# 第3章　汎用コーパス概観

## 3.1　はじめに

コンピュータ技術の発達に伴って，これまで数多くのコーパスが編纂されてきました。その数は近年急速に増加し，コーパスサイズもますます巨大化しています。Bookmarks for Corpus-based Linguists (http://tiny.cc/corpora) では，現在どのようなコーパスが提供されているかの情報一覧を見ることができますが，その数や種類の多さには驚かされます。著作権やその他の理由のため，研究者個人あるいは特定の研究機関でのみ使用が許されるコーパスもありますが，誰でも無料または有料で利用できるように広く一般のユーザーに公開されているコーパスもあります。この章では様々な研究目的に広く対応できるように編纂された「汎用コーパス」について概説します。それぞれのコーパスの特徴を知ることで目的に応じて使い分けることできます。

## 3.2　Brown 系コーパス

Brown コーパスは，1961 年当時のアメリカ英語の書き言葉の実態を反映したコーパスを作ることを目指して編纂された，世界初の電子化テクスト集です。様々なテクストジャンルからサンプルが収集され，総計 100 万語のコーパスが作成されました。これがいわゆる「第 1 世代コーパス」と呼ばれるコーパスの原型になり，その後このデザインに倣って，1961 年のイギリス英語版のコーパス Lancaster-Oslo/Bergen Corpus (LOB)，Brown/LOB の 30 年後の出版資料を用いた「クローン」とも呼ばれる Freiburg Brown (Frown) と Freiburg LOB (FLOB)，2005-2007 年のア



メリカ英語，イギリス英語を収集した AmE06，BE06 など様々なコーパスが作られてきました。これにより，同じ年代のアメリカ英語とイギリス英語の比較や，年代間の言語変化の調査が可能となりました。

Sketch Engine (3.6 参照) では Brown コーパスが無料で使えます。図 1 は Word Sketch という機能を使って，文法関係単位で動詞 make がどのような語と結びつきやすいか，すなわちコロケーションを表示したものです。例えば［object］という欄には実際に make の目的語として使われる語が頻度の高い順に並べられています。実用的な用例を選ぶのに役に立ちます。

［サイズ］各 100 万語

［英語の種類］アメリカ英語：Brown, Frown, AmE06／イギリス英語：LOB, FLOB, BE06

［利用形態］

The ICAME Corpus Collection (CD-ROM) →料金：NOK3,500, http://icame.uib.no/newcd.htm から注文。※ AmE06 と BE06 は含まない。1,700 万語を超える 20 のコーパスを収録。

Sketch Engine (https://www.sketchengine.co.uk/) →料金：£52.00 (年間)，※様々な言語の 100 以上のコーパスを提供。Brown の利用は無料。

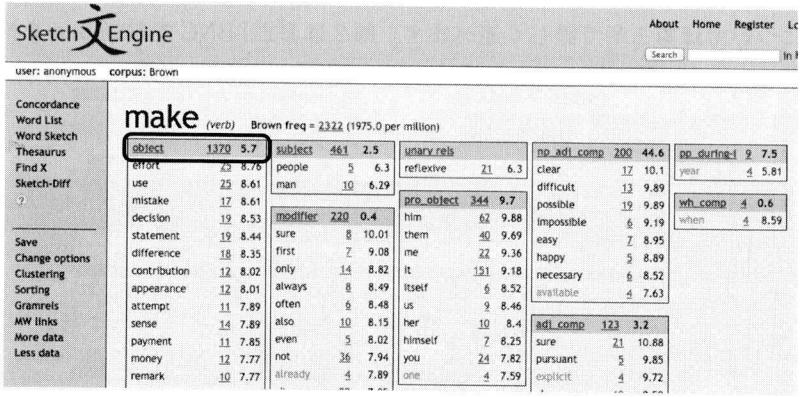


図 1　Brown コーパスにおける make のコロケーション

### 3.3 British National Corpus

1980年代以降，コンピュータ技術の飛躍的な発達により，Brown系コーパスより大きな，いわゆる「第2世代コーパス」と呼ばれる大規模コーパスが編纂されるようになりました。その中で今日もっとも多くのサイトで無料提供されているのがBritish National Corpus (BNC) です。BNCは，1億語を有するコーパスで，1990年代のイギリス英語の実態を反映する均衡コーパスを目指して作成されました。第1世代のコーパスと比べると，規模は100倍になりました。

BNCには10%の話し言葉 (1,000万語) が含まれています。その半分は，講義，説教，ラジオ番組などの公共のイベントからの録音，残りの半分は個人の日々の会話を録音したものです。残りの90%を占める書き言葉の部分は，25%のフィクションと75%のノンフィクションから構成されています。コーパスサイズが大きいことや，構成や出典情報に十分な配慮がされていることから，BNCは現在多くの言語学者に使用されています。

CORPUS.BYU.EDUではBNCを語，句，品詞レベルで検索することが可能で，ジャンル指定もできます。また，文法構造やコロケーションの検索，ジャンル間の比較も可能です。このサイトで提供されているその他のコーパス（後述のCOCAやTIMEなど）の検索結果と比較できるというのも魅力です。CORPUS.BYU.EDUの操作方法は第4章で，利用方法については第5章で詳しく述べます。図2はBYU-BNCで動詞banを検索し，ジャンル別の頻度を示したものです。圧倒的多数が新聞で使用されているのが分かります。

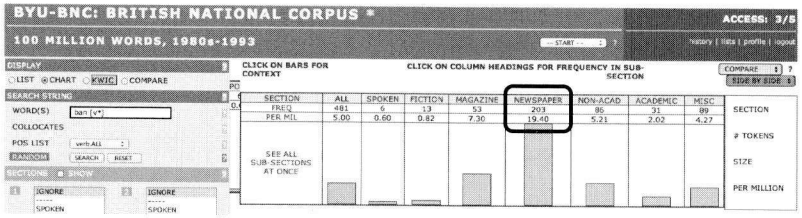


図2　BNCにおけるban（動詞）のジャンル別分布



BYU-BNC以外にBNCはBNC*web*やIntelliTextやPhrases in English (PIE) のサイトでも使用できます。それぞれ使い勝手が違うので用途に合わせて使い分けるとよいでしょう。BNC*web*は直感的に使用しやすいインターフェースで，非常に多機能です。共起関係を表すさまざまな統計値が用意されているので，特にコロケーションを調べる際に非常に有効です。BNC*web*の操作方法は第6章で詳しく述べます。

IntelliTextはLeeds大学のCentre for Translation Studiesによって開発されたコーパス検索用のインターフェースで，BNCの他にも様々な言語のコーパスを検索することができます。検索結果 (View Results) の表示機能では，通常のソート機能に加え，Kellyタブを開くとCEFRレベル（欧州評議会が提唱するヨーロッパ言語共通参照枠に基づく5技能6段階の習熟度レベル）別に単語を色分けして表示することができます (CEFRについては17.2.1参照)。図3はenvironmentを検索した結果に中級以上の語に色づけをして表示しています。テキストの難易度の判断に使ったり，レベルに合った語彙を選び，空欄補充問題などを作るのに役立ちます。

PIEではn-gramを検索することができます。「n-gram」とは，あるテクストから切り出した連続したn個の語の並びのことで，その連続した並びの頻度情報などから，意味のまとまりをもったフレーズを抽出してく

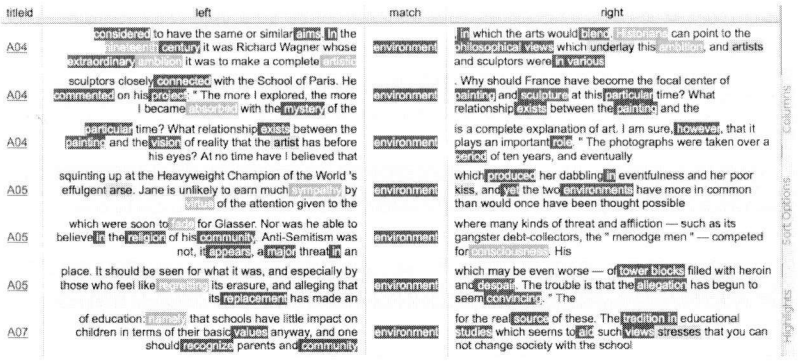


図3　語彙レベル別検索結果

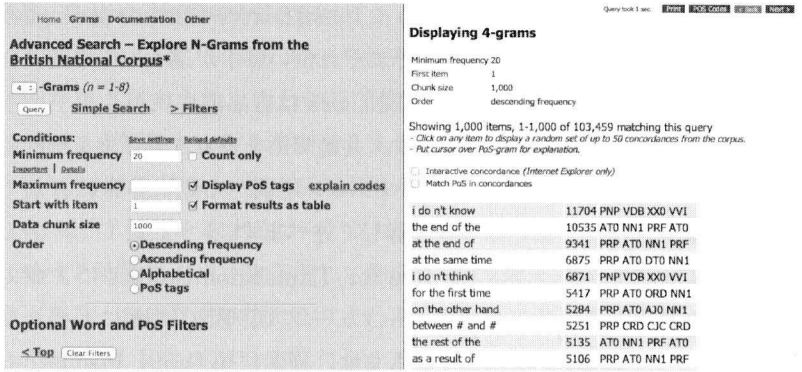


図4　BNC における 4-gram

れるので，実際にどのようなフレーズがよく使われ，重点を置いて教えるに値するのかを判断するのに有用です（n-gram について詳しくは 10.3.2 を参照）。図 4 は BNC の中の 4 語の並びを頻度順に並べた結果です。

［サイズ］1 億語

［英語の種類］イギリス英語

［利用形態］以下はすべて無料

BYU-BNC（http://corpus.byu.edu/bnc/）

BNC*web*（http://bncweb.lancs.ac.uk/bncwebSignup/user/register.php）

Intellitext（http://smlc09.leeds.ac.uk/itweb/htdocs/Query.html）

Phrases in English（http://phrasesinenglish.org/index.html）

## 3.4　WordbanksOnline

*Collins COBUILD English Dictionary* の編纂に使用するために 1980 年に始まったコーパスから発展してできたのが Bank of English（BoE）で，BNC と並んで第 2 世代コーパスの代表的な存在です。継続的にデータが追加される，いわゆる「モニターコーパス」として発展してきました。現在では 6 億 5,000 万語を有していると公表されていますが，だれでも利用できるわけではありません。BoE のうちから一部を抽出し再構成した



コーパスが WordbanksOnline の名称で一般の人にも利用可能です。WordbanksOnline では 1985 年から 1998 年までに収録もしくは公表されたイギリス，アメリカ，オーストラリアの書籍，新聞，雑誌，話し言葉，放送原稿など 12 のサブコーパスに分類された約 5,400 万語が検索対象となっています。このうち話し言葉は，放送原稿と自由会話の総計で約 1,400 万語（約 25%）となっています。

図 5 は WordbanksOnline で different の前後に来る語を共起関係の強いものから並べたリストです。上位には different が修飾する名詞 ways, things, types, something, kinds と共に検索語である different そのものが挙げられています。different のすぐ後に different がもう 1 度使われるのはどのような場合だろうかと興味を持ったなら，different をクリックして用例を確認できます。

［サイズ］5,400 万語

［英語の種類］イギリス英語，アメリカ英語，オーストラリア英語

［利用形態］

WordbanksOnline（http://www.collinslanguage.com/content-solutions/wordbanks）→料金：10 ユーザーまで £695（年間）

小学館コーパスネットワーク（http://scn.jkn21.com/wordbanks.html）→料金：1ID 21,000 円（年間）

**Collocation candidates**

| | Freq | T-score | MI | MI3 | log likelihood | min. sensitivity | salience |
|---|---|---|---|---|---|---|---|
| p/n **different** | 9424 | 96.658 | 7.854 | 34.258 | 85035.378 | 0.062 | 0.000 |
| p/n **ways** | 3847 | 61.780 | 7.987 | 31.806 | 35299.107 | 0.025 | 0.000 |
| p/n **very** | 8734 | 92.208 | 6.226 | 32.411 | 58830.070 | 0.020 | 0.000 |
| p/n **things** | 3807 | 60.824 | 6.137 | 29.926 | 25047.316 | 0.019 | 0.000 |
| p/n **quite** | 2864 | 52.957 | 6.581 | 29.548 | 20587.140 | 0.019 | 0.000 |
| p/n **something** | 3507 | 58.227 | 5.898 | 29.450 | 21914.044 | 0.016 | 0.000 |
| p/n **types** | 2446 | 49.356 | 8.938 | 31.450 | 25807.430 | 0.016 | 0.000 |
| p/n **each** | 3057 | 54.286 | 5.783 | 28.939 | 18608.408 | 0.015 | 0.000 |
| p/n **completely** | 2104 | 45.660 | 7.772 | 29.850 | 18637.799 | 0.014 | 0.000 |
| p/n **many** | 4624 | 66.597 | 5.599 | 29.948 | 27026.670 | 0.013 | 0.000 |
| p/n **kinds** | 1903 | 43.525 | 8.795 | 30.583 | 19668.565 | 0.013 | 0.000 |
| p/n **kind** | 1886 | 42.722 | 5.942 | 27.704 | 11878.562 | 0.012 | 0.000 |

図5　WordbanksOnline による different のコロケーション表示

## 3.5 Corpus of Contemporary American English (COCA)

COCA は上述の CORPUS.BYU.EDU で 2008 年に公開されたコーパスです。使いやすいインターフェースを持つ優れたツールで，様々な英語の調査が可能です。1990 年以降 1 年ごとに 2,000 万語のセクションで構成され，現在も毎年 2,000 万語ずつ拡張し続け，2013 年 8 月現在 4 億 6,000 万語を有します。無料で提供されるコーパスの中で最大，かつアメリカ英語コーパスとしては唯一の大規模均衡コーパスと言えるでしょう。データはサイズの等しい 5 つのジャンル（話し言葉，フィクション，大衆雑誌，新聞，学術誌）に分けられています。

図 6 では global warming のジャンル別，年代別頻度を示しています。主に雑誌で用いられ，2005 年以降頻度が飛躍的に伸びているのが分かります。また，話し言葉はラジオやテレビ番組の台本のない会話を文字起こししたもので，BNC 中の人前で話したスピーチなどの部分とは十分比較可能です。英米語間での話し言葉の比較に使うことができます。COCA の操作方法と活用事例については第 4 章と第 5 章を参照して下さい。

CORPUS.BYU.EDU には，1 億語を超えるコーパスが複数提供されており，これらのコーパス間での比較が可能なので，言語調査に役立つツールとしてますます広い支持を得ているサイトと言えるでしょう。

［サイズ］4 億 6,000 万語

［英語の種類］アメリカ英語

［利用形態］COCA（http://corpus.byu.edu/coca/）→無料（登録については 4.2）

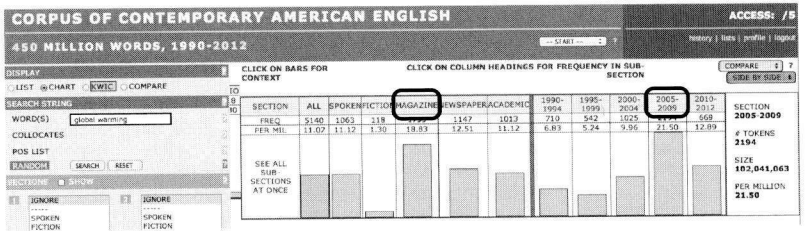


図 6　COCA における global warming の chart 表示



## 3.6 Sketch Engine

Sketch Engine はこれまで紹介してきた BNC や COCA などと違いコーパスそのものの名前ではありません。数多くのコーパスへのアクセスや多機能な分析ツール，コーパス作成支援などのサービスを提供する有料サイトです。変種を含め 50 以上の言語の，様々な目的に応じて作られた 100 以上のコーパスにアクセスすることができます。その中には前述の Brown 系コーパスや BNC も含まれます。ここで使用できるコーパスの多くは巨大で，最大の enTenTen12 は BNC の約 100 倍の 100 億語を突破しました。このような巨大コーパスの特徴は，すべてのテクストがコンピュータのウェブ自動巡回により特定のドメインから集められ，自動でタグ付けやレマ化がされていることです。このような膨大なデータから検索をして用例を抽出できるので，低頻度の語句の用例を見つけることも可能です。

Sketch Engine は非常に多機能で，Word Sketch 機能では，文法関係単位で検索語と他の語との結びつきの相性の良し悪し，すなわちコロケーションを表示することができます。それ以外にも，Thesaurus 機能では，対象語と同じように用いられる語を検出し，類義語の可能性のある語として提示してくれます。Sketch-Diff 機能では 2 語間の用法の違いが色付けして表示されます。図 7（次ページ）は日本語の「心」の訳語として使われる 2 つの英単語 heart と mind がどのように異なる使われ方をするのかを表しています。これら 2 語が主語のとき，どのような動詞が使われているのか（subject_of の欄），どのような形容詞が補語になっているのか（adj_subject_of の欄）を表しています。一方の語で特徴的に見られる共起語が，他方の語とは全く結合していないことが見て取れ，両語は異なる環境で使われることが分かります。

［サイズ］ukWaC は 15 億語，enTenTen12 は 100 億語など。多数のコーパスを提供。

［英語の種類］多種。ウェブ上で収集されたデータが多い。

［利用形態］Sketch Engine（https://www.sketchengine.co.uk/）→料金：

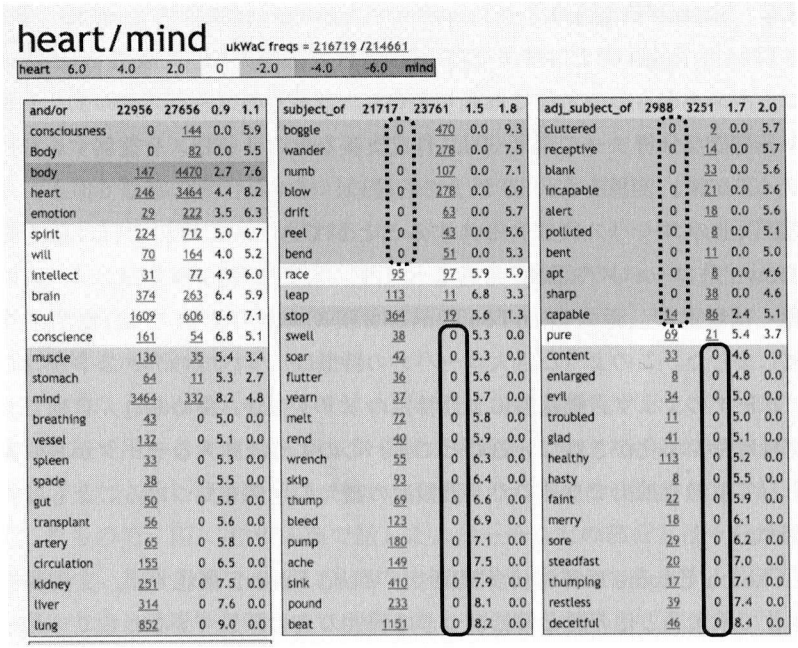


図 7　Sketch-Diff で heart と mind の類似点・相違点を表示

£52.00（年間）。ただし Open Corpora として公開されている British Academic Spoken English Corpus（BASE），British Academic Written English Corpus（BAWE），Brown は無料（3.2 参照）。

## 3.7　新聞・雑誌・その他のアーカイブ

英字新聞・雑誌をオンラインで無料で検索・閲覧できるアーカイブもいくつか存在します。語法の確認や適当な用例を探すのに役立ちます。

*TIME* のサイト（http://content.time.com/time/archive/）では，*TIME* の創刊以来の全記事が検索可能になっています（閲覧は一部無料）。また，上述の CORPUS.BYU.EDU では 1923 年から 2006 年までの Time 誌アーカイブに無料でアクセスすることができます（http://corpus.byu.edu/time/）。



*New York Times*（http://www.nytimes.com/）のサイトでは，創刊以来の全記事（1,300 万記事以上）を検索することができます（1923 年以前と 1986 年以降の記事は毎月 10 件まで無料）。

その他，印刷された書籍などを電子化したテクストアーカイブとして，The Oxford Text Archive（OTA: http://ota.ahds.ac.uk），Project Gutenberg（http://www.gutenberg.org）があります。これらのテクストをダウンロードし，ファイルに整形を施せばコーパスを自作することができます（第 11 章参照）。また自作コーパスを検索するにはコンコーダンサーが必要です。コンコーダンサーの 1 つである AntConc については第 12 章で詳しく紹介します。

## 3.8　コーパスとしてのウェブサイトと WebCorp Live

近年注目されている言語調査の資料として，ウェブサイトがあります。Google や Yahoo といった検索エンジンを使用して，インターネット上にある情報を丸ごと言語資料として利用するのです。人の手によって編纂されたどのコーパスよりも膨大なデータを持つウェブサイトは情報の宝庫だと言えると同時に，常に変化し続ける不確定なデータベースだとみることもできます。そのため，使用する際には注意しなければならないことがいくつかあります。

第 1 の問題は，ウェブサイトは常に変わり続けているということです。ある検索エンジンを使って今日検索した結果が明日も同様に得られるとは限らないのです。もう 1 度検索すれば同じ結果にたどり着くと過信せず，必要な情報・データは記録・保存しておくことが重要です。

また，データそのものの信頼性の問題もあります。インターネット上では書き手の情報が常に明らかとは限らないので，英語で書かれたデータをすべて英語の母語話者が書いた用例として扱うことはできません。この問題は検索するサイトの範囲を制限することで，ある程度解決できます。例えば，Google の「検索オプション」を使って検索結果をアメリカ合衆国やイギリスなどに絞り込んだり，ドメインを“ac.uk”や“edu”などの

教育機関用ドメインに限定したりすることで，ある程度のデータの制限ができます（14.2.3 参照）。例えば，日本人英語学習者がよく書く英文“My hometown has a lot of nature.”の妥当性を調べるとします。実際に has a lot of nature という表現を検索エンジンで調べると，ドメインの制限なしで検索すれば，約 11 万 8 千件が該当しますから，「使われる」という判断に至るかもしれません。ところが，“edu”にドメインを絞って検索すると，108 件しかヒットしません。しかも検索結果は，日本人留学生のコメントだったり，英文法の授業の資料で悪い例として取り上げられている場合だったりします。このことから，has a lot of nature は適切な表現ではないと判断することができるでしょう。

Google や Yahoo などの検索エンジンは，本来言語調査用に作られたものではなく商業目的で利用されることを前提としているので，検索表示件数に制約があったり，表示順などが操作されたりしていて，純粋な語句の検索結果として扱うには注意が必要です。以上の制約や問題点を理解した上で使用すれば，商用の検索エンジンもコーパスでの調査の結果を補足する情報が必要な場合や，非常に低い頻度の言語表現を扱う場合などの用例調査に有効に利用できます。なお Google の検索技法と留意点について詳しくは第 14 章を参照して下さい。

ウェブサイトをコーパスとして利用する際に不便な点としては，KWIC 表示などのコーパスツール機能がないため，検索結果を見渡しにくいことが挙げられます。しかし，例えば WebCorp Live (http://www.webcorp.org.uk/live/) で提供されている機能を用いれば，KWIC 表示やコロケーションの表示ができるようになります。WebCorp Live の検索画面では，使用する検索エンジン，言語，使用するサイトなどを選択し，検索語の後処理オプションでは，ソート方法やコロケーション表示の有無，使用するウェブページの期間などを指定できます。その結果，AntConc などのコンコーダンサーで検索したような KWIC 表示やコロケーション情報を得ることができます。WebCorp Live による検索は，まだ辞書にも登録されていないような，新語の使われ方を調べるのに有効です。図 8，9 では



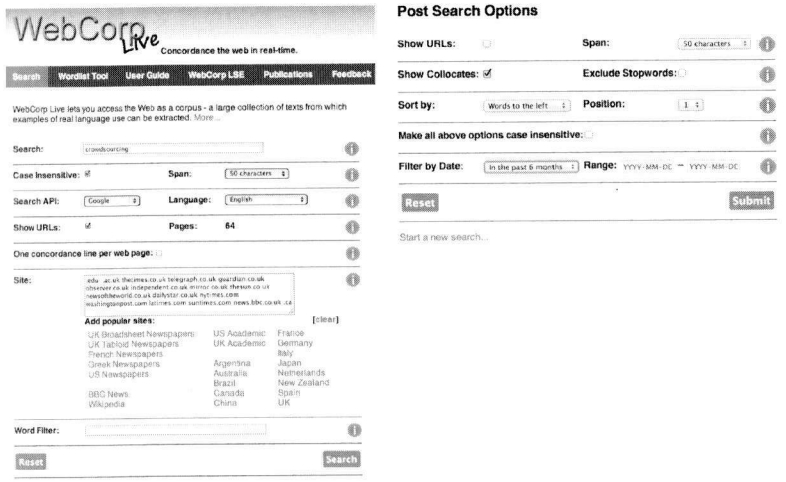


図 8　WebCorp Live の検索画面と後処理オプション

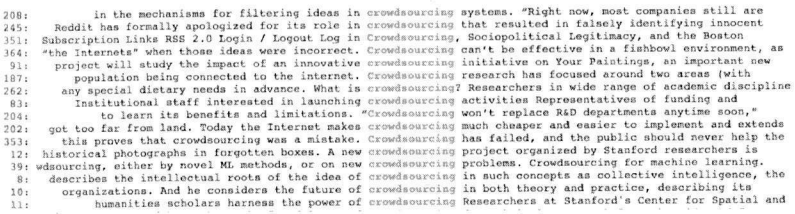


図 9　WebCorpe Live での crowdsourcing の検索結果

crowdsourcing という比較的新しい語を検索し，結果を KWIC 表示しています（いくつかの後処理を行っています）。これにより，この語がどのように使用されているか，この非常に新しいビジネス形態が新聞などでどのように取り上げられているのか，知ることができ，時事的な英語の解説に利用できます。

## 3.9　おわりに

一般に公開されている汎用コーパスを使用例と共に紹介してきました。汎用コーパスのサイズはますます大きくなる傾向にありますが，大きいも

のが常によいとは限りません。高頻度の語や一般的な構文を調べる場合にはそれほど大きなコーパスは必要ありませんが，頻度はそれほど高くないが，使用パターンやコロケーションを知りたい語や周辺的な構文を調べるときには大規模コーパスが必要です。調べたいことに対して正しいサイズのコーパスを使用することがコーパス検索成功の鍵となるでしょう。

# 第4章　汎用コーパス COCA の使用法

## 4.1　はじめに

この章では Corpus of Contemporary American English (COCA) の操作方法とその活用法を説明します。COCA は Brigham Young 大学の Mark Davies によって提供されているサイト，CORPUS.BYU.EDU で公開されている7つのコーパスの1つです。操作方法は共通ですので，COCA の検索方法を学べば他の6つのコーパスも同じように扱うことができます。詳細な説明や検索例は，初期画面の［Help/information/contact］のメニューから見ることができます。特に［GENERAL］の Brief tour (general) は簡潔にまとめられているので，COCA を使って何ができるのかを具体例を見ながら手早く学ぶことができます。

表1　CORPUS.BYU.EDU で利用可能な英語コーパス

| コーパス名 | 語数 | 国・地域 | 時期 |
|---|---|---|---|
| Global Web-Based English (GloWbE) | 1.9 billion | 20 countries | 2012–13 |
| Corpus of Contemporary American English (COCA) | 450 million | American | 1990–2012 |
| Corpus of Historical American English (COHA) | 400 million | American | 1810–2009 |
| TIME Magazine Corpus | 100 million | American | 1923–2006 |
| Corpus of American Soap Operas | 100 million | American | 2001–2012 |
| British National Corpus (BYU-BNC) | 100 million | British | 1980s–1993 |
| Strathy Corpus (Canada) | 50 million | Canadian | 1970s–2000s |