共起語率の分布からみるテキストの語彙的特徴

山崎 誠（国立国語研究所言語資源研究系）†

Lexical Characteristics of Text as Seen in the Distribution of Co-occurrence Rate

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

「現代日本語書き言葉均衡コーパス」（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ と略す）が 2011 年に完成し、それを利用した日本語研究のさまざまな展開が期待されている。BCCWJ の特徴として、多様な日本語を収録していることやアノテーションの充実が挙げられる。それらを生かした研究が今後発多く発表されることと思われる。本発表では BCCWJ のアノテーション情報を利用してテキストの結束性に関する特徴を捉える試みを紹介する。

２．テキストにおける結束性

結束性（cohesion）とは、文章をひとつの統一体としてまとめあげるために必要な性質のひとつとされる。結束性について最初に詳細に研究を行ったのは Halliday & Hasan(1976)である。それによると、結束性について次のように紹介されている。

「結束性が生じるのは，談話のある要素の解釈（INTERPRITATION）が別の要素の解釈に依存する場合である。一方を効果的に解釈するためには他方に頼らなければならないという意味で，一方は他方を前提（PRESUPPOSE）とする。こういうことが生じるとき，結束関係が成立する。その結果，前提語と被前提語という 2 つの要素が、少なくとも潜在的には，統合されて 1 つのテクストになるのである。」（邦訳 p.5）

庵(2007:12)によれば、結束性は推論にもとづくつながりである一貫性(coherence)の下位概念であるとされる。また、結束性には文法的結束性と語彙的結束性とがあり、前者の手段として「指示」「代用」「省略」が、後者には「再叙（reiteration）」と「コロケーション」がある[1] 。再叙には以下の 4 つのタイプがある。

(a)同一語（繰り返し）
(b)同義語（または近似同義語）
(c)上位語
(d)一般語

Károly(2002:162)によれば、英語の作文においては、(a)の同一語の繰り返しよりは(b)～(d)を合わせた「異なる語の繰り返し」の方が多く用いられるということだが、同義語（類義語）や上位語の判断を自動的に行うことが難しいため、本発表では(a)の同一語の繰り返しのみを観察対象とする。同一語の繰り返しは、本発表で用いた図書館書籍のデータでは、10,369 サンプル中同一語の繰り返し[2]が無かったサンプルは 17 個しかなかった。それらはいずれも延べ語数 22 語以下の小さなサンプルで、サンプルの短さがその原因である。ある程度の長さを持つテキストには必ず同一語の繰り返しがあると言ってよいだろう。

† yamazaki@ninjal.ac.jp

1 Halliday & Hasan(1976)では、文法的結束性と語彙的結束性の中間の性質を持つものとして「接続」が挙げられている。

2 ここでは同一語の繰り返しには、助詞・助動詞は含めていない。以下も同様。

３．データ

本発表では、2011 年 12 月にリリースされた『現代日本語書き言葉均衡コーパス』の DVD 版を使用した。Disk1 の M-XML フォルダに含まれる xml ファイルが対象である。この xml ファイルは可変長サンプルと固定長サンプルを統合したもので、短単位、長単位の形態論情報のタグのほか可変長部分には文章構造のタグを含んでいる[3]。

本発表ではこの xml ファイルにおいて<paragraph>というタグが付与された部分を対象にそこに含まれる短単位の形態論情報をもとに分析を行う。結束性を観察するには文も妥当な単位であるが、BCCWJ に付与された文を表すタグ<sentence>は見出しや図表のキャプションにも付与されており、通常の本文との区別をしなければならないため、今回の調査では確実に本文部分を表している<paragraph>タグを対象とした。<paragraph>タグを含むサンプル数は表 1 のとおりである。

表 1　対象サンプル数

媒体	全サンプル数	P サンプル数
出版書籍	10,117	9,742
雑誌	1,996	1,767
新聞	1,473	1,457
図書館書籍	10,551	10,369
白書	1,500	1,496
教科書	412	0
広報紙	354	354
ベストセラー	1,390	1,374
Yahoo!知恵袋	91,445	0
Yahoo!ブログ	52,680	0
韻文	252	0
法律	346	56
国会会議録	159	159
合計	172,675	26,774

教科書、Yahoo!知恵袋、Yahoo!ブログ、韻文は<paragraph>タグを用いていないため、対象サンプル数はゼロである。なお、<paragraph>タグの問題点については西部ほか(2011:232)を参照されたい。

表 2 は、対象となったサンプルの延べ語数、段落数、1 段落あたりの延べ語数、1 段落あたりの異なり語数のそれぞれの平均値である。1 段落当たりの延べ語数を見てみると国会会議録の値が大きい。これは国会会議録における段落の認定（1 発言が 1 段落）が影響しているものである。なお、語数には補助記号、空白、助詞、助動詞は含まれていない。

表 2　各媒体の延べ語数等の平均値

	サンプルの延べ語数	段落数	1 段落の延べ語数	1 段落の異なり語数
出版書籍	1,384.61	43.76	50.51	37.06
雑誌	891.17	29.81	40.05	33.27
新聞	334.33	9.28	38.78	33.33
図書館書籍	1,450.16	54.53	45.76	34.70
白書	1,793.10	29.32	64.74	44.33

[3] タグの詳細については小木曽ほか(2011)を参照。

広報紙	2,903.53	103.14	28.14	23.39
ベストセラー	1,404.46	69.30	29.52	24.28
法律	219.50	6.93	24.04	15.03
国会会議録	17,885.87	144.06	151.30	76.21

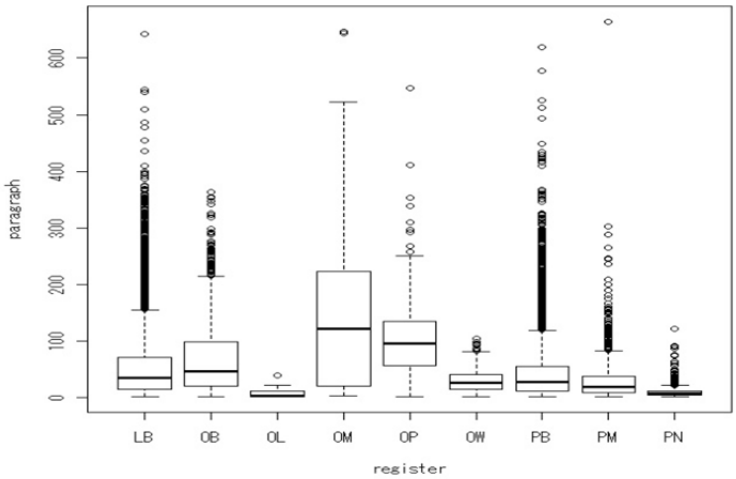

図 1 段落数の分布

図 1 は、サンプルあたりの段落数の分布の様子を媒体ごとに表したものである。全体的に分布が右に（大きい方に）かたよっていることが分かる。また、図書館書籍と出版書籍はほぼ似たような分布を示している。

４．結束性の算出方法

本発表では、ある段落とそれに隣接する段落との間で共通して現れる語の多寡に着目した。語の単純な繰り返しを扱うことのメリットは、他の結束性を表す現象と比べて正確な把握がしやすいこと、また、頻繁に起きる現象であるため、観察がしやすいことである。一方、デメリットとしては観察結果が「語」の単位認定基準に依拠してしまうこと及び同じ語か異なる語かだけの把握にとどまり、意味的な関係が把握できないことである。共通する語だけでなく、類義語等まで含めた計測方法として Hoey(1991)や Károly(2002)があるが、扱っているデータ量はさほど多くない。大量のデータを使って自動的に計測するには語の繰り返しがもっとも適していると思われる。

本発表では、以下の式により結束性の度合いを計り、共起語率と名付けた。

$$\mathrm{C(a,b)} = \frac{F(a,b)}{N_a}$$

a, b：段落番号(1～n)
$\mathrm{C(a,b)}$：段落 a の段落 b に対する共起語率。
$\mathrm{F(a,b)}$：段落 a と段落 b とで共通して現れる語の延べ語数を段落 a 内で数えた数。
N_a：段落 a の延べ語数。

共起語率は、水谷(1980)の非対称類似度を利用した指標である。そのため、連続する 2 つの段落の間の共起語率に 2 つの値が存在する。後続の段落に対する共起語率と前接の段落に対する共起語率である。上述の式では、b=a＋1 のとき、後続段落に対する共起語率とな

り、b=a－1 のとき、前節段落に対する共起語率となる。ただし、文章の冒頭の段落の前接段落及び最後の段落の後続段落は存在しないため、便宜的にその場合の共起語率は 0 とする。

この方法で共起語率を測るにはひとつ制約がある。それは、文章が 2 つ以上の段落から構成されていなければならないことである。そのため、表 1 で対象としたサンプルから 1 段落しかなかったサンプル 340 サンプルを除外した。

なお、計測対象からは言語表現とは見なさない補助記号、空白、及び文章の結束性には影響を及ぼさない助詞、助動詞を除外した。

５．結果

表 3 は、段落あたりの共起語の数と共起語率の平均値である。後続段落との共起語率と前接段落との共起語率とはほぼ等しい値を示している。このことは、どの媒体もそれぞれ同程度の依存関係でつながっていると解釈できる。個々に眺めてみると、法律、白書、国会会議録の共起語率が高く、新聞、ベストセラー、雑誌の共起語率が低いことが分かる。

表 3 共起語の数と共起語率

	後続段落との共起語数	後続段落との共起語率	前接段落との共起語数	前接段落との共起語率
出版書籍	12.98	0.22	12.74	0.22
雑誌	6.89	0.16	6.82	0.16
新聞	5.99	0.15	5.84	0.16
図書館書籍	10.49	0.19	10.36	0.19
白書	20.00	0.31	19.84	0.31
広報紙	5.19	0.18	5.13	0.17
ベストセラー	5.49	0.15	5.47	0.15
法律	12.16	0.48	12.31	0.47
国会会議録	40.45	0.30	39.01	0.30

表 4　NDC 別の共起語の数と共起語率

	後続段落との共起語数	後続段落との共起語率	前節段落との共起語数	前節段落との共起語率
0 総記	12.97	0.22	12.95	0.22
1 哲学	17.55	0.25	17.73	0.24
2 歴史	14.80	0.21	14.60	0.21
3 社会科学	15.02	0.24	14.84	0.24
4 自然科学	14.32	0.24	13.96	0.24
5 技術・工学	10.72	0.22	10.56	0.21
6 産業	11.03	0.21	10.82	0.21
7 芸術・美術	12.02	0.20	11.98	0.20
8 言語	10.40	0.21	10.17	0.20
9 文学	5.07	0.12	4.97	0.12
分類なし	3.46	0.13	3.45	0.13

表 4 は、図書館書籍のデータについて、NDC（日本十進分類法）別の共起語数と共起語率を算出したものである。図書館書籍全体では共起語率は 0.19 であったが、NDC 別に見ると「9 文学」と「分類なし」の値が他と比べて低いことが分かる。「分類なし」についてはデータを見ていないので理由は分からないが、「9 文学」は会話文のような短い段落が多いため、共起語率が低くなったと推測される（表 3 のベストセラーの値の低さもそれに起因しているであろう）。それを確かめるために、1 段落あたりの延べ語数の平均と共起語率の平均との相関を見てみよう。図 2 にその結果を示す。正の相関が認められ、決定係数は 0.799 と高い値を示した。

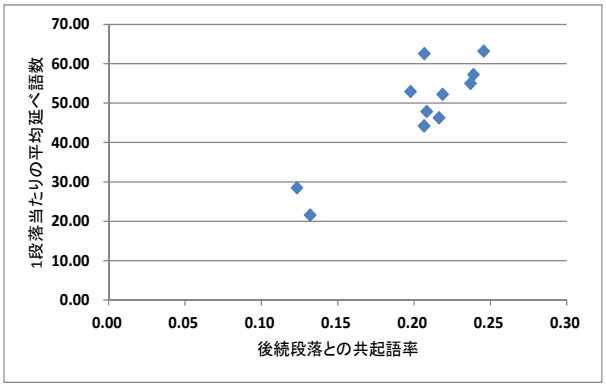

図 2　段落の延べ語数と共起語率との相関

６．文章中の共起語率の推移

共起語率の値はひとつの文章中でどのような変化を示すのだろうか。白書の例を見てみよう。図 3 は OW1X_00000（昭和 54 年版経済白書）というサンプルである。

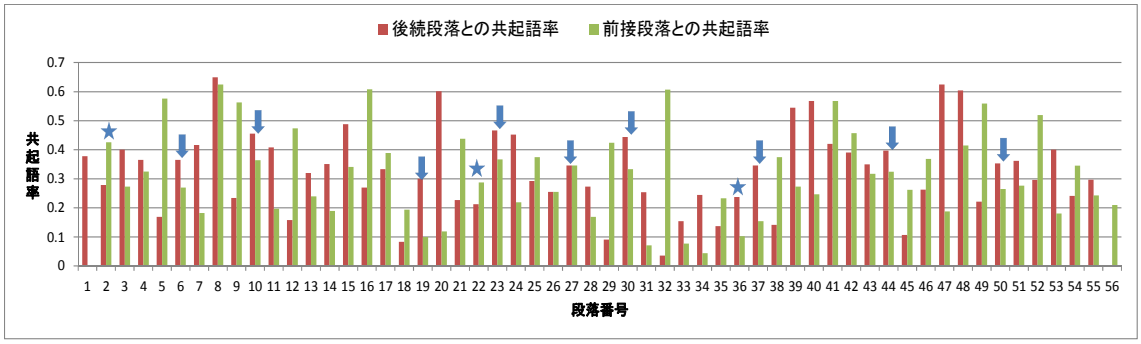

図 3　文章中の共起語率の推移

図 3 で、★を付けた 3 箇所は大きな節が開始する箇所、下向きの矢印を付した 9 箇所はその節の中で小見出しが立っている箇所である。矢印の部分における後続段落との共起語率（左側の棒）と前接段落との共起語率（右側の棒）とを比べてみると、9 箇所のうち 8 箇所が後続段落との共起語率が前接段落との共起語率を上回っている（残りの 1 箇所は同じ値）。このことは、新規の内容になった最初の段落は、新しい話題を展開させるため、その次の段落との結束性が高くなっていると言えるのではないだろうか。

逆に矢印の直前の段落は、あるまとまりの最後の段落を意味する。この部分の後続段落と前接段落の共起語率はどうなっているかというと、9 箇所中 6 箇所で前接段落との共起語率の値のほうが高い。これは一つの例にすぎないが、このような文章中での共起語率の推移を利用して段落のまとまりを自動的に推測することに応用出来る可能性がある。

７．まとめと今後の課題

本発表では非常に単純な指標である共起語率を用いて文章の結束性の度合いを観察した。その結果、法律、白書、国会会議録のように結束性の高い文章と新聞、ベストセラー、雑誌のように結束性の低い文章があることが分かった。NDC 別に観察したデータでは、文学の結束性が低いという結果になった。これは文学に会話文が多く、その会話が 1 段落と認定されているというデータの特徴の現れである。

また、文章中の共起語率の推移をみることにより文章のセグメンテーションへの応用が考えられることを示した。

今後の課題として以下の 3 点を挙げる。これらを通じて文章における結束性について客観的な記述を目指したい。

(1)西部ほか(2011:232)によると、サンプルを構成する文がすべて段落に分割される訳でないと指摘されている。また、<paragraph>の認定は行頭の空白をもとに自動的に認定しているとのことなので段落の実態を確認して分析に問題がないかどうか確認する必要がある。

(2)段落と文の両方を利用した結束性の測定の方法を探る。

(3)指示詞や接続詞など文法的結束性の手段との相関を調べること。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「テキストにおける語彙の分布と文章構造」による研究成果の一部である。データとして利用した BCCCWJ の書籍部分は、文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Halliday, M.A.K. and Hasan, R.(1976) *Cohesion in English*. Longman（邦訳『テクストはどのように構成されるか』、大修館書店、1997 刊）

Hoey,Michael.(1991) *Patterns of Lexis in Text.* Oxford University Press.

Károly,Krisztina.(2002) *Lexical Repetition in Text.* Peter Lang.

庵功雄(2007)『日本語におけるテキストの結束性の研究』、くろしお出版

小木曽智信、間淵洋子、前川喜久雄(2011)「『現代日本語書き言葉均衡コーパス』における形態論情報付き XML フォーマット」、言語処理学会第 17 回年次大会予稿集、pp.352-355.

西部みちる、大島一、間淵洋子、小林正行、田島孝治、高田智和、山口昌也(2011)『『現代日本語書き言葉均衡コーパス』における電子化テキストの構築』、国立国語研究所内部報告書(LR-CCG-10-03)

水谷静夫(1980)「用語類似度による歌謡曲仕分『湯の町エレジー』『上海帰りのリル』及びその周辺」「計量国語学」12(4)、pp.145-161.