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—

村田　年
(慶應義塾大学　国際センター)

Ⅰ. はじめに

筆者が現在携わる外国人学習者に対する上級段階の日本語教育においては、文章指導の際に、文章がジャンルによって異なる表現方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[1)]。特に、専門分野における学習・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習得することが要求されるため、その表現方式を効率的に指導することが求められる。

本研究では、表層表現[2)]である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し、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を目的とする。

Ⅱ. 分析に用いた資料

専門家によって執筆された教科書の論述的形式の文章は、専門分野を志向する学習者にとって、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の入門教科書と物理学論文を選んだ。比較のための資料としては、経済学教科書と物理学論文より論理的構成が弱いと思われる文学作品と新聞社説を選んだ。

(1) 経済学教科書

『はじめての経済学』岡田泰男・野澤素子・村田年編　慶応義塾大学出版会 (1995)

この教科書は、16名の経済学者がそれぞれ自分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。

(2) 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。

(3) 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、現代語版）を資料とした。総文数2528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』

(4) 新聞社説

日本経済新聞の1996年12月1日〜31日までの朝刊、夕刊の社説51編を資料とした。12月を選んだのは、他の月に比べて、その年1年の主な出来事に言及した記事が多く、総括的傾向が強かったためである。総文数1201文。

Ⅲ. 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その使用傾向がジャンルによって異なると考えられること、また、助詞相当句はその一部が接続語句と重なっている上、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られることによる。

(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（参考文献2）。

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) 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」「〜どころか」「〜はずだ」「〜ようにする」「〜ことになる」などがそれに当たる。こうした表現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。複合助辞は、語の枠を超えており、その基準についても定説がなく、定義については問題もあるが、本論文では、複合助辞のうち、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ほか (1989) を基準とした（参考文献3）。

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[3)]）。

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、～まま（で）、～に従って、～とすると、～とすれば、～としたら、～ために、～にもかかわらず、～のに対して、～とともに、～と同時に、～上に、～（た）結果

Ⅳ．分析資料の作成と分析方法

（1）分析資料の作成

異なる4つのジャンル（経済学教科書、物理学論文、文学作品、日経社説）の105編の文章を資料として、IIIで挙げた54の接続語句・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値を求めて、各語句の出現率とした。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）、語句の活用変化の形（例：～によって／により／によるN）を同一視して同じ語句として扱い、全数調査を行った。また、接続語句・助詞相当句の語句のうち、意味機能を二つ以上持つものについては細分化して（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）」）、各機能別に頻度を調べた。

（2）分析方法

1．Ⅳ（1）で作成した資料の54の接続語句・助詞相当句のうち、ジャンルの判別に特に有効な語句を確定するために、多変量解析の一手法である判別分析[4]のステップワイズ法[5]を用いて分析を行い、判別に寄与する語句を選択した。各ジャンルの資料数が異なるため、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。

2．個々の変量である接続語句・助詞相当句のジャンルごとの分布の違いを単変量的に検討するために、kruskal-Wallis検定（以後KW検定）[6]を用いた。

Ⅴ．分析結果とその考察

Ⅴ－1．判別分析の結果とその有効性

上記Ⅳの分析方法1により、量的データである54の接続語句・助詞相当句の出現率を説明変数とし、質的データであるジャンルを基準変数（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11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、4つのジャンルの判別に有効な11の変数（語句）が選択された。以下に選択された11語句を挙げる。

①「～によって/ により/ によるN（理由）」②「すなわち」③「ので」④「ながら」⑤「として」⑥「というのは（定義）」⑦「他方」⑧「～によって/ により/ によるN（方法）」⑨「～（の）もとで」⑩「にわたって/ にわたるN」⑪「にとって」

算出された判別関数の統計的有意性はp値によって判定される（判別関数の有意性の検定）。p値は、5％（あるいは1％）を境界として、p<.05（あるいはp<.01）の場合に統計的に意味のある関数であると判断される。本研究では、文章資料の所属するジャンルが4つなので、最大3つの判別関数が算出された。その3つの判別関数は、検定の結果、すべて高度に有意であった（関数1：$\Lambda=0.045$，$\chi^2=299.994$，p<．000，関数2：$\Lambda=0.228$，$\chi^2=142.803$，p<．000，関数3：$\Lambda=0.653$，$\chi^2=41.147$，p<．000，）。

ここで、選択された11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とジャンルの重心をプロットしたものを**図1**に示す。4つのジャンルのうち、経済学教科書、物理学論文は、他の2つのジャンルからきれいに分離され、文学作品と日経社説は近い位置ではあるが分離されていることが読み取れる。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。本研究の判別分析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1**に示す。表1を見ると、文学作品が日経社説に誤判別されることがあるものの、見かけの的中率は87％（7+15+49+20/105）という非常に高い値となり、この判別的中率の高さから、選択された11語句によって4つのジャンルは十分識別が可能であると考えられる。

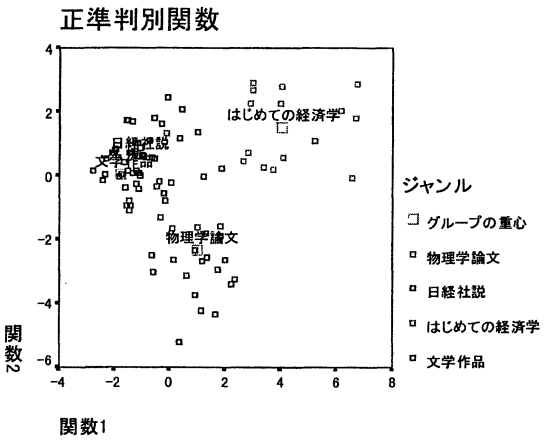

図1．判別分析による各資料とジャンルの重心のプロット

表1．見かけの的中率を算出するためのクロス集計表

ジャンル	判別分析に基づく予測グループ				
	文学作品	経済学教科書	日経社説	物理学論文	合計
文学作品	7		7		14
経済学教科書		15	1		16
日経社説	1		49	1	51
物理学論文			4	20	24
合　計	8	15	61	21	105

＊ 11語句を用いた判別関数による予測グループと実際のグループ（ジャンル）のクロス集計

次に、判別空間におけるジャンル間の関係について見ていく。ジャンル別の各グループ重心の関数を**表2**に示す。表2では、関数1によって、経済学教科書が他の3ジャンルから大きく分離され、関数2によって物理学論文が、関数3によって文学作品が分離されている。ただし、文学作品と社説については、図1においてそのジャンルの重心が非常に近い位置にあり、表1 のクロス集計結果からも分かるように、誤判別の主要な原因が文学作品と日経社説間における識別の誤りであることがわかる。ここで、Ⅳ（2）分析方法1によって選択された11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、構造係数と判別関数空間における各ジャンルの重心の関係を見る必要がある。構造係数とは、判別関数と各変数（各

語句）との間の相関係数であり、各変数（語句）の判別関数との関係の強さを表すものである（構造係数の利用法については参考文献5参照）。**表3**として11語句の構造係数を示す。関数1では、「すなわち」「として」「というのは」「他方」「にわたって/わたるN」が経済学教科書を他の3ジャンルから分離するのに非常に有効で、関数2では、「～によって／により／によるN（理由）」「ので」「～によって／により／によるN（方法）」「～[の]もとで」「～にとって」が、物理学論文を文学作品と日経社説から分離するのに有効で、関数3では「ながら」が文学作品を日経社説から分離するのに有効である。

表2．ジャンル別の各グループ重心の関数

ジャンル	関数1	関数2	関数3
文学作品	-1.783	-0.025	1.706
経済学教科書	4.060	1.507	0.244
日経社説	-1.261	0.625	-0.473
物理学論文	1.013	-2.348	-0.152

表3．選択された11語句の構造係数

	語　句	関数1	関数2	関数3
1	すなわち	*0.411	-0.041	0.087
2	として	*0.356	0.014	-0.058
3	というのは	*0.216	0.105	0.080
4	他方	*0.189	0.102	0.073
5	にわたって/にわたるN	*0.111	0.038	-0.084
6	によって/により/によるN（理由）	0.322	*-0.497	-0.260
7	ので	0.063	*-0.455	0.345
8	によって/により/によるN（方法）	0.148	*-0.377	-0.256
9	（の）もとで	0.122	*0.216	-0.129
10	にとって	0.169	*0.198	0.050
11	ながら	-0.037	-0.060	*0.785

＊有意な係数

最後に、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4**のようになる。この距離の差により、文学作品と経済学教科書に使用される接続語句・助詞相当句が最も大きく異なり、文学作品と社説で最も類似しているということがわかる。また、論述的形式の文章である経済学教科書と物理学論文の間でも、使用される語句に違いがあるということも判明した。

表4．ジャンル別各グループ重心間のユークリッド距離

	文学作品	経済学教科書	日経社説	物理学論文
文学作品				
経済学教科書	6.1727			
日経社説	2.3235	5.4411		
物理学論文	4.0514	4.9297	3.7567	

＊＼を境に対称である。

V－2．54全変数の単変量的分布の比較に基づく考察

次にⅣ(2)分析方法2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、54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。54語句中、KW検定結果が有意な43語句のジャンル別平均ランクをまとめたものを**表5**に示す。平均ランクが70以上の語句をそのジャンルに特徴的な語句と考えると、まず、経済学教科書に特徴的なものとして、「したがって」「～において/におけるN」「すなわち」「～とは」「～として」「たとえば」「～にとって」「ただし」「～によって/により/によるN（動作主体）」「～[の]もとで」「～を通じて/通じたN（媒介）」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、「したがって」「～によって／により／によるN（理由）」「～において／におけるN」「すなわち」「ので」「～によって／により／によるN（方法）」「～ため[に]（理由）」「～に対して/対するN（対象）」「～ため[に]（目的）」である。このように、文脈展開が明示的だと考えられる経済学教科書と物理学論文の文章では、「したがって」「～において／におけるN」「すなわち」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、「～ながら」「～まま」「～から」の3つで、そのうち2つが付帯状況を表す語句である。日経社説は、平均ランクが60以上のものがなく、他の資料との比較において特徴的と言える語句がないことがわかる。社説は、文字数の制約が厳しく1編当たりの平均文章数も23.6文で、他の3ジャンル（経済学教科書70.3文、文学作品180.6文、物理学論文93.5文）に比べて非常に少ないことから、接続語句の省略、助詞相当句の非用という可能性も考えられよう。このほか、物理学論文では理由を表すために、「～によって」「ので」「～ため[に]」が多く用いられているが、文学作品では「から」が用いられるというように、同じ意味機能を持つ語句でも、多用されるジャンルが異なることがわかる。

上記の単変量的に抽出した語句をV－1の多変量解析によって選択された11語句と比較すると、11語句中8語句が各ジャンルにおいても特徴的な語句として抽出されていることがわかる（経済学教科書：「すなわち」「～として」「にとって」「～[の]もとで」物理学論文：「～によって（理由）」「ので」「～によって（方法）」文学作品：「ながら」）。したがって、これら8語句は出現頻度も高く、ジャンルの判別に非常に有効な語句群であると言えよう。また、「～というのは（定義）」「他方」「～にわたって／にわたるN」は、そのジャンル別の平均ランクの比較から、文学作品、社説にはあまり出現せず、経済学教科書と物理学論文で使われ、経済学教科書でより多用されていることによって、経済学教科書と物理学論文を判別するのに効いていると考えられる。

以上の結果により、4つのジャンル（経済学教科書、物理学論文、文学作品、日経社説）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、文脈展開が明示的であると考えられる経済学教科書と物理学論文では、共通する特徴的な接続語句・助詞相当句が抽出できるとともに、多用される語句に差異があることもわかった。また、同じ意味機能を持つ語句でもジャンルによって使用傾向が異なることが明らかとなった。

Ⅵ．おわりに

ジャンルによって多用される接続語句・助詞相当句が違うということは、文章のジャンルによって、文の連接関係ならびに助詞の用い方についての表現方式が異なることを意味すると考えられる。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、比較していくことによって、文の連接関係ならびに助詞に関係する「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・助詞相当句を踏まえて、上級レベルの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。

表５．43語句のＫＷ検定結果

		語　句	検定統計量	p値	平均ランク			
					文学作品	経済学教科書	日経社説	物理学論文
	1	したがって	62.340	0.000	35.50	77.13	38.15	78.69
*	2	～によって/より/よるＮ（理由）	59.988	0.000	28.00	68.34	39.58	85.88
	3	～において/におけるＮ	54.104	0.000	37.71	80.91	37.78	75.65
*	4	すなわち	52.104	0.000	40.50	76.63	40.50	71.10
	5	～とは（定義）	48.684	0.000	43.00	84.09	44.18	56.85
*	6	ので	46.842	0.000	66.89	49.66	38.14	78.71
*	7	～によって/より/よるＮ（方法）	45.562	0.000	30.39	60.38	42.40	83.79
	8	～ため[に]（理由）	39.953	0.000	45.39	68.88	37.75	79.27
*	9	として	33.098	0.000	33.64	85.94	43.36	62.81
	10	たとえば	31.608	0.000	38.50	76.22	44.49	64.06
*	11	ながら	30.291	0.000	80.00	56.50	40.33	61.83
*	12	～にとって	28.444	0.000	48.36	76.97	48.91	48.42
	13	つまり	26.839	0.000	45.21	69.81	44.29	64.83
	14	ただし	26.741	0.000	46.00	72.13	47.12	56.83
	15	～[た]結果/Ｎの結果	25.738	0.000	39.50	69.28	45.31	66.35
	16	～によって/より/よるＮ（動作主体）	24.838	0.000	40.00	72.78	45.98	62.31
*	17	（の）もとで	24.687	0.000	43.50	74.94	52.22	45.58
	18	～まま	22.090	0.000	81.14	46.81	50.06	46.96
	19	～について/ついてのＮ	21.031	0.000	19.46	63.31	56.29	58.69
	20	～に対して/対するＮ（対象）	20.391	0.000	31.96	64.50	46.98	70.40
	21	～を通じて/通じたＮ（媒介）	19.918	0.000	45.50	71.56	49.89	51.60
*	22	というのは	19.007	0.000	49.00	65.63	49.00	55.42
	23	しかしながら	18.792	0.000	48.00	58.03	48.00	63.19
	24	なお	18.608	0.000	48.00	58.31	48.00	63.00
	25	[～から]～にかけて	18.531	0.000	49.50	65.97	50.57	51.56
	26	～に従って（根拠）	17.753	0.000	48.00	61.56	48.00	60.83
	27	～方[で]	17.448	0.001	34.00	64.81	48.65	65.46
	28	とともに	17.118	0.001	39.46	58.44	47.47	69.02
	29	～を通じて/通じたＮ（範囲）	17.010	0.001	51.50	61.34	51.50	51.50
	30	と同時に	15.375	0.002	60.04	65.03	46.12	55.50
	31	～[の]に対して（対比）	15.343	0.002	43.50	65.59	48.09	60.58
*	32	他方	14.830	0.002	49.50	62.94	49.50	55.85
	33	～に関して/関するＮ	14.199	0.003	38.00	63.78	48.92	63.23
	34	～ため[に]（目的）	13.492	0.004	45.39	68.88	37.75	79.27
*	35	～にわたって/わたるＮ	13.216	0.004	45.50	64.91	48.91	58.13
	36	よって	11.232	0.011	52.00	58.56	52.00	52.00
	37	反面	11.232	0.011	52.00	58.56	52.00	52.00
	38	とりわけ	10.604	0.014	49.50	62.25	52.71	49.50
	39	から	10.154	0.017	73.46	52.75	52.21	42.92
	40	～によって/より/よるＮ（対応）	10.122	0.018	45.00	61.88	49.53	59.13
	41	～としても/とすれば/としたら	9.074	0.028	62.57	57.94	45.38	60.31
	42	～をめぐって/めぐり/めぐるＮ	8.232	0.041	47.50	60.31	52.90	51.54
	43	～に基づいて/基づき/基づくＮ	7.918	0.048	43.50	49.56	52.56	61.77

本研究は、文部省科学研究費奨励研究Ａ（課題番号09780208 研究代表者 村田年「論述文を支える文型の研究とその日本語教育への応用」）の補助を受けて行った研究成果の一部である。

謝辞：統計分析については、統計数理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。また、資料作成のための検索プログラムに関しましては、慶應義塾大学理工学研究科原田賢一研究室の関洋平氏、九十九章之氏に御協力いただきました。御協力に感謝いたします。

＜注＞

1）表現方式の違いとしての文体の差異に関する指導の方針については、参考文献６の第2章（村田分担執筆）を参照されたい。

2）黒橋禎夫・長尾真（1992）の表現を用いた（参考文献１）。

3）例えば、「によって」は、「により」「による＋名詞」の形を持つが、「によって」一語で代表する。

4）判別分析法とは、基準変数がグループへの所属を表す質的データで、説明変数が量的データであるとき、各ケースの観測された特性に基づいて、所属グループの予測モデルを構築する際に有効な多変量解析の一手法である（参考文献４参照）。

5）ステップワイズ法とは、各変数毎にその変数が判別に寄与するかどうかをチェックしながら逐次的に変数を投入したり除去したりする手続きを繰り返して最終的な判別関数を求める方法である。変数の投入または除去に関する判定基準には、Wilksのラムダの統計量と呼ばれる指標を用いた。

6）ＫＷ検定とは、母集団における分布の位置の差を検出する統計的検定方法の一つで、各資料の全資料中の順位（ランク）に基づいて検定統計量が算出される。分布の位置の差を確認するためには、各群の平均ランクを参照すればよい。

＜参考文献＞

1．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』vol.1 No.1 pp.3-20

2．市川孝（1978）『国語教育のための文章論概説』教育出版

3．森田良行他（1989）『日本語表現文型』アルク

4．柳井晴夫他（1986）『多変量解析ハンドブック』現代数学社

5．前田忠彦（1995）「正準判別分析による学部の判定ー進路適性診断への応用ー」『HALBAUによる多変量解析の実践』現代数学社 pp.63-80

6．姫野昌子他（1998）『ここからはじまる日本語教育』ひつじ書房

7．村田年（1995）「上級日本語教育の方法を探る２ー文型教育から進める教材開発ー」『日本語と日本語教育』24慶応義塾大学日本語・日本文化教育センター紀要pp.17-37

8．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ー論述文を支える文型の確定を目指してー」平成10年度日本語教育学会秋季大会予稿集pp.165-171