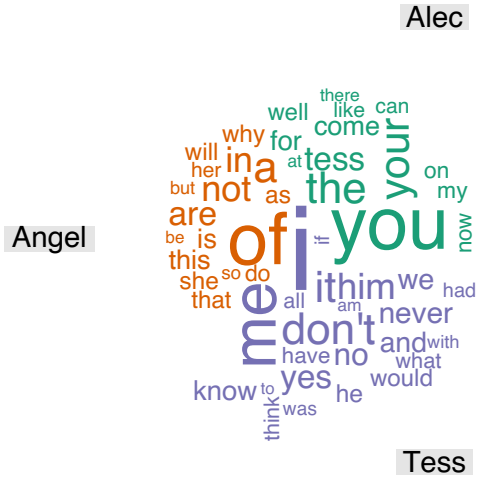

図 12 ワードクラウド

中央あたりに大きなフォントで表示されているのは*i*や*you*のような人称代名詞なので，まずは，この3人の頻出語にある人称代名詞の差を明らかにする。

図12でTessが*i*, *me*, *it*, *we*, *him*, *he*を他の2人より頻繁に使用しており, Angelは*she*と*her*を, Alecは人称代名詞*you*（と所有限定詞*your*, *my*）を多用している。コンコーダンスを考察すると, TessがAlecに対する発話での*him*と*he*はAngelを指すことが多く, TessがAngelに対する発話で*him*と*he*はAlecを指すことが多いのに対し，Angelの発話で*she*と*her*はほぼ全てTessを指す。会話で頻繁に三人称で互いのことを第三者に語ることが多いことから，この3人の感情が絡み合うことがわかり，物語の婉曲性・複雑性も反映されている。Alecの発話でも*tess*（Tess）が特徴語である。登場人物間の葛藤を発話部の語彙から読み取ることもできた。

また，TessとAlecの間で*i*と*you*の使用頻度で大きな区別があったが，その理由についても実際に使用されているコンテクストから分析してみた。結論から言うと, Tessは*i*を使って，自分の感情・考え・決断，自分の行動や状態に対する説明などを聞き手に伝えようとしている。*me*を使って，相手への依頼，他人の自分に対する言動やその言動が自分に与えた影響などを述べようとしている。さらに，Tessは*we*を多用し，他人と同じ立場に立ち,相手か第三者を自分の領域に受け入れる傾向にあり,より性格の優しい人物であることがわかる。それに対し, Alecが*you*, *your*, *my*のような相手と自分の間にきっぱりと境界線を引き，「私」と「あなた」を対立関係にするような言い方を好

む。より詳細にAlecの発話を見ると, Alecは主に*you*を使って, 自分の基準で聞き手に対して評価を下したり要望を伝えたりしている。性的に放埒で, Tessを凌辱した彼は, 会話部での用語のニュアンスもその攻撃的なイメージに相応しい。

次に注目したいのは, この3人の法助動詞の使用傾向である。Tessは*would*, Angelは*will*, Alecは*can*を相対的に高頻度で使用している。TessはAngelより法性の弱い法助動詞を使用しており, 性格がこの3人の中で最も従順的であることがわかる。最後は, ワードクラウドに現れたやや特殊な語について考えてみる。Tessの個人語で否定を表す語（*don't*, *never*, *no*）が上位に来ていることに注目に値する。Angelの個人語で*not*という否定辞が使用されているが, Alecの発話では否定の表現が相対的に稀である。Tessの個人語で否定の表現が数多くあることは, Tessの相手への抵抗感や彼女の性格で引っ込み思案のところを示している。

このように, 個人語の頻出語からでも登場人物の関係性や性格を観察することができた。

5. 終わりに

以上より, Hardyの代表作の一つである『テス』を対象として, TEIガイドラインに基づき, 文学作品テクストの構造化およびアノテーションを行った上で, 小説全体から会話部を抽出し, 登場人物間の対話ネットワークの考察および主要な登場人物の会話に用いられている語の頻度パターン解析を行った。このように, 本研究は技術的な制約を克服し, 会話部を対象とした量的研究の新しいアプローチを提案した。その結果, Hardyが『テス』の登場人物のキャラクタライゼーションを行う際に, 会話部に一定の工夫を入れたことがわかった。

RQ1（会話部の語彙による登場人物の分類の可否）については, 登場人物の語彙使用に違いがあると言える。キャラクターの性別や重要性による個人語の差が見られた。RQ2（登場人物の性格による個人語の差の有無）については, 会話で使われている語彙から, その登場人物の人物像が見えてくるという解答になる。法助動詞の使用に当たっては, 登場人物が会話部で*can*や*will*を好むか, *could*や*would*を好むかによってその人物の性格の強さや弱さを窺えることができた。また, 会話での頻出語から, その発話者は引っ込み思案の性格なのか, 自己中心のような人間であるのかもイメージがつくことが明らかになった。RQ3（人間関係による個人語の差の有無）については, 両者の強い結び付きがあると言える。クラスター分析と対応分析によって, 関係の近い人物同士や同じような設定の人物同士の個人語には似たようなパターンが潜在していることが確認できた。

本研究では, Hardyの代表作の『テス』を取り上げて会話部を抽出して分析したが, 地の文の語彙や小説全体の語彙的特徴についての検討は触れなかった。また, 試行的な研究でもあり, 品詞構成, 法助動詞と頻出語の3点のみに着目して考察を試みた。会話部分析の指標の汎用性が足りず, その他の分析指標の開発は途上である。今後は, より信頼性の高い計量手法の確立を目指している。さらに, Hardyの作品数を増え, 作品の出版年の変化につれて, Hardyの登場人物の特徴付けのパターンが終始一貫する特徴や傾向を示すものであるかどうかについての確認も行ってみると

良いであろう。

文献

Chapman R. (1990). *The Language of Hardy*. Basingstoke: Macmillan.

Ezawa K. (2015). Colours as Represented in *Tess of the D'Urbervilles* (1). 『東京成徳大学研究紀要—人文学部・応用心理学部』第 22 号, 129-137.

Ezawa K. (2016). Colours as Represented in *Tess of the D'Urbervilles* (2). 『東京成徳大学研究紀要—人文学部・応用心理学部』第 23 号, 141-152.

Johnson T. (1968). *Thomas Hardy*. London: Evans.

Hardy T. (1891). *Tess of the d'Urbervilles: A Pure Woman.* Toronto: Bantam Books.

Hardy T. *Tess of the d'Urbervilles: A Pure Woman*. https://www.gutenberg.org/cache/epub/110/pg110-images.html (参照 2022-05-26)

Howe I. (1967). *Thomas Hardy*. New York: Macmillan.

Imao Y. (2022). CasualConc (Version 3.0.4 beta) [Computer software]. Retrieved from https://sites.google.com/site/casualconcj/

Karsdorp F., Kestemont M., & Riddell A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Laird J. T. (1975). *The shaping of Tess of the d'Urbervilles*. Oxford: Clarendon Press.

Leech G. (1987). *Meaning and the English Verb*. Harlow: Longman.

Leech G., Short M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. New York: Pearson Longman.

Page N. (1973). *Speech in the English novel*. Longman.

Tabata T. (2002). Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English Corpus Linguistics in Japan*. Amsterdam: Rodopi, 165-182.

Tabata T. (2004). Differentiation of Idiolects in Fictional Discourse: A Stylo-Statistical Approach to Dickens's Artistry. In Hiltunen, R. and Watanabe, S. (Eds.), *Approaches to Style and Discourse in English*. Osaka: Osaka University Press, 79-106.

石川慎一郎(2005). 「詩の言葉・小説の言葉—コーパス処理に基づくトマス・ハーディのテクスト構成語彙の分析—」富山太佳夫, 加藤文彦, 石川慎一郎（編）『テクストの地平 ： 森晴秀教授古稀記念論文集』英宝社, 153-168.

石川慎一郎, 前田忠彦, 山崎誠［編］(2010). 『言語研究のための統計入門』. くろしお出版, 東京.

石田友梨, 大向一輝, 小風綾乃, 永崎研宣, 宮川創, 渡邉要一郎（編）(2022). 『人文学のためのテキストデータ構築入門　TEI ガイドラインに準拠した取り組みにむけて』. 文学通信, 東京.

田畑智司(2006). 「米国歴代大統領就任演説の言語を計る—多変量アプローチによるテキスト類型化試論—」大阪大学大学院言語文化研究科（編）『電子化言語資料分析研究 2005-2006』. 言語文化共同研究プロジェクト, 57-67.

橋本史帆(2019). 『トマス・ハーディの小説世界. 登場人物たちに描き込まれた国際事情と「グレート・ブリテン島」的世界』. 音羽書房鶴見書店, 東京.

深谷修代(2013). 「コーパスを用いた『白雪姫』の会話分析」. 『江戸川大学紀要』第 23 号, 67-90.

米倉綽(2009).「Thomas Hardy の英語と文体——その統語法と語形成——」. 『英語史研究会会報研究ノート』第 14 号, 15-30.