アカデミック・ライティングに見られる副詞に関する分析

阿辺川 武（国立情報学研究所）†
八木 豊（株式会社ピコラボ）
ホドシチェク・ボル（大阪大学言語文化研究科）
仁科 喜久子（東京工業大学名誉教授）

Analysis of Adverb in Japanese Academic Writing

Takeshi Abekawa (National Institute of Informatics)
Yutaka Yagi (Picolab Co., Ltd.)
Hodošček Bor (Osaka University)
Kikuko Nishina (Tokyo Institute of Technology)

要旨

我々は BCCWJ に科学技術論文を加えたコーパスを使用してレジスター誤り検出を行う日本語作文推敲支援システム「ナツメグ」を開発した。システムでは、アカデミック・ライティングの文体に近い準正用コーパスと、話し言葉を多く含む準誤用コーパスでの使用頻度の比を利用して、レジスター誤りと思われる表現を検出しているが、準正用コーパスでの頻度が高いにもかかわらず、システムが誤用と判定してしまう表現が存在する。本発表ではシステムの検出精度の向上をめざし、誤検出となる表現の中から、話し言葉と書き言葉のレジスターの異なりが顕著に見られる副詞に着目し、分析をおこなった。準正用コーパス中で頻度上位の副詞について、実際に用いられている文脈を参照し、書字形および語彙素別にまとめあげ、日本語教育の専門家の意見を参考にしながら、アカデミック・ライティングとしてふさわしい表現であるかを分析した。

１．はじめに

日本の大学で学ぶ理工系留学生は日本語での実験レポート、授業での課題レポート、卒業論文、学位論文、投稿論文が必要になることが想定される。これらをアカデミック・ライティングというジャンルの一部と考え、このジャンルの作文支援をすることを目的に作文支援システム「ナツメグ」の開発を進めている。「ナツメグ」は学習者が論文などの文章を入力すると、システムが入力された表現が適切か否かを判定し、不適切な表現の場合は、適切なヒントを提示することを目指している(八木ら 2014a)。

学生たちは初級から中級に至るまで、主として話し言葉を中心に学んでいるため、上級になって「である体」あるいは「だ体」の書き言葉による文章を学んでも、いざ書く場合になって、どのような用語を用いるかを習得できていないことがある。次の例文は我々が作成した学習者作文コーパス「なたね」の中にある理系学部１年生による１文である。

例 1：今日本では片仮名で書くのは ちょっと多いと聞いたことがある。意味は同じだが、片仮名で書き直したらなんだか新鮮でファッションな、おしゃれな 感じがするようになる。もし先生という言葉 は平仮名で書くとすぐ親切な先生が思い出す。ほんとに器用な言語と思う。

この文中で「ちょっと」「なんだか」「すぐ」「ほんとに」は話し言葉であり、アカデミックな文章では用いられない。「ちょっと」は「やや」に、「ほんとに」は「実に」などで言い換えることができる。

† abekawa [a] nii.ac.jp

本稿では作文推敲支援システムの開発にあたり学習者の文章を観察した結果、このような不適切な表現が見られる中で特に副詞に注目した。副詞を取り上げた理由として、他の品詞と比較すると論文などで用いられる副詞の数はかなり限られていること、また話し言葉と書き言葉のレジスターの異なりが顕著に見られること、そしてシステムの誤用判定と教育者の誤用判定結果が異なる表現が少なからず存在することからである。文末表現や句と句、文と文の接続などの機能語にも不適切な表現が見られるが、これらは共起関係や他の語との意味的関係を考慮しなければならないことも多く、定量的な分析が困難である。それ対して、副詞は独立した品詞として抽出しやすく、分析の緒としては適切だと判断した。

２．使用するコーパスと誤用判定の仕組み

話し言葉と書き言葉という対立、砕けた文章と堅い文章という対立、小説やエッセイなど主観や感性を重視する文章に対する学術的な客観性を重視する文章などのジャンルは多様であり、そこで用いられる言語表現も異なっている。このようにジャンルによって異なる表現のヴァリエーション（言語変異）を語のレジスターと呼ぶ(Halliday 1976)。本研究では理系留学生に必要とされるアカデミックなレポート・論文のための日本語表現をアカデミック・レジスターと定義し、開発中のシステムがその条件にふさわしい表現か否かの判定をすることで、目標とする文章を向上させることを我々は目指している。

システムのために用意するコーパスは国立国語研究所で開発した「現代日本語書き言葉均衡コーパス」（以下 BCCWJ と呼ぶ）および独自に収集した科学技術論文である。このコーパス中の副詞を分析対象とする。コーパスの中でアカデミックな文章に近いものを準正用データ、アカデミックな文章から遠いものを準誤用データとし、アカデミック・ライティングに適合した表現か否かを判定し、適切な表現に導くという手続きを取る。準正用データに含まれるデータは「科学技術論文」データと BCCWJ の中の「白書、法律」データである。これらの文書は、論文に準じる語彙と文体からなると判断した。一方、準誤用データは、同じく BCCWJ の中の「Yahoo!ブログ、Yahoo!知恵袋、国会会議録」である。Yahoo!ブログと知恵袋は、書き言葉であるが情緒的で口語的な表現が多い。国会会議録は、話し言葉を書き起したものであるため、話し言葉の要素が大きく、この 3 データはアカデミックな文章とは対称的なものであると判断し、準誤用データと位置づけた。その他の一般的な「書籍、雑誌、広報誌、新聞」など、どちらにも属していない中立のデータ群も有意差を決定するために用いている。これらのコーパスは、UniDic に基づいてデータが構成されおり、語は語彙素の下に語形があり、その下に書字形・発音形がある（伝ら 2007）。語彙素の下はさまざまな表記のヴァリエーションとしての書字形からなり、1 語彙素に対して、1 から十数個までの書字形が存在する。したがって、システムにおける語の頻度を計算するに当たっては、語彙素と書字形の関係に注意を払わなければならない。語彙素は意味用語を同一にする語形の集合で見る方が良い場合、語形はテキスト上でその語がどのような用字法で記載されているかを見分ける方が良い場合というそれぞれの観点で必要な単位であり、それぞれ分析時に使い分ける必要がある。日本語表記については、英語などのような一国の言語としての正書法が存在しないが、その補佐的なものとして文部科学省が公示した「公文書要領」があり、国の公文書はその指針に従って作成している(文部省 1960)。しかし、新聞、雑誌、その他の出版物は、それぞれの会社や機関が定めた文書作成規則に従って作成しており、強い拘束力はない。

ここで､我々が注目する副詞は書字形で約 7,400 項目存在する｡これらの書字形ごとに(ホ

ドシチェク 2011)の判定式を施すことで各項目の語についてレジスターとしての可否を判定する。例えば「良く」という語彙素は「よく、良く、ヨク、よーく」などの 15 種の書字形からなっている。全コーパスの語に対して頻度計算をした後、準正用データと準誤用データ間の使用頻度の差および有意差の有無によってアカデミック・レジスターとしての可否を示すことになる。

システムでは学習者によって入力された語の妥当性を判定式によって統計的に処理し、その語が有意に誤用と判定されれば、その語はアカデミックな文章としては適切でないため、学習者に注意が喚起される。学習者はこの喚起によって、不適切な用法に気づき、自ら適切な用法を検討するように導かれる(八木ら 2014b)。

表 1: 各コーパスで頻出する副詞（語彙素別、PPM:100 万形態素あたりの相対頻度）

	全体		準正用		準誤用			全体		準正用		準誤用	
順	語彙素	PPM	語彙素	PPM	語彙素	PPM	順	語彙素	PPM	語彙素	PPM	語彙素	PPM
1	そう	910.4	例えば	408.3	どう	1,537.4	15	最も	161.8	良く	62.6	直ぐ	245.8
2	どう	827.9	最も	262.7	そう	1,373.9	16	何故	161.6	予め	57.1	迚も	244.6
3	もう	434.2	特に	250.2	もう	690.8	17	全く	161.1	一層	55.8	可成	237.2
4	こう	411.3	先ず	249.3	こう	685.7	18	更に	157.9	極めて	54.6	何故	235.5
5	良く	300.5	より	227.5	矢張り	572.9	19	詰まり	157.4	余り	52.2	特に	227.6
6	未だ	259.4	どう	162.1	一寸	490.6	20	一番	153.1	可成	51.5	全く	203.4
7	例えば	251.1	更に	131.9	良く	404.0	21	余り	148.9	主に	51.5	宜しく	195.2
8	少し	243.0	略	119.2	少し	395.7	22	若し	145.6	未だ	47.9	勿論	194.7
9	先ず	230.6	こう	97.8	未だ	379.1	23	既に	145.4	もう	46.3	例えば	187.8
10	矢張り	227.3	詰まり	94.2	一番	300.0	24	勿論	143.7	全く	44.0	中々	186.0
11	特に	212.0	そう	84.7	又	282.7	25	迚も	130.3	十分	34.5	結構	176.9
12	又	208.0	必ず	79.5	余り	278.8	27	初めて	129.6	次いで	32.0	もっと	171.8
13	一寸	207.8	既に	78.7	色々	257.5	28	より	119.9	やや	31.9	初めて	163.2
14	直ぐ	186.1	直接	66.9	先ず	251.7	29	可成	117.6	若し	30.1	ずっと	130.7
15	最も	161.8	良く	62.6	直ぐ	245.8	30	もっと	115.5	何故	26.4	必ず	121.0

3．準正用データと準誤用データの比較

本システムが用いるコーパス全体および準正用データと準誤用データにおける語彙の構成について、その様相を概説する。表 1 はコーパス全体、準正用、準誤用データの語彙素別の副詞上位 30 位までを示している。全コーパスでは上位 30 位までで 53.29%をカバーしている。準正用データでは、30 位までで 71.54%、100 位では 91.26%をカバーしている。全コーパスにおけるカバー率と比較すると、テキスト中での副詞の使用が限られた高頻度語に集中していることがわかる。一方、準誤用データの上位 30 位までのカバー率は 58.70%であった。これにより、アカデミック・レジスターでは、他のグループより限られた副詞で文章が構成されていることがわかる。準正用コーパスと準誤用データの頻出副詞の異同を見ると、不一致語の中で準正用には存在せず、準誤用のみに見られる語は「一番、もっと、一寸、勿論、矢張り」など 17 語あり、これらの語が学習者コーパスの中でしばしば見られ、論文として違和感を与える一因になっている。

４．アカデミック・レジスターとして不適切とされた副詞の分析

システムの判定結果の妥当性を検証するために人手による判定と比較する観察実験を行った。その結果、システムが誤用と判定したものの中に日本語教育の専門家が科学技術論文のレジスターとして適切であると評価したものが少なからず存在した。両者の不一致の原因を知るために 1) 複数の書字形を有する副詞、2)高頻度副詞「こう」「そう」「どう」についての分析をおこなった。

４.１ 複数の書字形を有する副詞「矢張り」

先に述べたようにシステムが利用する語彙データは、BCCWJ で用いられている UniDic に依拠している。語彙素は書字形の異なる形を一つの概念としてまとめる語の抽象的な集合と言える。書字形を多く有し、システムが誤用であると判定した語として「矢張り」を例に問題点を述べる。

語彙素「矢張り」は語形「ヤハリ」「ヤッパリ」「ヤッパ」に分かれ、更に書字形「矢張り」「やはり」「やっぱり」「ヤッパリ」「矢っ張り」「やっぱ」などの話し言葉の発音に近い形として出現する。各コーパスにおける相対出現頻度を表 2 に示す。それぞれの書字形についてのシステムの判定は、「やはり」「やっぱり」が誤用となっている他は、低頻度のため判定不可(NA)となっている。「公文書要領」によると、語彙素「矢張り」は平仮名の「やはり」が推奨されているが、準正用データにおいては 78.5%、コーパス全体では 60.3%、準誤用データでは 58.2%であり、準正用データにおける表記法が他に比べて規範に沿っていることがわかる。なお準正用データでも「やっぱり」「ヤッパリ」のような砕けた口語を含んでいるが、これは論文中に引用した文芸作品などの引用と推測される。

表 2: 語彙素「矢張り」の相対頻度 (単位 PPM)

書字形	システムの判定	全体	準正用	準誤用
やはり	誤用	137.1	11.3	323.5
やっぱり	誤用	77.5	1.6	205.7
やっぱ	N/A	10.5	0.3	39.2
矢張り	N/A	0.8	0.5	0.6
やっぱし	N/A	0.7	0.1	1.7
ヤッパリ	N/A	0.3	0.4	1.0
やぱ	N/A	0.3	0.0	1.1
やつぱり	N/A	0.2	0.0	0.0
矢っ張り	N/A	0.1	0.1	0.0
矢つ張り、矢ッ張り、やッぱり、矢っ張、矢つ張	N/A	0.0	0.0	0.0

４.２「こそあど」語彙からなる副詞

「こそあど」語彙からなる副詞「こう、そう、ああ、どう」の占める割合は全ジャンルを通して非常に多く、準正用データにおいても「ああ」を除いて高頻度語に位置している。全体コーパス、準正用、準誤用の順で「そう」(1 位、11 位、2 位)「どう」(2 位、6 位、1 位)「こう」(4 位、9 位、4 位)である。科学技術論文では、「このように、そのように、どのように」という書き言葉の表現が併用されるため、「こう、そう、どう」の頻度が相対的に低くなっていると考えられる。しかし、システムによるレジスター判定では準正用データで高頻度であるにもかかわらず、これらの副詞が誤用となっている。このような様相

を科学技術に論文におけるレジスターの問題として検討する。なお、「ああ」については、準誤用データにおいて用例が存在するが、準正用データの中では、「ああ」が使用される例は極めて少なく、論文中に言語分析のための例文が入っているテキスト以外には見られない。一方、学習者コーパスにおける作文では「ああ」の使用がしばしば見られる。

４.２.１「こう」

全体コーパスで第 4 位、準正用データで第 9 位、準誤用データで第 4 位とどのコーパスにおいても高頻度であるが、3 データの比から計算すると判定式は誤用となる。しかしながら、準正用データにおける使用頻度は少なくはない。準正用データ中でどのような用法があるのか見るために、「こう」に続く連語をみると、「こうした」(74.8PPM)「こうして」(8.7PPM)が高頻度で出現し、これらの連語が準正用データにおける「こう」の 85.3%を占めている。これらの連語は文章中の前方照応の機能を果たしていることが多い。

例 2:こうして収集された日本語の用例文を翻訳家に英訳してもらう。(科学技術論文.自然言語処理.言語処理学会予稿集)

副詞句「こうして」、連体詞句「こうした」は話し言葉や砕けた文章にも見られる「こう」から派生した連語であり、改まった文章では「このようにして」「このような」という論文などでよく見られる形態に置き換えることができる。また、更に砕けた表現として「こんな（に）」との対照があるが、すべて準誤用での用法が多く、準正用ではほとんど見られない。これらの観察の結果として、アカデミック・レジスターとしては「こういう」は用いられることが少なく、「こうした」は準正用が準誤用より多いことがわかる。この観察から「このような／に」をアカデミック・レジスターとして認め、「こうした」もこれに準じて許容しでもよさそうである。

表 3: 副詞「こう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
こう	形態素	411.4	97.8	685.7
こうして	複合語	44.6	8.7	19.1
こうした	複合語	100.2	74.8	36.2
こういう	複合語	130.8	1.4	391.4
こう言う	複合語	2.7	0.1	4.5
こう云う	複合語	0.2	0.0	0.2
このような	複合語	148.9	255.1	71.1
この様な	複合語	1.6	1.6	4.0
このように	複合語	79.4	111.2	45.5
この様に	複合語	0.4	0.6	0.6
このようにして	複合語	7.3	12.6	0.8
こうやって	複合語	5.5	0.1	8.7
こんな	形態素	200.9	3.7	333.4

４.２.２「そう」

「そう」はコーパス全体で頻度 911PPM であり、副詞頻度の最高値である。準正用データでは 84.7PPM、準誤用データ 1,370PPM であり、システムの判定では誤用となる(表 4)。「こ

う」と同様に、判定結果が「誤用」であるにもかかわらず、準正用での出現頻度は低くない。そこで、「こう」の場合と同様に、後に続く語をみると、「A は B。そういう X は～」「A は B。そういった X（状態、状況）は～」などのような表現であり、科学技術文章の中では、慣用的な文型といえる。また、「A が B である場合～、一方 A がそうでない場合」というような前方照応の定型的な表現も多く見られる。（例：「ペアが含まれるなら真、そうでないなら偽である」）。これは、前文の内容を言い換えた代言（パラフレーズ）表現と言える。「そのよう」との対応を考えると、「そのようでないなら」という言い換えはできない。肯定表現では「そのような場合には」はとなり、「そう」は出現しない。

一方、「そう解釈できる」は「そのように解釈できる」と書き換えることが可能であり、「そういう」「そういった」「そう解釈できる」は、前記の用法より、科学技術文章の一般的な表現からやや遠い表現だと思われる。実例をみると、「そうして、そうした、そういった、そのように、そのような、そんなに、そんな」の連語において、準誤用データに圧倒的に多く用例があり、準正用データの例は少ない。「そうした」「そのような」は正用データ中でやや多く見られるが、いずれも全データ中の 10%以下である。結論として、「そう」の用法からすべての「そう」を科学技術論文レジスターから排除するのではなく、「A は B である。そうでない場合、A は～」「そういった」「そういう」などのように文脈上、前方照応の機能を担ったフレーズを正用として認めるなどの措置は有用であろう。

表 4: 副詞「そう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
そう	形態素	910.6	84.7	1,373.7
そうして	複合語	20.2	0.7	19.5
そうした	複合語	57.0	10.2	38.4
そういう	複合語	229.0	2.7	555.8
そう言う	複合語	8.7	0.2	9.5
そう云う	複合語	0.5	0.0	0.3
そのような	複合語	54.6	48.2	53.0
その様な	複合語	0.8	0.3	3.0
そのように	複合語	13.3	2.6	19.6
その様に	複合語	0.2	0.0	0.5
そのようにして	複合語	1.3	0.2	0.8
そうやって	複合語	6.6	0.1	6.4
そんな	形態素	316.2	5.1	443.0

４.２.３「どう」

「どう」は全コーパスで 828PPM(2 位)、準正用データ 162PPM(6 位)、準誤用データ 1,540PPM(１位)であり、準誤用データの中では最も多用される副詞である。その中で準正用データ中に顕著な句構造をみると、連語としての「かどうか」(148PPM)の頻度は高く、他の連語「どういう」、「どうすれば」、「どうしても」などと比較しても抜群に高頻度である。また、「どう考えるか」などのように「どう」の後に動詞が来て、「か」で結ぶ係り受け構造となるものがある。

書き言葉では「どう」は「どのように」とする方が、フォーマルな表現とされているた

め、その差を見ることにする(表 5)。準正用データでは「どのように」が「どう」の約 2.6 倍、一方、準誤用データでは「どう」の使用が「どのように」の約 21.3 倍となり、準正用では「どのように」の使用割合が高いことがわかる。「どういう」も「どのような」とフォーマルな表現に書き換えられるが、同様に相対頻度の比をみると、準正用では「どういう」:「どのような」を語彙素で比較すると、1:18、準誤用ではほぼ 2.6:1 と全く逆の使用頻度となる。従って、「どういう」を「どのような」へと書き換えすることを推奨すべきである。

さらに、「かどうか」と「か否か」の対比をみると、準正用では「かどうか」:「か否か」がほぼ 2.3:1、準誤用ではほぼ 27.4:1 となり、準正用では,準誤用のほぼ 12 倍になる。「かどうか」についても「か否か」への書き換えを推奨することが考えられる。また「どういった」は、やや書き言葉的な傾向があるが、これも「どのような」に書き換えられるものである。「どういった」と「どういうような」の用法は、例 3 のように執筆者個人の嗜好によることが多いように思われる。

例 3: このような箇所を読むことで，著者がどういった目的でその論文を参照したのかがわかる.（科学技術論文.自然言語処理）

以上「どう」についてまとめると、「どう」「どういう」「どんな/に」はアカデミックな分野で多用される「どのように/な」に置き換える指示を出し、アカデミック・レジスターとして書き換えを認めるべきであろう。また、「かどうか」は「か否か」への書き換える方が適切であるが、実際は「かどうか」が多用されているので、その許容の程度は検討する必要がある。

表 5: 副詞「どう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
どう	形態素	828.25	162.10	1,537.40
どうして	複合語	120.44	4.27	174.91
どうした	複合語	48.62	1.78	92.58
どういう	複合語	77.66	5.14	177.60
どう言う	複合語	0.78	0.00	2.78
どう云う	複合語	0.22	0.00	0.06
どのような	複合語	67.38	91.54	66.40
どの様な	複合語	1.02	0.71	3.69
どのように	複合語	59.48	61.18	72.91
どの様に	複合語	1.12	0.44	4.21
どのようにして	複合語	6.03	3.06	5.92
どうやって	複合語	22.02	1.18	47.21
どんな	形態素	158.63	12.57	251.99
かどうか	複合語	112.22	108.25	147.73
か否か	複合語	18.99	47.43	5.44

５．おわりに

高頻度副詞群の中にあって、システムの判定は誤用とされている語が存在し、その中で日本語教育の専門家の判定が正用となる語が少なからず存在した。専門家の判定は論文指

導者とも近いと考えられ、学習者がシステムを使用する際に、専門家が可とする語をシステムが誤用と判定すると、学習者に混乱を招く可能性が予測される。矛盾と思われる要因は、1）判別式の欠陥、2）データの偏りなどが考えられる。この矛盾を解消するために 1）については、頻度の閾値を人手の判断も加味しながら再検討し、比較的頻度の高いものは、論文執筆において使用が認められる語であるとすることも可能である。準正用の頻度が一定の値を超えていれば、正用とするという条件を加えることも検討の余地がある。ある程度高頻度であり、かつ専門家が可とする場合は、システム判定式に対して追加条件を設けることも一策であろう。

2）については、現時点で使用している各データに考慮すべき問題がある。準正用の論文データ中に言語処理、言語を扱うものがあり、その論文の中にかなりの割合で、話し言葉を含む例文が存在している。そのため、誤用データに属するような語が出現している。これを解消するためには、言語学・言語処理以外のさらに多くの論文データを投入することが考えられる。

以上、今回の副詞の分析を通して、判別式の問題、データ構造の問題とともに、語解析の問題も見えてきた。形態素を超えた連語・イディオムの扱い方、語彙素と書字形の問題などであり、副詞以外の語彙についても発展できる可能性が見られた。例えば、機能語、形容詞、形容動詞においても、同様の分析をすることで、判定式の精度をあげることも考えられる。また、学習者に対する対策として、混乱を防ぐためにも規範的な規則も導入し、リスト化したデータからヒントを提示する可能性があることを示した。これらをシステムに反映することで、精度の向上に努めることを今後の課題とする。

謝　辞

本研究は、文部科学省科学研究費補助金基盤研究 C「日本語作文支援システムにおける誤用の検出及び添削に有用な情報の提示法の研究」（平成 27～29 年度、代表者：阿辺川武）による補助を得ています。

文　献

八木豊、ホドシチェク・ボル、阿辺川武、仁科喜久子、室田真男（2014b）「作文推敲支援システムによる誤り指摘への学習者の対処に関する調査」日本教育工学会研究報告集 No.14(5)、pp.151-156.

八木豊、ホドシチェク・ボル、阿辺川武、仁科喜久子（2014a）「日本語作文推敲支援システム「ナツメグ」における誤用検出手法の評価」第 5 回コーパス日本語学ワークショップ予稿集、pp.167-170.

ホドシチェク・ボル、仁科喜久子（2011）「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会 異文化コミュニケーションのための日本語教育 2、pp.522-523.

伝康晴、小木曽智信、小椋秀樹、山田篤、他（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」日本語科学 22 号、pp.101-122.

Halliday, M. and Matthiessen, C. (2004) An Introduction to Functional Grammar (3rd Edition), Routledge

文部省(1960)、公用文の書き方―資料集―:
http://kokugo.bunka.go.jp/kokugo_nihongo/joho/series/21/21.html

関連 URL

日本語学習者作文コーパス「なたね」: https://hinoki-project.org/natane/

日本語作文推敲支援システム「ナツメグ」: https://hinoki-project.org/nutmeg/