統計数理 (2000)
第 48 巻 第 2 号 311–326

特集「“ ことば ”新研究」
［原著論文］

多変量解析による文章の所属ジャンルの判別 ―論理展開を支える接続語句・助詞相当句を指標として―

慶應義塾大学* 村田 年

(受付 2000 年 3 月 31 日；改訂 2000 年 7 月 18 日)

要 旨

専門日本語教育における学習者にとって，論文に代表される論述文の論理構造の理解は不可欠であり，その理解には接続語句・助詞相当句が指標として役立つと考えられる．本論文では，論述文の論理構造を支える接続語句・助詞相当句を抽出する研究の一環として，5 ジャンル (経済学教科書，物理学論文，工学論文，文学作品，新聞社説) 計 290 編（14134 文) の文章における接続語句・助詞相当句 62 項目の出現率を調査し，以下の分析を行う．

(1) 5 ジャンル計 108 編 (新聞社説は 222 編から単純無作為抽出による 40 編) の資料を対象に単変量的解析を行った後，正準判別分析 (多変量解析の一手法) を用いて分析を行う．
(2) (1) の分析で分離が明確でなかった文学作品と新聞社説の全資料 (総計 236 編) を対象に (1) と同様の分析を行う．

上記の結果より，文章の所属ジャンルが，12 の語句項目によって，正判別率 84%という高率で判別されるとともに，各ジャンルを分離する語句項目ならびに論述的形式を持つ文章に共通する語句項目が選択された．

以上，限定された資料内ではあるが，異なるジャンルの文章を判別するために，(i) 接続語句，(ii) 助詞相当句が有効な指標であることが明らかとなった．

キーワード：専門日本語教育 (JSP), 文章の論理構造，接続語句，助詞相当句，ジャンルの判別分析，出現率．

1. はじめに

私たちが日常使っている「ジャンル」という語は，ある文章を「小説」「論文」「社説」というように分類するときにも，いくつかの文学作品を「随筆」「日記」「手紙」というように分類するときにも使われている．このことからもわかるように,「ジャンル」という語が意味するものは一様ではない．しかし,「ジャンル」という語が，ある特徴パターンを共通して持っている文章グループを意味することに異論を唱える人はいないであろう．その意

*国際センター：〒108–8345 東京都港区三田 2–15–45.

味で,「ジャンル」は，書き手一人一人の個人的な文体の特徴を超えたところに存在する概念であると考えられる．

金 (1999) は，私たちが文体 (文章の書き手の識別に関する何らかの特徴パターン) に関する素養を持っていれば，読んだ文章が散文であるか，論文であるか，記事であるか，そのジャンルを見分けることは困難ではなく，その理由として，私たちがそれぞれの形式 (パターン) に関し学習を行っているからだと述べている．この形式 (パターン) の学習は，母語の場合にはもちろんのこと，外国語学習の場合にも重要であり，特に学習時間が限られている場合にはその効率的な習得が求められる．

筆者が現在携わる外国人学習者に対する日本語教育においては，専門分野での学習・研究を目的とする日本語学習者は，短期間に論述的な文脈展開を持つ文章の理解や作成能力を要求されるため，文章指導の際に，文章がジャンルによって異なる表現形式を持つことを客観的かつ具体的に理解させ，その知識を文章作成に活かせるように指導していく必要がある．そのためには，文章のジャンルによって異なる特徴パターンの抽出と提示が不可欠となる．

これまでの先行研究の中には，接続詞等の文の連接に関わる語句の文章中における出現数に着目したものもあるが，相対出現頻度に変えていないなど，実証的かつ定量的アプローチが不足していたため，真の意味での文章間の比較は難しかったと言える．今後，専門分野における教材開発の推進，教育方法の改善を考えるとき，その根拠の一翼を担う基礎的研究として，統計的手法を用いた計量文章分析は大きな意義を持つと考えられる．

本研究の目的は，表層表現[注1)]としての接続語句と助詞相当句が文章の文脈展開に重要な役割を果たすと考え，ジャンルによる文章の特徴が，これらの接続語句・助詞相当句の使用傾向の違いに反映されるということを実証することである．また同時に，論述的な文脈展開を持つジャンルの文章に特徴的な接続語句と助詞相当句を抽出することも目的とする．本論文は，同一ジャンルに分類された文章には共通した特徴パターンがあるという仮定のもとに，具体的にどのような表現形式によって，そのジャンルが分類され得るのかということを，多変量解析を用いて分析した結果である．

2. 分析資料

2.1 分析に用いた文章資料

専門分野を志向する学習者は，研究上の必要性から短期間で日本語の論文の読解ならびに作成の能力を習得することが求められる．その学習者が，各自の専門分野でまず触れる日本語の文章は，通常，専門家による教科書あるいは概説書で，研究の焦点が絞られてくれば，論文を読むことになる．専門家によって執筆された教科書や論文の論述的形式の文章は，学習者にとって専門分野の論述文の重要な文章モデルになると考えられる．そこで，専門分野の論述的形式の文章の資料として，学習者の多い分野の一つである経済学から入門教科書を，理工学から科学技術論文 (物理学論文，工学論文) を選んだ．比較のための資料としては，経済学教科書と科学技術論文より論理的構成が弱いと思われる新聞社説と文学作品を選んだ[注2)]．

(a) 経済学教科書

『はじめての経済学』，岡田泰男 他 編，慶應義塾大学出版会 (1995)[注3)]

この教科書は，16 名の経済学者が各自の専門について一つの章を執筆する形で構成された入門教科書で，理論と実践に関する記述の割合がほぼ半々になっている．16 編総文数 1124 文．1 編当たりの平均文数 70.3 文．

(b) 物理学論文
『日本物理学会誌』の 1997 年第 52 巻の No. 1〜12 までの「最近の研究から」に掲載された論文を各号から 2 編ずつ選び，合計 24 編を資料とした．総文数 2243 文．1 編当たりの平均文数 93.5 文 (但し，数式・記号は除く).
(c) 工学論文
近年の学術雑誌論文 14 編を資料とした．専攻分野別に電気工学 6 編，機械工学 4 編，計算機科学 2 編，管理工学 2 編．総文数 1725 文．1 編当たりの平均文数 123.2 文 (但し，数式・記号は除く)[注4].
(d) 新聞社説
1996 年 12 月 1 日〜31 日までの四大紙の朝刊，夕刊の社説 (読売新聞 58 編，朝日新聞 55 編，毎日新聞 58 編，日本経済新聞 51 編) の総計 222 編 (総文数 6514 文) から単純無作為抽出法で選んだ 40 編を資料とした [注5]．1 編当たりの平均文数 29.3 文.
(e) 文学作品
近代文学の文豪 3 人，森鷗外，夏目漱石，芥川龍之介の短編作品から，14 編 (新潮文庫) を資料とした．総文数 2528 文．1 編当たりの平均文数 180.6 文.
森鷗外：『余興』『杯』『普請中』『百物語』『二人の友』
夏目漱石：『初秋の一日』『三山居士』『子規の画』『日記』『手紙』
芥川龍之介：『仙人』『蜃気楼』『トロッコ』『好色』

2.2 指標としての接続語句と助詞相当句

分析の指標として用いたのは，接続語句と助詞相当句である．その理由は，まず，文章の文脈展開に重要な役割を果たすと考えられる接続語句は，ジャンルによって使用傾向が異なり，特に論理展開が明示的な文章 (論文はその代表例) における使用頻度が高いと考えられること，また，助詞相当句についてもその一部が接続語句と重なっており，経験的にその使用傾向がやはりジャンルによって違いがあると考えたことによる．以下にその具体的な語句を挙げる.

2.2.1 接続語句の定義

本論文で言う接続語句とは，接続詞を中核とし，接続詞的機能を持つ語句，接続助詞，接続助詞的機能を持つ語句の総称である．たとえば，副詞「つまり」「たとえば」「むしろ」等は接続詞的機能を持つ語であり，連語「そのため」「そのうえ」「その結果」も同様に接続語句に含まれることになる．接続語句については，市川 (1978) の文の連接関係の基本的類型を基準とした.

本論文で用いる接続語句の項目を以下に挙げる.

＜接続詞・接続詞的機能を持つ語句＞
順接型：したがって，(それ) ゆえに，よって，そのため (に)，とすると，とすれば，としたら，その結果
逆接型：しかしながら，それにもかかわらず
添加型：その上 (に)，その上 (で)，と同時に
対比型：それに対して，(その) 一方 (で)，他方 (で)，むしろ，(その) 反面
同列型：すなわち，つまり，たとえば，とりわけ
補足型：ただし，なお

＜接続助詞・接続助詞的機能を持つ語句＞
から，つつ，ながら，ながら (も)，ので，ものの，〜にもかかわらず，〜ため (に)，〜上 (に)，〜上 (で)，〜のに対して，〜一方で，〜反面，〜(た/の) 結果，〜と同時に

2.2.2 助詞相当句の定義

日本語を表現レベルから見たとき，文の連接や文末表現等において，形式化した語や助詞・助動詞が複合し，全体で一つの機能を持つ独自の表現形式を形作っていることが多い．例えば「〜にとって」「〜はずだ」「〜ようにする」「〜ことになる」「〜どころか」などがそれに当たる．こうした表現は複合辞 (あるいは複合助辞) と呼ばれ，日本語教育においては，「文型」として教育の重要な柱の一つとなっている．複合辞は，語の枠を超えており，その定義については問題が残ると言えようが，本論文では，複合辞のうち，助詞相当の機能を果たすものを助詞相当句と呼ぶことにする．なお，助詞相当句の分類については森田・松木 (1989) を基準とした．

本論文で用いる助詞相当句の項目を以下に挙げる (活用形は省略)[注6)]．

格助詞相当：〜として，〜にとって，〜について，〜に関して，〜に対して，〜をめぐって，(〜から)〜にかけて，〜 によって，〜によれば， 〜によると，〜を通じて，〜において，〜にあたって，〜をはじめ，〜にわたって
係助詞相当：〜とは，〜というのは
副助詞相当：〜に限らず，〜だけでなく，〜ばかりでなく，〜のみならず
接続助詞　：〜上で，〜上に，〜まま (で)，〜に従って，〜に伴って，〜とすると，〜とすれば，〜としたら，〜としても， 〜ために，〜にもかかわらず，〜のに対して，〜とともに，〜と同時に，〜(た) 結果

3. 分析方法

3.1 語句の使用頻度調査

異なる 5 つの文章資料 108 編 (経済学教科書，物理学論文，工学論文，四大紙社説，文学作品) を対象として，2.2 で挙げた接続語句・助詞相当句の各語句の出現頻度を調べた．そして，一文当たりの出現頻度に換算し直した相対出現頻度を求めて，各語句の出現率とした．接続語句・助詞相当句については，形態が同じで意味機能を 2 つ以上持つものについては細分化して別の項目とした（例：「〜ために(目的)」と「〜ために(理由)」,「〜を通じて(媒介)」と「〜を通じて(範囲)」)．機能名は語句の後に(　)付きで記した．このように，各語句の持つ一つまたは複数の意味機能の同定を並行して行い，機能別に頻度を調べた．

なお，出現頻度を調べるにあたっては，用字の差異 (例：〜に従って/にしたがって，なお/尚)，語句の活用変化の形 (例：〜によって/により/による N (N=名詞)，に関して/に関する N) を同一視して同じ語句として扱い，全数調査を行った．最終的に接続語句・助詞相当句の総項目数は 62 となった．

3.2 分析

2.1 で挙げた (a)〜(e) の 5 つの文章資料を各々一つのジャンルに所属すると仮定して分析を行う．

(1) まず，単変量的に，個々の変量である接続語句・助詞相当句 62 語句のジャンルごとの分布の違いを検討するために，Kruskal-Wallis 検定 (以後 KW 検定) を用いた．

(2) 次に，62 語句のうちでジャンルの判別に特に有効な語句を抽出するために，多変量解析の一手法である正準判別分析 [注7] のステップワイズ法を用いて分析を行い，判別に寄与する語句を選択した [注8].

(1) は，検討した変数に関する基礎的な情報の提示を目的とするものであり，日本語教育の実践家にとって有用な結果である．一方，(2) による判別可能性の検討及び，論述文を特徴付けるのに有用な情報の選択が本研究の主眼である．

4. 分析結果

4.1 62 全変数の単変量的分布の比較

分析方法 3.2(1) により，各ジャンルで使用頻度の高い語句にどのような違いがあるのかを調べるために，まず，62 全語句の単変量的分布を KW 検定の平均ランクによって見ていく．

62 語句中，KW 検定結果が有意な 35 語句のジャンル別平均ランクをまとめたものを表 1 に示す ($p < 0.01$).

ここで，便宜的に平均ランク 70 以上の語句をそのジャンルに特徴的な語句と考える [注9]

表 1. 35 語句のKW検定結果．p <0.01．なお, a は正準判別分析で選択された語句．b はジャンルの中での出現率の中央値.

	語句<代表項目>	検定統計量	p 値	経済学教科書		物理学論文		工学論文		社説		文学作品	
				平均ランク	中央値 b	平均ランク	中央値 b	平均ランク	中央値 b	平均ランク	中央値 b	平均ランク	中央値 b
1	～によって（方法）a	71.54	0.000	56.41	0.02	79.27	0.06	94.50	0.14	34.24	0.00	27.75	0.00
2	～において a	68.25	0.000	76.73	0.05	70.35	0.03	89.64	0.07	31.39	0.00	32.93	0.00
3	～によって（理由）a	67.46	0.000	67.81	0.04	88.81	0.08	66.11	0.04	35.38	0.00	23.50	0.00
4	したがって a	61.65	0.000	76.84	0.03	77.38	0.02	71.11	0.02	33.90	0.00	32.00	0.00
5	すなわち	49.57	0.000	76.13	0.03	68.75	0.01	72.86	0.01	37.00	0.00	37.00	0.00
6	なお	45.50	0.000	54.81	0.00	58.90	0.00	87.11	0.02	44.00	0.00	44.00	0.00
7	～とは（定義）a	43.63	0.000	86.03	0.02	57.02	0.00	54.29	0.00	44.47	0.00	43.00	0.00
8	～ため [に]（理由）	39.13	0.000	65.59	0.03	77.31	0.05	72.39	0.05	35.53	0.00	39.04	0.01
9	～に関して a	36.31	0.000	60.44	0.01	59.19	0.01	88.32	0.03	44.47	0.00	34.50	0.00
10	～にとって	34.35	0.000	80.47	0.01	50.58	0.00	46.50	0.00	50.61	0.00	50.64	0.00
11	ので a	33.30	0.000	44.16	0.00	75.79	0.03	67.36	0.02	38.15	0.00	63.68	0.02
12	～について	31.54	0.000	72.41	0.04	65.15	0.04	70.57	0.04	47.19	0.00	20.61	0.00
13	～まま a	29.36	0.000	49.22	0.00	49.19	0.00	53.00	0.00	48.70	0.00	87.71	0.02
14	たとえば	27.89	0.000	75.72	0.02	61.29	0.01	67.75	0.01	44.30	0.00	34.50	0.00
15	～に基づいて	27.87	0.000	48.50	0.00	61.63	0.00	80.21	0.01	48.00	0.00	42.00	0.00
16	～として	27.11	0.000	80.91	0.11	55.98	0.05	69.64	0.07	46.85	0.04	28.50	0.00
17	つまり	26.76	0.000	71.94	0.01	66.58	0.01	52.57	0.00	43.96	0.00	45.89	0.00
18	～ため [に]（目的）	26.61	0.000	68.59	0.05	45.23	0.03	82.32	0.07	53.80	0.03	28.46	0.00
19	～に対して（対象）	25.35	0.000	64.22	0.04	68.81	0.03	72.21	0.04	44.65	0.00	29.29	0.00
20	～から～にかけて a	24.45	0.000	68.53	0.00	53.65	0.00	51.50	0.00	51.50	0.00	51.50	0.00
21	～[の] に対して	23.65	0.000	67.69	0.00	62.13	0.00	66.21	0.01	44.58	0.00	43.00	0.00
22	～を通じて（媒介）	22.87	0.000	74.56	0.01	53.50	0.00	54.00	0.00	49.88	0.00	47.00	0.00
23	一方で	21.49	0.000	68.38	0.01	68.75	0.01	55.96	0.01	47.24	0.00	33.50	0.00
24	～に伴って	21.06	0.000	66.00	0.00	63.81	0.01	66.00	0.01	45.19	0.00	40.50	0.00
25	～によって（対応）	20.80	0.000	63.34	0.00	59.73	0.00	69.61	0.01	46.04	0.00	44.50	0.00
26	～[た] 結果	20.25	0.000	69.50	0.01	66.33	0.00	55.79	0.00	46.72	0.00	38.00	0.00
27	～というのは（定義）	19.75	0.001	67.59	0.00	57.10	0.00	50.50	0.00	50.50	0.00	50.50	0.00
28	ただし	18.51	0.001	67.81	0.01	52.19	0.00	71.00	0.01	49.34	0.00	41.50	0.00
29	～によって（動作主）	17.94	0.001	72.38	0.01	60.13	0.00	59.93	0.01	48.20	0.00	37.00	0.00
30	しかしながら	16.76	0.002	57.81	0.00	62.92	0.00	63.29	0.00	47.50	0.00	47.50	0.00
31	から a	16.67	0.002	56.78	0.02	45.25	0.01	38.50	0.00	55.90	0.01	79.75	0.05
32	～にわたって a	16.25	0.003	66.56	0.00	59.46	0.00	60.64	0.00	47.53	0.00	46.00	0.00
33	ながら（同時）	16.16	0.003	53.59	0.00	58.92	0.01	53.71	0.01	44.05	0.00	78.61	0.03
34	～とともに	16.10	0.003	58.88	0.00	69.60	0.01	59.21	0.01	47.79	0.00	38.07	0.00
35	よって a	14.62	0.006	58.94	0.00	52.00	0.00	63.36	0.00	52.00	0.00	52.00	0.00

と，まず，経済学教科書に特徴的なものとしては,「〜において/における N」「したがって」「すなわち」「〜とは」「〜にとって/にとっての N」「〜について/についての N」「たとえば」「〜として」「つまり」「〜を通じて/通じた N（媒介)」「〜によって/により/による N（動作主体)」の 11 語句が挙げられる．次に，物理学論文に特徴的と考えられる語句は,「〜によって/により/による N（方法)」「〜において/における N」「〜によって/により/による N（理由)」「したがって」「〜ため［に］(理由)」「ので」の 6 語句である．工学論文に特徴的なものは,「〜によって/により/による N（方法)」「〜において/における N」「したがって」「すなわち」「なお」「〜ため［に］(理由)」「〜に関して/に関する N」「〜について/についての N」「〜に基づいて/に基づく N」「〜ため［に］(目的)」「〜に対して/に対する N（対象)」「ただし」の 12 語句である．

上記のように，文脈展開が明示的だと考えられる経済学教科書，物理学論文，工学論文の文章では，いくつかの接続語句・助詞相当句が共通して多用されていることがわかる．

文学作品で 70 を超えるものは,「〜まま」「〜ながら(同時)」「〜から」の 3 つで，理由を表す「から」以外は付帯状況を表す語句である．社説は，平均ランク 70 以上のものがなく，他の資料との比較において特徴的な語句がない．社説は，紙面の都合上，文字数制約が厳しいと想像され，1 編当たりの平均文章数も 29.3 文となっていて，他の 4 ジャンル (経済学教科書 70.3 文，物理学論文 93.5 文，工学論文 123.2 文，文学作品 180.6 文) に比べて非常に少ないため，接続語句の省略，助詞相当句の非用という可能性が強く，どの項目もあまり使われないことに特徴があるとも言えよう．

次に，表現意図の観点から，論理展開に重要だと考えられる「理由・原因」「帰結」「定義」「例示」の表現を中心にジャンル間の比較を行う [注10)]．

「理由・原因」の表現として，物理学論文，工学論文では，他のジャンルに比べて,「〜によって」「〜ため［に］」「ので」が多く用いられ，文学作品では「から」「ので」がよく用いられている．経済学教科書では,「〜によって」「〜ため［に］」が多く用いられているが,「ので」はあまり使用されず,「から」がより多く用いられる傾向が見られる．社説は「〜によって」「〜ため［に］」「ので」の 3 語句ともあまり用いられず,「から」が使用されていると言えよう．

「帰結」の表現としては，経済学教科書，物理学論文，工学論文ともに,「したがって」が多用され,「〜［た］結果/N の結果」も社説，文学作品に比べ，よく用いられている.「よって」は，経済学教科書と工学論文に現れていて，経済学教科書では理論関係の 2 つの章で用いられており，いずれも国民総生産，国民所得の式 (あるいは計算) の説明に用いられている．工学論文では計算機科学と管理工学の 3 論文で用いられており，そのうち 2 つは式の証明に関わる所で用いられ，残りの 1 つは，論文全体の結論を述べる所で用いられている.「よって」は，一般的に数学の証明の帰結部分で用いられる語句であり，その延長で用いられている傾向が強いと言えよう．

「定義」の表現としては,「〜とは」「〜というのは」が経済学教科書で多く使われている．これは，教科書というジャンルでは，用語の定義づけが他のジャンルに比べて必要性が高いためだと考えられる.「すなわち」「つまり」は，ある事柄の説明を言い換えによって，より一層明確化する働きがあると考えられ，経済学教科書，物理学論文，工学論文ともに「すなわち」「つまり」が社説，文学作品に比べてよく用いられていることがわかる．ただし,「すなわち」と「つまり」では,「すなわち」の方が論理的構成の強いジャンルとそうではないジャンル間の差異をより明らかに示していて，経済学教科書，物理学論文，工学論文ではかなり多用されているのに対して，今回の社説，文学作品の資料中 (社説 222 編と文学作品 14 編の合計 236 編) には 1 回しか現れていない．

「例示」の表現としては,「たとえば」が経済学教科書，物理学論文，工学論文で社説，文学作品に比べてよく用いられていることがわかる．

そのほか，経済学教科書，物理学論文，工学論文で社説，文学作品と比較して，よく用いられる表現として,「〜(の)に対して」「一方」などの「対比」表現,「〜について」「〜に対して」などの「対象・関連」表現のほか,「相関」を表す「〜に伴って」がある．

4.2 判別分析の結果とその有効性

上記 3.2 (2) により，4.1 の単変量による分析結果から論述文の特徴を記述するのに有用な情報の抽出を行うため，多変量による分析を行った．

量的データである 62 の接続語句・助詞相当句の出現率を説明変数とし，質的データである文章資料グループ (以下ジャンルと呼ぶ) を基準変数 (判別目的であるグループ) として，ステップワイズ法を用いた判別分析を行った結果，逐次的に 12 個の説明変数が予測式に組み込まれ，その手続き内で削除された変数もなく，5 つのジャンルの判別に有効な 12 の変数 (語句) が選択された．以下に選択された 12 語句を挙げる．

①「〜によって/により/による N（方法)」②「〜によって/により/による N（理由)」③「〜とは(定義)」④「〜に関して/に関する N」⑤「〜から〜にかけて」⑥「したがって」⑦「〜にわたって/にわたる N」⑧「よって」⑨「〜において/における N」⑩「〜まま」⑪「ので」⑫「から」

本研究では，文章資料の所属ジャンルは 5 つなので，4 つの判別関数が算出された[注11)]．判別関数によるグループの分離の程度を示す記述的指標として，ウィルクスの Λ[注12)] を用いることができる．また，推測統計的立場から Λ に基づく χ^2 検定で，関数の有意性の検定を行うことができる．本研究の場合，正規性の仮定が満たされないことから，推測統計的指標は，あくまで目安程度にとどめるべきであるが，これらに関する指標を示すと，表 2 の通りである．

Λ の値を見ると，関数 2 と関数 3 の値の間の差が大きく，関数 1 と関数 2 に相対的に判別に大きく寄与する情報が含まれていることを示唆している．

次に，判別空間におけるジャンル間の関係について検討する．選択された 12 語句による判別分析の結果から求められる判別関数平面での各文章資料 (個体) の判別得点とジャンルの重心をプロットしたものが図 1 である．ここでは，関数 2 までに多くの情報が含まれることから，関数 1 と関数 2 の平面におけるプロットのみを示す．また，表 3 には各ジャンルの判別空間における重心を示した．

図 1 では，5 つのジャンルのうち，経済学教科書，物理学論文，工学論文が文学作品と社説から大きく分離される一方，文学作品と社説の分離は明確ではないことが読み取れる．

表 3 を検討すると，関数 1 によって，経済学教科書，物理学論文，工学論文と，社説，文学作品が分離されていることがわかる．関数 2 では，経済学教科書が物理学論文，工学論文から分離され，関数 3 では物理学論文が工学論文から，関数 4 では文学作品が他のジャ

表 2. 判別関数の固有値等.

判別関数	固有値	寄与率	累積寄与率	p 値	Λ	χ 2 乗	自由度
関数 1	5.427	56.8	56.8	0.000	0.015	414.543	48
関数 2	2.719	28.5	85.3	0.000	0.096	231.278	33
関数 3	1.005	10.5	95.8	0.000	0.355	101.893	20
関数 4	0.403	4.2	100.0	0.000	0.713	33.373	9

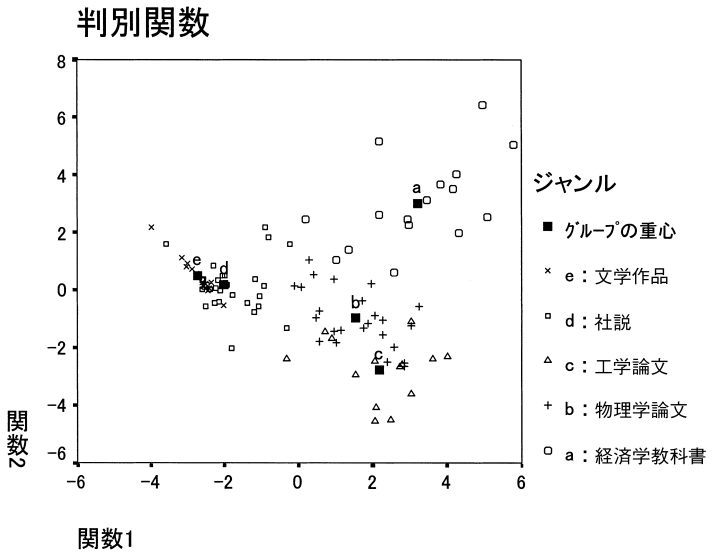

図 1. 判別分析による各文章資料とジャンルごとの重心のプロット.

表 3. 各ジャンルの判別空間における重心の関数.

ジャンル	関数 1	関数 2	関数 3	関数 4
経済学教科書	3.206	3.017	0.493	−0.002
物理学論文	1.553	−0.985	−1.586	0.117
工学論文	2.198	−2.758	1.649	0.070
社説	−2.025	0.176	0.089	−0.584
文学作品	−2.738	0.494	0.251	1.401

ンル，特に社説から分離されている．

判別分析においては，判別の可否は，普通，正判別率 (判別的中率) によって評価され，正判別率が高いほど，説明変数が基準変数の判別に有効に働くことを意味する．正判別率は，判別分析を行って判別規則を作成したその同じサンプルに対して判別規則を適用した場合に，サンプルの帰属する群がどの程度正しく判別されたかという割合を示す「見かけの的中率」によって簡便に評価することができる．

本研究における判別の可否を評価するために，見かけの的中率を算出するためのクロス集計表を表 4 に示す[注13)]．

表 4 のクロス集計結果から，正判別率は 84% $(14+19+13+37+8/108)$ という高い値となって，選択された 12 語句項目によって，5 つのジャンルの文章資料は十分判別が可能であることが検証された．また，誤判別は，文学作品と社説の間で起きている識別の誤りが主な原因であることもわかった．

次に，得られた判別関数に基づいて各ジャンル間の近さを総合的に評価するために，図 1 の各ジャンルのグループ間の重心の距離を求めると表 5 のようになる．

表 4. 正判別率を算出するためのクロス集計表 (12 語句を用いた判別関数による予測グループと実際のジャンルのクロス集計).

ジャンル	判別分析に基づく予測グループ					
	経済学教科書	物理学論文	工学論文	社説	文学作品	合計
a. 経済学教科書	14	1		1		16
b. 物理学論文		19	2	3		24
c. 工学論文		1	13			14
d. 社説				37	3	40
e. 文学作品				6	8	14
合計	14	21	15	47	11	108

表 5. 各ジャンルの重心間のユークリッド距離.

	経済学教科書	物理学論文	工学論文	社説
経済学教科書				
物理学論文	4.805			
工学論文	5.976	3.745		
社説	5.995	4.177	5.413	
文学作品	6.612	5.062	6.218	2.139

表 6. 選択された 12 語句の構造係数.

	語句	関数 1	関数 2	関数 3	関数 4
1	～において	∗0.363	−0.184	0.275	0.155
2	したがって	∗0.335	0.016	−0.135	0.070
3	～にわたって	∗0.147	0.077	0.021	−0.019
4	～によって（方法）	0.349	∗−0.513	0.179	0.097
5	～とは（定義）	0.232	∗0.324	0.147	−0.014
6	～から～にかけて	0.139	∗0.248	0.089	0.003
7	～によって（理由）	0.360	−0.110	∗−0.626	0.017
8	～に関して	0.153	−0.212	∗0.362	−0.177
9	よって	0.099	0.074	∗0.163	0.010
10	～まま	−0.138	0.047	0.093	∗0.684
11	ので	0.082	−0.196	−0.290	∗0.478
12	から（理由）	−0.142	0.147	0.066	∗0.444

∗ 有意な係数

この距離の差により，経済学教科書と文学作品に使用される接続語句・助詞相当句が最も大きく異なり，社説と文学作品で最も類似しているということがわかる．また，同じ論述的形式を持つ文章の間でも，経済学教科書，物理学論文，工学論文のそれぞれの間で使用される語句に違いがあるということも判明した．

最後に，正準判別分析で選択された 12 語句がどのジャンルの判別に有効かを見ていく．この考察のためには，構造係数と判別空間における各ジャンルの重心の関係を見ることが有効である．構造係数とは，正準判別関数と個々の変量との間の相関係数である．表 6 として，12 語句の構造係数を示す．

関数 1 では,「～において/における N」「したがって」「にわたって/わたる N」が経済学教科書，物理学論文，工学論文を他の 2 ジャンルから分離するのに有効であり，関数 2 では,「～によって/により/による N(方法)」「～とは(定義)」「～から～にかけて」が，経済学教科書を物理学論文，工学論文から分離している．関数 3 では,「～によって/により/に

よる N(理由)」「〜に関して/に関する N」「よって」が工学論文を物理学論文から分離するのに有効で，関数 4 では,「〜まま」「ので」「から」が文学作品を社説から分離していることがわかる．

以上の結果は，4.1 の単変量的検討の結果とも矛盾がないと言える．この点については 5. 総合考察でもう少し考察を加えることにしたい．

4.3 社説と文学作品の分離に関する分析結果

最後に，4.2 の分析で，分離が必ずしも明確ではない社説と文学作品のみを資料として，3.2 と同様に単変量，多変量の 2 つの方法を用いて，2 つのジャンルについて再分析を行い，検討を加える．

5 つのジャンルを前提とした 3.2 の分析では，便宜上，社説資料 222 編中 40 編を無作為抽出で選んで資料としたが，ここでは，分析結果が安定するように，4 種類の新聞の社説資料 222 編すべてと文学作品 14 編の合計 236 編を資料として分析を行った．

4.3.1 62 全変数の単変量的分布の比較

62 語句中，KW 検定結果が有意な 14 語句の資料別平均ランクをまとめたものを表 7 に示す．

5 つの資料間で平均ランクを比較したとき，一般に新聞四紙内での相互の差よりも，社説と文学作品の間での差が大きい場合が多い．4.2 の分析で選択された 12 語句以外に文学作品と社説を分離するのに有効な情報がこれらの語句項目に含まれていることを予期させる結果である．

4.3.2 判別分析結果とその有効性

62 の接続語句・助詞相当句の出現率を説明変数として，ステップワイズ法を用いた判別分析を行った結果，逐次的に 3 個の説明変数が予測式に組み込まれ，その手続き内で削除された変数もなく，5 つの文章資料グループの判別に有効な 3 つの変数 (語句) が選択された．以下に選択された 3 語句を挙げる．

①「ながら」②「ので」③「として」(順に表 7 の 2，1，9 の項目である)

表 7. 14 語句の KW 検定結果．p <0.01．なお，中央値とはジャンルの中での出現率の中央値である．

	語句<代表項目>	検定統計量	p 値	朝日		毎日		読売		日経		文学作品	
				平均ランク	中央値	平均ランク	中央値	平均ランク	中央値	平均ランク	中央値	平均ランク	中央値
1	ので	45.826	0.000	127.58	0.00	108.46	0.00	112.92	0.00	111.12	0.00	174.43	0.02
2	ながら	39.865	0.000	125.11	0.00	105.70	0.00	117.91	0.00	107.33	0.00	188.71	0.03
3	〜まま	24.221	0.000	122.12	0.00	103.59	0.00	118.06	0.00	115.58	0.00	178.50	0.02
4	〜のみならず	22.688	0.000	115.50	0.00	119.60	0.00	115.50	0.00	117.86	0.00	140.46	0.00
5	〜[と] 同時に	19.183	0.001	117.27	0.00	115.03	0.00	119.31	0.00	115.29	0.00	146.00	0.00
6	〜について	18.499	0.001	121.43	0.03	104.34	0.00	126.91	0.03	136.91	0.04	63.75	0.00
7	としても／とすれば／としたら	17.347	0.002	114.09	0.00	112.92	0.00	115.47	0.00	120.94	0.00	162.57	0.01
8	〜によって (理由)	16.064	0.003	111.36	0.00	136.87	0.00	108.50	0.00	123.54	0.00	93.50	0.00
9	として	15.007	0.005	123.15	0.03	101.10	0.00	142.34	0.04	114.34	0.03	88.64	0.00
10	〜ため [に] (目的)	13.652	0.008	114.64	0.03	105.92	0.03	137.16	0.04	127.46	0.05	75.86	0.00
11	つつ	12.979	0.011	131.24	0.00	111.97	0.00	121.22	0.00	108.29	0.00	121.43	0.00
12	から	12.192	0.016	111.30	0.02	130.12	0.03	104.20	0.00	117.09	0.00	163.04	0.05
13	〜だけでなく	11.327	0.023	107.92	0.00	116.62	0.00	131.09	0.00	122.95	0.00	99.50	0.00
14	〜によって (動作主)	10.379	0.034	126.33	0.00	115.48	0.00	128.86	0.00	108.02	0.00	95.50	0.00

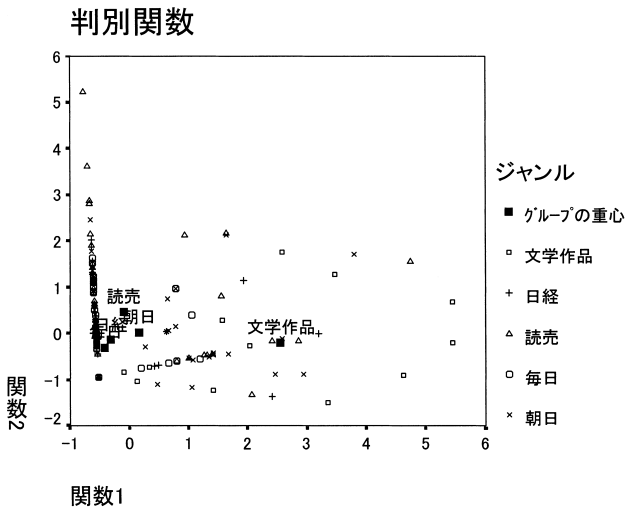

図 2. 判別分析による社説・文学作品の各資料と資料グループの重心のプロット.

表 8. 判別空間における社説グループと文学作品グループの重心.

文章資料	関数 1	関数 2
朝日	0.162	0.014
毎日	−0.411	−0.297
読売	−0.083	0.457
日経	−0.313	−0.141
文学作品	2.549	−0.203

表 9. 選択された 3 語句の構造係数.

	語句	関数 1	関数 2
1	ながら	∗0.713	0.096
2	として	−0.081	∗0.973
3	ので	0.658	−0.045

∗有意な係数

文章資料の所属するジャンルを仮に 5 つと考えたが，選択された変数が 3 であることから，ここでは 3 つの判別関数が算出され，検定の結果，判別関数 1 と 2 が有意であった (関数 1 : $\Lambda = 0.628,\ \chi^2 = 107.343,\ p < .000$，関数 2 : $\Lambda = 0.921,\ \chi^2 = 19.017,\ p = .004$，関数 3 : $\Lambda = 0.996,\ \chi^2 = 0.924,\ p = .630$)．また，累積寄与率は関数 1 のみで 84.5%であった．

ここで，選択された 3 語句による判別分析の結果から求められる判別関数平面での各資料の判別得点と文章資料グループの重心をプロットしたものを図 2 に示す．また，表 8 は，2 次元の判別関数平面における各文章資料グループの重心の値である．

表 8 から，関数 1 によって文学作品が社説から大きく分離されていることがわかる．関数 2 では，読売が若干他紙から分離されているが，あまり明確ではないことが読み取れる．

次に，構造係数と判別空間における各ジャンルの重心の関係から，選択された 3 語句がどの文章資料の判別に有効かを見ていく．表 9 に 3 語句の構造係数を示す．

関数 1 では,「ながら」が文学作品を社説から分離するのに有効であり，関数 2 では,「〜として」が，読売社説を他紙から分離していることがわかる．先の経済学教科書，物理学論文，工学論文を含めた判別分析では，関数 4 で「〜まま」「ので」「から」が文学作品を社説から分離するのに有効であったが，文学作品と社説のみでの分析では,「ながら」が効いていることがわかる [注14)]．

以上の 4.3.1 と 4.3.2 の結果は，分析の際に 5 つの文章資料 (朝日，毎日，読売，日経，文学作品) を 5 つのジャンルと仮定したことにそもそも無理があることを意味し，文学作品以外の 4 種類の社説資料は読売社説がわずかに分離されているものの，非常に近い位置に団

子状に固まっており，一つのグループを形成していることが確認できると言えよう．つまり,「社説」という「ジャンル」が，接続助詞・助詞相当句を説明変数として他のジャンルから確かに分離されていることを示していると考えられる．

5. 総合考察

まず，限られた資料の範囲内ではあるが，5 つのジャンル (経済学教科書，物理学論文，工学論文，四大紙社説，文学作品) を対象に，接続語句・助詞相当句の出現率を指標として，文章の帰属ジャンルの判別が可能であることが示された．また同時に，論述的な文脈展開を持つジャンルの文章 (経済学教科書，物理学論文，工学論文) に特徴的な接続語句・助詞相当句として,「〜において/における N」「したがって」をはじめ,「〜によって/により/による N（理由)」「すなわち」「〜ため［に］(理由)」「〜について/についての N」「たとえば」「〜に対して（対象)」「〜［の］に対して（対比)」「〜に伴って」などの語句が抽出された．

また，物理学論文，工学論文，経済学教科書に比べて論理的構成が弱いと思われる文学作品と社説では，5 ジャンルで判別分析を行った際には,「〜まま」「ので」「から」が判別に有効な語句として抽出され，文学作品と社説の 2 ジャンルだけで分析した際には,「ながら」が判別に効くことが明らかになった．文学作品に多用される語句として抽出された，付帯状況を表す「〜まま」「ながら」は，状況説明に用いられる語句であり，通常，論文では，論理展開の主要な流れにはほとんど現れない語句と言える．

これらの結果が示唆するところを，専門分野における日本語教育の実践面に照らして考えてみると，話し言葉を中心とした初級レベル以後の，書き言葉を中心とした中級レベル以上では，書き言葉としての文章のジャンルによる差異を，教材でより積極的に扱う必要があるのではないかと思われる[注15)]．従来の日本語教育の教科書の多くは，内容の多様性を重視したもの (随筆も社説も論文調の文章もというように様々なジャンルの文章が一冊の教科書に入っている) が多く，ある一つのジャンルに共通する表現形式 (パターン) の学習に焦点を当てたものはほとんどなかったと言える．しかし，専門分野での学習・研究を志向する学習者には，論文の読解・作成のための日本語力の習得が重要であり，社説や文学作品を学習していてもその能力が効率的に身につかない恐れがあるということである．したがって，専門分野に進むことを前提とした日本語教育では，多様な内容を持つ教材のほか，論文読解・作成に必要な表現形式を効率的に学習するため，論理的文章に共通する表現形式に重点を当てた教材の開発が必要だと思われる．

また，本研究の定量的な結果は，専門日本語教育の場で，専門分野の勉強だけでもかなり忙しい学習者に対する効率的な日本語コースデザインを作る際に，学習項目の選択，その導入・練習の順序等を決定するための有効な根拠になりうると考えられる．また同時に，学習者に対しては，日本語コースの最初の時点で，接続語句・助詞相当句に関する学習内容が定量的な成果を踏まえたものであり，研究活動に直結しているということを明らかにすることによって，学習者の学習意欲を高めることが期待できる．

本研究結果は，文章の内容に直接関わる語彙によってではなく，接続語句・助詞相当句という，文章構造そのものを支える，より汎用的な語句によって，文章の帰属ジャンルが判別され得ることを示唆しており，これらの情報の抽出に正準判別分析が有用であったと言える．

日本語教育の分野では，これまで接続語句・助詞相当句等を含む「文型」に関する先行研究の中で，文中における出現数に着目したものもあるが,「はじめに」で触れたように，相対出現頻度に変換していなかったり，文章の一部分のみを対象に検索が行われるなど，日

本語教育の専門家の教育経験や勘をたよりに，やや主観的に文型が抽出されていたと言えよう．従来の「文型」に関する知識が，こうした定量的かつ客観的な方法でも確認し得るということが，本研究のもたらした重要な知見である．

なお，本研究により，5 ジャンルの判別が可能であることは実証されたが，その内容が,「教科書」「論文」というジャンルの違いによるものか,「経済学」「物理学」「工学」という学問分野の違いによるものかは今後の検討課題である．

6. おわりに

本研究結果は，接続語句・助詞相当句の用い方がジャンルの特徴的パターンになり得ることを示している．各ジャンルで多用される接続語句・助詞相当句を具体的に抽出し，比較検討していくことによって，個人的文体を超えた，いわば社会的文体とも言えるジャンルによる文体の差異が明らかになるであろうという期待が強く持たれる．今後の課題としては，分析対象を広げて検証を行うとともに，日本語教育への応用として，ジャンル間で共通する接続語句・助詞相当句，共通しない接続語句・助詞相当句を踏まえて，専門分野に進む学習者の文章指導のために，論述文の文脈展開に必要な接続語句・助詞相当句の確定を目指したいと考えている．

本論文は，平成 11 年度言語処理学会年次大会で行った報告 (村田 (1999a)) をもとに，語句項目の再検討を行い，さらに工学論文と新聞社説 3 紙の資料を加えて新たに分析を行った研究成果である．なお，本研究は，文部省科学研究費基盤研究 (C)(2)(課題番号 11680317 研究代表者 村田年) の一部として行われたものである．

注

1) 黒橋・長尾 (1992) の表現を用いた．
2) 文章資料の選択については，社説のみが無作為性が高くなっているほかは，有意抽出である．これは，現存するすべての文章資料を母集団とすることが本研究では不可能であるという消極的な理由以外に，専門日本語教育における教材としての利用を考えた時，学習者のニーズに即した文章資料を対象とすることこそ教育的見地からは有意義だと考えられるからである．その意味で有意抽出は意味があるが，この対象から得られた結果の過度な一般化は慎まなければならない．
3)『はじめての経済学』(岡田泰男，野澤素子，村田年 編) は，留学生と日本人学生のために，慶應義塾大学経済学部の専門家 16 名と日本語教育の専門家 2 名との協力により編まれた入門教科書である．
4) 工学論文については，筆者が 1999 年度に担当した慶應義塾大学理工学研究科 (修士課程) の授業で，学生のニーズに従って取り上げた最近の論文 14 編を資料とした．具体的には，次の通りである．
   電気工学：『電子情報通信学会論文誌』4 編,『画像電子学会誌』2 編,
   機械工学：『日本機械学会論文集』2 編,『精密工学会誌』1 編,
   『精密機械工学会春季大会学術講演会講演論文集』1 編,
   計算機科学：『NICOGRAPH/MULTIMEDIA 論文コンテスト論文集』2 編,
   管理工学：『日本設備管理学会誌』1 編,『日本経営工学会誌』1 編
5) 結果の安定性を検討するため，本文中に示した資料とは別に，再度，単純無作為抽出で選び直した 40 編を用いて同様の分析を行い，最初の結果と比較したが，本質的に

変わらなかった．よって，以下の本文に述べる知見は，選出した 40 編に特殊の事情を反映したものとは言えない．なお，社説については，村田 (1999b) で日本経済新聞のみを対象として分析を行ったが，本結果と同様に文学作品と社説が近い位置で，その他のジャンルから分離され，社説ジャンルが安定していることがうかがえる．

6) 例えば,「によって」は,「により」「による＋ N (N=名詞)」の形を持つが,「によって」一語で代表する．

7) この用語は柳井・高木 (1986) による．今，多数の変数 $x = (x_1, x_2, \ldots, x_p)$ で特徴付けられる個体が g 個のグループに分かれているものとする．正準判別分析では，係数ベクトル a を用いて x の線型結合 $z = a'x$ を作る際，群の分離の程度を最大にするという基準で a を定める方法であり，求められた線型結合 z を本論文では，判別関数または単に関数と呼ぶ．z で各個体に与えられる得点，すなわち判別関数についての各個体の値を判別得点と呼び，これを判別関数の空間にプロットすることは，関数の意味を解釈する上で有用であり，図 1，図 2 で用いている．

8) 本研究で用いる変数は，いずれも特定の長さの文章中における単語の出現率であり，正規分布を仮定できる性格のものではない．なお，本研究で用いる正準判別分析は，解の導出自体に分布の仮定は入っていない (柳井・高木 (1986))．

9) ここでは，個体数 108 のデータに関するランクの総平均値は 54.5，標準偏差は 31.2 になるので，やや恣意的ではあるが，便宜上，当該グループの平均ランク 70 以上を特徴的語句と見なした．

10) ここで表現意図の分類に用いた「理由」「帰結」「定義」等の語は，日本語教育における「文型」の意味機能をグループ化するために，村田 (1998, 1999b) で用いたものである．

11) 判別関数は，変数の数 p，群の数 g としたとき，$\mathrm{Min}(p, g-1)$ だけ算出されるので，ここで求められる判別関数の数は 4 個となる．

12) ウィルクスの Λ の定義は柳井・高木 (1986) 参照．Λ は各判別関数ごとに算出される統計量で，記述統計的な観点からは，群の分離の悪さを示す指標である．0 と 1 の間の値をとり，1 に近いほど当該関数が群の分離に寄与する程度が低い，と解釈する．母集団分布に正規性を仮定できる場合には，Λ から判別関数の有効性に関する検定統計量を導くなど，推測統計的な指標として利用できるが，本研究では正規性を仮定することはできないので，表 2 に示した χ^2 検定の結果は目安程度の利用にとどめるべきである．

13) 各文章資料の数が異なるため，分析の際には資料の大きさに基づく事前確率を考慮に入れて，判別規則を構成する方法を用いた．

14) 単純無作為抽出による社説 40 編と文学作品 14 編の合計 54 編を対象として，ステップワイズ法による判別分析を行ったところ，ほぼ同様の結果が得られ,「ながら」がステップ 1 で選択されたことを付言しておく．

15) 姫野 他 (1998)，第 6 章 (村田)「ジャンルにおける型の違い」でも，この点の可能性について簡単に触れた．

謝　辞

統計分析については，統計数理研究所の共同利用登録制度により，同研究所の前田忠彦助手からご助言をいただきました．深く感謝いたします．

参 考 文 献

姫野昌子，小林幸江，金子比呂子，小宮千鶴子，村田　年 (1998). 『ここからはじまる日本語教育』，ひつじ書房，東京.

市川　孝 (1978). 『国語教育のための文章論概説』，教育出版，東京.

金　明哲 (1999). 日本現代文における書き手の特徴情報，人文学と情報処理，**20**，64–71.

黒橋禎夫，長尾　眞 (1992). 表層表現中の情報に基づく文章構造の自動抽出，自然言語処理，**1**(1)，3–20.

森田良行，松木正恵 (1989). 『日本語表現文型』，アルク，東京.

村上征勝，金　明哲 (1998). 『数量的分析編』，講座 人文科学研究のための情報処理，尚学社，東京.

村田　年 (1998). 異なるジャンルの文章における文型の出現傾向の相違──論述文を支える文型の確定を目指して──，日本語教育学会秋季大会予稿集, 165–171.

村田　年 (1999a). 接続語句・助詞相当句による文章の所属ジャンルの判別──多変量解析法を用いて──，言語処理学会第 5 回年次大会予稿集，213–216.

村田　年 (1999b). 論述文を支える文型の基礎的研究──多変量解析によるジャンル判別に有効な文型の抽出──，専門日本語教育研究，**1**，32–39.

柳井晴夫，高木廣文 (編) (1986). 『多変量解析ハンドブック』，現代数学社，京都.

Identify a Text's Genre by Multivariate Analysis —— Using Selected Conjunctive Words and Particle-phrases ——

Minori Murata

(International Center, Keio University)

It is quite important for advanced students of Japanese-language for specific purposes to understand the underlying logical structure of the text. Since the logical structure will enhance an ability to read and write technical papers.

Such items as the conjunctive words (i.e. the words which function as a conjunction in a sentence: Setsuzoku-goku) and particle-phrases (i.e. the phrases which function as a particle in a sentence: Jyoshi-sootoo-ku in Fukugo-ji) can provide important clues for understanding the logical structure of the text. The ultimate goal of this study is to clarify the logical structures of the technical texts in Japanese by focusing on the functions of conjunctive words and particle-phrases.

As a step toward achieving this objective, we chose 290 samples (14134 sentences in total) of five genres. Those five genres are (i) an introductory economics textbook, (ii) papers of the Journal of the Physical Society of Japan, (iii) papers of science and technology, (iv) editorial articles of 4 kinds of newspapers, and (v) modern novels. We counted the rate of appearance (per sentence) of the 62 selected conjunctive words and particle-phrases of each sample. The analysis was conducted in the following two steps,

(a) We first examined univariate distribution of the above 62 items and then applied the canonical discriminant analysis to 108 samples ((i) 16 samples (ii) 24 samples (iii) 14 samples (iv) 40 samples selected by random-sampling out of 222 (v) 14 samples).

(b) Secondly we applied the same method to 236 samples by use of all of the editorial articles of 4 kinds of newspapers (222 samples), and modern novels (14 samples) which were not well distinguished in the first step.

According to the result obtained in (a), these genres are classified with 12 conjunctive words and particle-phrases (out of 62) at a high apparent correct classification rate (84%). Following to the result obtained in (b), the words which distinguished 2 genres (i.e. (iv) and (v)) were clearly selected. These results indicate the existence of common conjunctive words and particle-phrases both in texts having an explicit logical structure (as in (i), (ii) and (iii)) and in texts having an implicit logical structure (as in (iv) and (v)).

Key words: Japanese for specific purposes, logical structure of a text, text's genre, canonical discriminant analysis, conjunctive words (Setsuzoku-goku), particle-phrases (Jyoshi-sootoo-ku in Fukugo-ji), rate of appearance per sentence.