玉岡賀津雄（2016)『レキシコンフォーラム』7, 239-264.

《チュートリアル》

# 共起表現研究のためのコーパス検索入門

玉岡 賀津雄

## １．はじめに

言語の研究では，ある表現が他のどのような表現とどのようなパターンで共起するかがテーマになることが多いようです。これは，コロケーション（collocation）と言われ，「連結語句」「連語」とか訳されます。ここでは「共起表現」と呼びます。たとえば，オノマトペは，どのような動詞とどのようなパターンで出現するのでしょうか。また，接辞は，どのような名詞と共起して，どのようなパターンを作るのでしょうか。こうした研究テーマには，コーパスの共起表現検索が活用できます。このチュートリアルでは，共起表現の検索の方法を紹介します。

まず，身近に使える Google の検索について，２つの動詞からなる複合動詞を例に検討します。次に,『茶漉』を使った共起表現の検索の方法を，例を挙げて詳しく説明します。その後で，国立国語研究所が開発した『少納言』と『中納言』の検索エンジン・ソフトについて簡単に触れます。さらに，コーパス作成・検索・出力の総合ソフトである『茶器』（ChaKi.NET）について短く説明します。最後にコーパス関連の情報が，分かりやすく整理された李在鎬先生のホームページを紹介します。

## ２．Google の事例および頻度検索

Google（https://www.google.co.jp/）を使うとインターネット上で簡単に大規模な検索ができます。たとえば，日本語では，２つの動詞を組み合わせて，１つの動詞を作ることがよくあります。これは，複合動詞といいます。典型的なのは，２つの他動詞の組み合わせです。たとえば,「ロケットを打ち上げる」であれば,「打つ」も「上げる」もいずれも他動詞です。これを，影山（1993）では,「他動性調和の原則」とよんでいます。ところが,「ロケットが打ち上がる」も正しい表現です。「打つ」は他動詞ですが,「上がる」は自動詞です。この複合動詞は他動性調和の原則にしたがっていません。

まず，用例検索をしてみましょう。Google の検索ボックスに，「打ち上がる」を入れると，いろいろなサイトの情報が表示されます。その中から，「花火が打ち上がる」という表現があることが分かります。また，「茨城でイルカ 150 頭が打ち上がりました。」という表現もあり，イルカが海岸に「打ち上がる」という場合にも使われることが分かります。これらの表現が使用されているサイトをみれば，どういう状況で，どのようなジャンルで使われているかも分かります。Google ではいろいろな用例検索ができます。

次に，どのくらいの頻度で使われているかを調べてみます。ふつうに考えれば，他動性調和の原則にしたがっている「打ち上げる」の方が，原則にしたがっていない「打ち上がる」よりも頻繁に使われていると思われます。そこで，Google の検索ボックスに「打ち上げる」を入れてみます（2015 年 4 月 30 日現在)。左上に「約 472,000 件（0.17 秒)」という結果が表示されます。「件」はヒット数で，頻度と考えられます。「約」がついているのは，1,000 までの単位で表示されているからです。つまり，「おおよそ」という意味です。また，0.17 秒というのは，Google のサーバーで検索に要した時間です。

一方，「打ち上がる」の件数は，なんと約 2,840,000 件（0.18 秒）となり，「打ち上げる」よりもはるかに多くなりました。このヒット数の結果を頻度としてそのまま使うと，他動性調和の原則にしたがわない「打ち上がる」の方が「打ち上げる」よりも 6 倍くらい頻繁に使われていることになります。しかし，これを信用してはいけません。なぜなら，他動性調和の原則にしたがっていない「打ち上がる」は，「打つ」と「上がる」が別々に検索されて，頻度が加算されている可能性があるからです。

そこで，Google の「検索オプション」（英語では，advanced search とよばれる）サイト http://www.google.com/advanced_search から検索をしてみましょう。「検索するキーワード」にはいくつかのオプションがあります。まず，「語順を含み完全一致」検索を行います。「検索対象と完全に一致するキーワードを二重引用符で囲んで入力します（例: “ヨークシャーテリア”)」という説明があります。そのとおりに入力して検索してみましょう。すると，“打ち上げる”は約 469,000 件で，“打ち上がる”は約 298,000 件という頻度が，左上に表示されます。先ほどの検索とは異なり，他動性調和の原則にしたがっている「打ち上げる」の方が「打ち上がる」よりも頻繁に使われています。二重引用符で囲んだ完全一致の「"打ち上げる"」（469,000 件）と，そのまま入力の「打ち上げる」（472,000 件）とのヒット数（頻度）は近いのに，「"打ち上がる"」（2,840,000 件）と「打ち上がる」（298,000 件）では，10 倍くらいの差があります。このように，単純な Google の検索ボックスからそのまま入力して頻度を調べると，一見うまく頻度が出たようにみえますが，そのままの頻度を使うことは危険であることが分かります。Google Japan は，一般向けに，活用形や表記の揺れをどう

扱っているのか説明していませんので，どのような検索が行われているか分かりません。

さて，二重引用符（“　”）を使った完全一致検索では，動詞の変化は考慮されていません。そのため, OR のコマンドを使って,「“打ち上がる”OR“打ち上がろう”OR“打ち上がれ”」の３つの活用で検索すると約 311,000 件と少し多くなります。このように，活用ごとに検索して，頻度を調べることもきます。もちろん，可能な活用を全部含むことで，特定の動詞のすべての活用を網羅した頻度を検索することもできます。

OR などは検索演算子とよばれ，検索の結果を絞り込むときなどに使います。「ウェブ検索 ヘルプ」の「検索での句読点、記号、演算子」に他の演算子について書かれています。URL は https://support.google.com/websearch/answer/2466433?hl=ja です。この検索方法だと，ある程度，信頼性のある頻度を出すことができると思います。なお，Google 検索エンジンの goo（http://www.goo.ne.jp/）のフレーズ検索を使ってコーパス研究をした例（杉村, 2013, 2014）があります。しかし，現在では，頻度が出なくなっていますので，頻度検索としては使用できません。

ここまで，Google 検索の方法を説明しました。ここで，Google 検索の頻度あるいはヒット数の問題点を３つ挙げておきます。第１に，語彙や表現の使用頻度は，ジャンルによって異なるといわれています。しかし，Google 検索では，コーパスの出典（リソース）がはっきりしませんので，コーパスの特性を説明することはできません。第２に，Google 検索には，誤りも含まれています。サイトに書き込まれた雑多な語彙や表現も一緒に検索されてしまうので，正しい表現だけを選ぶことはできません。そのため，Google 検索は，全体的なおおざっぱな検索結果と考えてください。第３に，検索の時間によって結果が多少変わることです。1 時間後に検索すると，すでに頻度が異なることがあります。そのため，科学的な研究に求められる再現性はありません。

ただし,「筑波ウェブコーパス」（http://nlt.tsukuba.lagoinst.info/）は，ウェブ上の html テキストから 11 億語のコーパスを構築して，検索できるようにしています。NINJAL-LWP for TWC という検索ツールで, 語彙の共起関係などを調べることができます。NINJAL-LWP は，国立国語研究所と Lago 言語研究所が共同開発したコーパス検索システムで，TWC（Tsukuba Web Corpus）は『筑波ウェブコーパス』を意味します。サイトに詳しい説明（2015 年３月 27 日版のユーザーマニュアル，バージョン 1.30）が書かれていますので，参照して下さい。これを使うと，ウェブ上の検索でも，結果に再現性があります。

さて，参考までに，毎日新聞の 2000 年から 2010 年までの 11 年間の記事（固有名詞を含んで，299,695,840 形態素数）で「打ち上がる」と「打ち上げる」を検索してみます。すると,「打ち上がる」または「打ちあがる」の２つの表記で，すべての活用形を含んで検索し

ても，わずかに 49 例しかみつかりませんでした。一方，「打ち上げる」または「打ちあげる」では，基本形だけでも 631 例が見つかりました。つまり，ほぼ誤って使われることはないと思われる新聞というジャンルのコーパスでは, 少なくとも 12 倍以上の頻度の差があることになります。これは，誰がいつ検索しても同じ結果になりますので，再現性があります。新聞記事でみる限りでは，他動性調和の原則にしたがっている「打ち上げる」のほうが，原則にしたがわない「打ち上がる」よりもはるかに頻繁に使われていることが分かります。なお，語彙的複合動詞のデータベース（約 2,700 語）が，国立国語研究所の基幹型共同研究プロジェクト「日本語レキシコンの文法的・意味的・形態的特性」（プロジェクトリーダー 影山太郎）によって，作成されており，Lago 言語研究所（赤瀬川史朗）がオンライン検索システム（http://vvlexicon.ninjal.ac.jp/）を構築していますので，参照して下さい。以上のように，Google と新聞記事の検索結果を比べると大きな違いがでてきます。コーパスの出展や特性の違いで頻度に違いがあることをよく考えて，注意して研究に活用しましょう。

### ３．簡単にできる共起表現検索エンジン：『茶漉』

意外に知られていないのですが，パデュー大学の『茶漉』と呼ばれるコーパス検索エンジン（Linux 上でＣ言語を使用）は，共起頻度が検索し易いように設定されています（深田, 2007)。コーパスの出典がはっきりしているので，結果についての説明がジャンルごとにできます。2015 年６月からは，最新の Linux に搭載した『茶漉』専用の 64 bit サーバーで起動するようになり，検索が迅速にできるようになりました。『茶漉』の製作者は，パデュー大学外国語外国文学学科教授および同大学先端技術言語学研究所所長の深田淳先生です。当初，言語データ（コーパス）を検索可能なデータファイルに変換する段階で形態素解析システム『茶筌』(奈良先端科学技術大学院大学自然言語処理学講座で開発）を用いていました。茶筌を用いて立てたお茶（データ）から必要な情報のみを漉し取るシステムということで,『茶漉』と命名したそうです。ただし，現在の『茶漉』の形態素解析には和布蕪（メカブ, MeCab）が使われています。なお，この検索エンジンは，科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成 13 年度～15 年度，研究代表者: 大曾美惠子）の一環として作成されたものです。このサイトは，2002 年（平成 14 年）にすでにできていたのですから，極めて画期的です。

私はこの検索サイトのおかげで, コーパス研究の主要論文を３つ書くことができました。まず，2004 年に，統語的・語彙的複合動詞の２つの動詞の共起頻度のパターンを分析した論文を，Taylor & Francis 出版の国際ジャーナルの *Journal of Quantitative Linguistics* に掲載

することができました。その後，2011 年に日本言語学会が発行する学術誌『言語研究』に，新聞と小説のコーパスにおけるオノマトペと動詞の共起パターンを分析した論文（玉岡・木山・宮岡, 2011）と，新聞コーパスの用例に基づく自他両用の「－化する」における自動詞用法と他動詞用法の比較の論文（木山・玉岡, 2011）を掲載することができました。これらはすべて共起表現に関する研究です。

『茶漉』は，用例検索はもちろん，共起表現の検索が，簡単にできるのが強みです。共起表現は，少なくとも２つ以上の語の組み合わせです。その際，ターゲット語が頻繁に見られる表現であればよいのですが，１語の頻度そのものが低いと，他の語との共起頻度はさらに大きく落ち込みます。そのため大規模コーパスが必要になります。『茶漉』はまさに大規模コーパスでの頻度検索が可能です。『茶漉』はあまり知られておらず，説明がサイトのマニュアルくらいしかないので，ここで詳しく紹介します。

### 3.1 『茶漉』のコーパスとその規模

『茶漉』のサイトは，http://tell.cla.purdue.edu/chakoshi-wiki/です。このサイトに入ると，まず，左端の上に「茶漉 WiKi」と書かれており，その下に「茶漉一般公開サイト」がでてきます。これをクリックすると，ユーザーネームもパスワードも入力することなく検索サイトに入れます。ただし，この検索サイトで検索できるのは，「青空文庫」と「名大会話コーパス」のみです。「青空文庫コーパス」は，青空文庫（http://www.aozora.gr.jp）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，許諾の上，選んでコーパス化したものです。コーパスに収録されている作品のリストは，サイトの説明にすべて挙げられています。もちろん，著作権のなくなった小説のコーパスですので，現代語だけを選んでいるとはいえ，やや古い日本語表現になります。

一方，「名大会話コーパス」は，2001 年から 2003 年頃に集めた会話を文字化したデータです。日本語の話しことばのコーパスは少なく，このコーパスは自由に使用できるので貴重です。２名から４名の話者による約 100 時間の雑談が収録されています。会話は大半が親しい者同士の雑談です。女性が 161 名，男性が 37 名です。年齢，出身地は様々で，詳細はサイトの参加者情報をみてください。

一方，左上の「茶漉 WiKi」に「パデュー大学茶漉サイト」というのがありますが，これはアクセス制限があり，ユーザーネームとパスワードが必要です。このサイトへのアクセスについては，深田淳先生（afukada@purdue.edu）に直接メールで問い合わせてください。条件を満たせば，無料で使用できるはずです。また，検索の詳細についてメールで質問しても，すぐに優しく教えてくださいます。

「パデュー大学茶漉サイト」にアクセスできれば,「名大会話コーパス」と「青空文庫」に加えて,「講談社ブックス」, 松竹映画『男はつらいよ』シリーズのシナリオ集である「寅さんシナリオ」での検索ができます。さらに, 2015 年 6 月からは, 1991 年から 2010 年までの 20 年間の「毎日新聞」のコーパスでの検索が可能になりました。毎日新聞の記事は, １年ごとにファイルにまとめられているので, 特定年, 複数年, あるいは 20 年分全部を検索することができます。

『茶漉』で検索できるコーパスの出典とそれらのサイズは表１に示したとおりです。表１の数値は, MeCab で解析した形態素数です。MeCab では, 品詞のタグが付けられるものすべてを検索対象の文字列としており, それらを「語」と呼んでいます。語と形態素の区別は難しく, MeCab では, 厳密に区別していません。たとえば,「国会で」であれば, 名詞の「国会」と助詞の「で」で２つの形態素あるいは語となります。また,「食べさせられる」という述部を解析すると, 動詞の「食べる」と動詞接尾の「させ」と動詞接尾「られる」の３つに分けられ, ３つの形態素あるいは語と数えます。また, 記号や句読点, 空白もカウントされます。そのため, 表１の数値は, 厳密な形態素数あるいは語数を示しているわけではありません。あくまで, 概算として考えてください。

表１ 『茶漉』で検索できるコーパスとサイズ

| コーパスの出展 | 形態素数 |
|---|---|
| 名大会話コーパス | 2,025,113 |
| 青空文庫(小説集) | 8,589,963 |
| CASTEL-J( 講談社ブックス等) | 3,061,913 |
| CASTEL-J(寅さんシナリオ) | 767,746 |
| 毎日新聞20年分 | 650,821,986 |
| 合　計 | 665,266,721 |

*注*: 表の数値は MeCab で解析された形態素数で, 記号類が含まれています。

一般公開サイトの「名大会話コーパス」が 2,025,113,「青空文庫」が 8,589,963 で, これで, 形態素数が 1,000 万を越えます。パスワードが必要なパデュー大学茶漉サイトの「CASTEL-J（講談社ブックス等)」の形態素数は, 3,061,913 で,「CASTEL-J（寅さんシナリオ)」は, 767,746 です。さらに, 凄いのは, 1991 年から 2010 年までの 20 年間の毎日新聞記事のコーパスです。総形態素数は 650,821,986 です。6.5 億を超える大きさです。『茶漉』は, すべてのコーパスを一緒に検索することができます。その場合の形態素数は

665,266,721 になります。もちろん，全部のコーパスを使う必要はなく，適当にコーパス規模を選択できるのも『茶漉』の強みです。『茶漉』は，出典がはっきりしており，素人にもウェブ上で簡単に検索ができますので，共起表現の事例検索および頻度検索で，その能力が十分に発揮されるはずです。

3.2 『茶漉』での検索方法

「茶漉一般公開サイト」でも，アクセス制限のある「パデュー大学茶漉サイト」でも，使えるコーパスに違いがあるだけで，検索の方法は同じです。いずれであっても「『茶漉』日本語用例・コロケーション抽出システム」というサイトに入ります。画面の上から順番に指定すればよいように設計されています。

まず，コーパスを指定します。「一般公開サイト」だと，図１のように２種類のコーパスが使えます。コーパスは，チェックボックスをクリックすることで，複数のコーパスを指定することができます。指定を解除したい場合は，再度クリックすると解除されます。図１の場合には，「名大会話コーパス」が選ばれており，このコーパスについての検索だけが行われます。

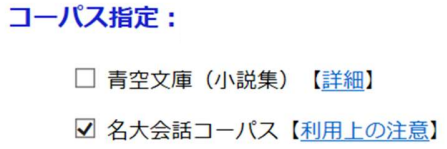


図１ 「一般公開サイト」のコーパス指定

「パデュー大学茶漉サイト」のコーパスは，図２のようになっています。名大会話コーパス，講談社ブックス，寅さんシナリオ，青空文庫，毎日新聞のコーパスがあります。図２では，毎日新聞の 2008 年，2009 年，2010 年の３年間がチェックされています。この条件だと，毎日新聞３年分だけが検索されます。目的に応じて，コーパスの大きさを適当に調整することができます。

次に，検索のパターンを設定します。『茶漉』のマニュアルに紹介された例（例３）を使って説明します。これは，「誘う」という動詞の目的語にどんな名詞が来るかを検索する例です。「誘う」は，たいてい「人が人を誘う」という使い方が一般的だと思いがちです。しかし，有生名詞ばかりでなく，無生名詞が出現する可能性もあるのではないでしょうか。


図２　「パデュー大学茶漉サイト」のコーパス指定

まず，検索ターゲット語（kw: keywords）の語形に「誘う」を入力します。平仮名で書かれることもあるので，半角の「|」を間にいれて「誘う|さそう」としておけば，２つの表記で検索できます。品詞は「+動詞」と入れます。これで動詞だけが検出されます。すべての活用形について検索したいので,「全活用形」にチェックを入れます。品詞のところに「含」「除外」の選択があります。初期設定は「含」です。この条件であれば，品詞が動詞であれば，どんな活用でも検索することになります。「除外」を選択すれば，動詞を除外することになります。ここでは，動詞だけを調べたいので，動詞を指定して「含」とし，動詞だけの出力に設定します。検索ターゲット語の「誘う」の目的語を対格の「を」と指定するために，－１のところに,「を」を入力します。「を」の助詞には活用はないので，後は何も設定しません。以上のこの例の検索設定は，図３に示したようになります。

検索サイトは，上から順番に設定していけば良いようにデザインされています。説明が前後しますが，初めにスパン（span）の設定があります。スパンとは，検索ターゲット語の共起表現を探す範囲です。初期設定は，前後で３になっています。必要と思われる検索ターゲット語の前後の語数を半角文字の整数で入力します。「誘う」の目的語検索では，検索ターゲット語の対格の「を」とさらにその前の名詞の２つに絞りますので，図３のように，「前」の欄に半角で 2 と入力します。「後」は考慮しなくてよいので，半角で 0 と入力します。

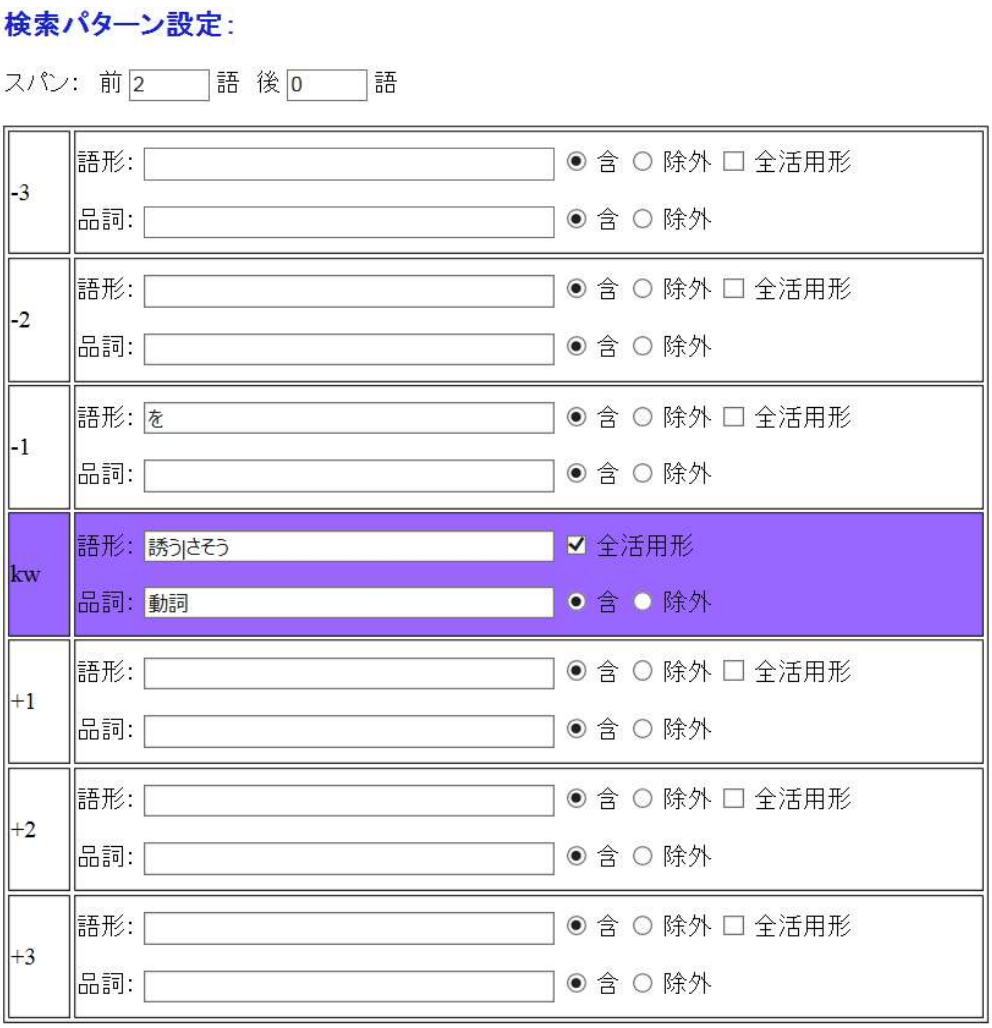


図３　「誘う」の目的語を検索する「検索のパターンを設定」画面

図３で示した「検索パターン設定」のさらに下には，図４で示した「コロケーション出力設定」と「kwic 出力設定」があります。今回は，事例検索なので，「コロケーション出力設定」は無視します。kwic とは，keyword in context の略です。ここの設定で，ターゲット語の前後の語数の設定と，前後の文数の設定が選べます。初期値は，文指定になっており，前後の文がそれぞれ 0 ですので，検索で選ばれたすべての用例が文単位で出力されます。検索ターゲット語の含まれていない１つ前の文も一緒に出力したい場合には，「前」を 1 とします。後の文を出力したければ，「後」を 1 とします。「誘う」の目的語の場合には，文は必要ではないので，「語数指定」をチェックします。語数は２つでよいので，2 を入力します。そうすると，検索ターゲット語の前の２語の名詞と「を」が表示されます。「形態素

区切り文字列」というのは，形態素ごとになんらかの区切りを入れたいときに使います。たとえば，「/」を入れたければ，これを入力します。初期設定は，半角スペースになっています。そのまま空欄にしておきます。「誘う」の入力は図４のようになります。これで，2010年の１年間の毎日新聞の記事（総形態素数は，29,032,329）で検索することができます。258件が見つかりました。有生名詞の「親類を誘う」もありますが，驚いたことに，無生名詞を目的語とする場合が多いことが分かります。

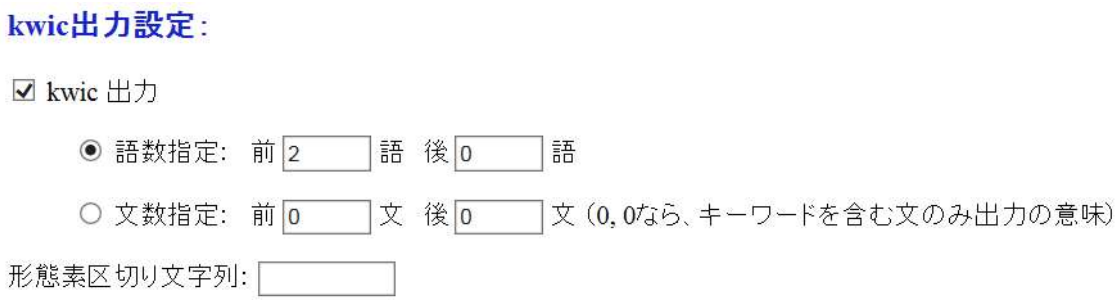


図４　「誘う」の目的語検索のための「kwic 出力設定」

より詳細にみるために,「目的語＋を＋誘う」の全出力を Excel にコピーします。その際，Excel の「テキストファイルウィザードを使用」します。「形態素区切り文字列」は初期設定の半角スペースなので,「カンマやタブなどの区切り文字によってフィールドごとに区切られたデータ」を選び，「次へ」をクリックすると，Excel は自動的に「スペース」を認識して，「名詞」「を」「誘う」の３つに分けてくれます。そして，「完了」をクリックして，Excel で 258 件を整理することができます。そうすると，「笑いを誘う」がもっとも多く，65 件で，全体の 25.2%を占めます。次に「ミスを誘う」が 32 件で，全体の 12.4%です。「失策を誘う」が９件,「オウンゴールを誘う」が７件です。有生名詞の人を誘うという表現はほとんどでてきません。青空文庫の小説と比べてみるとおもしろいでしょう。やってみて下さい。

### 3.3 『茶漉』のコロケーションの出力

複数の共起表現の検索をしてみましょう。玉岡・木山・宮岡（2011）の論文では，オノマトペと共起する動詞の種類と頻度を調べて，エントロピーを使って共起パターンを指標化しています。この場合は，検索ターゲット語の kw の欄の語形に検索したいオノマトペを入れます。たとえば，「どんどん」を入力します。「どんどん」の品詞は副詞ですが，品

詞は空欄にしておきます。この語の前に来る語は検索の必要がないので，スパンの「前」は半角で 0 を入力します。次に，「どんどん」の後にくる動詞を調べたいので，「後」は３語までとして，3 を入力します。ここまでが検索の手続きです。

『茶漉』には，コロケーション（共起）強度計算の対象となる語彙を制限する「コロケーション出力設定」がついています。たとえば，「どんどん」がどういう動詞と共起するかを知りたければ，図５のように，品詞の蘭に「+動詞」を入力し，他の品詞を共起計算から除外できます。逆に「-動詞」とすると，動詞のみを除外することもできます。語形の入力によって，特定の動詞（「働く」とか）を指定することもできます。「活用形をまとめて集計する」をチェックすると，共起する動詞について，さまざまな活用形を同一の語として，合計して集計し，辞書形のみが出力表示されます。これをチェックしていないと，各活用形が別々に集計されるので，集計された表が分かり難くなります。

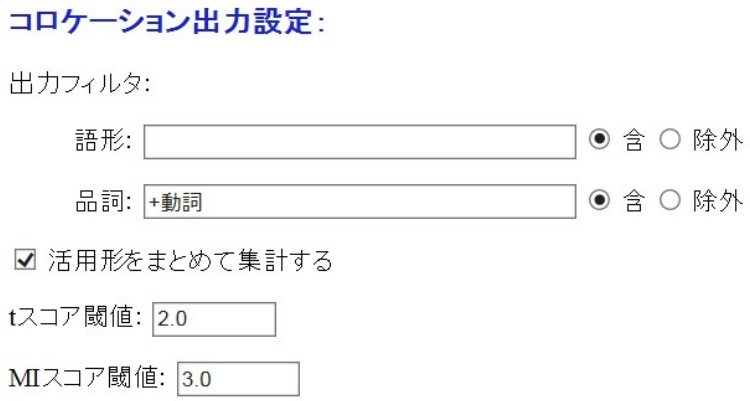


図５　「どんどん」と共起する動詞のコロケーション情報を出力する例

さらに，*t* スコアと MI スコアが個々の共起表現ごとに一覧表で出力されます。これらは，共起の強度や有意性を表す指標です（詳細は，石川, 2006, 2012 および李・石川・砂川, 2012 を参照）。神戸大学国際コミュニケーションセンター教授の石川慎一郎先生のウェブサイトの「コーパス言語学」「言語統計学」のタグから各種情報が入手できます（http://language.sakura.ne.jp/s/）ので，参照してください。『茶漉』では，「*t* スコア閾値」「MI スコア閾値」は，共起（コロケーション）出力に表示する際の，最低限必要な閾値として使用されています。初期設定は，図５のように，*t* スコアが 2.0 で，MI スコアが 3.0 になっています。これより小さい値の共起表現は出力されません。今回は，この閾値をそのまま使います。

『茶漉』で 1991 年から 1995 年までの５年間の毎日新聞の記事(総形態素数が 109,246,846)で検索すると,「どんどん」と動詞の共起を検索すると，1,941 件が検出されます。また，動詞ごとに集計された指標の一覧表が出力されます。たとえば,「どんどん」と共起する「膨らむ」の場合は，$t$ スコアが 3.60，MI スコアが 10.02，G スコアが 154.84，コーパス頻度が 843 (「膨らむ」の頻度)，スパン頻度が 13 (「どんどん」との共起回数)，期待頻度が 0.015 と標示されます。同様に,「近づく」「広がる」「伸ばす」「攻める」「失う」「買う」「進む」「変わる」「売れる」などについても指標が標示されます。これらの指標は，個々の動詞が「どんどん」と共起する強さを示すものです。

「どんどん」と複数の動詞との共起パターンを一括して示すためには，別の指標が必要です。エントロピー (entropy) と冗長度 (redundancy) という情報量の尺度 (Shannon, 1948) を用いると，個々のオノマトペと多様な動詞との共起パターンを１つの数値で表し，比較することができます（尺度の詳細は，有本, 1982; 堀, 1979; 海保, 1989; 木山・玉岡, 2011; 玉岡, 2011; 玉岡・木山・宮岡, 2011 を参照)。エントロピーは，何が起こるか予測がつかないという乱雑さの増減を示す指標です。オノマトペと動詞の共起関係については，共起する動詞の種類とその種類ごとの共起頻度に基づいて１つの指標が算出されます。エントロピーが大きい場合は，オノマトペは多様な動詞と共起しており，エントロピーが小さい場合には特定の動詞のみ共起していることを示します。この指標を使うと，オノマトペがどの程度多様な動詞と共起するかという意味的拡散の程度を比較検討することができるわけです。

指標について，もう少し詳しく説明します。図６に示したように，$t$ スコア, MI スコア, G スコアは,「どんどん」というオノマトペと特定の動詞についての指標です。そのため，動詞ごとに異なる指標が算出されます。「どんどん」と「膨らむ」との共起であれば，$t_1$, $MI_1$, $G_1$ で,「近づく」であれば $t_2$, $MI_2$, $G_2$ というわけです。特定のオノマトペについて，どの動詞が特に頻繁に共起するかを判定するには，これらの指標は役に立ちます。

一方，エントロピーは，特定のオノマトペと共起するすべての動詞を含んで，どのような共起パターンになっているかを総合的に示す指標です。オノマトペと動詞の共起パターンを１つのセットとしていますので，複数のオノマトペの動詞との共起パターンを直接比較することができます。たとえば，1991 年から 1999 年までの 9 年間の毎日新聞の全記事のコーパス（総形態素数 273,514,662）だと（玉岡・木山・宮岡, 2011)，「どんどん」のエントロピーは 7.81 と大きく,「だらだら」が 4.72,「ぐらぐら」が 2.87 です。新村 (編) (1998)『広辞苑』（第５版）によると,「だらだら」の意味は３つで,「どんどん」と「ぐらぐら」は２つの意味を持っています。動詞との共起パターンからみると意味の数と関係なく,「ど

んどん」は非常に多様な動詞と共起しているようにみえます。このように，エントロピーを使うと，個々のオノマトペと動詞の共起パターンを直接比較することができます。

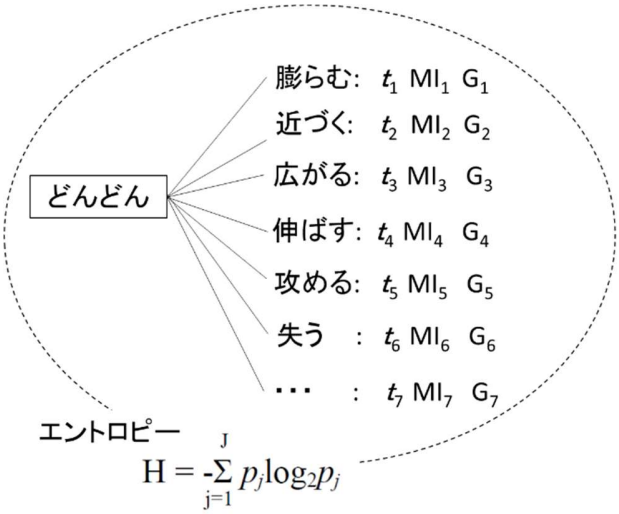


図６　$t$スコア，MIスコア，Gスコアとエントロピーが示す指標の違い

さて，「ごくごく」は，意味が１つしかない単義のオノマトペです。「飲む」など特定の動詞とのみ共起し，共起パターンに多様性がありません。予想通り，エントロピーは低く，1.19 です。意味が１つしかない単義の 14 語（きらきら，ゆらゆら，とぼとぼ，すらすら，じろじろ，ぴょんぴょん，すやすや，ぼうぼう，げらげら，ごくごく，めそめそ，ぽたぽた，しとしと，ずきずき）と意味が複数ある多義の 14 語（どんどん，ばたばた，ばりばり，だらだら，がんがん，ころころ，かんかん，ぐらぐら，ぱちぱち，ぷんぷん，ちょろちょろ，ことこと，ぐうぐう，しくしく）のオノマトペのエントロピーを $t$ 検定で比較することもできます。両者には，有意な違いがみられました（具体的な数値は，玉岡・木山・宮岡, 2011 を参照）。「どんどん」「だらだら」「ぐらぐら」の例のように，多義のオノマトペの中では，さらに多様性があるようですが，単義と多義との比較では，多義のオノマトペのほうが, 単義のオノマトペよりも, 動詞との共起パターンが多様であることが分かります。単義・多義のようなグループに分けることができれば，エントロピーの値を使って，統計的に解析することもできます。

なお，エントロピーの統計解析の $t$ 検定（コーパス研究の $t$ スコアとは異なる）は，通常の統計で扱う $t$ 検定とは微妙に公式が違います。エントロピーの標準偏差の計算が，微妙に一般の連続変数の場合と異なります。エントロピーの平均と標準偏差が算出できれば，

*t* 検定の公式に数値を代入すると結果が計算できます。Miyaoka and Tamaoka（2005）の 52 ページに書かれた一連の公式を参照してください。厳密に言えば，統計専用のソフトはエントロピーの値の解析には使えないのですが，２つのグループの有意差が微妙でない限り，通常の *t* 検定とほぼ同じ結果になります。

3.4 ワイルドカード「*」とミニ和布蕪（メカブ）

『茶漉』の検索パターンの設定では，語形と品詞の指定ができるようになっています。語形としては，「てきぱき」「実際」「さっぱり」「を」「面子」など完結した語を入力することができます。それに加えて，「駄菓子」「駄洒落」など「駄」で始まる複合語あるは合成語（compound words）を検索したい場合は，検索ターゲット語（kw）に「駄*」と入力し，品詞は「名詞」としておくだけで検索できます。「*」はワイルドカードといわれ，「駄」に後続して結合されるすべての複合語を抽出することができます。前後の共起表現はありませんので，スパンは前後ともに半角の 0 を入れておきます。2010 年に出版された毎日新聞の記事では，240 件が検出されました。「駄目」が非常に多いのですが，「駄菓子」や「駄目押し」もでてきました。

ところが，ワイルドカードが常に正しい検索をしてくれるわけではありません。たとえば，Miyaoka and Tamaoka (2005)は，漢語の接尾辞と接頭辞の共起パターンを調べています。「安全性」「主体性」などの「性」の接尾辞を持つ熟語を検索したい場合は，語形として「*性」とターゲット検索（kw）に入力し，品詞は「名詞」としておくだけで検索できそうに思えます。ところが，「安全性」は「安全」と「性」の名詞と名詞語尾という２つの語として記録されており，別々に扱わなくてはなりません。こういう問題を避けるために，『茶漉』には「ミニ和布蕪」のタブがあります。「和布蕪」はメカブと読みます。通常は，「ワカメの付着器の上にある葉状部の中で厚く折り重なってひだ状になった部分」を意味します。しかし，ここでは形態素解析プログラムの MeCab を意味します。これをクリックすると，「日本語の文やフレーズを入力して「解析」ボタンを押してください。」という説明がでてきます。空欄に「安全性」を入力し，「解析」をクリックすると，MeCab による形態素解析の「解析結果」が出力されます。それによると，「安全性」は「安全」が「名詞-形容動詞語幹」で，「性」は「名詞-接尾」と示されます。これで，「安全性」は２つの語として処理されることが分かります。なお，MeCab は 京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所が共同で開発した形態素解析ソフトです（詳細は，Kudo, Yamamoto and Matsumoto, 2004 を参照してください）。無料で公開されており，http://osdn.jp/projects/sfnet_mecab/releases/からダウンロードして，個人のコンピュータにイ

ンストールすることもできます。

以上のように，「性」については，検索は２語としなくてはなりません。「性」の接尾辞が作る熟語は，「前」にくる名詞との共起表現になります。そのため，『茶漉』では，スパンは「前」に 1，「後」に 0 を入れます。検索ターゲット語（kw）は，語形が「性」で，品詞は「+接尾」を選びます。「+」をつけておくと，これだけを抽出してくれます。さらに，この前に名詞がくるものを選びますので，－１の欄の品詞を「+名詞」としておきます。「kwic 出力設定」は，名詞と「性」の共起なので，「語数指定」の「前」を 2 とします。「後」はないので 0 とします。これで検索すると，ほぼ問題なく「性」を接尾辞とする熟語が検出できます。出力されたものは，いったん Excel に取り込んで，頻度を数えたりすることができます。さらに，「性」に先行する名詞の *t* スコア，MI スコア，G スコアなども出力されます。

検索したい接頭辞と接尾辞と決めて，「ミニ和布蕪」で語構成をチェックしながら接辞の熟語の種類と頻度を検索してください。Miyaoka and Tamaoka (2005)では，接頭辞として「大」「不」「無」「新」「初」「非」「全」「再」「超」「反」「未」「毎」の 12 種類，接尾辞として「的」「者」「性」「学」「化」「論」「家」「式」「界」「風」「状」「用」の 12 種類を選らんで，熟語の種類と共起頻度によるパターンをエントロピーで比較しています。両者のエントロピーを *t* 検定で比較した結果，「右側主要部の規則 (right-hand head rule)」（影山, 1982, 1999: Williams, 1981）にしたがって，接尾辞のほうが接頭辞よりも多様な熟語を作ることを示しました。ただし，Miyaoka and Tamaoka（2005）の研究で使用した頻度データは，『茶漉』で検索したものではありません。

さらに，こうした接辞のコーパス検索は，大和・玉岡・初（2015）の人を表す接尾辞の日本語派生語の「勉強家」「写真家」の「家」，「会社員」「銀行員」の「員」，「経験者」「科学者」の「者」など，人を接尾辞とする熟語の検索にも応用できます。もちろん，検索結果からエントロピーを算出して，名詞との共起パターンを接尾辞ごとに指標化して比較することも可能です。

## ４．『現代日本語書き言葉均衡コーパス』の検索：『少納言』と『中納言』

国立国語研究所と文部科学省科学研究費特定領域研究「日本語コーパス」プロジェクトは，共同で『現代日本語書き言葉均衡コーパス』（BCCWJ: Balanced Corpus of Contemporary Written Japanese）を開発しました。「均衡コーパス」といわれるように，現代日本語の書き言葉の全体像を把握するために，書籍，雑誌，新聞，白書，広報紙，知恵袋，ネット掲示板，教科書，国会会議録，法律関係などのジャンルを広くカバーして 104,612,423 語のデー

タを集めています。出典ごとの検索もできるようになっています。こうした大規模コーパスの必要性や意義については，国立国語研究所言語資源研究系の教授である前川喜久雄先生の記事（前川, 2007; Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, and Den, 2014）を参照してください。また，書き言葉均衡コーパスのデータの集計方法などの詳細は，同研究系の准教授である山崎誠先生が書かれた解説があり，http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html（2013 年２月 26 日付け）からダウンロードできます。また，品詞構成表や語種構成表なども，このサイトからダウンロードできます。さらに，国立国語研究所の作成あるいは関係したデータベースの情報およびサイト一覧は，https://www.ninjal.ac.jp/database/を参照してください。日本語の研究に役に立ちそうなさまざまな情報が掲載されています。

国立国語研究は,『少納言』という検索サイト（http://www.kotonoha.gr.jp/shonagon/）を作っています。これは，登録が不要で，自由にアクセスできる検索サイトです。上記の書きことば均衡コーパスを使って検索ができます。このサイトの最後に「利用条件を読んで少納言を使う」があります。それをクリックすると，利用条件が提示されます。同意すればOK を押して，検索サイトに入ります。『少納言』は，ウェブ上で動く事例検索サイトです。「可愛げがある」という表現を「書籍」で検索してみます。メディア/ジャンルとして,『現代日本語書き言葉均衡コーパス』の中から「書籍」だけを選びます。「書籍」は 1971 年から 2005 年までの作品が集められており，総語数は約 6,270 万語です。「可愛げがある」を入力して検索すると，１件見つかりました。検索結果は，左から横長に以下の順番に表示されます。ここでは，出力例を縦に記します。

[表示番号] 1
[前文脈] 立派な環境破壊だ。それに比べれば「庚申の夜明かし」は日時が決まっているので、まだ
[検索文字列] 可愛げがある
[後文脈] 。 「まあ、気心の知れた者たち同士で集まって夜通し騒ぐというのは、なかなかに楽し
[執筆者] 藤原眞莉(著)
[生年代] 1970
[性別] 女
[メディア／ジャンル] 書籍/9 文学
[タイトル] 姫神さまに願いを

玉岡賀津雄 (2016)『レキシコンフォーラム』7, 239-264.

[副題] 鏡語りの森
[巻号] （該当なし）
[編著者等] 藤原眞莉|著
[出版者][注] 集英社
[出版年] 2000

*注*: 上記の例では，集英社が[出版者]となっています。コーパスは，出版社だけでなく官公庁や個人なども含まれており，より広い概念として[出版者]という用語が使われています。そういう理由で，検索結果の列名は，[出版社]ではなく，[出版者]となっています。

[後文脈]は，「可愛げがある」の後にくる表現です。検索結果では，文の途中で切れています。これは，表示文字数は，前文脈、後文脈ともに表示文字数は 40 字までだからです。作品および著者に関する情報が詳細に示されますので, 言語研究の事例検索には最適です。もちろん, [表示番号]は最後の番号がそのまま頻度になりますので, 検索表現のジャンルごとの出現頻度も知ることができます。

さらに，国立国語研究所は，詳細の検索ができる『中納言』という検索エンジンを作成しています。これも,『現代日本語書き言葉均衡コーパス』を使って検索することができるウェブ上のアプリケーションです。『中納言』を利用するためには，申し込みをしなくてはなりません。まずは，コーパス開発センター（http://pj.ninjal.ac.jp/corpus_center/）にアクセスしてください。このサイトには，国立国語研究所のコーパス関係のさまざまな情報が掲載されています。右上に「コーパス利用申込」というタブがありますので，そこをクリックしてください。『中納言』を使うには,『現代日本語書き言葉均衡コーパス』の申込書と利用許諾契約書のそれぞれをダウンロードし，申込書は１部，利用許諾契約書は２部を両面印刷にして，所定の住所に送ります。契約が成立すれば，無料で,『中納言』による『現代日本語書き言葉均衡コーパス』検索ができます。『現代日本語書き言葉均衡コーパス』を DVD で入手することも可能ですが，こちらは有料です。申込書も利用許諾契約書も記入例を PDF ファイルでダウンロードできますので，参考にすればすぐに書きあげられます。契約期間はオンラインが１年，DVD が２年ですが，それぞれ自動更新されます。『中納言』を使うと，さまざまな検索ができますので，ぜひご活用ください。

### ５．コーパス作成・検索・出力総合ソフト：『茶器』（ChaKi.NET）

『茶器』あるいは ChaKi.NET（Matsumoto, Asahara, Kawabe, Takahashi, Tono, Ohtani and

Morita, 2005; 松本・浅原・橋本・投野・大谷・森田, 2006）にも，お茶に関係した名前がついています。このソフトは，奈良先端科学技術大学院大学および株式会社総和技研が中心になって開発した最先端のツールです。『茶器』は，コーパスを独自に作成し，それ使って各種の検索をし，さらに作成したコーパスの修正もできるという総合的な機能を備えたソフトです。http://osdn.jp/projects/chaki/releases/から無料で自由にダウンロードできます。これは，上級者向けのソフトです。

『茶器』は，MeCab と CaBoCha の２つのソフトを利用して，平文に付加情報（アノテーション）を加えて，コーパスを自動的に作成する機能を備えています。つまり，自前のテキストを使って，形態素解析，文節区切り，文節係り受けを施したコーパスを作成することができます。たとえば，独自に集めた作文を，アノテーション付きのコーパスにして，検索することもできます。独自にコーパスを作成するには，最高のツールです。

『茶器』を使って，付加情報を平文に追加するためには，MeCab と CaBoCha をダウンロードしてインストールしておかなくてはなりません。MeCab は，このチュートリアルですでに説明したとおりで，ダウンロードサイトも紹介しました。CaboCha（「南瓜」）は，日本語係り受け解析ソフトで，コーパスに係り受け情報を付加することができます。詳細は，工藤・松本（2002, 2004）および松本・高岡・浅原・工藤 （2004）を参照して下さい。CaBoCha も無料で http://taku910.github.io/cabocha/からダウンロードできます。

『茶器』は，自前のコーパスあるいは購入したコーパス（新聞コーパスなど）の平文に形態素および係り受け情報が付加しますので，これらの情報に関係したさまざまな高度な検索ができます。また，MI スコアなどのコーパス情報も計算することができます。このように，ChaKi.NET は，コーパスのための高度な総合的ソフトです。興味のある方は，ぜひ『茶器』を使ってさまざまなコーパスを独自に作成して，検索してみてください。

## ６．李在鎬先生のホームページ（ウェブサイト）

最後に，これまで説明してきた検索エンジンを含んで，さまざまなコーパス関連の情報がほぼ洩れなく簡潔に紹介しているホームページを紹介します。

筑波大学人文社会系准教授の李在鎬先生は，日本語の研究に使えそうなコーパスの紹介，入手方法，分析ツールなどの各種情報を網羅したウェブサイト「コーパス日本語学の情報館」を作成しています。2013 年 5 月にリニューアルされ，http://jhlee.sakura.ne.jp/からアクセスできます。図７のように，ウェブサイトに入ると５つのタブがあります。

玉岡賀津雄（2016)『レキシコンフォーラム』7, 239-264.

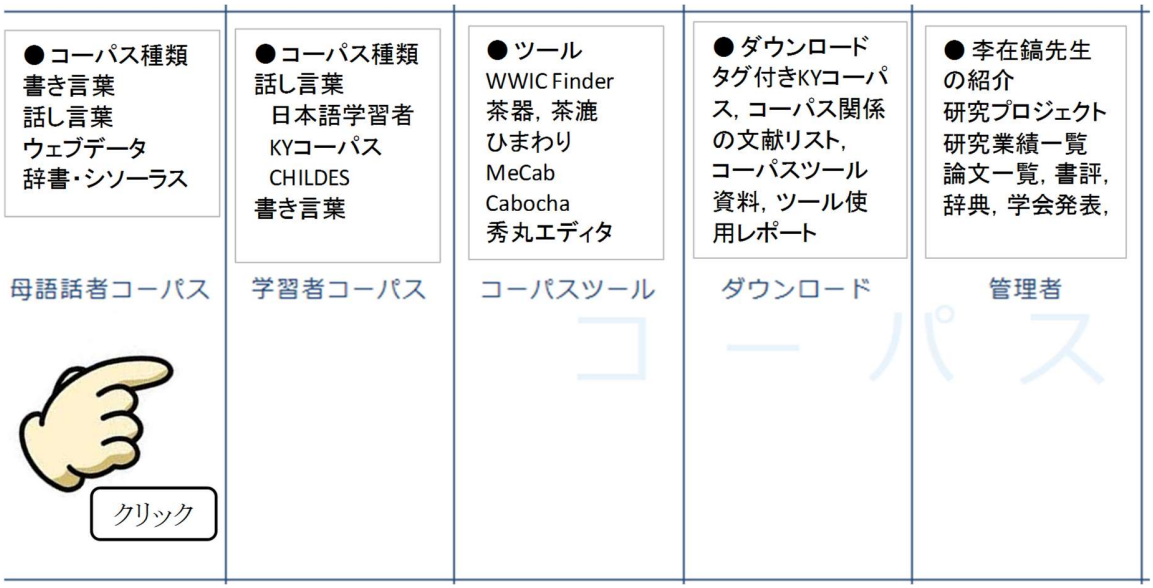


図７　李在鎬先生の「コーパス日本語学の情報館」ウェブサイト

もっとも左にある「母語話者コーパス」のタブをクリックします。次に見えるサイトには，「書きことばコーパス」「話しことばコーパス」「ウェブ（データの）コーパス」「辞書・シソーラス」の４つの分類があり，各種の情報が掲載されています。「書きことばコーパス」では，このチュートリアルでも紹介した『少納言』と『中納言』で検索できる現代日本語書き言葉均衡コーパスの情報もあります。さらに，青空文庫，新潮文庫の 100 冊，日英新聞記事対応付けデータ，日英対訳文対応付けデータ，京都大学テキストコーパス(Version 4.0), 高額ではありますが, 毎日, 朝日, 読売新聞記事のデータ版の新聞記事データベース，戦後の帝国議会や国会議事録を検索することができる国会会議録検索システム，裁判所の判例情報を検索することができる判例検索システム，形態素解析済みコーパスの必要な情報が分かりやすく紹介されています。

「日本語学習者コーパス」のタブをクリックすると日本語学習者関連のコーパスの情報があります。情報は，「話し言葉コーパス」と「書き言葉コーパス」に分かれています。「話し言葉コーパス」では，日本語学習者会話コーパス，第２言語としての日本語学習者の話し言葉コーパスである KY コーパス，日本人幼児の日本語発話コーパスも含まれた国際的な言語獲得コーパスである CHILDES（Child Language Exchange System）の情報，C-JAS（Corpus of Japanese as Second Language）が紹介されています。また，「書き言葉コーパス」では，アノテーション付きで，ウェブ上での検索エンジンも準備されている日本語学習者作文コーパス（JC コーパス），日本語学習者による 1,565 件の日本語作文と母語の対訳デー

タベース（作文対訳 DB），東京外国語大学の伊集院郁子先生が構築した日本語を母語とする大学生（134 名）と日本語を学ぶ大学生（台湾 57 名，韓国 55 名）によって書かれた日本語意見文を収録したデータベースが紹介されています。

「コーパスツール」では，ツールを５つに分類して，紹介しています。「データ抽出系ツール」としては，このチュートリアルで詳細に紹介した『茶漉』も紹介されています。その他，チュートリアルで紹介した『茶器』，KWIC Finder，KH Coder，国立国語研究所が開発し，言語研究用に設計された全文検索システムの『ひまわり』などの検索エンジンが紹介されています。「データ解析系」としては，やはりチュートリアルで紹介した MeCab と Cabocha も紹介されています。近代文語文を解析できる形態素解析辞書の近代文語 UniDic，形態素解析ソフトの Juman，係り受け解析ソフトの KNP なども紹介されています。「テキスト検索系ツール」としては，秀丸エディタが紹介されています。このソフトは，高機能テキストエディタであり，コーパス専用のソフトではありませんが，大規模ファイルの編集やコーパスデータの処理にも使える優れものです。「テキスト変換系ツール」としては，テキストファイルの文字コード（Shift-JIS, UTF-8, UTF-7, Unicode, EUC）を相互に変換できる文字コード変換ツール for .NET2.0 が紹介されています。このソフトは，複数のファイルをまとめて変換できるので便利です。「テキスト加工系（置換など）ツール」としては，Excel のデータを一括検索し，置換することができる EXCEL 検索置換が紹介されています。

「ダウンロード」では，コーパス分析に基づく日本語学・日本語教育の文献リスト Ver06 がダウンロードできます。ただし，2007 年 11 月 6 日までの文献なので，それ以降の論文は含まれていません。また，『日本語科学』2007 年の第 22 号，『月刊言語』2007 年 7 月号，『日本語教育』2006 年の 130 号，『日本語学』2003 年 4 月臨時増刊など，コーパス日本語学の特集号の目次が紹介されています。ただし，これらはあくまで紹介だけで，ダウンロードできるわけではなりません。コーパスツール資料としては，格フレーム辞書，KWIC Finder と正規表現，形態素解析によるコーパスツール，統計解析ソフト R があり，PDF ファイルでダウンロードできます。また，茶まめ，KWIC Finder，格フレーム辞書，KH Coder，本語語彙大系のコーパスツール使用レポートもダウンロードできます。このタブの紹介は，今後アップデートされることを期待しています。

最後に「管理者」のタブでは，李在鎬先生の科研などの研究プロジェクトが書かれています。さらに，書籍一覧には，李在鎬先生が書かれたあるいは編集された書籍が，表紙の写真付きで並べて紹介されています。コーパス研究について書かれた記事もたくさんありますので，参照して下さい。論文一覧の中には，PDF ファイルでダウンロードできるように設定された論文もあります。さらに，書評，辞典，学会発表，講演などの一覧が紹介さ

玉岡賀津雄（2016)『レキシコンフォーラム』7 , 239-264.

れています。

以上のように，このチュートリアルで紹介したコーパスや検索エンジンもすべてこのサイトで簡潔に紹介されています。また，クリックするだけで，各関連サイトへも飛ぶことができるようになっていますので，非常に便利です。参照して，皆さんの研究に活用してください。なお，リンクが切れているサイトを発見した方，あるいはこのウェブサイトに掲載されていない研究リソースの情報をお持ち方は，ホームページ管理者の李在鎬先生（jhlee.n@gmail.com）にご連絡下さい。すぐに対応してくださいます。

**謝辞**

本チュートリアルの執筆にあたり、パデュー大学の深田淳先生、筑波大学の李在鎬先生、神戸大学の石川慎一郎先生、国立国語研究所の山崎誠先生、名古屋大学の大学院生の于劭贇さんからコーパス検索や分析についての貴重なコメントを頂きましたことを、ここに感謝いたします。

**引用文献**

有本卓（1982)『確率・情報・エントロピー』東京：森北出版.

深田淳（2007）「日本語用例・コロケーション情報抽出システム『茶漉』」『日本語科学』22, 161-172.

堀淳一（1979)『エントロピーとは何か』東京：講談社ブルーバックス.

石川慎一郎（2012)『ベーシックコーパス言語学』東京: ひつじ書房.

石川慎一郎（2006)「言語コーパスからのコロケーション検出の手法―基礎的統計値について―」『統計数理研究所共同研究レポート』190, 1-14.

海保博之（1989)「第 1 講：情報をはかる―エントロピー・伝達情報量・冗長度」海保博之（編)『心理・教育データの解析法 10 講―応用編』14-26. 東京：福村出版.

影山太郎（1993)『文法と語形成』東京: ひつじ書房.

影山太郎（1999)『形態論と意味』東京: くろしお出版.

木山幸子・玉岡賀津雄（2011)「自他両用の「－化する」における自動詞用法と他動詞用法の比較: 新聞コーパスの用例に基づく多変量解析」『言語研究』139, 29-56.

工藤拓・松本裕治（2002)「チャンキングの段階適用による係り受け解析」『情報処理学会論文誌』43(6), 1834-1842.

工藤拓・松本裕治（2004)「カーネル法を用いた言語解析における高速化手法」『情報処理学会論文誌』45(9), 2177-2185.

玉岡賀津雄（2016)『レキシコンフォーラム』７, 239-264.

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *The proceedings of 2004 conference on empirical methods in natural language processing*. http://chasen.org/~taku/index.html.en からダウンロード.

李在鎬・石川慎一郎・砂川有里子（2012)『日本語教育のためのコーパス調査入門』東京: くろしお出版.

前川喜久雄（2007）「コーパス日本語学の可能性—大規模均衡コーパスがもたらすもの—」『日本語科学』22, 13-28.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation, 48,* 345-371.

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurita Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). Chaki: An annotated corpora management and search system. *Proceedings of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

松本裕治，高岡一馬，浅原正幸，工藤拓（2004）「茶筌と南瓜による日本語解析--構文情報を用いた文の役割分類」『人工知能学会誌』19(3), 334-339.

松本裕治・浅原正幸・橋本喜代太・投野由紀夫・大谷朗・森田敏生（2006）「タグ付きコーパス管理/検索ツール『茶器』」『言語処理学会第 12 回年次大会論文集』460-463.

Miyaoka, Yayoi, and Katsuo Tamaoka (2005). An investigation of the right-hand head rule applied to Japanese affixes, *Glottometrics, 10,* 45-54.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379-423 (Part I) and 623-656 (Part II).

新村出（編）（1998)『広辞苑』（第５版）.

杉村泰（2013）「コーパスを利用した複合動詞「V1-抜く」の意味分析」『名古屋大学言語文化論集』(名古屋大学大学院国際言語文化研究科), 35(1), 49-63.

杉村泰（2014）「コーパスを利用した複合動詞「V1-抜く」と「V1-抜ける」の意味分析」『名古屋大学言語文化論集』（名古屋大学大学院国際言語文化研究科）, 35(2), 55-68.

玉岡賀津雄（2011）「コーパス分析の研究例２: 複合動詞の計量的解析」 中本敬子・李在鎬・黒田航 (編) 『認知言語学研究の方法』(pp. 181-195)，東京: ひつじ書房.

Tamaoka, Katsuo, Hyunjung Lim, and Hiromu Sakai (2004). Entropy and redundancy of Japanese lexical and syntactic compound verbs. *Journal of Quantitative Linguistics, 11(3),* 233-250.

玉岡賀津雄・木山幸子・宮岡弥生（2011）「新聞と小説のコーパスにおけるオノマトペと動詞の共起パターン」『言語研究』139, 57-84.

玉岡賀津雄（2016)『レキシコンフォーラム』７, 239-264.

大和祐子・玉岡賀津雄・初相娟（2015）「中国人日本語学習者の人を表す接尾辞の派生語習得に影響する要因」『小出記念日本語教育研究会』23, 33-45.

Williams, Edwin (1981) On the Notions ‘Lexically Related’ and ‘Head of a Word’. *Linguistic Inquiry, 12,* 245-274.

――――――― [玉岡　賀津雄（たまおか　かつお）名古屋大学・教授]