Webコーパス概観

投野由紀夫

1. はじめに

Webコーパスの加速度的な発展は、インターネットの爆発的進化と無縁ではない。Kilgarriff & Grefenstette (2003) によれば、1999年7月に登録されていた世界のネットワーク総数（ホスト・コンピューターの数）は約560万であったが、それが2001年1月には1億2500万、2003年1月には1億7200万に激増し、同年のGoogleの試算では地球上のアクセス可能な静的Webページの総量は約20テラバイト、語彙数にして約2兆語（英数字換算）と言われた（ibid: 337)。

その後、Googleのブログでは2008年時に1兆の独立したURLが存在する、といわれ、Yahooが2009年には英語だけで196億ページが登録されていると発表している。英語に関してはXu (2000) が検索エンジンExciteから試算して英語のWebページは全Webの約70％であるとしていたが、最近の南出 (2008) の研究によると、2004年1月～2006年7月までの30ヶ月で144億ページを収集し、そのトップレベルドメイン（TDL）の言語解析をしたところ、約42％が英語であった、と報告している。収集方法に日本で行った調査という影響が若干出てはいるものの、Webページは英語中心から多言語化へ急速に移行している模様である。

現在の英語Webページに出現する英単語の総数に関する推定はさまざまな方法があるが、ここではKilgarriff & Grefenstette (2003) の推定方法を用いてみる。表1がその結果である。Medical treatment, deep breathという2つの句に関してBNC, 2003年のAltaVistaでのヒット数をもとにテキスト・サイズの推計をしている。これを2010年のAltaVistaのヒット数を求めて推計してみると、現在のWebのテキスト量は2兆6500億語となる。2003年時点でのGoogleが推定した世界中の全Webテキストの総量が2兆語であったから、英語だけで2兆6500億語は十分可能性のある数字であるといえよう。

『英語コーパス研究』第18号（2011), pp.83-95

表 1：英語 Web のテキスト量

調査対象の語句	BNC	2003 (AltaVista)	2010 (AltaVista)[1]
medical treatment	414	1,539,367	32,600,000
deep breath	732	868,631	19,400,000
テキスト総量	100,000,000	118,665,000,000	2,650,273,000,000
BNC による推定量		1186 億語	2 兆 6500 億語

2. Web コーパス：分野の進展

次に Web コーパスのコンピューター言語学の分野における歴史的変遷を、前述の Kilgarriff & Grefenstette (2003) を基に解説する。まずコーパス (corpus) がコンピューター言語学の世界に紹介されたのは 1989 年、Association for Computational Linguistics (ACL) で最初にコーパスが公に議論される。当時はコーパスの標本抽出などの概念は、かえって雑多なテキストが混在している印象を与え、また規則ベース（rule-based）のコンピューター言語学の世界は未だに Chomsky らの生成文法の知見から生まれた文法規則（文脈自由文法など）を機械に学習させる、という発想が強かったために、自然言語の実際に使用されたデータはノイズが多すぎる（□too dirty□）と考えられていた。コーパスが本格的に自然言語処理の分野で論じられるようになったのは、1993 年 Ken Church らが ACL で Using large corpora と題する special issue を刊行した時であった。Web がコンピューター言語学に登場するのはもう少し後のことで、1999 年 Mihalcea & Moldovan (1999) が Web の検索エンジンに工夫を加え、ヒット数をもとに単語の曖昧性解消（word sense disambiguation）を行った。また Resnik (1999) は当時 Canadian Hansard に限定されていたパラレル・コーパスが Web 上で大量に入手できる可能性を示した。そして、2003 年 Web as Corpus と題する特集が *Computational Linguistics* に掲載され、その後は Web をソースとする研究が一般的になったのである。

Web as Corpus のコーパス言語学分野での牽引力となったのは WaCky (Web-as-Corpus kool ynitiative) という研究グループであった。[2] Adam Kilgarriff, Marco Baroni, Sylvia Bernardini, Stefan Evert, Sebastian Hoffman, Serge Sharoff らが中心で 2005 年から workshop が開催された。彼らは後述する一般的な Web コーパス作成工程に関するさまざまな技術を研究・公開し、個人レベルでも Web コーパス作成が可能なツール群を用意した。Web as Corpus のワークショップは計 5 回（2005-2009）開催され、それぞれオンラインの予稿集も刊行されている。これらの活動と並行して、コーパス言語学の分野においてもWeb利用は急速に進み、サーチエンジンの結果をそのままコンコー

ダンサーで利用するWebCorp[3] (by A.Renouf), 独自のクローリング（インターネットを自動巡回してWebテキストを集める作業）による大規模コーパスを無料で提供するWebAsCorpus.org[4] (by Bill Fletcher) のようなサイトが現れ、商用ではAdam KilgarriffのSketch Engine[5]が20億語規模の英語コーパスはじめ数十の多言語コーパスをサービスし、かつユーザーがクローリングの技術を使って自作Webコーパスの自動構築ができるシステム（WebBootCAT）を提供している。また多言語環境で消滅危機言語などのWebコーパス収集をしているCrúbadán[6] (by Kevin Scannell) などもユニークな試みである。またWordSmith（by Mike Scott）のようなスタンドアローンのコンコーダンサーにもWebクローリングの機能が実装されるなど、Webコーパス作成そのものが比較的身近なものとなってきている。

3. Googleology is Bad Science

Webをコーパスとして使うという発想を耳にすると、「Googleで事足りるのではないか」という意見が多く聞かれる。Webコーパス研究に携わる人々もこれにはさまざまな理論的経験的反論をしている。そのいくつかのポイントをKilgarriff (2007) にしたがって、まとめておく：

(1) Search engineの結果では十分な用例が収集できない。
結果表示は数百万件とあっても、実際にそれだけの用例をダウンロード出来るわけではない。
(2) 各例文について十分なコンテクストを得られない。
Googleではせいぜい前後10語程度
(3) 抽出条件に歪みがある。
タイトルや見出しによく出てくるものが通例リストのトップに来たり、ランクを上げるために人工的なキーワード埋込みのサイトなどがあり、信頼性に疑問がある。
(4) 言語分析に必要な言語注釈付け（基本形検索、品詞検索など）が利用できない。
(5)「…を含むページ」という表示が示すように、サーチのヒット数は単語頻度ではなくページ頻度である。
(6) APIによる自動実行で検索できる回数に上限（1000件まで）があり、Webコーパスの構築用には不適。

以上のような問題点から、Webコーパスの研究者たちはGoogleそのものをコーパス検索用に用いる危険性を警告している。

4. Web コーパスの一般的な作成工程

詳しい自然言語処理技術は今号の田中氏の論考に譲るとして、ここでは一般的な Web コーパス作成工程に関して概説する。Web コーパスは一般に以下のようなプロセスで生成される：

(1) Seeds／keywords の選択
URL を指定してクローリングを行ってもよいが、たいていの場合、自分が収集したいテキスト群に出現して欲しい単語を「種語（seed word)」としてリスト化する。それを数語ずつ組み合わせて「n-語の組、n-タプル（n-tuple)」を作成する。

(2) URL の自動取得
N-タプルを検索エンジンにフィードして、自動で URL を取得する。数年前までは 1 日 1000 件という制限付きで Google API が利用できたが、現在では非公開なので、Yahoo! や Bing などを利用する。

(3) URL からのページ取得
これには自動巡回ソフト（crawler とか spider と称される）を用いる。WaCky の研究グループでは wget, heritrix などを用いていた。

(4) クリーニング
通常の自動巡回ソフトでは収集結果に大量のゴミが混じる。そのために HTML フォーマット以外のページを除外したり、ファイル（テキスト）・サイズの制限をかけて、クローリングの際よりテキストとしてふさわしいページだけに絞り込む工夫を行う。かつ、取得したページのリンクや広告などのいわゆるネットの決まり文句部分（boilerplate）を除去するスクリプトを走らせたり、事前に bad word list という含めたくない文言を用意しておいて猥褻文書などを排除したりする。最後に Web ページに頻繁に見られる重複部分（duplicate）の除去も大事なクリーニングのプロセスである。

(5) 形態素解析・品詞タグ付与・インデックス化
最後に言語データとしての形態素解析や品詞タグ付与などの言語注釈付けを行う。さらに大量文書を高速に検索できるようにインデックス化する。これには IMS-CWB, Manatee, Lucene, Sphinx などの全文検索エンジンを利用する。

5. 主要な英語 Web コーパスと関連する研究書

ここでは現在利用可能な英語 Web コーパスの代表的なものを紹介する。

(1) UKWaC
WaCkyのグループが英語、イタリア語、ドイツ語の3つのWebコーパスを作成公開しているが、そのうちの1つ。現在は、WaCkyのサイトからダウンロードできるほか、Adam Kilgarriff作成のSketch Engineでも利用可能。イギリス英語のサイトに特化して約15億語が収集されている。

(2) Web Corpus 2007
Bill Fletcherのサイト（http://webascorpus.org/）で利用でき、2007年度公開のWebページから約5億1800万語の英語データを収集したもの。彼はWaCkyのワークショップの発表の一環で複数のWebコーパスを構築、オンラインで公開している。

(3) New Model Corpus
Adam KilgarriffのSketch Engineのサイトで利用出来る1億語のWebコーパス。ユニークな点は、British National Corpus (BNC)のコーパス・デザインをできるだけ踏襲してクローリングが実施されており、BNCが90年代初めまでのデータということで古さは否めないため、Webコーパスでどれだけ BNCのような均衡コーパスのイメージを再現できるか、を目標に作成された。Kilgarriffらによればこれに多様なアノテーション技術をほどこし、1億語規模のデータでWebコーパスとしてどのように精密な言語注釈付けが可能かを検証するテストベッドの役割を果たさせるということである。

(4) Oxford English Corpus
Oxford University Pressが所有する社内用のWebコーパス。2000年から2009年までの10年間に出版・使用された約20億語からなり、大部分がWeb素材から構築されている。特徴としてはこれだけ大規模なデータであるにもかかわらず均衡コーパスの概念を保っていること。米語、英語、オーストラリア英語、など約10の英語変種、話し言葉vs書き言葉などのモード、フォーマルvsインフォーマルなどのレジスター情報、20の専門分野区分などでサブコーパス検索が可能。

(5) SEKAI Corpus
小学館が社内研究開発用に構築した20億語規模のWebコーパス。経済、法律、政治、コンピューターの4分野および科学技術コーパスのPERCを加えて、それらとBNCの出現頻度を相互比較できるインタフェースを備えている。作成のための工夫として、ネット上に存在する専門用語辞書をseed wordとして利用し、それらをもとに極めて正確に該当する専門分野テキストを自動収集している。専門分野英語をこれだけの規模でBNCと相互比較できるようにしたコーパスはまだ世界でも珍しく、辞典編纂への応用が期待される。

Webコーパスを扱った専門書としては、Hundt et al. (2007), Fairon et al. (2007) などが参考になる。またWeb as Corpusのワークショップの第4回、第5回はオンラインで閲覧が可能。[7]

6. Webコーパスへの一般的な疑問

シンポジウムでは各講師の発表の前に以下のようなWebコーパスに関する一般的な質問リストを用意した。このうちのいくつかに応える形でシンポジウムの質疑応答が行われた。ここでは質問を挙げながら、討議内容を簡単に要約して紹介する。

6.1. コーパスの定義・サイズに関する問題

Webコーパスに関する疑問の1つは、従来のコーパス言語学の概念からすると、Webから作成したテキストの集合は果たしてコーパスといえるのか、という点である。コーパスを設計する際、対象となるテキストの標本抽出（sampling）と代表性（representativeness）はきわめて重要であり、その点で自動巡回で収集されたWebコーパスを疑問視する見方もあろう。しかし、Kilgarriff and Grefenstette (2003) ではWebからのテキスト集合体はコーパスである、と明言している。彼らによればMcEnery & Wilson (1996) のような厳密なコーパス定義は「コーパスとは何か？(What is a corpus?)」という質問と「よいコーパスとは何か？(What is a good corpus?)」という質問を混在させており、コーパスの定義は「複数のテキストの集合体」という広義の意味合いでとらえるべきだとしている。

Webコーパスが信頼できるかどうか？という疑念も実はこの定義の問題と密接に関係している。Webコーパスが信頼できるかどうかは、それで何を調べたいか、という研究目的によるのであり、それはBNCのような優れたデザインに基づいて構築されたと考えられている均衡コーパスにもまったく同様の質問をすることができる。そして同時にコーパス・サイズの問題に関しても「大きければいい、という議論は本当か？」という疑問がある。これは逆にいうと小さいコーパスの利点は何かということになろう。Biber (1993) の研究のように2000語単位のテキストでも高頻度品詞の研究などでは安定した統計が得られる。逆に辞書の用例を得たい場合にどのくらいのコーパス規模が適切かを考えてみよう。筆者が今回のシンポジウムのためにBootCatを用いて作成した4つの2000万語クラスのWebコーパスのデータを表2に示す：

表2：Webコーパスの規模と用例数

	Corpus A	Corpus B	Corpus C	Corpus D
Tokens	22,889,948	23,574,444	24,588,570	23,918,232
Types	321,718	338,197	327,724	331,523
4コーパス全てにある単語	105,264 (97.97%)			
各コーパスにしかない単語	15,376	13,542	13,170	13,127
500件以上用例がある単語	4,238	4,736	4,880	4,394

COBUILDやMacmillanの辞書編纂者が辞書記述に参照するコンコーダンスの用例数が平均500件と言われている。表2で500件以上用例がある単語を見てもらえばわかるとおり、2000万語クラスのコーパスでは4000語余りの単語群しかこの基準を満たさない。逆算して1万語の学習語彙の記述をコーパスを用いてある程度満足いくようにすると仮定しよう。このデータでは1万語レベルの単語の平均用例は187例なので、500例を集めるには約7000万語のコーパス規模が必要という計算になる。では平均的学習英語辞典がカバーする約5万語の記述になると、実に表4のコーパスでは平均12例しか用例が得られない。500例を見るためには、約10億語のコーパス規模が必要となる。「サイズは重要」ということはこの結果から見ても明らかだが、一方で巨大コーパスを作成する手間や方法とのトレード・オフが当然ある。サイズが大きければ大きいほど、収集するテキストの内容やバランスに関する統制はしにくくなる。このへんを勘案して賢く選択する必要が利用者（作成者）にはあるといえよう。

6.2. Webコーパスのデータの中身に関する疑問

第2の大きな疑問はいくつかの技術的な問題である。以下のような疑問点が挙げられる：

(1) 母語話者以外の書いた英語が混入している可能性があるのではないか？
(2) 文書情報（作成日時）などは特定できるのか？
(3) 著作権はどのようになっているのか？
(4) 従来のコーパスと比較可能なコーパスをWebで作れるのか？

疑問点(1), (2)に関してはシンポジウムの席上でも必ずしも明確な答えが出なかった。Webの文書から母語話者の書いたものを特定するという作業は、極めて困難である。URLを指定して、新聞や雑誌などの校正が加えられたテキストのみを集めていればまだしも、一般のクローリングでは非母語話者の書い

たものが混入している可能性は大いにある。技術的にはスペリング・エラーや文法エラー等からある程度判定は可能であろうが、実際は母語話者でもスペリングや文法のミスを犯すから、確信をもって判別はし難いであろう。またWebを使った時系列データのコーパス化にも取り組んでいる研究者がいるが、Webそのものは特定の時間軸を指定しても、それ以前に書かれた古いテキストに関しては厳密に制御することはできない。

著作権に関しては、今号の田中氏の論考を参照していただきたいが、最近ではCreative Commons (CC) ライセンスというインターネット時代の新しい著作権ルールの普及が著しく、web上のテキストにもこのCCライセンスの付与されたものが大量に利用可能になってきている。[8] これらのテキストは非営利で改変しないなどの条件をクリアすれば著作権許諾を得なくとも利用可能だ。

最後に(4)の従来のものと比較可能なWebコーパスに関しては、前述のNew Model CorpusがBNCをモデルにしたWebコーパスの試みとして注目に値する。Brown CorpusからLOB Corpusを作った際にも米語とイギリス英語の違いにより若干ジャンル構成が変更になったりした。

また小学館コーパス・ネットワークが構築したSEKAI Corpusについても、シンポジウムでの中村・星野両氏の発表にあったように、従来の均衡コーパスと比較可能な設計を施すことによってWebコーパスの使い勝手が飛躍的に向上することがあろう。図1はSEKAI Corpus 20億語によってbloodという単語を検索し、その分野ごとの内訳を相互比較するチャートを出力したものである。この画面は通常のbloodを検索したコンコーダンサー出力ウィンドウから集約結果としてワンクリックで見ることができる。bloodという単語がPERC Corpusのような科学技術分野、またLaw-Criminalという法律－犯罪部門で多用されていることが一目瞭然となる。それをBNC-written, BNC-spokenといった汎用コーパスとの頻度の相対的位置で解釈できるため、どの程度専門用語化した用法であるか、あるいは一般語としての用法であるか、といったことを考察する際に極めて強力なツールとなる。

このようにWebコーパスに関しても設計および比較の方法を適切に考え、かつ、クローリングの精度が向上すれば、テキスト領域やレジスターなどに配慮の行き届いたWebコーパスの構築も十分可能になってくるであろう。

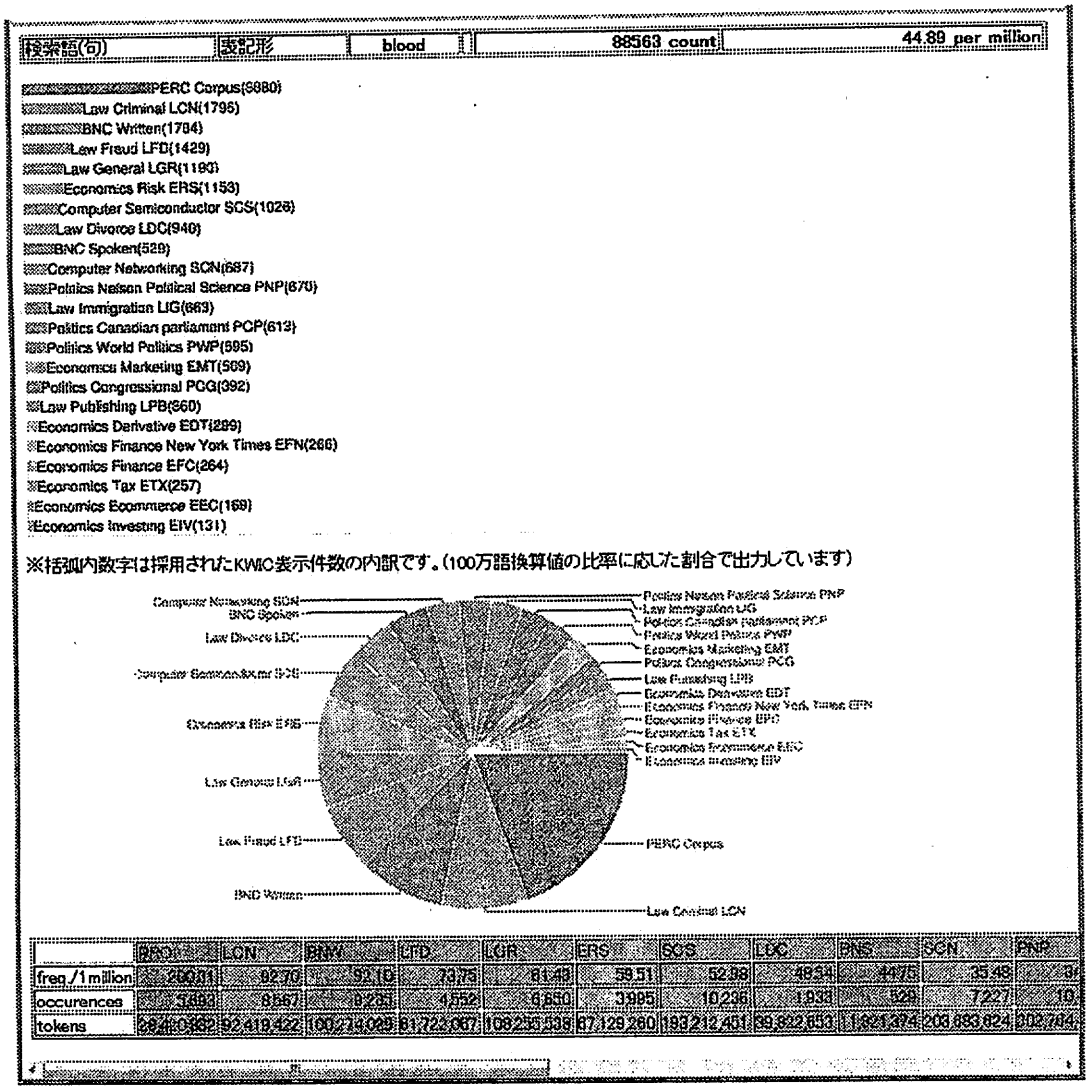

図 1：小学館 SEKAI Corpus のサブコーパス相互比較画面

6.3 Web コーパスを扱うコンコーダンサーの処理能力

Web コーパスは通例数千万語から数十億語という規模になる。この場合、従来の検索ソフトは使用可能なのであろうか？この疑問に応えるべく、筆者は簡単な実験を行った。WaCky で開発された BootCaT front-end で seeds としては UkWaC からの高頻度 5000 の内容語（機能語を省く）を使用、3 個組（3 tuples）を 1000 セットで各 10URL、合計 1 万 URL でクローリングを行った。これで収集すると 2200 ～ 2400 万語クラスのコーパスなら 1 セット 4 － 5 時間で出来上がる。このうちクローリングに成功した 2000 万語クラスのコーパスおよび 8000 万語のコーパスを用いて、Windows ベースのスタンドアロンのコンコーダンサーでコーパス全体の頻度リストを作成させ、実力を比較してみた。結果は表 3 のとおりである。

表 3：大規模データによるコンコーダンサーの処理能力[9]

Corpus ID	総語数	AntConc	MonoConc Pro	WordSmith 5
001 (15MB)	2,626,891	40s	23.7s	5.5s
002 (200MB)	28,468,571	動作停止	1m 44s	1m 03s
003 (474MB)	80,231,744	動作停止	3m 34s	2m 40s
005 (134MB)	22,889,948	動作停止	1m 06s	43s
006 (138MB)	23,574,444	動作停止	1m 01s	44s
007 (144MB)	24,588,570	動作停止	1m 10s	45s

表 3 で示したように、フリーのコンコーダンサーとして人気の高い AntConc は 200 万語規模のものでは動作しているが、2000 万語のデータでは途中で動作が停止してしまった[10]。一方、商用コンコーダンサーの MonoConc Pro 2.2, WordSmith Ver.5 に関しては 2000 万語クラスでは 40 秒から 1 分程度で処理できる。全体的に WordSmith の方が処理速度は高速であった。

6.4. Web コーパスは偏っていないのか？

最後に、Web はどうしても Web である、という性格から抜けられないのではないか、という疑問を呈しておこう。表 1 で Kilgarriff たちが Web の規模を推定するために検索した 2 つの語句、medical treatment と deep breath を比べてみると、面白いことに BNC においては medical treatment よりも deep breath の方が頻度が高いのに、Web ではその反対の現象が起こっていた。SEKAI Corpus で deep breath を検索した結果が図 2 である：

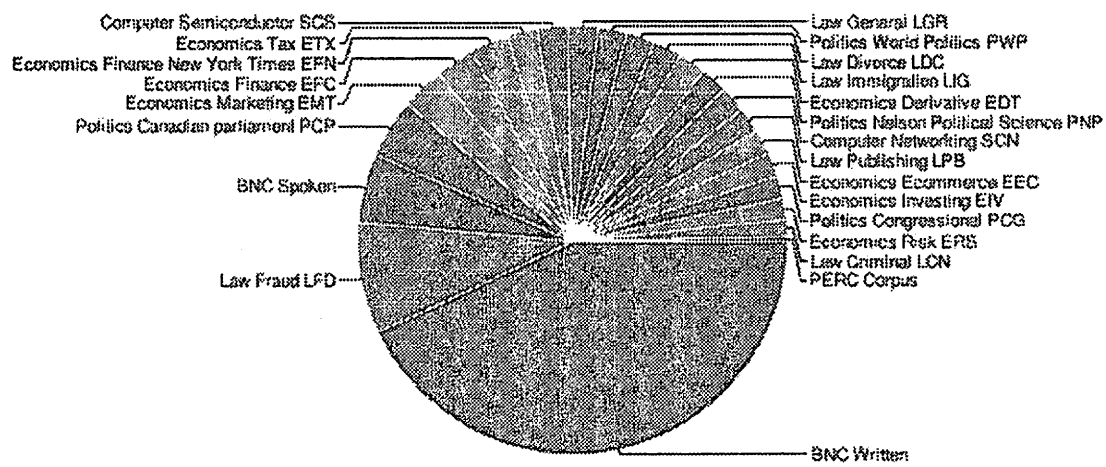

図 2：「deep breath」の検索結果（SEKAI Corpus）

deep breathという表現は「深呼吸をする、深く息をする」という意味だが、これがBNC Writtenに非常に多いのは、BNCが書き言葉部分にかなりの割合で文学作品を含んでいることがあろう。Web上には我々が通常目に触れるような文学作品は著作権上の制限でアップされていない。著作権切れの古い文学作品は現代英語のクローリングの対象には好ましくない。そういう理由で、全体にmedical treatmentのような単語が出やすい、という性質をWebコーパス自体が持っているのかもしれない。これは別にWebコーパスの価値を一概に損ねるものではないが、しかし、BNCが考慮した「テキストの消費者」としての人間が通常目にするテキスト群のバランスを考えた場合、Brown Corpusの時代から存在する情報散文(informative) vs. 創作散文(imaginative)の配分は重要なのではないか。その点で、Webコーパスには何かが足りない、という指摘もあながち間違いではないかもしれない。この点に関しては、もう少し詳細に大羽氏の今号での論考が参考になるかもしれない。

7. まとめ

本稿ではWebコーパスのシンポジウムでの論点を筆者なりに整理し、その全体像を歴史的な観点を紹介しながら示し、かつ続く田中氏、大羽氏の論文の序章となるようなWebコーパスへの期待と疑問を投げかけてみた。

Webコーパスは否応なく今後のコーパス言語学のリソースの1つの主流となるであろう。その際に、上述したようなさまざまな疑問点を踏まえて研究者自身が目的に応じて吟味し、Webコーパスの長所・短所を認識しつつ、この膨大な言葉の海を操る術を身につければ、コーパス研究の地平に新しい光が見えてくるかもしれない。このシンポジウムおよびまとめの論考がその一助となれば望外の喜びである。

注

本稿は英語コーパス学会第35回大会シンポジウム「進化するWebコーパス：現状と課題」での発表に加筆修正したものである。シンポジウムの参加者、および編集委員長、匿名の査読者からの貴重なコメントにこの場を借りて感謝の意を表したい。

[1] 2010年3月10日現在
[2] http://wacky.sslmit.unibo.it/
[3] http://www.webcorp.org.uk/
[4] http://webascorpus.org/
[5] http://www.sketchengine.co.uk/
[6] http://borel.slu.edu/crubadan/

[7] WAC4: ttp://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf
WAC5: http://www.sigwac.org.uk/attachment/wiki/WAC5/WAC5_proceedings.pdf

[8] Sketch Engine の WebBootCAT ではこの CC ライセンスの文書のみをダウンロードする機能もある。

[9] Corpus ID は 10 回クローリングを行って所要時間の計測とコーパス作成が成功か失敗かを調べたために通し番号をつけてある。ちなみに 10 回の試行のうち、3 回クローリングが途中で停止し、200 万～ 600 万語程度の規模しか収集できなかった。Corpus ID: 003 だけがが 8000 万語と大規模なのは、クローリングのセッティングを URL100 件と 10 倍にしたため。参考データとしてご覧頂きたい。

[10] この点に関しては、その後、実験データと実験に使用したコーパス・データを Laurence Anthony 氏に提供したところ、氏のパソコンでも処理が停止する事実を確認できた（personal communication)。氏によれば、AntConc はファイル処理の際にテキスト・サイズの上限を設けているので、その条件を変更しさえすれば動作するようになる、ということであった。ヴァージョンアップを望みたい。

参考文献

Biber, D. (1993). Representativeness in corpus design. *Linguistic and Literary Computing* 8 (4): 243-257.

Fairon, C., Naets, H., Kilgarriff, A. & de Schryver, G-M. (2007). Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, ncorporating Cleaneval. Presses univ. de Louvain.

Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). Corpus Linguistics and the Web. Amsterdam: Rodopi. Kilgarriff, A. and G. Grefenstette (2003) Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3): 333-347.

Kilgarriff, A. (2007). Googleology is Bad Science. *Computational Linguistics* 33 (1): 147-151.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3): 333-347.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mihalcea, R. and Moldovan, D. (1999). A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Meeting of ACL*, pp. 152-158, College Park, MD, June.

Resnik, P. (1999). Mining the Web for bilingual text. In *Proceedings of the 37th Meeting of ACL*, pp. 527-534, College Park, MD, June.

Xu, J. L. (2000) Multilingual search on the World Wide Web. In *Proceedings of the Hawaii International Conference on System Science (HICSS-33)*, Maui, Hawaii, January.

南出健吾 (2008)『全世界のWebページTLD・言語分布解析』早稲田大学大学院理工学研究科情報ネットワーク専攻2007年度修士論文．2008年2月4日．