コーパスを用いた類義表現研究

小西　円（東京学芸大学）

日本女子大学文学部学術交流企画主催
機能後用例文データベース「はごろも」研究会共催
公開シンポジウム　コーパスを使った類義語・多義語研究
2018.12.22

1

（1）これまでの類義表現研究

2

これまでの類義表現研究

大きな研究の進歩

例）　逆接の接続助詞「けれども」「のに」「ても」

（1）形（形態や統語のルール）

例）「終止形＋けれども」、「連体形＋のに」、「て形＋も」

例）○「だろうけれども」　／　×「だろうのに」、「だろうても」

（2）意味

例）「薬を飲んでも治らなかった」　←「飲めば治る」という予測が実現しなかった

例）「薬を飲んだのに治らなかった」　←「それはおかしい」「驚き」というニュアンス

例）「けれども」には独自のニュアンスがない　　（前田1995、2009）

3

これまでの類義表現研究 ―日本語教育から見ると―

（1）　内省による分析が中心

異なる内省判断？

??「雨が降っているのに、外で遊びなさい」

??「雨が降っているのに、外で遊びたい」

??「雨が降っているのに、外で遊びますか」（前田1995、2009）

（2）　意味記述が中心

（3）　レンマによる記述が中心

意味記述の捉え方が学習者によって異なる

レンマ：辞書の見出し語　　出現形：実際に使う形

「けれども」には「けど」「けども」「けれど」が含まれる

実際に使うのは個々の出現形

（小林・小西・砂川・清水・奥野2016）

4

(2)コーパスを用いた類義表現研究

5

コーパスを用いた類義表現研究

出現数の比較・出現レジスター(ジャンル)の比較

例)接続詞「しかし」「でも」「だが」「ところが」(清水2017)

検索対象:BCCWJ

	しかし	でも	だが	ところが
出現数	全68,609例	全36,706例	全17,871例	全11,394例
文体	常体(70%) 敬体(30%)	常体(10%) 敬体(90%)	常体(100%) 敬体(0%)	常体(60%) 敬体(40%)
レジスター(PMW)	図書館・書籍(1075.7) 国会会議録(780.0) 教科書(7116)	ブログ(1095.3) 知恵袋(879.8) 白書(ほぼなし)	新聞(356.9) 図書館・書籍(344.0) 白書、知恵袋、国会会議録(ほぼなし)	図書館・書籍(208) 国会会議録(204.8)

6

コーパスを用いた類義表現研究

前接語や前接品詞の比較

前文脈や後文脈に共起する語(副詞や文末表現)の比較

例)逆接条件「としても/にしても」

『日本語文型辞典』

「XとしてもY」: 仮に事実がXであっても/成立していてもYの成立や阻止に有効に働かない

「…にしても」: …で述べられているような事態であることを仮に認めた場合でも、

7

コーパスを用いた類義表現研究

前接語や前接品詞の比較

前文脈や後文脈に共起する語(副詞や文末表現)の比較

例)逆接条件「としても/にしても」(中俣2017)

検索対象:BCCWJ

	～としても	～にしても
出現数	全6,520例	全961例
前接語(機能語)	た(4,668)だ(460)Vない(176) ている(135)である(85)	た(186)Vない(140)れる・られる(33) である(23)てもらう(9)
「Vた」となる場合のV	ある(457)なる(156)する(59)	

8

コーパスを用いた類義表現研究の特徴

（1） 実際の使用例に基づいている

（2） 計量的（量的）な研究 → 意味記述の可視化

（3） 量的 かつ 質的な分析

（4） 規則の抽出だけではなく、傾向の抽出

9

（3）文体情報を用いた類義表現研究の可能性
―日本語教育の観点から―

10

レジスター（ジャンル）とは

BCCWJにおけるレジスター

- 書籍（出版書籍、図書館書籍）
- 新聞
- 雑誌
- 政府刊行白書
- 教科書
- 広報誌
- ベストセラー
- Yahoo!知恵袋
- Yahoo!ブログ
- 韻文
- 法律
- 国会会議録

出版サブコーパス
図書館サブコーパス

特定目的
サブコーパス

「書籍」にはいろんな
タイプの本がある
→ ここを可視化したい

11

『BCCWJ 図書館サブコーパスの文体情報』の活用

- BCCWJに収録されている図書館サブコーパス（LB）10,551サンプルに文体情報を付与
- 9名の判定者
- ① 文体判断が可能かそうでないかを判断
- ② 文体判断が可能であれば以下の5点について判定
  - （a） 専門度 1 専門家向き ← → 5 小学生・幼児向き
  - （b） 客観度 1 とても客観的 ← → 4 とても主観的
  - （c） 硬度 1 とても硬い ← → 4 とても軟らかい
  - （d） くだけ度 1 とてもくだけている ← → 3 くだけていない
  - （e） 語りかけ性度
    1 とても語りかけ性がある ← → 3 特に語りかけ性はない

12

文体情報の判定例(柏野2013より)

専門度:1　専門家向き　(LBi4_00021『がんと遺伝子』)

E2F以外のRB結合タンパク質としては，転写因子RAX，T細胞が活性化するときに誘導されるIL－2，GM－CSF，HIV－2などの転写を活性化する転写因子EIF－1や先に述べた細胞周期を制御するサイクリンDなどがある．おもしろいことに，EIF－1やサイクリンDのRB結合ドメインにはlargeT抗原やE1Aタンパク質と同じようにLXCXEというアミノ酸配列が存在する．また，RBタンパク質は骨格筋分化を支配する重要な遺伝子群My oDファミリー(MyoD，myogenin，MRF4，myf－5)の産物とも複合体を形 成し筋分化にも関与しているらしい．

専門度:4　中高生向き　(LBf9_00090『超魔炎獄変』)

白く薄い空気のヴェールが、漂うように揺らめいている。

シャ…ァアン、シャラ…ァアン…。

闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。

響きに道を開けるかのようにすぅ…っと霧が左右に分かれた。

それは。霧の中にたたずむそれは。闇。…いや。闇ではない。

13

文体情報の判定例(柏野2013より)

客観度:1　とても客観的　(LBo3_00158『行政法要論』)

たとえば委員会の開催が「急施を要する場合」にあたるかどうかとか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、覊束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適当と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政 治的判断に委ねらるべきであり(以下略)

客観度:4　とても主観的　(LBo3_00132『教師をめざす若者たち』)

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みを一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。

14

文体情報の判定例(柏野2013より)

硬度:1　とても硬い　(LBi3_00033『現代法社会学入門』)

取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。

硬度:4　とても軟らかい　(LBa4_00010『恐竜の世界をたずねて』)

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」(図86)という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

15

文体情報の判定例(柏野2013より)

くだけ度:1　とてもくだけている　(LBf9_00067『男はオイ！女はハイ…』)

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとついってやれとばかりすぐ電話にとびついた。

「ハイ、こちら－です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

語りかけ性度:1　とても語りかけ性がある　(LBt1_00013『5分間集中力トレーニング』)

精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。

動物病院にいらっしゃる飼いまさんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼いまさんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。

それでは、どうすればいいのでしょうか。

16

調査の対象と方法

対象：　逆接の機能表現16種（接続助詞に限定）

機能表現一覧はtutujiの分類による（松吉他2007、2008）

方法：（1）図書館・書籍サブコーパスから対象となる語を検索

（2）用例を確認し、ごみ取りを行う

（3）『BCCWJ図書館サブコーパスの文体情報』の判定例とマッチングさせる

（国語研：浅原正幸氏の協力を得る）

（4）サンプルの重なりを除きサンプル異なり数を算出

例）機能語Aが同一サンプルに複数回出現している場合、「1回」とカウントする

（5）サンプル異なり数をもとに、判定例の平均を算出

17

調査の対象

	調査対象語	調査対象数	サンプル異なり数		調査対象語	調査対象数	サンプル異なり数
カラトイッテ類	からといって	841	706	トシテモ・ニシテモ類	としても	2718	1962
クセニ類	くせに	503	434		としましても	3	3
	にしては	535	502		としたって	26	24
	わりに	294	270		にしても	1811	1309
ケレドモ類	けれども	2019	927		にしましても	7	7
	けれど	4155	1742		にしたって	94	86
	けども	181	98		にせよ	1123	803
	けど	11347	2673		にしろ	628	405

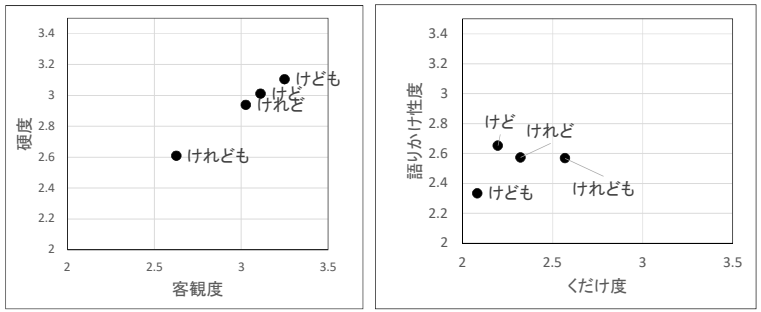
先行研究：馬場俊臣（2018）

・『BCCWJ図書館サブコーパスの文体情報』データを用い、語の文体差を数値化

・専門度、客観度、硬度、くだけ度の4指標は相互に強い相関がある。

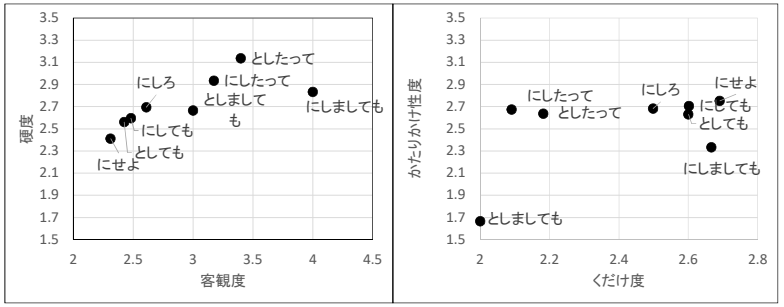
18

「けれども」「けれど」「けども」「けど」



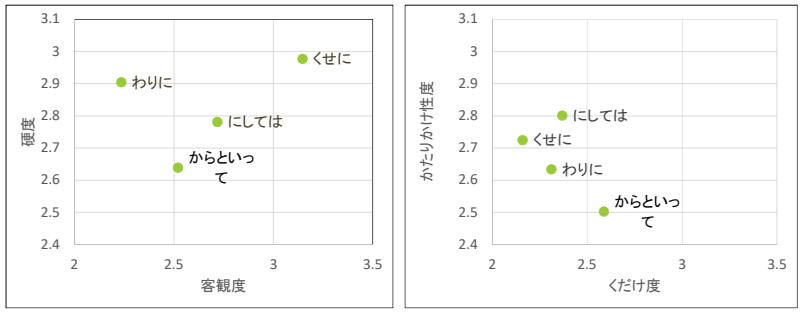
19

トシテモ・ニシテモ類



20

「からといって」「くせに」「にしては」「わりに」



21

硬度からみた機能表現と語

語に関しては馬場(2018)より

形式	にせよ	としても	にしても	けれども	からといって	としましても	にしろ	にしては	にしましても	わりに	にしたって	けれど	くせに	けど	けども	としたって
硬度	2.41	2.56	2.60	2.61	2.64	2.67	2.69	2.78	2.83	2.90	2.94	2.94	2.98	3.01	3.10	3.14

モデル(外来語)
シナリオ(外来語)
キーワード(外来語)
ビーク(外来語)
システム(外来語)

元気(漢語)
魔法(漢語)
洗濯(漢語)
暢気(のんき)(漢語)
元気(漢語)
頂戴(ちょうだい)(漢語)
ベランダ(外来語)
キッチン(外来語)

ちゃう(助動詞)
ふうん(感動詞)
本当(漢語)
すうっと(副詞)
ずーっと(副詞)
こら(感動詞)

22

文体情報を用いた類義表現研究の可能性

・硬度や客観度のような指標から識別できる類義表現の場合、有効。

・意味の差が大きく、硬度や客観度では識別できない類義表現の場合は、有効性が低い。

・語と機能表現を、同様の指標で並べたときに、その機能表現をもちいる場合の典型例が作成できるのではないか。

23

引用文献

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4-1

小林ミナ・小西円・砂川有里子・清水由貴子・奥川育子(2016)「類義表現分析の可能性」砂川有里子編『講座日本語コーパス5 コーパスと日本語教育』朝倉書店

清水由貴子(2017)「逆接を表す表現」中俣尚己編『現場に役立つ日本語教育研究5 コーパスから始まる例文作り』くろしお出版

中俣尚己(2017)「条件・逆接条件を表す表現」中俣尚己編『現場に役立つ日本語教育研究5 コーパスから始まる例文作り』くろしお出版

馬場俊臣(2018)「『BCCWJ図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性」『言語資源活用ワークショップ2018発表論文集』

松吉俊・佐藤理史・宇津呂武仁(2007)「日本語機能表現辞書の編纂」『自然言語処理』14-5

松吉俊・佐藤理史(2008)「文体と難易度を制御可能な日本語機能表現の言い換え」『自然言語処理』15-2

前田直子(1995)「ケレドモ・ガとノニとテモ」宮島達夫・仁田義雄編『日本語類義表現の文法(下)』くろしお出版

前田直子(2009)『日本語の複文 条件文と原因・理由文の記述的研究』くろしお出版

本研究は科学研究費助成事業(課題番号18K12420)の研究成果の一部である

24