推量的副詞の共起情報に基づいた genre 別の文末表現の分析

Bor Hodoscek[1]　Andrej Bekes[2]　仁科　喜久子[1]

[1]東京工業大学　{nishina.k.aa, hodoscek.b.aa}@m.titech.ac.jp
[2]リュブリャーナ大学　andrej.bekes@guest.arnes.si

1　はじめに

本研究では，非母語話者が適切な表現を獲得することを可能にする作文支援システム「なつめ」(Yoshihashi and Nishina)の開発の一環として，副詞と文末表現の呼応の習得に関する分析を行う．

推量的副詞が文末モダリティと呼応し，様々なコーパスでその呼応の様相の異なることが研究されてきた(工藤，木田ら，Bekes，Srdanović ら 2008)．工藤は推量的副詞と述部モダリティの共起関係を調査し，結びつきの強いものを呼応と見なした．そして，その呼応の様相の異なることに基づいて，18 個の推量的副詞を四つのモダリティタイプに分けた(表 1)．Srdanović ら(2007)は，工藤の調査をウェブコーパスに拡大し，クラスタ分析で四つのモダリティタイプの存在を確認しながら，コーパスによって呼応が変化するという結果を得ている．さらに，スルダノヴィッチら(2008)は，推量的副詞の分布に基づいてコーパスをクラスタ分析によって分類した．その結果，副詞の代表的なモダリティタイプがコーパスの genre を特定する有効な指標であることを明らかにした．

表 1: 工藤のモダリティタイプによる推量的副詞の分類[1]

確信	きっと，必ず，絶対に
推測	どうやら，よほど
推定	恐らく，多分，さぞ，大方，大抵，大概
不確定	もしかすると，ひょっとすると，ことによると，案外，必ずしも

非母語話者が特定の genre において文章作成をマスターするのは困難であり、これに関して Halliday（1993）は理工系の英語について分析している。この分析から、文章の複雑さが語彙密度という測定方法で数量的に計ることができるとしている．この方法は日本語への応用としては実装された(佐野)．さらに Halliday（2004）は genre の異なるコンテクスト情報を見極める方法としてシステミック文法の中で特定の genre の中に特徴的にみられる語彙やテキスト構造を REGISTER として研究を発展させている.推量的副詞の分布が REGISTER によって変化する傾向は Srdanović ら(2007)で取り上げられている．その変化の様相によってコーパスの genre を判定できることが期待される．

本稿では，推量的副詞と文末表現の呼応を様々な genre のコーパスにおいて計量的に分析し，genre 毎の文末表現に着目してその形式を明らかにする．

2　使用コーパス及び推量的副詞の分布

2.1　分析対象のコーパス

本稿の調査に用いたコーパスは，『現代日本語書き言葉均衡コーパス』(BCCWJ)を初め，新聞，教科書，技術論文である．BCCWJ は書籍，白書，Yahoo!知恵袋，国会会議録の四つのサブコーパスからなる現代日本語の書き言葉をバランス良くとらえたコーパスである．BCCWJ から対象としたのは，白書と国会会議録(国会)のサンプルデータである．書籍と特に Yahoo 知恵袋はくだけた日本語の表現があり，日本語学習者にも有益な情報となり得るが，時間的な制約から今回はフォーマルな表現のコーパスを対象として限定した．BCCWJ 以外のコーパスとして，毎日新聞

1　工藤にはさらに「どうも」と「あるいは」があるが，モダリティ以外の用法が多いので，今回の調査から外した．

2002 年の 1-3 月分(毎日)，16 冊の大学理系基礎科目教科書(教科書)，自然言語処理学会発表論文集 2005-2008 年分[2](論文)を調査に使用した．

2.2 共起データ作成

木田らは，副詞の「多分」「恐らく」「きっと」「決して」の呼応を得るために，人手作業と自動抽出を両方行って，それらを比較，評価した．その結果，大規模コーパスからの抽出がバリエーションを得るために有意義であることを明らかにしたが，本稿の一つ一つのコーパスの量は少ないため，n-gram などで文末表現の自動抽出はしていない．人手作業の能率を上げるために，CaboCha[3]の解析結果をもとに推量的副詞と文末表現のペアを抽出した．CaboCha の解析結果において，推量的副詞が文末に位置する文節に係れば，リストに追加することにした．さらに，リストを人手作業で修正し，推量的副詞に対して文末表現をまとめる．

2.3 推量的副詞の分布

コーパス毎の推量的副詞の出現割合を表 2 に示す．また，コーパス毎の推量的副詞を四つのモダリティタイプに分けて形態素 100 万に対する出現割合を表 3 に示す．表 2 から，政府系の白書の推量的副詞の分布は他のコーパスと隔たりがあり，特殊コーパスであるといえる．「必ずしも」がその 9 割以上を占めていることは，婉曲表現による断言の回避を示していると考えられる．また，論文と教科書は白書と同じく「必ずしも」が多いが，その次に「必ず」が現れることから，どちらも専門的なコーパスで，確信のモダリティを持つことが期待されるコーパスであると考えられる．国会では，他のコーパスと違って推定が 3 割近くあり，フォーマルな会話の特徴を表している．

表 2:各コーパスの推量的副詞の分布(出現割合)

副詞	コーパス名				
	毎日	国会	教科書	論文	白書
きっと	**12.79**	1.09	0.00	0.55	0.00
必ず	**33.56**	12.01	**36.54**	**21.31**	3.31
絶対に	9.82	8.52	1.92	0.55	0.83
どうやら	4.57	0.87	0.00	1.09	0.00
余程	2.74	0.87	0.00	0.00	0.00
恐らく	9.36	**28.17**	5.77	8.74	0.00
多分	9.59	6.99	5.77	1.64	0.00
さぞ	0.68	0.00	0.00	0.00	0.00
大方	0.23	0.00	0.00	0.00	0.83
大抵	2.97	0.22	1.92	2.73	0.00
大概	0.00	0.22	0.00	0.00	0.00
もしかすると	2.51	1.75	0.00	0.55	0.00
ひょっとしたら	0.68	0.66	0.00	0.00	0.00
ことによると	0.00	0.00	0.00	0.00	0.00
案外	2.28	1.75	0.00	0.00	0.83
必ずしも	8.22	**36.90**	**48.08**	**62.84**	**94.21**
形態素数	688万	467万	52万	311万	467万

また，国会会議においては教科書，論文と毎日よりも確信の「必ず」が少なく，確信のモダリティより推定の方が多用されている．一方毎日新聞は「きっと」と「絶対に」の使用で確信が出現割合の 56.17%を占めている．新聞の社会的な役割は事実を正確に伝えることなので，その特徴として不確定が少なく確信が多くなっていると考えられる．

表 3:モダリティタイプとコーパスによる推量的副詞の形態素 100 万に対しての出現割合

モダリティタイプ	コーパス名				
	毎日	国会	教科書	論文	白書
確信	35.76	21.16	38.27	13.16	1.07
推測	4.65	1.71	0.00	0.64	0.00
推定	14.54	34.84	13.40	7.70	0.21
不確定	8.72	40.18	47.84	37.24	24.61
合計	63.67	97.89	99.51	58.75	25.89

各コーパスで，どの程度推量的副詞が文章に含まれているかを把握するにあたって，コーパスを頻度が高い順から並べると，

教科書＞国会＞毎日＞NLP＞白書

となる．教科書の確信と不確定の用法が他のコーパスより高頻度ということは，確信の可

2 PDF からテキストを抽出するプログラムは http://poppler.freedesktop.org/の pdftotext を使用した．また，その中から日本語に変換できなかったものは除外した．

3 http://chasen.org/~taku/software/cabocha/

能性を含意するためであろう．

3　文末表現の分析

推量的副詞は陳述副詞に含まれ，呼応する文末表現が話者の陳述となっている．これに関して，南は日本語の構文を ABCD の四階層の入れ子構造として説明した．

例 1：{Ø{恐らく{多くの国民は{納得し}$_A$ない}$_B$のではなかろうか}$_C$}$_D$。(毎日)

例 1 は，推量的副詞の「恐らく」と文末モダリティの「のではないだろうか」が C 層にあり，意味的に呼応するとしている．用言の「納得しない」の後に様々な機能表現が加えられ，その中にも文末モダリティが含まれている．その文末モダリティを部分的にみると「の/では/ない/だ/ろう/か」のように分けられる．D 層は話者の相手に対する何らかの働きかけを意味し，「ね」「よ」などの終助詞がその一つである．しかし，モダリティと相手に対する働きかけを含意した文末表現の方が，非母語話者の適切な文末表現の習得に必要と考えられる．そこで，述語の後の文末モダリティとその他の文末までの語を合わせて文末表現であるとする(例 1 の下線部分に相当)．本稿では，このような文末表現の連なりの様相を観察し，各コーパスの特徴を明らかにする．

表 4:各コーパスの「必ずしも」と共起する文末表現の出現割合(２%以下を除く)

必ずしも									
国会	%	毎日	%	論文	%	白書	%	教科書	%
ない	54.38	ない	80.56	ない	50.00	ない	67.54	ない	68.00
ものではない	3.75	とはいえない	5.56	わけではない	15.45	とはいえない	10.53	とは限らない	12.00
のではないか	3.13	とは限らない	5.56	とは限らない	14.55	ものではない	8.77	わけではない	8.00
わけではない	3.13	わけではない	5.56	とはいえない	4.55	とは言い難い	4.39	必要はない	4.00
ないわけだ	2.50	とは思わない	2.78	必要はない	4.55	わけではない	3.51	ものではない	4.00
				ものではない	4.55			とは言い難い	4.00
				Ø	3.64				

表 5:各コーパスの「必ず」と共起する文末表現の出現割合(２%以下を除く)

必ず									
国会	%	毎日	%	論文	%	白書	%	教科書	%
Ø	52.73	Ø	76.47	Ø	87.18	なければならない	25.00	Ø	89.47
のだ	7.27	だろう	4.41	のだ	2.56	てください	25.00	というわけでもない	5.26
わけだ	5.45	のだ	2.94	わけではない	2.56	はず	25.00	だろうか	5.26
だろう	3.64	はず	2.94	てください	2.56	てしまう	25.00		
なければならない	3.64			なければならない	2.56				
わけだね	3.64			ない	2.56				

表 6:各コーパスの「恐らく」と共起する文末表現の出現割合(白書を除いて２%以下を除く)

恐らく							
国会	%	毎日	%	論文	%	教科書	%
だろう	9.52	だろう	29.27	Ø	40.00	Ø	33.33
Ø	7.94	Ø	19.51	だろう	13.33	だろう	33.33
と思う	6.35	のだろう	19.51	と思われる	13.33	のではないだろうか	33.33
のではないか	5.56	に違いない	4.88	ものだろう	6.67		
だろうというふうに思う	2.38	はず	4.88	ものだと推測される	6.67		
のではないかと思う	2.38	ない	4.88	だろうね	6.67		
と思うのだね	2.38	のではないだろうか	4.88	と考えられる	6.67		
のではないかというふうに感じている	2.38	と思われた	2.44	ない	6.67		
のではないだろうか	2.38	ではないだろう	2.44				
のではないかと	2.38	ではないか	2.44				
のだろう	2.38	ではないかと思う	2.44				
だろうと思う	2.38	まい	2.44				
のだね	2.38						

全コーパスに現れ，かつ高頻度の推量的副詞は限られており，確信，推定，不確定のモダリティタイプからそれぞれ「必ず」，「必ずしも」，「恐らく」の三つ組を比較対象にし，その呼応の出現割合が 2%以上の文末表現を表 4-6 に示す[4].

国会には他のコーパスと比べて文末表現の多様性が見られ，モダリティを伴わない述語の出現割合も少ない．推定の「恐らく」は不確定の「必ずしも」より頻度が少ないにもかかわらず，文末表現が非常に多様である．そして，文末表現が他より長く，話し言葉のコーパスの特徴を示している．

一方，確信を表す「必ず」は，長い文末表現と文末モダリティの組み合わせが比較的少ない．この違いは，副詞の機能として，「恐らく」が不確実な予測表現を文末に要求するのに対して，「必ず」が断定的で明解な陳述を要求するからだと考えられる．

「恐らく」の場合，モダリティを伴わない述語が論文に 40%，教科書に 30%，毎日に 29.27%あり，書き言葉の国会と比べ，文末にモダリティを控えている傾向が見られる．白書は推量的副詞の中で「必ずしも」以外のものが比較的少ない傾向をすでに第２節で示したため，現状の白書のデータ量では比較が難しい．

4 まとめと今後の課題

本稿では，『現代日本語書き言葉均衡コーパス』を初め，新聞，教科書，技術論文などのそれぞれの genre に見られる推量的副詞と文末表現の共起情報によって文末表現を計量的に分析し，分布の傾向について論じた．その結果，この共起情報が genre の有効な指標であることが分かった．今後の課題としてコーパスの種類と量を拡大し，推量的副詞の数を増やすことでさらに検証を進める．

さらに，副詞とそれに呼応する文末表現に関して，非母語話者が適切な表現を獲得することを可能にする作文支援システム「なつめ」への応用を検討する．具体的な応用としては入力した推量的副詞と文末表現が目的の genre にどの程度ふさわしいか，またはそうではない場合，類義の推量的副詞及び文末表現の提示もできる機能をシステムに追加することを検討する．

4「Ø」はモダリティを伴わない述語

参考文献

Bekes, Andrej. (2006). “Japanese suppositional adverbs in speaker-hearer interaction.” Proceedings of the third conference on Japanese language and Japanese language teaching, Rome 2005. Venezia: Libreria editrice cafoscarina. 34-48.

Halliday, M.A.K, Matthiessen, Christian M.I.M. (2004). “An introduction to Functional Grammar.” 3rd edition. London: Hodder Education.

Halliday, M. A. K. (1993). “Some grammatical problems in scientific English.” In M.A.K. Halliday & J.R. Martin (Eds), Writing Science: Literacy and Discursive Power. London: Falmer Press.

木田敦子，山本英子，神崎亮子，井佐原均. (2004). コーパスからの呼応表現自動抽出のための正解データ作成. 自然言語処理学会第 10 回年次大会発表論文集. D1-03.

工藤浩．(2000)． 副詞と文の陳述のタイプ．『日本語の文法３モダリティ』(森山卓郎，仁田義雄，工藤浩)．岩波書店．161-234.

南不二男．(1974)．『現代日本語の構造』．大修館書店．

Srdanović Erjavec, Irena, Bekeš, Andrej, Nishina, Kikuko. (2007). “Cluster analysis of suppositional adverbs and clause-final modality.” In Asian and African Studies: Languages and Realities of China and Japan. Volume XI. Issue 3. Ljubljana: University of Ljubljana, Faculty of Arts. 21-31.

スルダノヴィッチ・イレーナ，ベケシュ・アンドレイ，仁科喜久子．(2008)複数のコーパスに見られる副詞と文末モダリティの遠隔共起関係．特定領域研究，「日本語コーパス」平成 19 年度公開ワークショップ(研究成果発表会)予稿集．223-230.

佐野大樹，丸山岳彦．(2008)．システミック文法に基づく書きことばの複雑さ測定−日本語大規模コーパスを用いた語彙密度計測−．自然言語処理学会第 14 回年次大会発表論文集. 1097-11

Yoshihashi Kenji, Nishina Kikuko. (2007). “Japanese Composition Support System Displaying Co-occurrences and Example Sentences.” In Proceedings of the International Symposium on Large-Scale Knowledge Resources March 2007. 119-122.