# 第9章 コーパス分析の研究例2

## 複合動詞の計量的解析

### 9.1 共起頻度の分析のための指標

クロード・シャノン(Claude E. Shannon)は、『通信の数学的理論』(1948)という論文において、「エントロピー」と「冗長度」という2つの概念を発表した。これらは、情報量の尺度である。エントロピーは、あいまいさや乱雑さの増減を示す指標である。言語研究においては、表現の種類とその使用頻度に基づいて1つの値を算出し、不規則性を示すことができる。また、冗長度は、エントロピーとエントロピー最大値(最も不規則な状態)を利用して得られる無駄の程度(規則性とも考えられる)を示す指標である。エントロピーと冗長度の尺度を組み合わせることで、ある表現の多様性と規則性とを簡単な数値で表すことができるのである。

近年、大規模コーパスが整備され、「共起頻度」が容易に算出できるようになってきた。これに伴い、複数の基準で選択された表現の「異なり頻度」や「延べ頻度」を解析する手法が求められるようになってきている。このような、それ自体はノンパラメトリック・データである共起頻度に対して、通信の数学理論で扱われるエントロピーと冗長度を適用させパラメトリック・データの指標に変換することによって、様々な多変量解析を行うことができるようになる。また、これら2つの指標をもとに各表現を二次元上でプロッティングし、個々の共起表現パターンを記述的に考察することも可能である。さらに、新聞、小説、雑誌など、異なるタイプとサイズのコーパスから得られた共起頻度のパターンについて、エントロピーと冗長度の指標に基づい

て、各コーパスの特徴を比較検討することもできる。本章では、エントロピーと冗長度の算出方法とそれらの指標を用いた多変量解析の例を紹介する。

## 9.2　語彙的・統語的複合動詞の統語構造

計量言語学の分野には、*Journal of Quantitative Linguistics* という英語の学術誌がある。この学術誌に、Tamaoka, Lim, & Sakai(2004)の「日本語の語彙的・統語的複合動詞のエントロピーと冗長度」というタイトルの論文が掲載されている。この論文は、エントロピーと冗長度の2つの指標に基づいて、語彙的／統語的複合動詞の特徴を新聞と小説のコーパスで比較している。以下、この研究を例に、コーパスから得られる共起頻度の分析法を紹介する。

日本語では、動詞を2つ組み合わせることで複合動詞を作ることができる。影山(1993、1999)は、日本語の複合動詞を2種類に分けている。1つは、語彙的複合動詞で、初めにくる動詞($V_1$)と次にくる動詞($V_2$)の組み合わせに「語彙的な習慣化が見られる」(影山 1999: 189)としている。語彙的複合動詞には、2つの動詞の間に「語の形態的親密性」を示すサエやモを挿入することができず、1つの語となっていると指摘している。もう1つは統語的複合動詞である。これについて影山(1999: 189–190)は、その名の示す通り「統語的な構造」に由来し、補文構造という形で捉えられるとしている。たとえば、「彼は昼食を食べ始めた。」と「子供は手紙を投函し忘れた。」における補文構造は、「彼は［昼食を食べ］始めた。」と「子供は［手紙を投函し］忘れた。」という「$V_1$ することを(が)$V_2$」の部分を指す。この構造を基にして、「食べ」を「始める」に、「投函し」を「忘れる」に結合させると、表面的には、「食べ始める」と「投函し終わる」という複合動詞になると説明している。つまり、統語的複合動詞では、2つの動詞は基本的に別々の語であるという考え方である。

両者を分かりやすく区別するために、統語構造から違いを説明する。語彙的複合動詞は、$[_{vp}$ NP-$o$ $[_{v}$ $V_1$ $V_2]]$ という構造になり、先にくる動詞($V_1$)である「飲む」が、次にくる動詞($V_2$)の「比べる」と結合して「飲み比べる」



という1つの動詞($V_1 + V_2 = V$)を作ると考える。この複合動詞に対して、「コーヒーを」という対格の名詞句(NP, noun phrase)が結びついて動詞句(VP, verb phrase)を作るという構造を持つ。それに対して統語的複合動詞は、$[_{vp2}$ $[_{vp1}$ NP-$o$ $V_1]$ $V_2]$ という構造になる。「コーヒーを」という名詞句(NP)が直接に「飲む」の動詞($V_1$)と結合して動詞句($VP_1$)を作る。そして、この動詞句が後にくる「終える」の動詞($V_2$)に結びつき、さらなる動詞句($VP_2$)を作るという構造と考える。このように、両者の複合動詞は、統語的に異なる構造を持つと考えることができる。語彙的複合動詞の場合は、「コーヒー」というものを複合動詞が受けて「飲み比べる」ものと解釈され、「飲む」と「比べる」の2つの動詞を切り離すことができない。これに対して、統語的複合動詞の場合は、「コーヒーを飲む」という動詞句で示す行為を「終える」という動詞で受けることになる。つまり、コーヒーを飲むことを「終える」のであり、統語的複合動詞においては、2つの動詞が別々に機能している。

## 9.3　語彙的・統語的複合動詞およびコーパスに関連した2つの仮説

Tamaoka, et al. (2004)は、語彙的／統語的複合動詞に関する2つの仮説について、新聞と小説のコーパスから得られる共起頻度を用いて実証した。まず、複合動詞の種類に関する仮説である。語彙的複合動詞は、慣用句のように強く結合していると考えられる。したがって、これら2つの動詞の組み合わせは限定的であり、それほど多様な動詞が組み合わされることはないであろうと予想される。一方、統語的複合動詞は、多様な第1動詞($V_1$)を受けることができると考えられる。これらの構造の異なる2種類の複合動詞について、通信の数学理論におけるエントロピーと冗長度の指標の違いから予測してみる。まず、語彙的複合動詞は、ある特定の2つの動詞の共起パターンは偏っていると考えられるので、エントロピーは低く、冗長度が高いであろう。それに対して、統語的複合動詞は、多様な動詞が結合して共起すると考

えられるので、エントロピーが高く、冗長度は低いと予想される。これが第1の仮説である。

次に、コーパスの種類に関する仮説である。新聞は、複数の新聞記者が一般大衆に情報を伝達するために、簡潔で分かりやすい一定の表現スタイルが採られる。一方、小説は、特定の作家の個性に応じた多様な表現が駆使されている。両者は、書き手の意図と目的において大きく異なっている。仮に、語彙的／統語的複合動詞が一般的な統語構造を持っているといえるなら、語彙的／統語的複合動詞は、新聞か小説かに関係なく類似した共起パターンを示すことが予想される。これが第2の仮説である。Tamaoka, et al. (2004)では、新聞と小説の2種類のコーパスを比較しているが、新聞のコーパスとして、1991年から1994年までの4年間に刊行された毎日新聞を利用しており、その延べ総語数は88,454,573語である。一方、小説のコーパスとしては、青空文庫コーパスに収録されている文学作品のうち、現代語で書かれているものを選んでコーパス化したものを利用しており、延べ総語数は8,370,720語である。毎日新聞は、10年間以上の新聞記事を使用することもできたが、小説のコーパスとのサイズのバランスをとって両コーパスを比較しやすくするために、4年間分に限定し、小説のコーパスの約10倍の大きさにした。

## 9.4　エントロピーと冗長度の公式と計算法

シャノンによると、エントロピーは以下の式で定義される。

$$H = -\sum_{j=1}^{j} p_j \log_2 p_j$$

複合動詞については、まず第2番目にくる動詞($V_2$)を基準にしてエントロピーを計算する。たとえば、「歩く($V_2$)」を含んだ統語的複合動詞を構成する第1動詞($V_1$)を考えてみると、毎日新聞の1991年から1994年までの4年間の記事においては、18種類の第1動詞($V_1$)と結合する。これが異なり頻度



である。これら18種類の動詞と「歩く」とが結合する総頻度は44回であるが、これが延べ頻度である。最も多いのが「売り歩く」で7回、次に「尋ね歩く」で6回であった。上記の公式の$p_j$に相当するのが、「売り歩く」など個別の複合動詞の頻度が「歩く」($V_2$)を基準につくられる全複合動詞の総頻度に占める割合である。具体的には、「売り歩く」の頻度の7を、総頻度の44で割った0.159が$p_j$の値となる。$\log_2 p_j$は、「売り歩く」の場合は$\log_2 0.159$であり、−2.652の値が得られる。次に、$p_j \log_2 p_j$として、$0.159 \times (-2.652) = -0.422$と計算される。「歩く」($V_2$)と結合してつくられる複合動詞は18種類あるので、同様の計算を個々の複合動詞について18回行い(Σのjの部分)、それらをすべて積算して、−1を掛けると、「歩く」(V2)を含む複合動詞について3.780というエントロピー値が算出される。

エントロピーとともにシャノンが提示した有名な概念は、冗長度である。冗長度とは、無駄の程度を表す指標である。ただし、Tamaoka et al. (2004)の研究の場合は、これを、2つの動詞から成る複合動詞の組み合わせと共起頻度の偏りに相当するものであり、同じような2つの動詞の組み合わせが繰り返し使われる度合いを示していると捉えている。シャノンによると、冗長度は、以下の公式で得られる。

$$R = (1 - H/H_{max}) \times 100(\%)$$

Hはエントロピーであり、*Hmax*はエントロピー最大値を意味する。エントロピー最大とは、すべてが等しい確率で生起する場合である。つまり、いずれが起こっても不思議ではない混沌としたまったくの無秩序の状態を意味する。ある第2動詞を基準とした複合動詞の種類、すなわち異なり頻度を$J$とすると、以下の式でエントロピーの最大値が得られる。

$$H_{max} = \log_2 J$$

例えば、「歩く」($V_2$)が作る複合動詞は、総頻度が44回で18種類の複合動詞を作るので、異なり頻度の$J$は18である。エントロピーの最大値は、$H = \log_2 18$で4.170となる。この数値は、18種類の複合動詞がすべて等しく生起する場合のエントロピーである。つまり、どの「歩く」($V_2$)から作られる複合動詞も同数回だけ出現するので、もっとも規則性の無い状態であるという

わけである。さて、冗長度は、得られたエントロピーをそのエントロピー最大値で割り、その数値を1から引いて100倍して、パーセントで示したものである。「歩く」($V_2$)についていえば、エントロピー値が3.780であるので、これをエントロピー最大値の4.170で割って、1から引いた値である0.09348に100を掛ける(すなわち、%で示す)と、9.348%となる。これが冗長度である。

## 9.5 Microsoft Excelによるエントロピーと冗長度の計算

これらのエントロピーと冗長度の計算は、Microsoft Excelを用いて簡単に行うことができる。表1は、毎日新聞の4年間のコーパスにおいて、第1動詞($V_1$)が第2動詞「歩く」($V_2$)と結びついてつくられる複合動詞の共起頻度の一覧である。表1をExcelのワークシートであると考える。1行目はタイトル行として、#(複合動詞の種類の連続番号)、複合動詞、$V_1$(第1動詞)、共起頻度と続く。2行目から各複合動詞の情報が記入される。番号のすぐ右に位置する複合動詞の「売り歩く」(Excelとしては B2 のセル)は、$V_1$の「売り」(C2のセル)が$V_2$の「歩く」と7回(D2のセル)共起していることを示す。次に、D20のセルに共起頻度の合計を入力する。そして、積算を示す「SUM」を入力して、その範囲をD2からD19と指定する(「=SUM(D2:D19)」)。これで、D20のセルに、総共起頻度の44回が算出される。

次に、$p_j$の計算式をE行に入力する。E2は、「売り歩く」の頻度が、「歩く」が作る複合動詞の総共起頻度に占める割合であり、「売り歩く」の共起頻度7回(D2)を44回(D20)で割った数値が入ることになる。そこで、E2のセルに、式の入力を意味する「=」に続けて、D2のセルを総共起頻度のセルであるD20で割るという式を入力する(「=D2/$D$20」)。D20のセルは、この数式を3行目以下にコピーしてもセル・アドレスが変わらないようにするために、$を列番号と行番号の前に挿入する。もちろん、総共起頻度が44であることが分かっているので、「=D2/44」と入力しても同じである。これで、



**表1　毎日新聞の4年間のコーパスで「歩く」($V_2$)が作る複合動詞**

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | # | 複合動詞 | $V_1$ | 共起頻度 | $p_j$ | $\log_2 p_j$ | $p_j \log_2 p_j$ | | | | |
| 2 | 1 | 売り歩く | 売り | 7 | 0.159 | −2.652 | −0.422 | | | | |
| 3 | 2 | 訪ね歩く | 訪ね | 6 | 0.136 | −2.874 | −0.392 | | | | |
| 4 | 3 | 捜し歩く | 捜し、探し | 5 | 0.114 | −3.138 | −0.357 | | | | |
| 5 | 4 | 連れ歩く | 連れ | 5 | 0.114 | −3.138 | −0.357 | | | | |
| 6 | 5 | さまよい歩く | さまよい | 3 | 0.068 | −3.874 | −0.264 | | | | |
| 7 | 6 | 食べ歩く | 食べ | 3 | 0.068 | −3.874 | −0.264 | | | | |
| 8 | 7 | 泳ぎ歩く | 泳ぎ | 2 | 0.045 | −4.459 | −0.203 | | | | |
| 9 | 8 | 泊まり歩く | 泊まり | 2 | 0.045 | −4.459 | −0.203 | | | | |
| 10 | 9 | めぐり歩く | めぐり | 2 | 0.045 | −4.459 | −0.203 | | | | |
| 11 | 10 | 訴え歩く | 訴え | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 12 | 11 | 踊り歩く | 踊り | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 13 | 12 | 楽しみ歩く | 楽しむ | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 14 | 13 | 伝え歩く | 伝え | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 15 | 14 | ねり歩く | ねり | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 16 | 15 | 触れ歩く | 触れ | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 17 | 16 | 見歩く | 見 | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 18 | 17 | 呼び歩く | 呼び | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 19 | 18 | わたり歩く | わたり | 1 | 0.023 | −5.459 | −0.124 | | | | |
| 20 | | | | 44 | | Σ | −3.780 | | | | |
| 21 | | | | エントロピー（H） | | | 3.780 | | | | |
| 22 | | | | 生産性（J） | | | 18 | | | | |
| 23 | | | | 累積生産性 | | | 44 | | | | |
| 24 | | | | エントロピー最大値（Hmax） | | | 4.170 | | | | |
| 25 | | | | 冗長度（%） | | | 9.348 | | | | |

=E2*F2
=LOG(E2,2)
=D2/$D$20
=SUM(G2:G19)
=−G20
=SUM(D2:D19)
=LOG(G22,2)
=(1−G21/G24)*100

E2に0.159の値が得られる。

さらに、$p_j$を底が2の対数($\log_2 p_j$)に換算する。Excelに準備されている「log(数値, 底)」という対数の計算式(関数)を利用する。F5のセルで、E2のセルの値を底2の対数に変換するよう「=log(E2, 2)」を入力すると、−2.652の値が得られる。G2のセルに、E2($p_j$)とF2($\log_2 p_j$)を掛けた値を出すため、「=E2*F2」と入力する。これで、−0.422の値が得られる。18種類の複合動詞があるので、ここまでの「売り歩く」に関するE2からG2までの3列の式を19行目までコピーする。表1のように、18種類の複合動詞に関する計算が一度にできあがる。

エントロピーと冗長度を算出するために、まずG20のセルで$\Sigma p_j \log_2 p_j$の計算を行う。すでに個々の複合動詞の$p_j \log_2 p_j$がG2からG19のセルに得られているので、これらの値を積算すればよい。「=SUM(G2:G19)」を入力すると、−3.780が得られる。さらに、G20で得られた値を正の値にするため

に、G21のセルでこれに－1を掛ける。「=-G20」と入力して得られた3.780の値が、「歩く」($V_2$)が作る複合動詞の共起パターンを示すエントロピーである。

ここからは、冗長度を算出する手続きである。まず、「歩く」($V_2$)が作る複合動詞が18種類あるので、「歩く」($V_2$)の生産性(J)は18、その累積生産性は総共起頻度に相当する44であることを確認する。G22のセルに生産性の18、G23に累積生産性の44を入力しておくと間違いが少なくなる。次に、G24のセルに、エントロピーの最大値(Hmax)を算出するための式を入力する。エントロピーが最大とは、18種類の複合動詞が同じ頻度だけ起こる場合のことであり、生産性(J)の18(G22)の底が2の対数であるので、「=LOG(G22,2)」を入力する。これで、「歩く」($V_2$)のエントロピー最大値4.170が得られる。最後に、これらを用いて冗長度(R)を計算する。エントロピーをエントロピー最大値で割った値を1から引いて、その値を百分率で示したものが冗長度である。エントロピーはG21のセル、エントロピー最大値はG24のセルに計算されているので、これらを利用した計算式「=(1-G21/G24)*100」をG25のセルに入力する。もちろん、「=(1-G21/G24)」と入力し、「%」のアイコンをクリックして百分率を表示する方法もある。これで、「歩く」($V_2$)の冗長度9.348%が得られる。このExcelのフォーマットを用いて、「歩く」($V_2$)以外の第2動詞が作る複合動詞のエントロピーと冗長度を計算するとよい。

## 9.6　仮説1の検証：語彙的／統語的複合動詞のエントロピーと冗長度による比較

仮説1は、語彙的複合動詞と統語的複合動詞の特徴の違いである。本研究でエントロピーと冗長度の指標を用いる最大の魅力は、第2動詞を基準として見られる複合動詞の特徴を、語彙的複合動詞と統語的複合動詞との間で直接に比較できることである。

共起頻度とは、複合動詞をつくる2つの動詞が共起して出現する回数である。たとえば、新聞のコーパスでは「売り歩く」は7回、「訪ね歩く」は6回、「連れ歩く」が5回出現した。このような共起頻度をコーパスで検索すると、多くの複合動詞は数回の頻度に留まっており、頻繁に出現する複合動詞の種類は少ない。したがって、複合動詞の母集団分布に正規分布を仮定することはできない。また、語彙的／統語的複合動詞の2つの母集団分布の分散が等しいと仮定するのも難しい。このような尺度は、正規分布を仮定しないノンパラメトリック(以下、ノンパラ)のデータとして分析するのが普通である。しかし、様々な多変量解析を適用できるパラメトリックのデータに対して、ノンパラのデータに適用できる解析手法には限界がある。



共起頻度は1つずつ増えていく変数であるため、等間隔の尺度として捉えることができる。それならば、数学的な規則にしたがって共起頻度を他の変数に変換することで正規性や等質性を確保することが可能である。ここで、共起頻度、エントロピー、冗長度の3種類の変数の数値をすべてパラメトリック・データであると想定してそれぞれの特徴を比較してみる。図1は、これら3つの指標の、新聞のコーパスから得られた複合動詞48種の度数分布をヒストグラムで描いたものである。

共起頻度の平均は164.042、標準偏差は334.905である。歪度(わいど)は3.219であった。歪度とは、分布の非対称性の度合いを示す指標である。正規分布の場合は左右対称で、歪度は0である。正の歪度を持つ分布は、右に長く尾を引き、負の歪度を持つ分布は、左に長く尾を引く。一般に、歪度の絶対値が1以上の場合、その分布は正規分布と有意な差があることを示す。共起頻度の分布は正規分布でないことが示される。また、共起頻度の尖度(せんど)は10.524であった。尖度とは、データの分布のスソの広がりぐあいを示す指標である。正規分布の尖度は0で、尖度が正の場合、データの分布は正規分布よりもスソが長くなり、尖度が負の場合は、正規分布よりも分布のスソが短くなる。ここでの共起頻度は非常にスソが長いことを示している。

肝心の共起頻度の分布の正規性であるが、Kolmagorov-Smirnovの正規性の検定の結果は0.352で有意であり($p$<.001)、この正規性は確保されない。また、このデータは48種類の複合動詞であるが、サンプル数が50以下の場

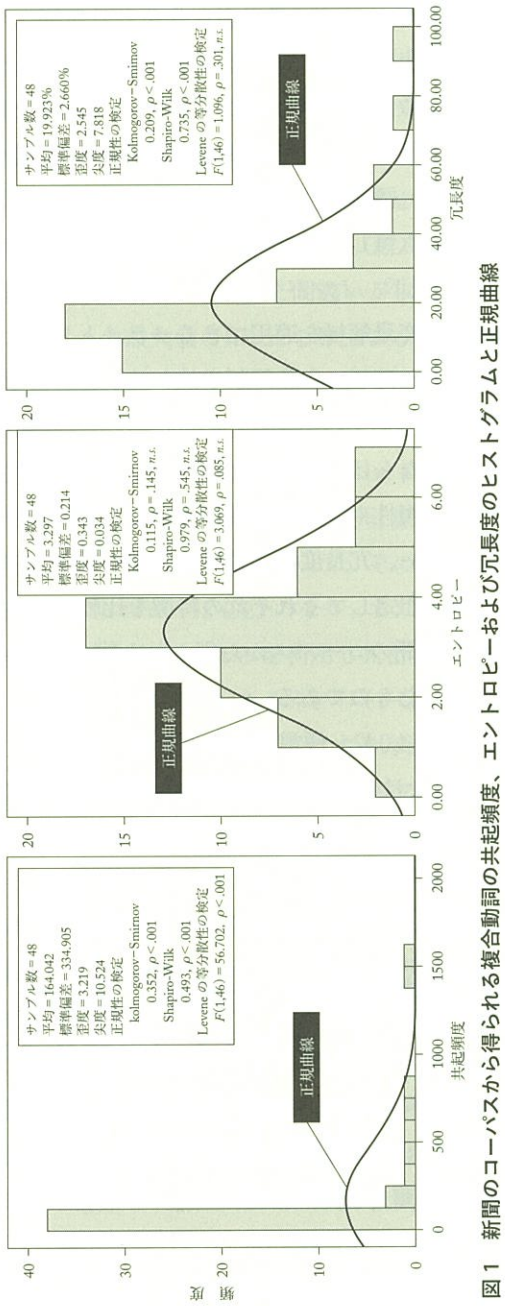


**図1　新聞のコーパスから得られる複合動詞の共起頻度、エントロピーおよび冗長度のヒストグラムと正規曲線**



合に参照すべきとされるShapirno-Wilkの正規性の検定を行っても、0.493でやはり有意となり($p$<.001)、これが正規分布ではないことを示している。さらに、2つの群の等分散が仮定できるか否かについて、Leveneの検定を用いて、48種類の複合動詞のうち37種類の語彙的複合動詞と11種類の統語的複合動詞の差を比較すると、$F(1,46) = 56.702, p$<.001で有意になり、この等質性は保証されないことが分かる。以上の結果から、共起頻度の数値に対してパラメトリック・データの統計解析を適用することはできないことが分かる。

エントロピーの平均は3.297、標準偏差は0.214である。歪度は0.343で、尖度は0.034である。Kolmagorov-Smirnovの正規性の検定の結果は0.115で有意ではないので($p$ = .145, *n.s.*)、これを正規分布とみなすことができる。サンプル数が50以下の場合に参照するShapirno-Wilkの正規性の検定の結果も、0.979で有意ではなく($p$ = .545, *n.s.*)、正規分布とみなしてよいことを示している。さらに、語彙的複合動詞と統語的複合動詞の2つのグループの分散の等質性をLeveneの検定で確認すると、$F(1,46) = 3.069, p$ = .085, *n.s.*で有意ではなく、等質性が保証されている。以上のように、エントロピーは複合動詞の共起頻度を見事に正規分布に変換していることが分かる。したがって、パラメトリック・データに適用できる様々な統計解析を行うことが妥当であることが確認できた。エントロピーは、コーパスから得られる共起頻度を分析するのに適切な指標であるといえよう。

冗長度の平均は19.923%、標準偏差は2.660%である。歪度は2.545で、尖度は7.818である。冗長度のほうは、正規分布とはみなせないことが分かる。さらに、Kolmagorov-Smirnovの正規性の検定でも0.209で有意であるので($p$<.001)、正規性は保証されない。また、サンプル数が50以下の場合に参照するShapirno-Wilkの正規性の検定を参照しても、0.735でやはり有意となり($p$<.001)、これが正規分布ではないことを示している。しかし、分散の等質性については、Leveneの検定によって語彙的複合動詞と統語的複合動詞の2つのグループを比較すると、$F(1,46) = 1.069, p$ = .301, *n.s.*で有意ではなく、これらが等質であるとみなすことができる。以上の結果から、冗長

度については、母集団が正規分布しているという条件を満たしていないので、$t$検定を用いるには問題がある。しかし、冗長度は、基本的に比率尺度であるため、パラメトリック・データとして分析することが可能であり、さらに2つの母集団分布の分散が等しいという条件を満たしているので、分散分析で比較することができる。

ここまで、共起頻度、エントロピー、冗長度の3つの指標を新聞のコーパスについて検討してきた。小説のコーパスについても同じことができるが、検討は省略する。

さて、本題の2種類の複合動詞の比較に入る。2つのグループを比較する場合、独立したサンプルの$t$検定を行うのが普通である。しかし、エントロピーは母集団の正規分布という条件を満たしているものの、冗長度は満たしていないので、冗長度について$t$検定を行うのは不適切である。一方、語彙的／統語的複合動詞の2つのグループの分散の等質性は両指標ともに保証されている。そこで、2グループの比較ではあるが、エントロピーと冗長度の両指標について一元配置の分散分析を行う。ただし、論文ではこうした詳細の分布についての検討は報告しないのが普通である。

語彙的／統語的複合動詞の2つのグループについての一元配置の分散分析を、新聞と小説のコーパスとで個別に行った。その結果、新聞のコーパスについては、統語的複合動詞のエントロピー($M$ = 4.138, $SD$ = 1.949)の方が、語彙的複合動詞のエントロピー($M$ = 2.974, $SD$ = 1.160)よりも有意に高いことが示された［$F(1,46) = 8.946, p<.01$］。しかし、冗長度については、統語的複合動詞($M$ = 24.963%, $SD$ = 25.880%)と語彙的複合動詞($M$ = 18.425%, $SD$ = 15.721%)の間に有意な違いはなかった［$F(1,46) = 1.069, p = .307$, *n.s.*］。このエントロピーの分析結果は、統語的複合動詞の$V_2$が、語彙的複合動詞に$V_2$に比べて多様な$V_1$と結合して複合動詞を作っていることを示している。冗長度については、両複合動詞を作る2つの動詞の結合頻度のパターンの規則性に違いがないことが示された。

小説のコーパスについては、まず語彙的／統語的複合動詞の分散の等質性を確認する。Leveneの等分散性の検定は、エントロピーも［$F(1,35) = 0.363, p = .551$, *n.s.*］、冗長度も［$F(1,35) = 0.128, p = .722$, *n.s.*］共に等質性を保証する結果であった。そこで、小説のコーパスについても同じ一元配置の分散分析を行った。その結果、統語的複合動詞のエントロピー($M$ = 4.864, $SD$ = 1.124)の方が語彙的複合動詞のエントロピー($M$ = 3.717, $SD$ = 0.901)よりも有意に高いことが示された［$F(1,35) = 9.144, p<.01$］。しかし、冗長度については、統語的複合動詞($M$ = 10.593%, $SD$ = 6.504%)と語彙的複合動詞($M$ = 10.117%, $SD$ = 11.679%)の間に有意な違いはなかった［$F(1,35) = 0.012, p = .913$, *n.s.*］。新聞のコーパスと同様に小説のコーパスでも、エントロピーの分析結果は、統語的複合動詞の$V_2$が、語彙的複合動詞に$V_2$に比べて多様な$V_1$と結合して複合動詞をつくっていることを示した。エントロピーの指標は、仮説1の結果を支持している。



## 9.7　仮説2の検証：語彙的／統語的複合動詞の共起パターンについての新聞と小説との違い

第1の仮説は、語彙的複合動詞と統語的複合動詞という2種類の違いを検討するものであったが、第2の仮説は、新聞のコーパスと小説のコーパスとで、$V_2$を基にして$V_1$と結びつく複合動詞の特性に違いがあるかどうかである。まず、毎日新聞の4年間のコーパスは延べ総語数が88,454,573語で、青空文庫コーパスは延べ総語数が8,370,720語であり、両コーパスのサイズには10.57倍の違いがある。しかし、表1と表2で示したエントロピーと冗長度は、$V_2$を基準とした多様な$V_1$との結合による共起頻度のパターンを示したデータサイズに左右されない指標であるので、コーパスのサイズに関係なく、2つのグループの複合動詞のパターンを比較することができる。新聞と小説のコーパスに共通して出現する共起頻度の述べ頻度が10回以上の複合動詞は、34種類である。エントロピーは正規分布しているので$t$検定を行うこともできるが、仮説1との一貫性を考えて、これら34種類の複合動詞について反復測定による分散分析を行う。反復測定である理由は、同じ複合動詞が、新聞のコーパスと小説のコーパスで算出されていると考えるからであ

る。

エントロピーについて新聞と小説のコーパスの違いを比較してみると、新聞($M$ = 3.422, $SD$ = 1.581)の方が、小説($M$ = 4.080, $SD$ = 0.991)よりも有意に低かった[$F(1,33) = 6.898, p<.05$]。これは、新聞よりも小説のコーパスの方が、$V_2$を基準として$V_1$と結合して作られる複合動詞のパターンが、多様性に富んでいることを示している。冗長度についても同じ分析をした結果、新聞($M$ = 20.958%, $SD$ = 19.864%)の方が、小説($M$ = 10.727%, $SD$ = 11.006%)よりも有意に高かった[$F(1,33) = 7.358, p<.05$]。これは、新聞の方が小説よりも複合動詞の2つの動詞の結合関係が規則的であったことを示している。以上のように、エントロピーと冗長度の両方において、新聞と小説のコーパスの複合動詞の共起頻度パターンに有意な違いが見られ、新聞のコーパスの複合動詞は、小説のコーパスほどの多様性は無く、より規則的なパターンを示すことが分かった。当然ながら、新聞は、一般大衆に情報を伝達するために、簡潔で分かりやすい一定の表現スタイルを採っているので、小説よりもエントロピーが小さく、冗長度が大きくなったのであろう。一方、小説は、作家の個性に応じた多様な表現が現れるために、新聞よりもエントロピーが大きく、冗長度が小さくなったのであろう。

エントロピーと冗長度は、両者のコーパスの書き手の意図と目的を反映した結果を反映しており、両指標がコーパスの違いを比較するのに有効な方法であることが分かる。以上のように、エントロピーと冗長度の指標は、仮説2を支持していた。

## 9.8　最後に(エントロピーと冗長度を用いたその他のコーパス研究)

最後に、ここで紹介したのと同様の方法で研究した論文を3つ紹介する。第1に、玉岡・宮岡・林(2003)であり、韓国語を母語とする日本語学習者による敬語表現についてエントロピーと冗長度の指標を用いて検討した研究である。第2に、Miyaoka & Tamaoka(2005)は接頭辞と接尾辞の共起頻度について、接尾辞の方が接頭辞よりもエントロピーが有意に高いことを見出し、右側主要部の規則を支持した研究である。第3に、玉岡・木山・宮岡(2008)はオノマトペと動詞の共起パターンを分析した結果、ヒトの言語産出と新聞のコーパスは類似しているが、小説のコーパスとは異なっていることを示した研究である。



**参考文献**

影山太郎(1993)『文法と語形成』ひつじ書房.
影山太郎(1999)『形態論と意味』くろしお出版.
Miyaoka, Y., & Tamaoka, K. (2005) A Corpus investigation of the right-hand head rule applied to Japanese affixes. *Glottometrics*, 10 (RAM Verlag, Lüdenscheid, Germany) 45–54.
Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* 27: pp. 379–423 (Part I) and pp. 623–656 (Part II).
玉岡賀津雄・木山幸子・宮岡弥生(2008)「ヒトの言語産出とコーパスの頻度はどのくらい類似しているか」『日本言語学会第136回大会予稿集』122–127.
Tamaoka, K., Lim, H., & Sakai, H. (2004) Entropy and Redundancy of Japanese Lexical and Syntactic Compound Verbs. *Journal of Quantitative Linguistics* 11: pp. 233–250.
玉岡賀津雄・宮岡弥生・林炫情(2003)「エントロピーと冗長度で表現の多様性と規則性を表す試み—韓国語系日本語学習者の敬語表現を例に」『日本語科学』14(国立国語研究所)98–112.