共起語集合の頻度分布と語の属性との相関

山崎　誠（国立国語研究所言語資源研究系）†

Correlation between Frequency Distribution of Collocational Set and Key Word's Attribute

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

本稿は、コロケーションを計量語彙論的な観点から記述することを目的とする。コローションの定義はさまざまであるが、本稿では、文脈においてある語と共起する別の語と組み合わせのことと広く捉える。Halliday & Hasan（1976 : 374[1]）では、コロケーションは再叙と並んで語彙的結束性のひとつとされ、「コロケーションによる結束性がテクストに及ぼす効果は，微妙なもので評価しにくい。」（同前 : 379）とされている。本稿では、Hallidayらが取らなかったアプローチ、すなわち、コロケーションという現象を集合としての語彙を量的に観察した場合にどのような特徴が見えてくるかについて考察するものである。

２．共起語集合

本稿で利用する概念「共起語集合」について説明する。計量語彙論では集合としての語彙をもとに、延べ語数や異なり語数、類似度などを利用して分析を進める。本稿ではコロケーションのキーとなる語の前後の一定の距離に現れる語の集まりを考える。例えば、(1)のような語の連続による文脈があった場合、t_i がキーとなる語、t_{i-1} がキーの１語前の語、t_{i+1}がキーの１語後の語などとなる。

(1)　…, t_{i-3}, t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}, t_{i+3},…

対象表現域において、ある単位語 t が、同一の見出し語 m に対応するすべての場合において、t との距離 d の位置にある語の作る集合を $V_{m(d)}$と書くことにする。距離は、相手となる語の相対的位置からキーとなる語の相対的位置を引いた値で表す。したがって、m 自身との距離は 0 であり、前文脈方向がマイナス、後文脈方向がプラスとなる。定義により、$V_{m(0)}$は要素の異なりが見出し語 m のみである集合となる（ただし、m の延べ語数は 1 とは限らない）。このように定義した $V_{m(d)}$ を考えたとき、d の値の変化によって、$V_{m(d)}$の計量的指標がどのように変化するか、また、その変化は見出し語 m の持つ属性とどのように関係するかが本稿の興味の中心となる。

３．データと方法

本稿で利用するデータは『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）である。BCCWJ には 13 のレジスターがあるが、本稿ではそのうち主として図書館書籍（LB）を利用する。分量が多く、結果の安定性が得られるためである。なお、本稿で用いる言語単位は、短単位である。

上で定義した $V_{m(d)}$を求める見出し語の選定は、使用頻度の多い語から品詞を異にするものを適宜選んだ。d の範囲は、文を越えないものとする[2]。したがって、見出し語 m を持つ単位語 t が文末にある場合、後続の文脈がないため、t_{i+1}は存在しない。なお、距離の測定

† yamazaki@ninjal.ac.jp

1 ページは邦訳による。以降の同書からの引用も同じ。

2 文の認定は BCCWJ の DVD に含まれる短単位 TSV ファイルの文頭ラベルを使用した。文頭ラベルが B(=文頭）である語から、次の B が出てくるまでを１文とした。

の対象からは空白と補助記号を除いている。

４．分析１

４．１　概観

図 1～図 8[3]は、適宜選択した見出し語 8 語について、キーとなる語の前後 20 語について延べ語数と異なり語数の推移を示したものである。調査対象は BCCWJ 全体である。図 1 の「思う」の例では、延べ語数はキーのマイナス側は、－1 語まで増え続け、＋1 語以降は下降に転じる。この傾向は図 2 の「見る」、図 3「関係」でも同じである。一方、図 4「人間」、図 5「新しい」、図 6「すごい」、図 7「しかし」、図 8「なお」は、＋1 語まで延べ語数が増え続け、＋2 語以降は下降に転じる。延べ語数の推移は、キーから文頭ないし文末

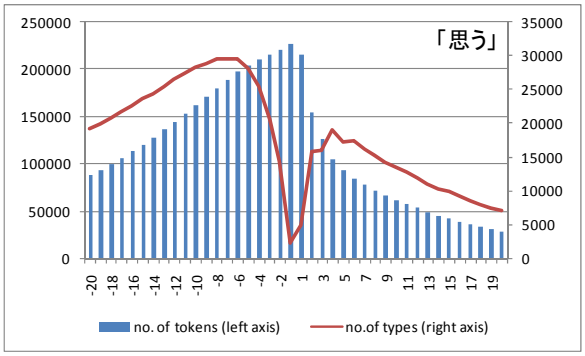

図 1　計量的指標の推移：「思う」

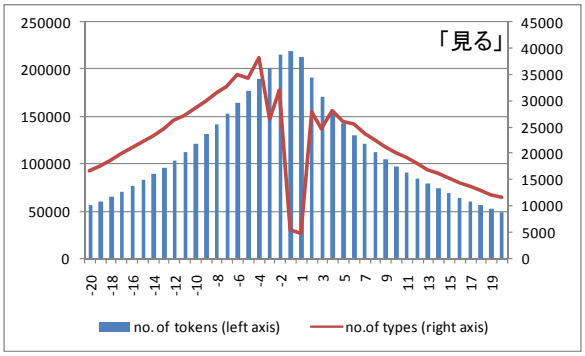

図 2　計量的指標の推移：「見る」

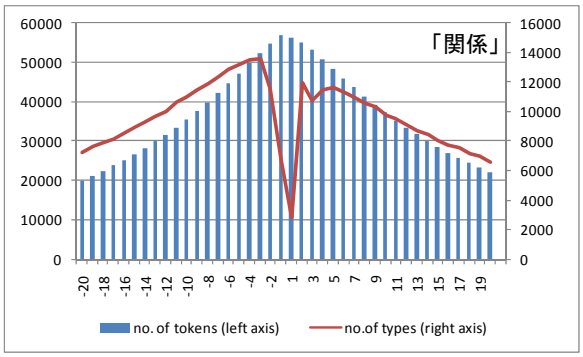

図 3　計量的指標の推移：「関係」

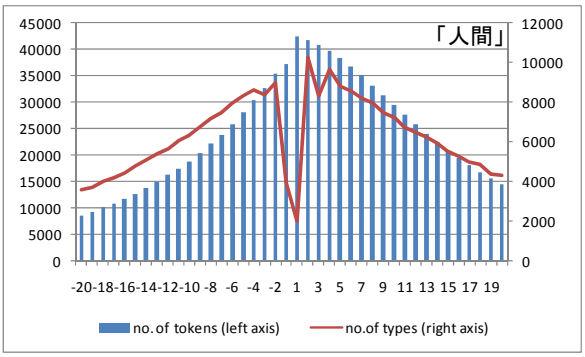

図 4　計量的指標の推移：「人間」

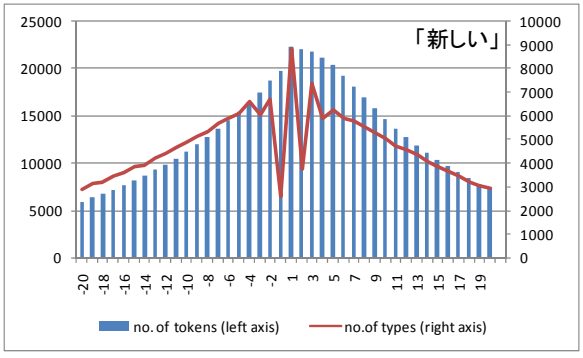

図 5　計量的指標の推移：「新しい」

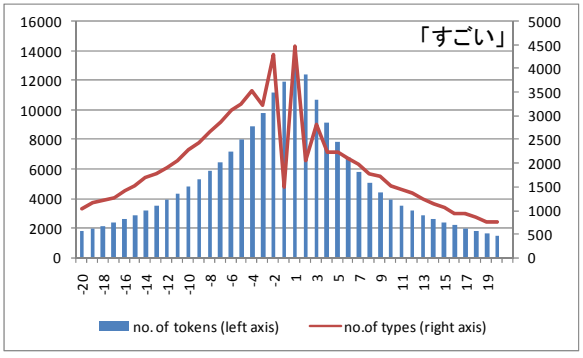

図 6　計量的指標の推移：「すごい」

[3] 図 1～8 のいずれも棒グラフが延べ語数（左軸）、折れ線グラフが異なり語数（右軸）を表す。横軸はキーからの相対的位置である。これは離散的な値をとるため、折れ線グラフにするのは妥当ではないが、見やすさのため、便宜的に使用した。以降のグラフも同様である。

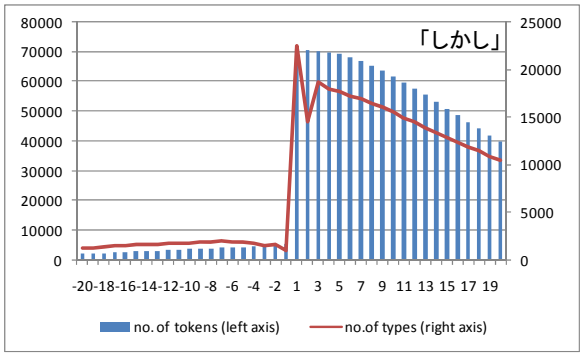

図 7　計量的指標の推移：「しかし」

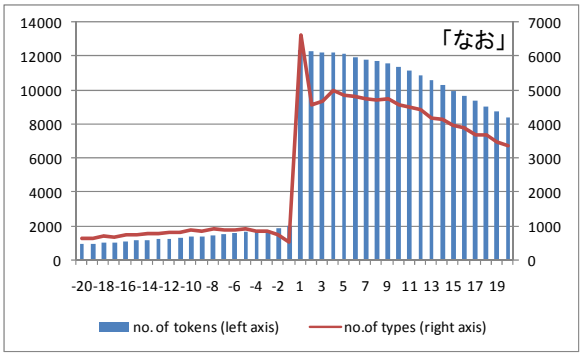

図 8　計量的指標の推移：「なお」

まで何語あるかの分布を意味していることから、その語が平均して文のどの辺に位置しているかを表していることになる。「思う」が「見る」に比べてプラス側の延べ語数の減少が大きいのは文末によく現れることを意味している。また、接続詞である「しかし」「なお」はマイナス側が非常に少ない非対称的な形をしているのもその品詞性の表れである。

図で折れ線で示した異なり語数の推移は、延べ語数の推移とは違って、やや複雑な様相を示している。図 1「思う」では、キーのマイナス側は、－7 語まで上昇し続け、－1 語まで下降し、＋4 語まで再上昇し、＋5 語で若干下降し、＋6 語で上昇、＋7 語以降は下降する。個別に異なる部分はあるものの、「思う」「見る」「関係」「人間」ではキー付近に谷ができる形の分布となっている。図 5「新しい」と図 6「すごい」は－1 語と＋2 語との 2 か所に谷ができる分布であり、図 7「しかし」と図 8「なお」はマイナス側は語数が少なく推移はほぼ一定しているようであるが、プラス側は＋1 語で急に上昇し、いったん小さな谷を作り、下降するという分布になっている。

大局的に見ると、異なり語数の推移は延べ語数の増減に伴う自然増・自然減となっていると見られる部分と、その傾向に反し、延べ語数が増えても減少する、あるいは、延べ語数が減っても増加する部分とに分けられる。前者は語彙の量的な特徴として一般的な現象と考えられるが、後者は当該のキーとなる語の持つ、コロケーションとしての特徴が現れているものと解釈できる。すなわち、キーとなる語の影響によって特定の語の出現が多くなったため、延べ語数の値と異なり語数の値の関係にも影響したものであろう。このようなコロケーションの影響を受けていると思われる部分をマイナス側からの自然増の傾向が破られる箇所（すなわち減少に転じた箇所）、同様にプラス側を値の大きい方から見た場合の自然増の傾向が破られる箇所で特定すると、「思う」が－6 語から＋5 語の範囲、「見る」が－3 語から＋3 語、「関係」が－2 語から＋4 語、「人間」が－3 語から＋3 語、「新しい」「すごい」が－3 語から＋4 語、「しかし」が＋2 語、「なお」が＋3 語となっている[4]。このプラス側の転移箇所、およびマイナス側の転移箇所にはさまれた部分をコロケーションの影響を受けている範囲と考えることができる。

図 9、図 10 は、「思う」について、レジスターごとに延べ語数と異なり語数の推移を見たものである。図 9 の延べ語数では、13 のレジスターのうち 12 個が－1 語目が最大になり、以降下降する傾向を取っている[5]。ちなみに－1 語目における延べ語数がいちばん多いのは、Yahoo!知恵袋(OC)で、以下図書館書籍(LB)、出版書籍(PB)、Yahoo!ブログ(OY)、国会会議録(OM)と続く。国会会議録のマイナス側のカーブは他のレジスターと比べてゆるやかであるが、これは一文が長いということの現れであろう。表 1 は、図 10 の異なり語数の推移について、法律(OL)を除くレジスターごとにキーとなる語に向かってプラス・マイナスそれ

4　接続詞についてはマイナス側は語数が少ないため、評価は行わない。

5　残りの一つのレジスターは法律(OL)で、「思う」が 5 回しか現れないため、傾向を把握することは難しい。

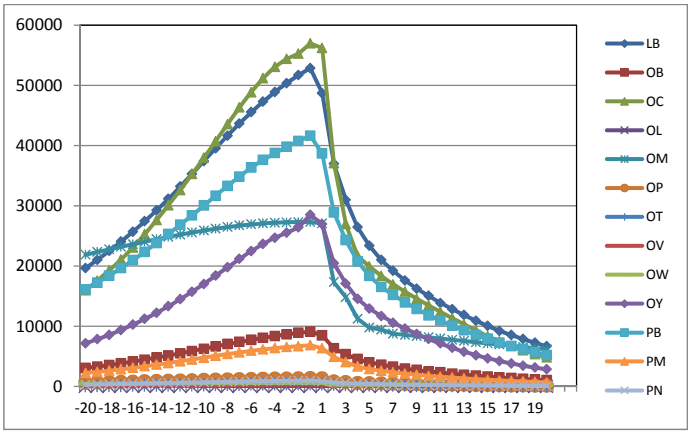

図 9　レジスター別延べ語数の推移「思う」

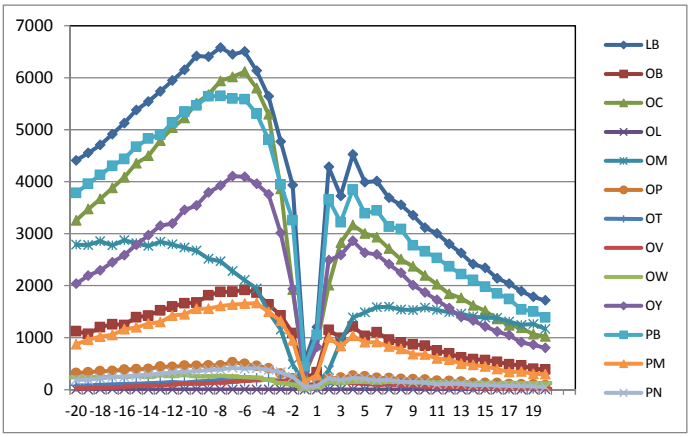

図 10　レジスター別異なり語数の推移「思う」

表 1　レジスターごとの転移箇所

レジスター	マイナス側の転移箇所	プラス側の転移箇所
図書館書籍(LB)	−7	+5
ベストセラー(OB)	−5	+3
Yahoo!知恵袋(OC)	−5	+3
法律(OL)	＊[6]	＊
国会会議録(OM)	−19	+18
広報紙(OP)	−10	+3
教科書(OT)	−11	+17
韻文(OV)	−10	+17
白書(OW)	−18	+16
Yahoo!ブログ(OY)	−6	+3
出版書籍(PB)	−6	+5
出版雑誌(PM)	−4	+17
出版新聞(PN)	−6	+14

6　法律(OL)は用例数が少ないため、転移箇所を判断できない。

ぞれの方向から自然増の傾向が破られる転移箇所を示したものである。BCCWJ 全体では前述のようにこの範囲は－6 語から＋5 語であったが、レジスターで見ると、ベストセラー(OB)、Yahoo!知恵袋(OC)はプラス方向にもマイナス方向にも範囲が狭くなっている。また、Yahoo!ブログ(OY)と広報紙(OP)はプラス方向のみ、出版雑誌(PN)はマイナス方向のみ範囲が狭くなっている。範囲が広くなった主な理由は異なり語数が少なく、値が安定していないためであろう。例えば、白書(OW)のマイナス側の数値の推移は、次のようになっている。「234、242、237、244、244、262、243、275、258、282、256、253、262、255、238、233、186、124、101、12」下線を施した 5 箇所が上昇から下降に転じた点である。

４．２　ＴＴＲによる観察

図 11～14 は $V_{m(d)}$の異なり語数をその延べ語数で割った値、Type/Token Ratio（以下、TTR とする）の推移を示したものである。TTR は語彙の豊かさを表す指標とされ、語彙の計量的な分析や文章の評価によく用いられている。TTR の値が高いほど集合における見出し語の種類が多く、語彙的に豊かであるとされる。本分析でのデータは、キーとなる語から等距離にある語を集めた集合であるため、文脈を有していない見出し語の集合という特徴がある。したがって、そのような集合における TTR の値が意味するものは、データ中に同一文脈がどれだけ複数回使用されているかということの観察になるだろう。

図 11～14 により、TTR の動きはキー付近に谷を形成することから、図 1～8 の異なり語数の推移にやや似ているが違う点もある。図 11 の動詞ではマイナス方向プラス方向ともに相対位置の絶対値が大きくなると TTR の値も高くなる傾向がある。これは図 12 の名詞、図 13 の形容詞でも同じである[7]。図 14 の接続詞ではマイナス方向には TTR が高くなる傾向があるが、プラス方向ではそれがなく、フラットになっているのが特徴的である。図 7、8 からプラス方向で延べ語数の減少が見られることから、延べ語数が一定のためこのようにフラットになったわけではない。キーから離れるにしたがって、TTR の値が大きくなって

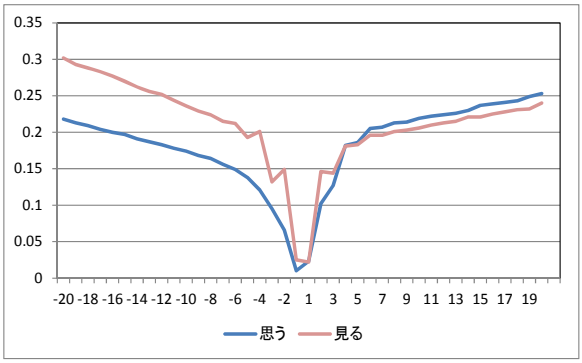

図 11　ＴＴＲの推移：動詞

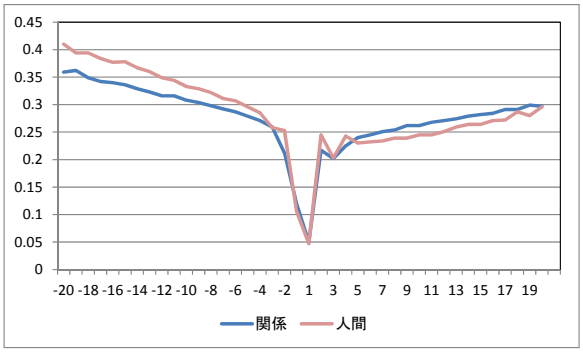

図 12　ＴＴＲの推移：名詞

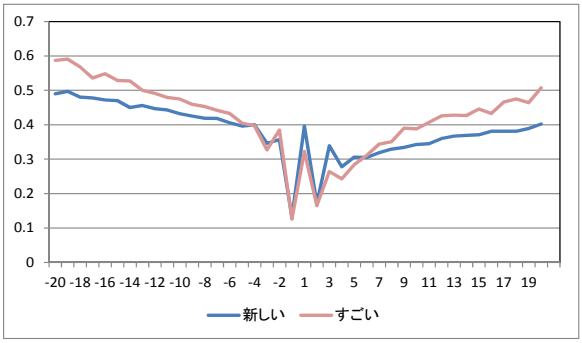

図 13　ＴＴＲの推移：形容詞

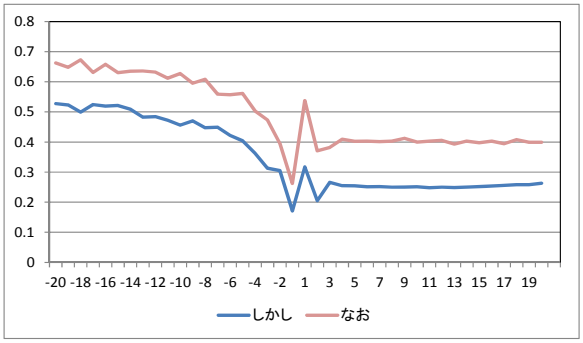

図 14　ＴＴＲの推移：接続詞

[7] 図 13 は谷が二つある型であるが、その外側は絶対値の大きい方向に対して単調増加の傾向が見て取れる。

いくということは、コロケーションの影響がどの辺まで届いているかの判断にも関係する。

TTR の値が一定になるまでコロケーションの影響があるとすると、少なくとも図 11～13 に挙げた語群についてはキーから前後 20 語までコロケーションの範囲ということになる。これは４．１節で述べた異なり語数の推移から見たコロケーションの範囲（ほぼ一ケタ台前半の値）とはずいぶん違っている。どちらがコロケーションの範囲として妥当かは本稿では決めがたいが、その検証方法のひとつとして、延べ語数を一定にしておいて TTR を測ることを次の課題としたい。

５．分析２

５．１　動詞の場合

この節では、対象を図書館書籍(LB)に、$V_{m(d)}$の距離の範囲を±５にした場合について考察をする。図 15、16 は動詞を対象にして TTR の値を観察したものである。図 15 は UniDic の品詞体系で「動詞－一般」を、図 16 は「動詞－非自立可能」を品詞に持つものである。いずれも似たような動きを示しているが、特徴的なのは、－１語と＋１語の TTR の値が低くなっており、その部分を谷として両側に開いた形を作ることである。また、動詞一般の「考える」「出る」「使う」「聞く」「書く」には－３語目に小さな谷が出来ている。図 16 の非自立可能の方でも「見る」「掛ける」「終わる」の－３語目に小さな谷ある。「始める」「続ける」「切る」は－３語目に谷はないが、－４語目、－５語目まで観察すると値が減少している箇所が認められる。また、－３語目ほど顕著ではないが、＋３語目にも TTR の値が鈍化する部分がある。図 15 では「使う」「聞く」、図 16 では「見る」「掛ける」「切る」である。TTR の値が単調に推移しない理由は、キーから±３語目によく出現する語があることを想定させる。この場合、キーの前後３語目までがコロケーションとして注目すべき範囲であると推測される。

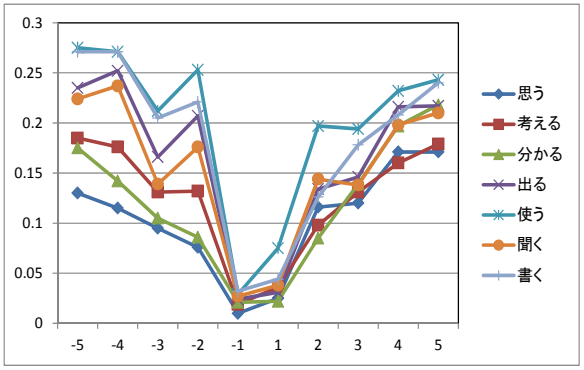

図 15　TTR の推移：動詞・一般

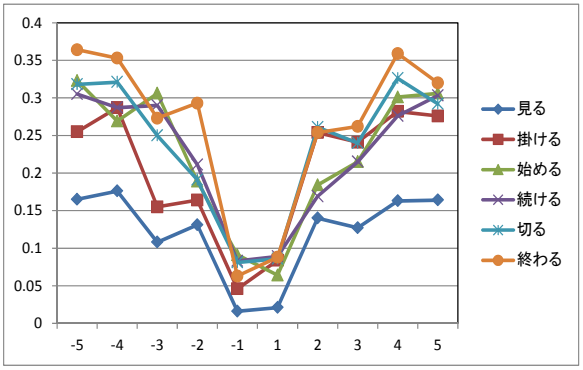

図 16　TTR の推移：動詞・非自立可能

図 15、16 からはキーを挟んで TTR の値が対称的になっているのではなく、全体的にキーの前の方が値が高いように見受けられる。特に図 15 のプラス側は折れ線が混み合っているのに対し、マイナス側はばらけている印象がある。このことを確かめるために、プラス側の TTR の値からマイナス側の TTR の値を引いた値を表 2、3 に示した。この値が 0 より小さければマイナス側の TTR の方が大きいということになる。表 2、3 の網掛けの部分はその差が 0 より小さい部分である。表 2 では 35 箇所中、20 箇所のセルが 0 より小さく、TTR の値に関しては対象でなく、マイナス側のほうが値が高いことが分かる。このことは、キーとなる語の前５語以内に現れる語彙のバラエティーの方が、後の５語以内に現れる語彙のバラエティーよりも多いことを意味する。ただし、表 3 ではその傾向は確認されず、網掛けの箇所は 30 箇所中 16 箇所にとどまるが、キーから１語目の部分を除くと若干傾向が高まる（24 箇所中 15 箇所）。

表 2　キーから等距離はなれた語集合の TTR の差：動詞・一般

キーからの距離	1 語	2 語	3 語	4 語	5 語
思う	0.015	0.04	0.025	0.056	0.041
考える	0.017	-0.034	0	-0.016	-0.006
分かる	0.001	-0.001	0.034	0.055	0.043
出る	0.007	-0.073	-0.02	-0.036	-0.018
使う	0.046	-0.056	-0.018	-0.039	-0.032
聞く	0.011	-0.032	-0.001	-0.039	-0.014
書く	0.012	-0.095	-0.027	-0.063	-0.031

表 3　キーから等距離はなれた語集合の TTR の差：動詞・非自立可能

キーからの距離	1 語	2 語	3 語	4 語	5 語
見る	0.005	0.009	0.019	-0.013	-0.001
掛ける	0.038	0.09	0.086	-0.005	0.021
始める	-0.027	-0.005	-0.091	0.032	-0.017
続ける	0.006	-0.042	-0.075	-0.011	-0.002
切る	0.005	0.07	-0.01	0.005	-0.026
終わる	0.025	-0.039	-0.011	0.006	-0.044

５．２　名詞の場合

名詞における TTR の分布を見てみよう。図 17 は普通名詞、図 18 は固有名詞及び普通名詞だが助数詞としても使うもの（時間、パーセント）を選んだ。名詞は動詞と違い、TTR の谷に相当する部分が 1 例を除いては 1 箇所（＋1 語目）である。この違いは、それぞれ動詞、名詞の前後にくる助詞助動詞の影響ではないかと思われる。前節で述べたように、－3 語目、＋3 語目に TTR　の値が鈍化する箇所があることも同様である。

次に、意味的な違いは TTR の推移にどのように関係しているかを見てみよう。図 17 から「女」と「男」の TTR の値がほぼ重なるくらいによく一致していることが分かる。意味的な類似性のためとも解釈できるが、対比する意味で挙げた「人間」と「子供」もその分布はかなり似ているため、必ずしも意味的な類似が理由とは言い切れないようである。図 18 の「日本」と「アメリカ」は値の大きさは異なるが値の推移の様子は似ている。助数詞にもなる「時間」「パーセント」は谷の位置がずれており、推移が似ているとは言い難い。ちなみに動詞の場合と同じように、キーから等距離にある TTR の値をプラス側からマイナス側を引いた値は、図 17 の 4 語では、20 箇所中 18 箇所が、図 18 では 20 箇所中 13 箇所がマイナスの値をそれぞれ示した。名詞においてもキーの前の語彙の種類の方が、後ろに

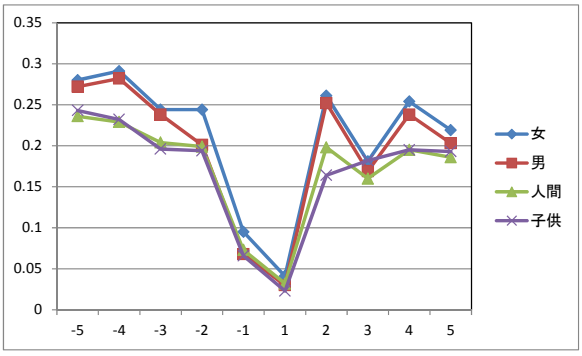

図 17　TTR の推移：普通名詞

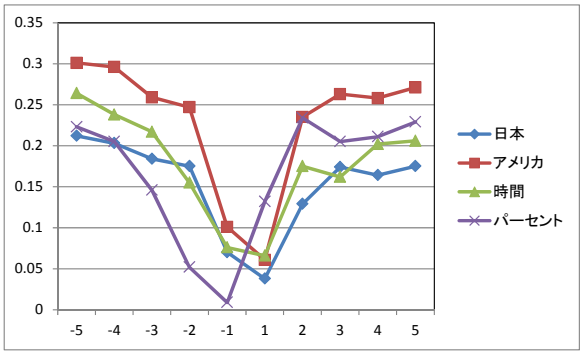

図 18　TTR の推移：固有名詞、助数詞可能

くる語彙の種類よりも多い傾向があることが確認された。

５．３　形容詞の場合

形容詞は図 19 の活用の場合と図 20 のシク活用の場合とに分けた。１例を除いて－１語目に谷を作る分布を示している。動詞、名詞の場合とはやや異なり、TTR の値が鈍化する箇所がマイナス側は－３語目であるが、プラス側が＋２語目と＋４語目の２箇所あるのが特徴的である。図 19 では意味的に関連の深い「良い」「悪い」と「大きい」「小さい」の値の推移がそれぞれ類似していることが分かる。図 20 では「嬉しい」の谷の位置が＋１語目にずれているが、今のところこれを説明する解釈は持ち合わせていない。プラス側とマイナス側の値の差は、図 19 で 20 箇所中 12 箇所が、図 20 で 25 箇所中 16 箇所が 0 より小さく、名詞、動詞と類似の傾向を示すことが確認された。

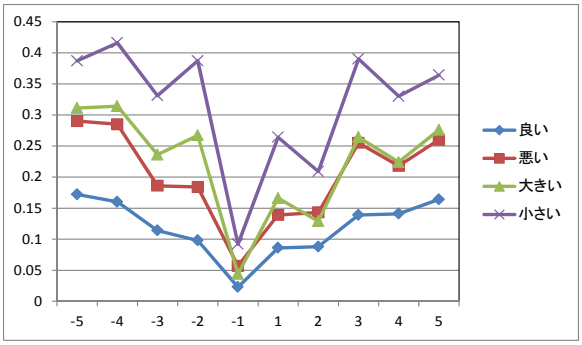

図 19　TTR の推移：形容詞ク活用

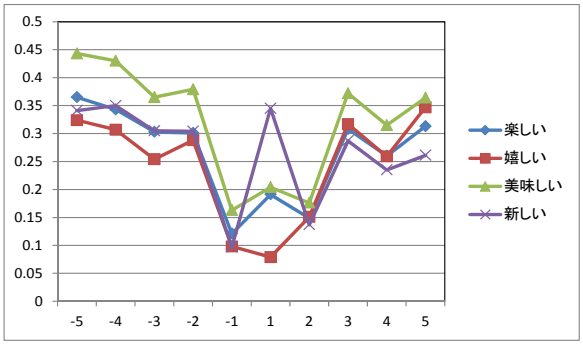

図 20　TTR の推移：形容詞（シク活用）

６．まとめと今後の課題

本稿では、共起語集合（キーとなる語の前あるいは後ろの特定の位置に出現する語の集合）という考えを用いて、BCCWJ においてコロケーションが現れる様子を計量的な指標の観察から記述した。用いた指標は、延べ語数、異なり語数、TTR である。得られた知見をまとめると次の 3 点になる。(1)異なり語数の推移からは異なり語数が自然増ではなくなる範囲をコロケーションとして位置付け、「思う」「見る」などの語について個別の記述を行った。(2)TTR の推移については±20 語目でも値が一定しないことから TTR によりコロケーションの範囲を定めるのは、別の工夫が必要であることが示唆された。(3)図書館書籍（LB）に限ってキーの前後 5 語の TTR の動きを観察した場合、動詞、名詞、形容詞それぞれ特徴的な推移があること、また、キーからマイナス側の方がプラス側よりも値が高い傾向にあることが分かった。

今後の課題としては、調査語の範囲を広げること、共起語集合同士の類似度を用いた分析、特に、類似度を用いた語の分類を試みたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCCWJ は、国立国語研究所のプロジェクト及び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）