国語・国文学論文におけるアカデミック性判断の指標

大妻国文　第43号　2012年3月

国語・国文学論文における アカデミック性判断の指標

中　尾　桂　子

概要：文系の論文ではどのような表現があることでアカデミックな性質が高まると言えるのか。村田（2007）は，理系と文系，論文と新聞・小説という文章ジャンルの区別に接続表現が寄与し，文系のみに特徴的に使用される接続表現を特定しているが，接続表現以外にも論述展開に影響すると考えられる語は多い。また，そのような語の中には，文系の下位分野の差が明確に現れる語があるのではないか。文系論文のうち，国文学系と国語学系の論文の共通利用語やその中の有意差のある語を比較して特定した，接続詞，動詞，助詞相等語句，文末表現等を指標に，国文学系と国語学系の論文の判別分析を行なった。結果，村田（2007）で文系特定に寄与するとされた接続表現は文系内部の下位類には大きく寄与するものではないが，文末表現のうち，「～か。」「～になる。」「～う。」は区別の指標になると考えられた。このことから，アカデミックな表現を用いて文系論文の下位分野を区別する可能性が考えられる。

キーワード：国文学，国語学論文，アカデミック性，χ^2検定，判別分析

1. はじめに

本稿では，国文学系と国語学系の論文の語彙で共通利用語彙の中の使用差のある語を比較し，文系論文の下位分類の指標となりそうな語を調べ，アカデミック性判断の指標として学生の論文指導への応用の可能性について考察する。

また，先行研究で文系ジャンルの特定に寄与するとされた接続詞（いわゆる品詞として「接続詞」とされる語のみ）や，接続助辞等の語句や文の接続を行なう助詞相当語句（以下，両方合せて「接続表現」）が，文系論文の下位分類においても判別の指標にできるか。さらに，接続表現以外にも分野の差が確認できそうなものがないかについて調べ，それを通して，ジャンル差特定につなが

る語からアカデミック性判断の観点について考察してみたい。以下2観点を今回の課題とする。

RQ1：文系論文のうち，国語学（現代語）と国文学（近現代）にも，村田（2007）で検出された接続表現の差が見られるか

RQ2：接続表現の他にジャンル差判断に利用できる語句がないか（文末形態，品詞比率，格助詞などはどうか）

2. 本研究の目的と方法

2.1. データ

本稿では，文系論文のうち，日本の文学を扱う分野と，日本語の文法現象を扱っている分野を題材として取り上げる。ここでは，便宜上，それぞれを，国文学系論文，国語学系論文と呼び分ける。

今回，国文学系論文とするものは，時代区分やカテゴリー等を問わず，日本の文学を分析するものを全て対象とする。また，国語学系も同様に，日本語の分析を行なうものを全て対象とし，語彙，意味，統語，音声といったカテゴリーや言語学，日本語学等の差を区別しない。

各ジャンルの論文は，国立情報学研究所の学術情報ナビゲータ［サイニィ］（以下CiNii）で公開されている無料アクセスが可能な学術論文PDFをランダムに採取して利用する。発表年度は考慮せず，一著者一論文を優先採用する。

調査対象論文の内訳は，国文学系が49本，国語学系が49本で，文の数は，それぞれ10,546文と10,401文である。

国語・国文学論文のそれぞれの語の総量は，扱われる題材次第で，引用や漢字，図表の含有率が異なるものの，可能な限り，論述する文章の分量が同程度になるように配慮して，データ容量1.2M程度を目安として採集した結果，次のような語数と文や段落数のデータとなった。

使用データ：国語学と国文学の学術論文

・**国文学**（近現代－明治以降の文学を扱うもの―49本：1.24MB）

総抽出語数	401,266語
異なり語数	14,453語
文	10,546文
段落	10,487段落（KH Coderの標準出力）

・**国語学**（現代日本語文法について扱うもの―49本：1.20MB）

総抽出語数	410,969語
異なり語数	21,640語
文	10,401文
段落	10,352段落（KH Coderの標準出力）

また，分析対象とする文章は，次の条件にあてはまる箇所のみとする。

・**古文の引用箇所や他人の引用文は，著者の文とは異なるため，分析対象から除外する**

・**原則「。」のある文のみを対象とし，タイトルやページ数，文章末注記，参考文献は対象外とした。ただし，頁脚注で「。」を含む文は本文相当となっている**

・**一単位を「。」で区別して処理する都合上，引用符（「」）内の句点（“。」”）は消去したため，“」”のみとなるが，「」を伴ったままの形で一単位の文として扱う**

収集した論文は10ページから15ページ程度のページ数のものが多いが，わずかながら，20～30ページの長い論文も含まれている。全ての論文の長さが一定になっているわけではないが，それは，論文の論理展開の構成要素であるパーツの大きさの違いであると捉え，論理的な文章構造自体には差がなく，論理展開に関連する語句には差異はないと考える。

収集した論文は，文章を語彙単位で分析するために，データ化し，語彙リストを作成した。データ化の際，PDFファイルをtextへ読み替えたため，主に漢字において文字コード変換が不可能な部分があった。平仮名表記の部分や接続表現等にかかる部分に問題がなかったことから，読み込みや変換不可能な文字

は原則「記号」で処理した。また，2, 3部の論文は，表示状態が悪いためにペーパーをスキャンしてからtextへ変換するという方法でデータ化したが，その際の読み取り不可能な漢字も「記号」として処理している。

以上のように抜粋したテキスト部分は，形態素解析処理に基づき，主要品詞別の語彙リストにまとめる。形態素解析から語彙リスト作成に至る一連の作業は，データマイニングシステムKH Coderを利用している。このシステムの日本語形態素解析は，単語認定に，新情報処理開発機構（RWCP）のIPA品詞体系（THiMCO97）を修正して作成されたIPA辞書を利用している。そのため，システム内で一つの単語としては認識されない複合語，例えば，「とし（て）」，「について」，「にもかかわらず」などのようなものがある。これらは，形態素解析結果を無視して，複合語としてのまとまりで1つの単語として処理するために，KH Coderで言う「強制抽出語」として指定して単語単位で扱っている。

2.2. 分析の指標として利用する語

国文学，国語学の各49編の論文には，それぞれ共通して出現する語が含まれている（図1）。また，ただ単に高頻度であるだけでなく，高頻度，かつ，出現回数の多い語は，内容的な特異性があるというよりは，ある特定の働きを持つ実質的な語である可能性が高い（図2）。共通出現の高頻度語は，文系論文というジャンルでの特徴的な語の可能性があると考える。

国文系論文と国語学系論文との使用語の頻度数を比較し，どのような語に相違が現れるかを見るために，まず，国文系論文49編と国語系論文49編の使用頻度の高い語のうち，実質的な語が含まれる名詞，動詞，形容詞，副詞，接続詞といった品詞別語彙リストと，機能的な語である助詞類の語彙リストを概観してみる。手始めに，出現頻度の上位から10語の普通名詞とサ変名詞を表1に抜き出してみた。表1の「文学」は国文学，「語学」は国語学，”N”は名詞を略したもので，頻度は素頻度である（以後も同様）。

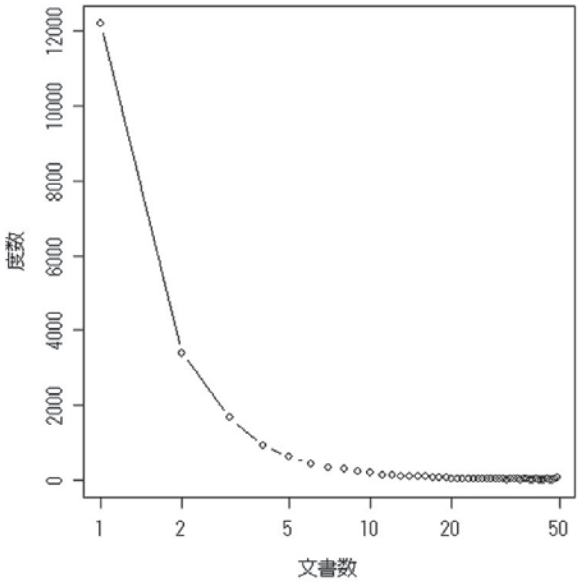

図1：度数と文書数の関係（文学系論文）

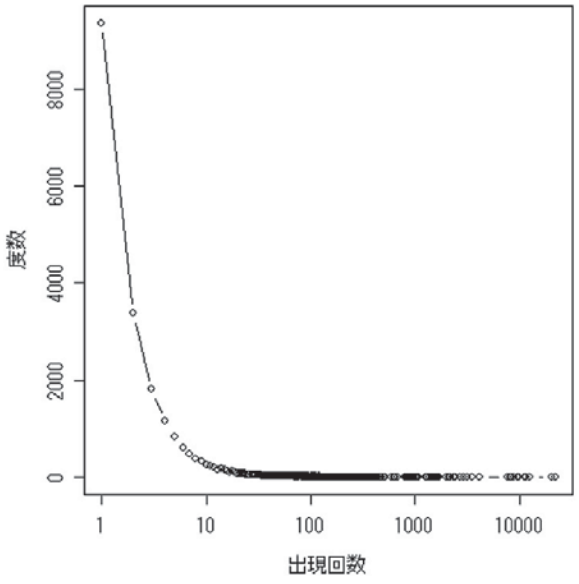

図2：度数と出現回数の関係（文学系論文）

表1の普通名詞（N）には，文学，あるいは，語学といったジャンル特有の表現だと考えられる「坊ちゃん」「文学」や「文法」「動詞」などの語が散見され，使用頻度の高い上位の普通名詞（N）の中には，国文学系論文・国語学系論文で共通する語は少ない。

ただし，サ変名詞は，語学系論文でヴァリエーションも頻度数も多い。サ変名詞は文学系，語学系に共通して利用されている語も多いが，内容やテーマに特化した語の出現が比較的少ないようにも見える。出現数が100回以上のサ変名詞の中で，国文学系，国語学系に共通して出現する13語を抜き出して差を確かめてみる。

国文学系，国語学系の総語彙数には差があることから，素頻度を補正（イェーツの補正）しながら，より詳しく検定するために，石川（2010）の付属マクロを利用して，共通利用されている13語のサ変名詞（表2）の使用頻度に有意な差が見られるか，カイ2乗検定を行なった。有意差ありとされる語が多く，それらは主に国語学系での使用数が多いという結果であった（表2）。

表1：国文学 —— 国語学の頻出名詞上位30語

	文学-N	頻度	語学-N	頻度	文学-サ変N	頻度	語学-サ変N	頻度
1	坊っちゃん	632	文法	956	存在	394	表現	1160
2	文学	625	動詞	920	表現	347	意味	1096
3	作品	587	名詞	581	関係	294	関係	802
4	自分	445	形式	547	意味	284	研究	466
5	人間	437	言語	538	生活	251	変化	453
6	近代	419	助詞	455	意識	233	使用	444
7	小説	392	構造	454	運動	198	存在	395
8	世界	375	構文	440	指摘	186	認知	386
9	主義	347	対象	434	認識	181	説明	367
10	物語	308	主語	409	行動	167	機能	318

表2：国文学 —— 国語学論文に共通利用のサ変名詞のχ^2検定結果

個別の語	カイ二乗値	*p*値	自由度(df)	個別の語の頻度の差の有意性判定	頻度が高い論文
存在	0.07	0.7914	1	有意差なし（χ^2 =0.07, *p*=.791）	
表現	419.17	0.0000	1	有意水準0.1%で有意差あり（χ^2 =419.17, *p*=.000）	語学
関係	222.91	0.0000	1	有意水準0.1%で有意差あり（χ^2 =222.91, *p*=.000）	語学
意味	458.27	0.0000	1	有意水準0.1%で有意差あり（χ^2 =458.27, *p*=.000）	語学
指摘	0.83	0.3610	1	有意差なし（χ^2 =0.83, *p*=.361）	
研究	145.13	0.0000	1	有意水準0.1%で有意差あり（χ^2 =145.13, *p*=.000）	語学
話	7.79	0.0052	1	有意水準1%で有意差あり（χ^2 =7.80, *p*=.005）	語学
記述	21.21	0.0000	1	有意水準0.1%で有意差あり（χ^2 =21.21, *p*=.000）	語学
行為	25.58	0.0000	1	有意水準0.1%で有意差あり（χ^2 =25.58, *p*=.000）	語学
評価	19.18	0.0000	1	有意水準0.1%で有意差あり（χ^2 =19.18, *p*=.000）	語学
説明	127.47	0.0000	1	有意水準0.1%で有意差あり（χ^2 =127.47, *p*=.000）	語学
確認	12.27	0.0005	1	有意水準0.1%で有意差あり（χ^2 =12.27, *p*=.001）	語学
変化	220.91	0.0000	1	有意水準0.1%で有意差あり（χ^2 =220.91, *p*=.000）	語学

しかし，サ変名詞は，「する」がついて動作性名詞を表すものであるが，その文法的な性質から，論文のテーマに直接関係するというよりは，陳述や解説といった論述のために利用される動作性の概念語である。国文学系でも国語学系でも同様に論述に利用しているはずである。

表2をよく見ると，「表現」「意味」「関係」「研究」など，術語や理論の固有名詞に関連する語があり，これらのP値がほぼ0%に近いことから「有意差あり」という結果が出されている。これらの語の使用状況を論文本文に返って確かめてみたところ，国語学系の指標頻度の高いサ変名詞は，専門的な術語として利用されている場合が多く含まれていた。そうすると，文や特定の語ということになるため，差が出るのは当然である。ということは，有意差の有無や差として判断できない。当然ながら，国語学系論文でもサ変名詞を概念説明用に利用する場合も含まれているが，名詞類を文体的な有意差検定に利用する場合は術語での使用か一般語としての使用かを区別した上で利用したほうがいいことには違いなく，今回は，判別の指標としては考えないことにする。

さて，名詞では，内容やテーマとの関連性には触れずに，論述用に利用されているであろう語として，サ変名詞で比較したが，他の品詞である動詞，形容詞，副詞，接続詞，助詞類については，名詞に比べると，内容やテーマとの直接的な関係は反映され難い。続けて，名詞同様に検定してみる。なお，名詞以外の品詞では，普通名詞やサ変名詞といった品詞の下位分類を区別せず，KH Coderで出力される品詞に準じるものとする。

動詞，形容詞，副詞，接続詞，助詞類にも，サ変名詞の場合と同様に，共通して利用されている語が15〜30語程度ずつ存在していた。それらを石川(2010)の付属マクロを利用して検定した。今回の調査対象とした論文データ数が49編と少ないことから，検定の結果，1%水準〜0.1%水準の範囲で有意差があるもののみを表3に抜き出してみる。自由度は1である。

表3：共通利用語で有意差が1%～0.1%水準の語

	個別の語	カイ二乗値	*p*値	頻度が高いコーパス
動詞	現れる	12.96	0.0003	語学
	見える	11.31	0.0008	語学
	見る	7.40	0.0065	語学
	認める	6.64	0.0100	語学
サ変名詞	意味	458.27	0.0000	語学
	表現	419.17	0.0000	語学
	関係	222.91	0.0000	語学
	変化	220.91	0.0000	語学
	研究	145.13	0.0000	語学
	説明	127.47	0.0000	語学
	行為	25.58	0.0000	語学
	記述	21.21	0.0000	語学
	評価	19.18	0.0000	語学
	確認	12.27	0.0005	語学
	話	7.79	0.0052	語学
形容詞	長い	20.73	0.0000	文学
	少ない	19.77	0.0000	語学
	広い	15.33	0.0001	語学
	遠い	8.74	0.0031	文学
	無い	8.59	0.0034	文学
副詞	後に	14.16	0.0002	文学
	必ず	6.77	0.0092	語学
接続詞	および	11.97	0.0005	語学
	たとえば	11.91	0.0006	語学
	次に	11.79	0.0006	語学
	ただ	11.06	0.0009	文学
	では	10.37	0.0013	文学
	ところが	8.58	0.0034	語学
	実は	8.27	0.0040	文学
	したがって	7.87	0.0050	語学
	または	6.81	0.0091	語学
	従って	7.76	0.0053	語学
接続表現（助詞類）	から－理由	12.3582397	0.0004	文学
	のみ	9.319966106	0.0023	語学
	が－接助	7.466506411	0.0063	語学

統語的な使用差を反映すると考えられる助詞類は，もちろん共通して利用される語が多く，また，ほぼ同様のものがほぼ有意差なく利用されていたが，この中で，少ないながらも，「から－理由」「のみ」「が－接助」の3語に有意差（カイ2乗検定の結果）があった。「から－理由」「が－接助」はどちらも複文を構成する統語的性質を持った接続形式である。また，「のみ」は副助詞ではあるが，文が埋め込まれる場合もあることから，広い意味での接続表現であると考えれば，これらに見られる有意差は，松岡（1995）を念頭におけば，興味深いことである。すなわち，論述スタイルの文体では，文を接続する「接続表現」も，論理的な文章の展開を行なっているということを示すと考えられ，それが，アカデミック性の高い論文ジャンルの下位分野の差に影響するものだと考えられることによる。

表3を見る限りでは，共通利用の語の中で「有意差有り」と考えられる語は少ない。データとなる論文の数という問題もあるだろうし，また，予測した語があてはまるものではなかったこともあるだろう。どの品詞も同様に，共通性と高頻度，有意性を判断するのに利用できるものでもないということも関係するだろう。したがって，この検定結果から伺えそうなことは，質的な分析の指標としてこれらの使用状況を考えるということである。両方に共通して利用するものであるにもかかわらず，有意差がある語は，その語の使用状況を質的に分析していくことで，利用法から分野別の文体差を担う使用法が見出せると考えられるからである。

では，判別の目的となる国文系論文と国語系論文をどのような語を指標に判別すればいいだろうか。指標となりそうなものを探りながら，いくつかみていくことになるだろうが，術語の影響を受けないことを優先し，両方のグループで共通して利用されているということにはあまりこだわらずに検討していく。次章では，国文学系論文と国語学系論文からそれぞれ作成した品詞別語彙リストを利用し，高頻度であるが，偏りの少ないものを検定した結果に基づいて，それらの語がどの程度，下位分野の差を判断するのに利用できるのかについて確認する。

3. 文系下位分野論文の分析方法

一般に，個々のデータが所属するグループを自動的に判定する手法として判別分析が利用されるが，「コーパス研究では，著者推定や習熟度推定，ジャンル推定といったテクスト分類に広く用いられる」という（石川，2010）。

また，村田（2007）は，接続表現は論述的な文章の文脈展開において重要な役割を果たすと言う松岡（1995）を受け，接続表現を指標に，経済，工学，物理学，文学の分野の論文の論述形式の違いを，65の接続表現を変数として，370編の論文中でその分布を比較し，84.6%の精度でジャンルに分けられること，さらに，判別に特に有効であった19の接続表現を65の接続表現から明らかにしている。村田はこの分析に，ノンパラメトリック検定（nonparametric test）[1]のクラスカル・ウォリス検定[2]を用いている。また，65の接続表現の出現率を説明変数とし，論文分野を基準変数として，判別目的のグループに据え，三つ以上のグループの判別を行なうことから正準判別分析のステップワイズ法を用いて判別分析し，文系論文に特有の19の接続表現を明らかにしている。

ジャンルによる特徴的な語がどの程度寄与しているかを見ることが目的であることから，本稿でも，判別分析を用いて，接続表現等を指標に，国文学系論文と国語学系論文を判別する。

判別分析の手法には，全変数で判別関数を作る線形判別法，変数増減法（ステップワイズ法），3群以上を判別する正準判別法の3種類ある。村田（2007）は65の接続表現からジャンル判別に特に有効な語句を選択するために，正準判別分析の変数増減法を利用して判別に寄与する語句を特定しているが，本稿で対象とする論文が国文学系論文，国語学系論文の2群だけであること，ならびに，指標とする，接続詞，接続表現，文末表現，品詞がせいぜい30語以内であることから，全変数を利用する線形判別法を利用する。

3.1. 判別分析の指標1 —— 村田（2007）で指摘された文系指標の接続表現

村田（2007）は，理系や経済から，文系論文を判別するのには，選別した19

の接続表現で十分に判別が可能であることを検証している。文系論文の下位分野分類に応用できるのかを見るために，本稿でも，村田（2007）の19の接続表現を指標にした国文学系，国語学系の判別分析を行なって見る。

ただし，この19の接続表現のうち，漢字，平仮名の区別や形態的に重複するものは，同じものとして統合し，全部で16個を指標とする（表5）。

以下，基礎統計量，相関行列，判別関数係数表，係数検定結果表，判別結果表，判別得点表の誤判別の状況の7観点ごとに見ていく。

1）基礎統計量と平均

16の接続表現を指標に，石川（2010）に付属のマクロ，Segal Statを用いて判別すると，表4のような全群の基礎統計量が得られ，1群の方で値が高いものとして**〈について，による系，から、，ために，ものの，むしろ，とともに，つつ，にもかかわらず，ながらも，うえで〉**の11項が，また，2群の方で値が高いものとして**〈として（する），ので、，まま，ただし，によれば〉**の5項目が判別された。

表4：第1群・第2群。群間平均値表（接続表現）

	変数	1群平均値	2群平均値	群間平均値
1	として（する）	56.06612096	57.12549211	56.59580654
2	について	20.77769531	12.2913234	16.53450936
3	による系	23.50361778	21.89276866	22.69819322
4	から、	9.382400106	8.35054968	8.866474893
5	ので、	5.867053829	7.667270645	6.767162237
6	ために	4.12684725	2.944664046	3.535755648
7	ものの	2.530889328	1.992225273	2.261557301
8	まま	1.891607403	2.245890114	2.068748758
9	ただし	2.440890506	2.595380546	2.518135526
10	むしろ	1.385680201	0.957223193	1.171451697
11	とともに	1.351891143	1.122103813	1.236997478
12	つつ	1.144499091	0.776354573	0.960426832
13	によれば	0.896318979	0.897891444	0.897105211
14	にもかかわらず	0.848015181	0.740113104	0.794064143
15	ながらも	0.454236492	0.222313057	0.338274774
16	うえで	1.075989315	0.90398553	0.989987422

2）相関行列（上方）及び，分散共分散行列（下方）表

「上で」と「ただし」に中程度の相関（0.541）が表5で伺えるが，それ以外に相関の高いものが認められないため，このまま分析を進めてみる。

表5：第1群　相関行列（上方），分散共分散行列（下方），分散（対角線上太字）

変数	として（する）	について	による系	ので、		にもかかわらず	ながらも	うえで
として（する）	**514.389**	0.133678	0.067714	0.02122		−0.08462	0.137863	0.103473
について	46.77263	**237.998**	0.184132	0.008675		0.076154	0.009073	0.122987
による系	19.84095	36.69877	**166.906**	−0.2267		−0.13118	−0.08932	−0.1401
から、	24.65503	13.94175	18.62156	0.010745		−0.02605	−0.27666	−0.151
ので、	3.113189	0.865742	−18.9453	**41.843**	中	−0.25865	−0.23837	0.123607
ために	11.75314	7.424659	10.06632	−7.08091		0.281162	−0.04198	0.009256
もののの	9.213022	−6.12462	−0.83581	−0.42039		−0.0624	0.303283	0.063671
まま	−2.60948	−10.0869	−2.85644	3.845165	略	0.013113	0.000419	0.043178
ただし	−0.19281	9.823256	−1.2911	4.497944		−0.00374	0.013885	**0.541283**
むしろ	−0.59009	−4.24981	−4.13621	−1.18628		0.287317	−0.04879	0.106668
とともに	7.464751	3.932166	3.232848	−1.25561		−0.15089	0.141849	0.337038
つつ	3.362621	6.962829	-0.33919	−2.05644		−0.00636	−0.16402	−0.05753
によれば	1.338459	−2.03108	1.856715	−1.29995		0.015715	−0.05549	−0.14149
にもかかわらず	−2.82971	1.732268	−2.49887	−2.46697		**2.17405**	−0.04016	−0.09435
ながらも	3.091157	0.138374	−1.14086	−1.52437		−0.05853	**0.97736**	0.277286
うえで	4.646022	3.756251	−3.58334	1.583		−0.27541	0.542705	**3.91941**

3）判別関数係数と係数検定結果から

表6の判別係数はいずれも低い。強いて言えばという程度で「むしろ」が0.698で，判別得点に寄与しているようであるが，マハラノビス平方距離は0.807と低く，誤判別率が0.326，つまり，67％程度の精度での判別ということである。

また，係数検定の結果（表7），マハラビノス汎距離が0.426である。「〜について」の偏F値が7.118で，有意水準1％のF分布の統計量6.96を超えていることから，強いて言えばという程度で，当該係数の有意性が確認できる。

判別の成功率が低いことから，「〜について」の影響はさほどないだろう。

表6：判別関数係数

変数	1群－2群
として（する）	−0.008
について	0.054
による系	0.009
から、	0.021
ので、	−0.03
ために	0.061
もののの	0.063
まま	0.020
ただし	−0.041
むしろ	0.21
とともに	−0.006
つつ	0.063
によれば	0.011
にもかかわらず	−0.035
ながらも	0.151
うえで	0.008
定数項	−1.243
マハラノビスD^2	0.808
誤判別率	0.327

表7：係数検定結果

	1群－2群	
変数（j）	D^2（-j）	偏F値
として（する）	0.781	0.471
について	0.426	7.118
による系	0.799	0.148
から、	0.795	0.225
ので、	0.786	0.384
ために	0.775	0.577
もののの	0.773	0.596
まま	0.806	0.030
ただし	0.786	0.373
むしろ	0.698	1.921
とともに	0.808	0.002
つつ	0.802	0.108
によれば	0.808	0.005
にもかかわらず	0.806	0.032
ながらも	0.792	0.277
うえで	0.808	0.003

F（1,81,0.01）= 6.96

4）判別結果と判別得点表から

正判別率は64.3%であった（表8）。1群と2群間で，各々49例中，半分程度が入れ替わっているため，相違があると考えられはするものの，明確に判別されるという程度ではない。

表8：「接続表現」判別結果

前 ＼ 後	1群	2群	正判別率
1群	27	22	55.1%
2群	13	36	73.5%
総合			64.3%

以上のような判別分析の結果から，村田（2007）で他ジャンルの文章から文系論文ジャンルの特定に寄与した接続表現は，文系論文の下位分類の判別には，さほど大きく寄与するものではないことが確認できた。ただし，これには，対象とした論文のサンプル数の問題や，文系，理系という大きなジャンルの差の

特定用に検証された接続表現であることから，想定外のことでもない。

3.2. 判別分析の指標２――「接続詞」

２章で検定の結果有意差が見られた他の語はどうか。村田（2002）や松岡（1995）に倣い，まず，接続詞を指標にして判別分析を行なおう。

2章の表3にまとめられた接続詞は，100回以上使用されているものであるが，このうちで有意差が見られる接続詞はさらに少ない。そこで，対象論文データのどちらかで，出現頻度20回以上の接続詞29種を指標に判別分析を行なってみる。

この結果から判別に寄与している接続詞を見つけ，さらに，先の概略調査で見られた高頻度共通使用の接続詞がどの程度かを確認することで，文系論文の下位分類に利用できそうな接続詞を特定できるのではないか。同様に判別分析を試みる。

1）基礎統計量　平均

29個の接続詞を指標に石川（2010）付属のマクロ，Segal Statを用いて判別すると，表9のような全群の基礎統計量が得られ，1群の方で値が高いものとして**〈たとえば（例えば），または（又は），すなわち（即ち），したがって（従って），および（及び），次に，一方，また，なお，ところが，つまり，だから，だが，それでは，それで，そこで，しかしながら，さて〉**の17項が，また，2群の方で値が高いものとして**〈実は，ゆえに，では，ただ，だが，そもそも，そして，しかも，しかし，こうして，かつ，あるいは〉**の12項目が識別された。

相関行列と分散共分散行列表を見ると，「さて」と「ところが」，「実は」と「つまり」，「では」と「ところが」の３組で中程度の相関（0.461～0.599）が見られたものの，特に，相関の高い接続詞はなかったため，このまま進める。

2）判別関数係数・係数検定結果から

判別関数を見ると，いずれの得点も同程度に低く，マハラノビス平方距離は，4.8439で，誤判別率が0.1355である（表10）。また，偏F値で，有意水準1％の

F分布の統計量7.02を超えるものはなく，有意性が確認できるものはない（表11）。5%水準での基準等計量においても，限界値3.981を超えるF値のものはなく，有意性が確認できるものはない。強いて言えば，「たとえば」(3.7696)，「なお」(3.4757)，「ただ」(3.1380)，「だが」(3.8790)，「そこで」(3.6734）が挙げられる。

表9：第1群・第2群。群間平均値表（接続詞）

変数	第1群平均値	第2群平均値	1群－2群間平均値
たとえば例えば	5.989923	3.533049	4.761486
または又は	1.591614	0.913403	1.252509
すなわち即ち	4.012568	3.302855	3.657712
したがって従って	2.669855	1.75169	2.210772
および及び	2.177045	1.17257	1.674807
実は	0.547275	1.365131	0.956203
次に	1.822381	0.591467	1.206924
一方	3.63197	2.031543	2.831756
ゆえに	0.223882	0.50515	0.364516
また	14.27251	10.39324	12.33288
なお	3.076569	1.006461	2.041515
ところが	1.061565	0.594533	0.828049
では	0.797174	1.727096	1.262135
つまり	4.702862	4.611569	4.657215
ただ	1.202825	2.437654	1.820239
だから	1.943569	0.190038	1.066804
だが	0.745312	2.670268	1.70779
それでは	0.606675	0.479834	0.543255
それで	1.956052	0.042579	0.999315
そもそも	0.874542	0.948026	0.911284
そして	4.471122	7.595251	6.033187
そこで	3.959224	0.920177	2.4397
しかも	0.649954	1.231588	0.940771
しかしながら	1.094839	0.600506	0.847673
しかし	7.681311	9.531002	8.606156
さて	0.997707	0.626454	0.812081
こうして	0.135565	0.765642	0.450604
かつ	0.880992	1.056009	0.968501
あるいは	2.793285	3.469284	3.131285

表 10：判別関数係数

変数	1群－2群
たとえば例えば	0.2014462
または又は	0.0467427
すなわち即ち	0.0572185
したがって従って	0.1622973
および及び	0.0370815
実は	−0.4013
次に	0.0402374
一方	0.0771456
ゆえに	−0.301965
また	0.0708817
なお	0.3324999
ところが	0.2575334
では	−0.309991
つまり	−0.012089
ただ	−0.28279
だから	−0.4207
だが	−0.274582
それでは	−0.16656
それで	−0.351828
そもそも	−0.102394
そして	−0.083693
そこで	0.5800262
しかも	−0.074108
しかしながら	−0.014869
しかし	−0.074397
さて	0.0881011
こうして	−0.595621
かつ	−0.084669
あるいは	−0.088733
定数項	−0.40283
マハラノビスD^2	4.8439429
誤判別率	0.1355685

表 11：係数検定結果

	1群－2群	
変数（j）	D^2（-j）	偏F値
たとえば例えば	4.383704	3.769689
または又は	4.827855	0.125082
すなわち即ち	4.799245	0.34866
したがって従って	4.683487	1.268448
および及び	4.837454	0.050396
実は	4.59329	2.002479
次に	4.837983	0.046287
一方	4.796976	0.36645
ゆえに	4.788037	0.436645
また	4.655718	1.492785
なお	4.417841	3.475789
ところが	4.763739	0.628172
では	4.544521	2.405877
つまり	4.841738	0.017115
ただ	4.457416	3.138078
だから	4.794397	0.386684
だが	4.371074	3.879043
それでは	4.821142	0.177409
それで	4.803863	0.312469
そもそも	4.832149	0.091651
そして	4.680934	1.289009
そこで	4.394852	3.673446
しかも	4.834846	0.070671
しかしながら	4.843296	0.005019
しかし	4.652979	1.514993
さて	4.835099	0.068705
こうして	4.628091	1.71743
かつ	4.834584	0.072706
あるいは	4.783314	0.473791
	F（1,68,0.01）= 7.02	

3）判別結果表

総合正判別率が88.8%で，1群は91.8%，2群で85.7%であった（表12）。相互検証後の正判別率は71.4%で，F分布の統計量が3.98%であった。

表 12：「接続詞」判別結果

前 ＼ 後	1群	2群	正判別率
1群	45	4	91.8%
2群	7	42	85.7%
総合			88.8%

同ジャンル内の下位分類に，文系論文内で利用されている一般的な接続詞のうち，「たとえば（例えば)」「なお」「ただ」「だが」「そこで」が利用できそうにも見えるが，接続詞は文系で共通した利用傾向のものだとも考えられ，下位分野判別にはより明確な結果が出るような別の指標を考えたほうがいいのではないだろうか。

3.3. 判別分析の指標 3 —— その他の品詞

個人の好みや分野の習慣が反映されやすそうな品詞には，副詞や形容詞が上げられるが，論文というジャンルでは，使用される副詞に偏りがある上に，使用率が少ない。

論述に関係するものとしては，構文を形成し，展開を進める文末表現が考えられる。特定の文末表現は次章で取り扱い，ここでは，構文の要素となる助詞，動詞についてみておく。

まず，動詞であるが，動詞は，高頻度上位語であっても，特定の概念に特化した語彙が少ないと考えられることから，100 回以上の使用頻度が見られる上位 34 語を利用して国文学系と国語学系論文を判別しようとしたが，約半数程度の語が国文学系か国語学系かのいずれかでしか 100 回以上利用されていないことから，共通して利用されている上位 15 語を用いて判別した。

助詞は，動詞と同じく，高頻度上位語であっても，特定の概念に特化した語彙が少ない。共通して利用される助詞のうち，100 回以上の利用頻度がある上位 37 語を指標に国文学系，国語学系論文を判別した。

1）動詞上位 15 語の基礎統計量と平均

15 語の高頻度動詞を指標に石川（2010）付属のマクロを用いて判別し基礎統計量を確認する。全群をまとめた表 13 から，国文学群の方で若干ながらも値が

高いものとして2項が，また，国語学群の方で値が高いものとして13項目が識別された。

しかし，相関行列と分散共分散行列表を見ると，文学において「持つ」と「与える」の相関が高い（0.707）。他にも中程度の相関（0.409～0.582）を取る語が散見しており，かつ，マハラノビス平方距離が**1.9533**と，判別率もよくない。

そこで，石川（2010）付属のマクロ，Segal StatのVIFという多重共線性をチェックする機能を利用して，相関の高さに見られる妥当性低下の問題が生じていないか，重回帰分析を行なって確認して見ると，多重共線性については問題がないが，分散分析の結果，自由度が1のため，自由度二重調整済重相関係数を見ると0.5708であった。対象となるケース論文で特異な語彙使用を行なっている論文を抜いてみたが，より良くなることはなく，調査対象のデータに不適切性が伺えるが，これは，動詞を指標として分野別の展開を見るという考えが雑駁であり，データ量の少なさによる個々の論文の不均等性が分析に影響すること，さらに，動詞とテキストの性質との関連性が低いことを表すものでもあると考えられる。

表13：動詞15語を指標とした文-語間の基礎統計量

変数	文郡平均値	語群平均値	文－語群平均値
異なる	2.06122449	4.326530612	3.193877551
見る	12.18367347	14.53061224	13.35714286
言う	7.183673469	8.265306122	7.724489796
考える	6.836734694	14.20408163	10.52040816
行う	2.755102041	3.285714286	3.020408163
思う	5.469387755	5.755102041	5.612244898
持つ	6.408163265	7.836734694	7.12244898
示す	4.755102041	9.734693878	7.244897959
述べる	5.836734694	8.469387755	7.153061224
書く	5.653061224	2.12244898	3.887755102
捉える	3.244897959	4.367346939	3.806122449
知る	4.428571429	3.510204082	3.969387755
得る	2.571428571	2.918367347	2.744897959
認める	2.12244898	3.040816327	2.581632653
与える	2.183673469	2.224489796	2.204081633

表 16: 判別関数係数

変数	1群－2群
異なる	−0.0999
見る	−0.01322
言う	0.027337
考える	−0.10387
行う	−0.02244
思う	0.011992
持つ	0.006373
示す	−0.05518
述べる	−0.00722
書く	0.15754
捉える	0.018794
知る	0.001815
得る	−0.09075
認める	−0.11294
与える	0.118452
定数項	1.372152
マハラノビスD^2	1.953314
誤判別率	0.242337

表 17: 判別関数係数検定結果

	1群－2群	
変数（j）	D^2（-j）	偏F値
異なる	1.854241	1.407333
見る	1.93171	0.302815
言う	1.928396	0.349459
考える	1.494455	6.951346
行う	1.944583	0.122108
思う	1.950963	0.032835
持つ	1.951848	0.020472
示す	1.810614	2.042481
述べる	1.951176	0.029866
書く	1.321936	9.879765
捉える	1.937207	0.225558
知る	1.95313	0.00256
得る	1.872252	1.147905
認める	1.810533	2.043672
与える	1.865898	1.239242

F（1,82,0.01）= 6.95

表 18：判別結果（1 群－2 群）

前 ＼ 後	1群	2群	正判別率
1群	40	9	81.6%
2群	9	40	81.6%
総合			81.6%

相互検証結果
n = 検体数
正判別率 = 75.5%（74／98）

2）助詞上位 15 語の基礎統計量と平均

動詞同様，100 回以上の利用頻度がある上位 37 語を指標に国文学系，国語学系論文を判別したところ，相関行列表に 0.8 を超える係数値が多くみられたため，100 回以上，200 回未満の助詞を利用して，再分析した。基本統計量として国文学系，国語学系の平均値で差が明確にある語は 26 語中 10 語程度であまり大きな差がないものがほとんどである。マハラノビス平方距離は 4.9452 で，限界値F=7.01 を上回る語はないが，正準判別率は 85.7%，相互検証の結果，すなわち，判別精度は 72.4%である。

助詞は文構造の基本的な差を示すのではないかと考えられるが，明確に判断がつくほどの語が特定されたとは言えず，文系論文の下位分類には利用できないと考えられた。

4. 高頻度文末表現（「。」の直前の語）を指標にした判別分析

文末表現で文系のジャンル識別ができないか確認する。形式的に文末として捉えられるように，語彙リスト作成の際に，句点「。」をつけた形のものを1単

表 17：「文末表現」基礎統計量

変数	第1群平均値	第2群平均値	1群－2群間平均値
である。	53.16674156	51.73339274	52.45006715
いる。	29.48263841	34.55453842	32.01858841
ない。	22.49347878	22.25906423	22.3762715
する。	14.69173029	8.821401343	11.75656582
ある。	13.05588705	9.014978306	11.03543268
なる。	13.3081036	5.919847641	9.613975619
う。	12.59914468	10.76406204	11.68160336
れる。	11.76762055	8.780941093	10.27428082
られる。	10.88913792	5.024098356	7.956618139
った。	6.633715431	14.45432605	10.54402074
できる。	5.769935821	2.211702854	3.990819337
だろう。	3.927600488	5.735265644	4.831433066
か。	4.717683747	11.0698336	7.893758673
たい。	4.240600959	4.497055337	4.368828148
考える。	1.875984621	0.36516424	1.120574431
示す。	2.408352804	0.28681087	1.347581837
言える。	1.6701457	0.847231095	1.258688397
思う。	1.26752446	0.771917505	1.019720983
だ。	1.165590077	4.662709918	2.914149997
表す。	1.352809292	0	0.676404646
おく。	1.151805539	0.457540542	0.804673041
わかる。	1.467515673	0.890357024	1.178936348
みる。	1.377879605	0.20834561	0.793112607
しまう。	0.429429685	0.873934945	0.651682315
言う。	0.190388646	1.0430574	0.616723023

位とし，100回以上使用されているものから，国文系，国語系で共通して利用されている25語を用いた。指標となった文末表現は表17のものである。

表18の判別関数の係数検定表を見ると，1%水準の限界値が7.00で，この値を上回るF値の表現は「〜か。」がある。他にもかろうじて上回っているものに「〜になる。」「〜う。」がある。一定の信頼性が得られたものとすると，この3種の文末表現が国文学系論文と国語学系論文で使用差があり，どちらの系統の論文かの判別に寄与していると考えられる。

表18：判別関数係数

変数	1群－2群
である。	0.003091
いる。	−0.02655
ない。	0.060626
する。	0.125128
ある。	−0.02809
なる。	0.227711
う。	0.162983
れる。	0.007639
られる。	0.149796
った。	−0.07417
できる。	0.003736
だろう。	0.209384
か。	−0.39291
たい。	−0.13967
考える。	0.485819
示す。	0.120603
言える。	0.259149
思う。	0.14623
だ。	−0.24856
表す。	0.454165
おく。	−0.0802
わかる。	0.110027
みる。	0.195995
しまう。	−0.84466
言う。	−0.37056
定数項	−3.93264
マハラノビスD^2	10.09137
誤判別率	0.056104

表19：判別関数係数の検定

	1群－2群	
変数（j）	D^2（-j）	偏F値
である。	10.08918	0.011227
いる。	9.930686	0.835367
ない。	9.784612	1.611793
する。	9.649083	2.34712
ある。	10.06749	0.122921
なる。	8.659182	8.198522
う。	8.740757	7.681727
れる。	10.08913	0.011519
られる。	9.152638	5.170872
った。	9.666161	2.253654
できる。	10.09119	0.000909
だろう。	9.236146	4.680969
か。	6.433017	25.44599
たい。	9.839679	1.317157
考える。	9.133097	5.286412
示す。	9.977229	0.591405
言える。	9.782797	1.621544
思う。	9.980535	0.574137
だ。	9.081049	5.595855
表す。	9.792008	1.572082
おく。	10.07221	0.09861
わかる。	10.05032	0.211554
みる。	10.00082	0.468381
しまう。	9.474763	3.314795
言う。	9.596513	2.636313

F（1,72,0.01）= 7.00

表 20：「文末表現」判別結果

前 ＼ 後	1 群	2 群	正判別率
1 群	45	4	91.8%
2 群	3	46	93.9%
総合			92.9%

正判別率は，国文学系（1 群）で 91.8%，国語学系（2 群）で 93.9% であり，全体では 92.9% である。もちろん，誤判別と位置づけられているように，両者でもそれぞれの有意差のある文末表現を使うことには違いないが，使用頻度差があることでの文体的特徴を担っていることには違いない。

相互検証結果を行った結果は，正判別率 81.6% であることから，文末表現の使用頻度の差は文系論文の国文系，国語学系という下位分類に利用できるのではないかと考えられる。もちろん，質的に詳細な検討が必要であるが，国語学系には，疑問を投げかける方法の論述形式が特長的に利用されている可能性があり，興味深い。

5. 品詞比率による判別分析

品詞は文体差の分析に用いる（樺島・寿岳 1965）ことから，名詞，動詞，形容詞，形容動詞，助詞，助動詞，接続詞といった品詞別使用頻度を指標に線形判別分析を行なってみた。1% 水準の限界値が 6.93 で，それを上回る F 値の品詞はなかった。正判別率は国文学系（1 群）で 71.4%，国語学系（2 群）で 73.5%，全体では 72.4% であった。念のため，判別が最もうまくいくように説明変数の組み合わせを変更することで一部の変数でも判別ができる変数増減法でも見てみたが，導入変数を形容動詞としたときにのみ，助動詞が限界値を上回る 16.35 の F 値となった。

しかし，誤判断の内容を見ると，形態素解析システムで処理した品詞分類に依存した語彙リストであることによる問題も多く含まれる。若干，形容動詞と関連する助動詞の使い方に違いが見られるのかもしれないが，品詞で判別するほど特徴的な違いとして見るのではなく，やはり，形式的な個別の差を総合的

にコーディングする際に利用する方がよいと考え，品詞含有率は，別ジャンルの差を見るのには利用するが，同ジャンル内の下位分類の考察には加えない。

6. 考　察

RQ1 は，文系論文における国文学系と国語学系の下位分類に，村田（2007）が他ジャンルの文章から文系論文ジャンルの特定に寄与すると特定した接続表現が利用できそうか確かめるという課題であったが，全体的に見ると，文系論文の下位分類判別には，さほど大きく寄与するものではないことが確認できた。

しかし「むしろ」「について」は，再度調査してみた方がいいと考えられた。対象とする文系論文のサンプル数を増やして，再度，検証し，その上で考察する必要がある。

RQ2 は，接続表現以外にテキストの差を判断するのに，なにか影響するものがないか（文末形態，品詞比率，格助詞などはどうか）ということで，村田（2007）の接続表現以外の接続詞と，文末表現を見たが，接続詞では寄与するものがなかった。しかし，文末表現のうち，「〜か。」「〜になる。」「〜う。」は区別の指標になると考えられそうであった。ただし，これらの文末表現が具体的にどのような意味で使われているかといった詳細な分類と，同時に，コーパスの規模を大きくして，再検証しておくことも必要だろう。

接続表現や文末表現，両方のいくつかの併用で文系論文の下位分野特定に寄与するものが得られそうだとは考えられる。併用の場合の組み合わせや，最も，大きく影響するものの特定に向けて，次のような反省点を今後の課題につなげたい。

今回，類似分野のコーパスに対して，それぞれの品詞別語彙リストを作成し，上位語，共通語，有意差のある語という3つの観点で抜き出した語句を用いて，文系下位分野の国文学系，国語学系論文の区別が可能かということを検討してみたわけであるが，有意差のある語が，真に有意差のある語であるかという観点からの見直しと確認に，十分気を配ったとは言えないこと，また，有意差のある語が，すなわち，判別に寄与するわけではなかったこと，共通して利用さ

れる語という変数の特性が，データに対してどのような関係にあると位置づけられるかについてはもう少し深く考察しておく必要があることが，反省点として確認された。ジャンル差を識別する手法として，検定，判別分析の併用を試みたが，その方法，また，ほかの方法を比較し，手法自体の特性を十分に生かせるように，併用する場合の手法上の意味に対する理解を深める必要がある。

また，品詞別語彙リストを用いたことについてであるが，全品詞語彙リストにおけるある一定の基準以内の語という指定で変数として利用する方法と比較することを検討してみたい。品詞別語彙リストの中で，変数を，拡張，または，縮小して判定精度向上を検討したが，どの語を利用するかという特定方法自体についての準備段階の検討と工夫が必要であると考えられた。

さらに，用いた文系論文の数が少ないことから，調査データの個人差が大きく反映されることになったため，論文を統合して利用しても，調査対象の論文の性質が明確になるように整理して採集する方法を工夫する必要がある。

7. まとめ

国文学系の論文ではどのような表現があることでアカデミックな性質が高まるか考察するために，判別分析を用いて，文系論文の中での国文学系と国語学系の論文で，下位分類に応用可能な表現があるか確かめ，国文系論文特有の語句のうち，どの接続表現がどの程度関与しそうか確かめようとした。

指標とした語は，文末表現，接続詞，接続表現，副詞，形容詞，助動詞といった品詞や形態別の語彙リストから抜き出したものを利用したが，判別分析の結果，文末表現の「か。」「になる」「う。」等，いくつか，文系論文の下位分類に影響しそうな語句が指摘できたことから，アカデミックな表現を用いて文系論文の下位分野が区別できる可能性が伺えた。

ただし，学生のアカデミック性を高めるための指導項目としての文系論文の下位分野特定語句を考えるなら，論文記述指導等に応用できる観点として，たとえば，学生のレポートや卒業論文などの研究者の論文とは若干異なるであろう文章との比較を検討することも必要であろう。

今回のケーススタディーを通して得られた反省を踏まえ，国文学系のアカデミック性判断の観点となる表現と，その出現度の高さから文系のアカデミック性判断の指標として利用できそうな語句の特定を今後の課題としたい。

本稿は，統計数理研究所共同研究合同発表会「言語研究と統計2011」(2011年3月14日　於)統計数理研究所（東京都立川市）での報告をまとめたものである。

注

1　母集団分布に関して，正規分布などのある特定の分布を仮定しないで統計的検定を行う方法である．この手法の利点は，多少の制約がある場合もあるが，どのような母集団分布からのデータであっても適用可能なことである(http://lbm.ab.a.u-tokyo.ac.jp/~omori/kensyu/nonpara.htm)。

2　分析に利用する量的変数が正規分布に従うとは言えないことから，一元配置分散分析に相当するノンパラメトリック検定で、3群以上における差の検定を行なっている。これは，代表値として「平均値」ではなく、中央値で「分布の位置」の差をみるもので，データ値を小さい方から順番にランクデータに変換し（同順位がある場合は、平均順位をわりあてる)、各群ごとに順位を足し合わせ（順位和)、各群のケース数で割って平均ランクを求める。この値をもとに検定統計量を計算して、有意差かどうかを調べるが，結果、有意差が認められた場合、中央値を代表値として述べることが多い。本来は，平均ランクに差があるかどうかを調べることが目的である。

参考文献

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.

樺島忠夫・寿岳章子（1965)『文体の科学』綜芸舎.

中尾桂子（2009)「語彙の統計量と総合評価の関係―作文評価の基準特定にむけて―」大妻女子大学文系紀要41，pp129-146.

村田年（2002)「論理展開を支える接続語句―接続語句，助詞相等句による文章ジャンル判別を通じて―」『計量国語学』23 (4)，pp311-328.

村田年（2007)「専門日本語教育における論述指導のための接続語句・助詞相当語句の研究」統計数理研究所特集第66巻第2号，pp269-284.

メイナード，泉子・K（2004)『談話言語学：日本語のディスコースを創造する構成・レトリック・ストラテジーの研究』くろしお出版.

松岡弘（1995)「文学的文章をモデルにした文章の論理的構成の指導について」『日本語と日本語教育：阪田雪子先生古稀記念論文集』三省堂.

KH Coder：http://khc.sourceforge.net/