カテゴリ情報とジャンル情報を利用したBNCコーパス検索システム開発のための予備的考察

中島浩二

A Preliminary Study for Developing a Concordance Program of British National Corpus Using Category and Genre Information

Kohji NAKASHIMA

Abstract

This paper begins with a brief survey of several epoch-making language corpora. The British National Corpus (BNC), which was completed in 1994 after three years of development, is one of the most representative and reliable corpora in terms of both quality and quantity. The first version of the BNC was limited to EU countries, but the World Edition has been available on CD-ROM since 2000.

In order to develop an efficient concordance program for the BNC World Edition, it is crucially important to conduct a fundamental analysis of the structure of corpus data. The CDIF (Corpus Data Interchange Format), which was strongly influenced by TEI Guidelines, has been adopted in the BNC. This provides ample contextual information as well as grammatical information about the corpus data. In order to develop my Web-based BNC concordance program, I will refer to this enriched tagged information including *Text Classification Codes* and David Lee's *Genre Classification Scheme*.

This paper concludes with a basic strategy to develop my BNC concordance program named 'BNCfinder+'.

1. BNC コーパスについて

A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (通称 Brown Corpus) が、Brown 大学の Henry Kučera と W. Nelson Francis の手によって 1964 年に完成した。これは、1961 年に刊行されたアメリカ英語の資料から約 100 万語を電子テキストとして抜粋編集したものであり、15 ジャンル、おのおの 2000 語余りのサンプル 500 個によって構成されている。100 万語の言語コーパスというのは現在からしてみると規模的にはそれほど大きくないが、当時の高価で限られたコンピュータリソースを考慮すれば、画期的な成果であったと言えるだろう。

それに触発されて 1978 年に Brown Corpus のイギリス英語版 Lancaster-Oslo/Bergen Corpus of British English (通称 LOB Corpus) が完成し、1980 年代から 90 年代にかけて ACE (Australian Corpus of English)や Wellington Corpus of Written New Zealand English (WWC) などの英米以外の英語変種コーパスも誕生した。さらに 1990 年からは ICE (International Corpus of English)という英語を母語あるいは第 2 言語としている地域の言語資料をコーパス化するプロジェクトも継続的に行われている。その他、Brown Corpus と LOB Corpus の約 30 年後の言語資料をコーパス化した Freiburg-Brown Corpus of American English (通称 Frown) と Freiburg-LOB Corpus of British English (通称 FLOB) も、英語の通時的変化を比較研究する道を開いたという点で注目すべき成果と言えるだろう。

上述の言語コーパスに共通した特徴は、いずれも 100 万語規模のコーパスであるという点である。それは、最初に開発された Brown Corpus が約 100 万語のコーパスであったことから、それとの比較対象研究をする際の利便性を考慮した結果あえてその規模にあわせたという側面もあるのだろうが、作業量や予算規模、当時のコンピュータ資源の制約も大きな要因ではないかと考えられる。規模が小さければ統計的な信頼性は減るのが一般的な傾向なので、コーパスサイズの大規模化は必然的な方向となっていく。1990 年代から 2000 年代にかけての British National Corpus (BNC)や Bank of English (BoE)などの大規模コーパスの開発はそのような時代の要請によって誕生したと言えるだろう。

British National Corpus (BNC) は、Oxford University Pressを主幹として1991年に始まったプロジェクトであり、3年後の1994年に完成した。1975年以降のイギリス英語の書き言葉9000万語と話し言葉1000万語から構成される約1億語のコーパスで、サンプルコーパスとしては世界最大規模である。1995年にEU (European Union)内の研究者に公開され、2000年にはBNC World Editionのリリースによって世界中に公開された。BNCはplain textのコーパスではなく、品詞タグやジャンル情報も付与されており(annotated corpus)、英語研究の基盤的道具立てとして活用しやすいよう配慮された設計デザインになっている。次章では、BNC World Edition (2000)をもとにその設計内容の詳細を探っていく。

2. BNCコーパスの構造とヘッダ情報解析

コーパスに品詞や文法情報などの2次的言語情報を各コーパス作成者が恣意的に付与するとコーパスデータを共有する際の非互換性が問題になってくる。互換性がないだけなら非互換性を吸収するプログラムを作ることによってある程度対応することは可能かもしれないが、そもそもその2次的情報の言語学的妥当性についての合意がないところでの情報付与は、第3者がそれを利用する場合かえってマイナス要素になりかねない。そこで、コーパスに情報を付加する際の世界的規格の必要性からTEI (Text Encoding Initiative)プロジェクトが推進された。元々は軍事技術を目的に開発されたSGML (Standard Generalized Markup Language)に準拠した形でTEI P3 (TEI public proposal number 3)ガイドラインが発表されたのが1994年で、BNCはこのガイドラインに強く影響を受けたCorpus Document Interchange Format (CDIF)規格を採用している。なお、2007年には最新版となるP5 Guidelinesがリリースされている。

検索システムを開発する際に重要なことは、そのデータ構造をよく分析することである。やみくもに力技で検索する仕組みを作ったとしても、現実的な時間内に結果が返って来ないのでは実用にならない。データの規模が大きくなればなるほど、そのことに留意しなくてはならない。

BNCは現代のイギリス英語の書き言葉9000万語と話し言葉1000万語の計約1億語の言語コーパスなので、そのデータ量も膨大なものになる。

プログラムを作って調べてみたところ、BNC World Edition は、書き言葉のファイル数は 3144 ファイル、話し言葉のファイル数は 910 ファイル、総ファイル数は 4054 ファイルで構成されていた。ディレクトリの中も構造化されており、多数の深いサブディレクトリの中にデータファイルが配置されている。BNC World Edition の CD-ROM 版を使って Unix の HDD にインストールした場合、BNC コーパスのデータ部分は次のようなディレクトリ・ファイル構成になっていた。

BNC-world/Texts/ というサブディレクトリの下に A, B, C, …, H, J, K (‘I’ は ‘1’ と紛らわしいので避けているのだろう)の 10 のサブディレクトリが配置され、その配下が次のようにさらに深く階層化されている。なお、最上層の ‘BNC-world’ というディレクトリ名はインストール時に任意で決められる。

```
BNC-world/Texts/A/A0/A00
BNC-world/Texts/A/A0/A01
BNC-world/Texts/A/A0/A02
          (中略)
BNC-world/Texts/A/A1/A10
BNC-world/Texts/A/A1/A11
BNC-world/Texts/A/A1/A12
          (中略)
BNC-world/Texts/F/F7/F71
BNC-world/Texts/F/F7/F72
BNC-world/Texts/F/F7/F73
          (中略)
BNC-world/Texts/K/KS/KSU
BNC-world/Texts/K/KS/KSV
BNC-world/Texts/K/KS/KSW
```

図 1. BNC World Edition のファイル及びディレクトリ構成

図 1 にあるように、ファイル名はアルファベットおよび数字の 3 文字によって構成され、ファイル名の末尾の英数字を除く一文字一文字がサブディレクトリの path を表している。つまり、ファイル名が分かればルートディレクトリからの相対的位置も同時に分かるというわけだ。し

かし、位置が分かったとしても、検索が実行されるたびに4000を超えるファイルを開いて中の文字列を調べるのは、いくら以前に比べて劇的に高速になったとは言え、一般的なパーソナル・コンピュータにはかなり負荷のかかる処理作業であることに変わりはない。

コーパスを使った言語研究をする場合、多くの場合、どのような文脈(register)で使用された言語資料なのかが重要になることが多い。書き言葉と話し言葉で、言葉の使われ方に違いがあることは誰でもすぐ気づくが、同じ書き言葉でも新聞英語と散文では文体が異なるし、話し言葉でも一方的に発信するレクチャー的な発話と双方向的な会話でスタイルに違いが出るだろう。また、同じ文脈においても男女や年齢層の違いによっても差異が出る。いや、そうではない。むしろ、言語を客観的に研究する者は、状況や文脈によって言葉遣いに違いが現れると先験的に決めつけてはいけないのだ。調べてみて結果的に違いが出る場合もあれば出ない場合もあるだろうから。少なくとも言語コーパス検索システムで重要なことは、ある状況や文脈下での言語資料を他と区別して検索可能な仕組みを保証しておくことだろう。BNC World Editionのファイルにはそれを可能にする情報がtagの形で記録されている。一例として、BNC-world/Texts/A/A0/A03 ファイルをテキストエディタで開いてみると、ヘッダ部に次のような情報が記されている。

```
<catRef target="alltim3 allava2 alltyp3 wriaag0 wriad0 wriase3
wriaty2 wriaud3 wridom5 wrilev2 wrimed2 wripp5 wrisam1 wrista3
writas3"/><classCode scheme="DLee">W pop lore</classCode>
```

図2. TEI Headerタグ情報の一部

まるで呪文のような記号の羅列だが、<catRef>(=category reference)タグ内の属性名targetの値として指定されているalltim3, allava2, alltyp3, wriaag0, wriad0, wriase3等々が、そのファイル内にあるコーパステキストのTEIによる定義にそったカテゴリ情報(Text classification codes)を表しているのだ。図2の例に出てくる値の意味するところを一部挙げてみる。

```
alltim3 - Publication Date: 1985-1993
allava2 - Text availability: Worldwide rights cleared
```

alltyp3 - Text type: Written books and periodicals
wriaag0 - Author age band for written material: Unknown
wriad0 - Author domicile: Unknown
wriase3 - Author sex: Mixed

このようなテキストのカテゴリに関する情報が BNC コーパスを構成している 4000 以上のファイルの TEI ヘッダ部に書き込まれているのである。つまり、alltim3 という値を持ったファイルのコーパスデータから検索すれば1985年から93年の間に発行された文章だけからの検索になるし、alltyp3 という値を持ったファイルからのコーパス検索は書籍や定期刊行物の文章だけからの検索になる。さらに、その両方の値を持っているファイル群からの検索をおこなえば、両者の連言、すなわち「1985 年から 93 年の間に発行された書籍や定期刊行物」からの絞り込み検索になるという訳である。このように BNC で使われているカテゴリ情報は、ある特定のカテゴリ条件にあてはまる文からの検索を可能にしてくれるという点で言語研究者にとってデータ収集の強力な手がかりとなりうる。なお、BNC で用いられている全てのカテゴリ情報の値の定義は、インストールされるファイル BNC-world/Doc/HTML/cdifhd.html をブラウザで読み込めば参照可能である。

図 2 を見ると、<catRef>タグに続いて、次のように、<classCode>タグが記述されていることに気づく。

<classCode scheme="DLee">W pop lore</classCode>

ここにある scheme="DLee" は、David Lee's Genre Classification Scheme [1] を指している。 David Lee は、Written texts を 46 genres に、Spoken texts を 24 genres に分類しているが、BNC コーパスでは、先に挙げたカテゴリ情報<catRef>とは独立に David Lee のジャンル情報も<classCode>タグの中に記述されている。 'W pop lore' は、Written text の popular magazines を意味している。

先に述べたように、コーパステキストの分類は、この分類法が正しいとか正しくないとか誰もが合意できるものではない。そこには恣意性が入り込む余地が否定できないし、ある特定の分類分けが未来永劫有効あるいは妥当であるという保証もない。BNC の編纂者たちが冗長性をおそれず複数の分類法を並立させたことは、独善性を排してより公平な視点をコーパス編纂に取り入れていこうという姿勢を表していると考える。

3. 効率的な検索プログラム作成のための方略

前章で述べたように、BNC World Editionのコーパステキストファイルの総数は4054もある。それらのファイルが格納されているディレクトリBNC-world/Texts/のサイズを見ると、全てテキストファイルであるにも関わらず約1.55GBもの容量（ディスク使用量）を要していた。1ファイルあたりで換算すると平均380KB程度要していることになる。現在はコンピュータの処理速度が上がったとはいえ、ファイル入出力に関わる部分は依然負荷の大きな処理である。検索の度に大量のファイルを開いて該当ファイルを見つけ、そこから検索にヒットする情報を取り出した後でファイルを閉じるという処理を行っていては、実用的な時間内で検索を行うシステムを構築することは困難だ。もう少し効率的なやり方はないだろうかとインストールされたファイルをあれこれ調べていると、BNC-world/SGML/というディレクトリ内にbncfinder.datというファイルがあることに気が付いた。このファイルには次のような情報が記述されている。

```
A00 107 112 423 6673 6894 alltim3 allava2 alltyp5 wriaag0 wriad0
wriase0 wriaty2 wriaud3 wridom4 wrilev2 wrimed3 wripp5 wrisam5
wrista2 writas3
A01 127 167 597 7882 8115 alltim3 allava2 alltyp5 wriaag0 wriad0
wriase0 wriaty1 wriaud3 wridom4 wrilev1 wrimed3 wripp5 wrisam5
wrista1 writas3
A02 56 50 223 3347 3430 alltim3 allava2 alltyp5 wriaag0 wriad0
wriase0 wriaty1 wriaud3 wridom4 wrilev2 wrimed3 wripp5 wrisam1
wrista1 writas3
A03 298 360 1051 19255 19972 alltim3 allava2 alltyp3 wriaag0 wriad0
wriase3 wriaty2 wriaud3 wridom5 wrilev2 wrimed2 wripp5 wrisam1
wrista3 writas3
```

図3 BNC-world/Texts/bncfinder.datファイルの一部

各行の構成は、

three-character identifier
size of text in Kbytes
number of <p> or <u> elements
number of <s> elements
number of <w> elements
number of orthographic words
all classification codes assigned to this text

図 4 bncfinder.dat の各論理行の構成

となっている。[2] つまり、3 文字からなるファイル名と 5 つの数値情報、その後に先の章で述べた TEI の Text classification codes が続いているのだ。これをうまく利用すれば、あるカテゴリにターゲットを絞って検索をしたい場合、全てのファイルを逐一開いてヘッダ内の<catRef>タグを見てそのファイルが目的のカテゴリに属すか否かを判断する必要はなく、bncfinder.dat ファイルを見て、Text classification codes に検索対象のカテゴリ情報を持っているファイルがどのファイルであるか検索し、該当するファイル群のみを実際に開いて検索するコンピュータプログラムを作成可能であることが分かる。bncfinder.dat ファイル内で該当のファイルを収集するのは現代のコンピュータにとっては一瞬でできる処理だし、特に複合的なカテゴリの条件積で検索対象をしぼった場合は実際に開く必要のあるファイル数を大幅に少なくすることが可能になり、検索処理システムを飛躍的に速く軽いものにできるのだ。

ただし、bncfinder.dat では David Lee の Genre Classification Scheme の情報が扱われていない。せっかく BNC コーパスの本体であるテキストファイルのヘッダ情報に入っているジャンル情報なのに、それを活用しないのはもったいないと考える。そこで、全てのファイルヘッダにある<classCode scheme="DLee">...</classCode>部分からジャンル情報を機械的に抜き出し、各論理行末にその情報を追加する形で次のように bncfinderPlus.dat ファイルを作成した。

A00 107 112 423 6673 6894 alltim3 allava2 alltyp5 wriaag0 wriad0 wriase0 wriaty2 wriaud3 wridom4 wrilev2 wrimed3 wripp5 wrisam5 wrista2 writas3 **W_non_ac_medicine**

A01 127 167 597 7882 8115 alltim3 allava2 alltyp5 wriaag0 wriad0 wriase0 wriaty1 wriaud3 wridom4 wrilev1 wrimed3 wripp5 wrisam5 wrista1 writas3 **W_non_ac_medicine**
A02 56 50 223 3347 3430 alltim3 allava2 alltyp5 wriaag0 wriad0 wriase0 wriaty1 wriaud3 wridom4 wrilev2 wrimed3 wripp5 wrisam1 wrista1 writas3 **W_institut_doc**
A03 298 360 1051 19255 19972 alltim3 allava2 alltyp3 wriaag0 wriad0 wriase3 wriaty2 wriaud3 wridom5 wrilev2 wrimed2 wripp5 wrisam1 wrista3 writas3 **W_pop_lore**
図5. bncfinderPlus.datファイルの一部 [3)]

このbncfinderPlus.datをもとに検索システムを作成すれば、TEIのカテゴリ情報、あるいはDavid Leeのジャンル情報、さらにそれらを組み合わせた条件下で、効率的なBNCコーパス検索システムを構築することが可能になるだろう。

4. 運用システムBNCfinder+ 開発に向けて

ここまで、BNC World Editionを効率的に検索するシステムの開発に向けて構成ファイルおよびデータ構造の基盤的分析をおこなってきた。また、David Lee の Genre Classification Scheme 情報を利用してbncfinderPlus.datという拡張データベースの構築をおこなった。

ここでは、実際に稼働するBNC World Editionのコーパス検索システムをどういう方向性で構築していくのか、その概略を検討する。

＜システムの概要＞

- OSに依存しない汎用性と管理の容易さを考慮して、Web上で動作するCGI-basedなシステムにする。
- ユーザーインターフェイスは、現在広く使われている一般的なWebブラウザで動作するHTML 4.0 Traditionalと基本的なCSSを利用する。中長期的にみて将来も互換性が保たれるようmarkup言語等の記述をすることが肝要である。
- CGIとやり取りをするプログラムの開発言語としては、文字列処理に便利なPerlを採用する。

・ユーザーインターフェイスは、マニュアル等を見なくても操作できるようシンプルなものにする。ただし、上級ユーザの高度で複雑な検索をしたいという要求にも応えるため、regular expression (Perl 5.8で利用可能な正規表現と同等) が利用可能なものにする。
・BNC World Edition の提供するデータは品詞情報などが埋め込まれたtagged corpus なので、それを活用可能なものにする。
・TEI のカテゴリ情報と David Lee のジャンル情報を利用した検索が可能なものにする。
・検索結果は、コーパス検索システムで一般的に使われている KWIC (KeyWord In Context)形式にする。また、どのファイルを検索対象にしたのか分かりやすく提示するシステムにする。

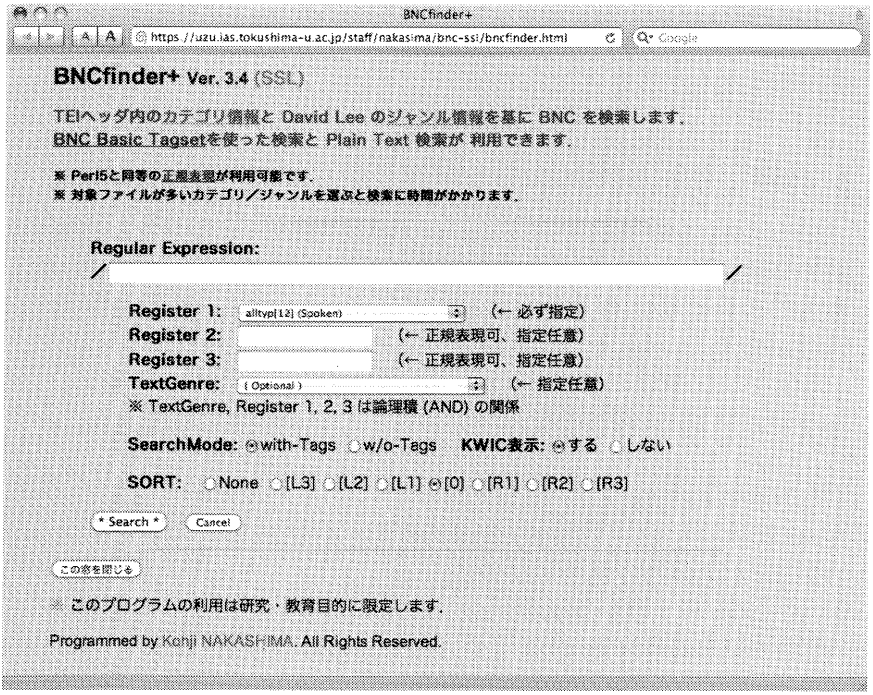

図 6. BNCfinder+ のトップ画面

これらの諸要求・条件を考慮したシステムを具体的に構築していくこ

とになるが、システム開発の技術的詳細については別の稿に譲ることにする。

注

1) David Lee's Genre Classification Scheme の全ての Code は、http://homepage.mac.com/bncweb/manual/genres.html に記載がある。
2) BNC User Reference Manual - Software for the BNC http://www.natcorp.ox.ac.uk/docs/userManual/bncsmop.xml を参照のこと。
3) 実際は、<classCode scheme="DLee">W pop lore</classCode>のように値の中にスペースが含まれているが、一つの文字列として扱いやすくするため、空白部分を underscore（'_'）に変換して bncfinderPlus.dat に書き入れた。
* 'BNCfinder+' は、中島のホームページ：
https://uzu.ias.tokushima-u.ac.jp/staff/nakasima/ よりアクセス可能。ただし、著作権の関係上、制限的に運用している。

参考文献

齊藤俊雄・中村純作・赤野一郎（2005）『英語コーパス言語学 -基礎と実践- 改訂新版』研究社
武藤健志・トップスタジオ（2004）『独習 Perl 第2版』SHOEISHA
結城 浩（1998）『Perl で作る CGI 入門 -応用編-』SOFTBANK
Aston G. and Burnard, L. 1998. *The BNC Handbook.* Edinburgh University Press.
Lee, David Y.W. 2001. *Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle.* Language Learning & Technology, Vol.5(3): 37-72. [Available at: http://llt.msu.edu/vol5num3/lee/default.html]