日本医学英語教育学会 第9回総会
【特別講演】

コーパスの医学英語教育への貢献：
PERC Corpus プロジェクトを中心に
How Corpus Linguistics Helps EMP Education, with Particular Reference to the PERC Project

演者
投野由起夫
明海大学外国語学部助教授

それでは最初に，コーパス言語学の基礎的な概念をお話しします。ここでお話しするのは，まずコーパスの定義(definition)です。それから，コーパスにはいろいろな情報を言語的に入れ込むということをします。1つはmark-up，もう1つはannotationといいますが，そこにはどのような情報が入るのかということをご紹介します。それから，英語のさまざまなコーパスがすでにかなりできあがっているのですけれども，どんなものがあるかをザッと紹介します。最後に，ESPとか，私はここで初めて知ったのですが，EMPというような言い方をするそうですね，そのEMP Corporaはどうかというようなことを見たいと思います。

コーパスの定義と言語情報

まずdefinitionということですけれども，コーパスというのはいろいろ条件があるのです。1つ目は*New Oxford Dictionary of English*(NODE)で，いまは2版が出ていますが，その98年のdefinitionです。"a collection of written or spoken material in machine-readable form, assembled for the purpose of studying English structures, frequencies, etc."ということで，collectionであると。それから，書き言葉でもあるし，話し言葉でもあるよということです。またmachine readable formということで，コンピュータが読めるようになっているということです。そして「特定の目的のために集められている」ということです。

もう1つ，別のdefinitionがありまして，これは私がランカスターの同じ仲間たちと一緒に書いたtextなのですけれども，これにもいくつか似たようなことが繰り返されています。"Corpus is a collection of machine-readable authentic texts, which is a sample to be representative of a particular language or languages variety"ということです。ここでもいくつかポイントになるような要素が挙げられています。

コーパスというのは単純な寄せ集めではないのです。例えば，Googleか何かで検索したような結果とは違います。コーパスの場合は，中身が何かがわかっているということです。そして，収集には目的があるということです。特定の言語の編集なりバラエティ，あるいは医学英語なら医学英語というような，特定の目的をまず頭に入れて集められているということです。

続いて，コーパスにはいろいろな特徴があります。まず，言語的な特徴，corpus mark-upというものを見てみましょう。コーパスにはさまざまなtext的な情報を入力するのですけれども，普通はコーパスはただのplain textなのです。そこにさまざまな情報を入れる入れ方があります。大きく分けると，mark-upとannotationとに分かれています。"mark-up"というのは，textに関する情報を入れるのです。

演者紹介：投野由起夫氏（とうの・ゆきお）

東京学芸大学卒業。同大学大学院英語科教育専攻(英語教育学)，英国ランカスター大学大学院言語学科博士課程修了，Ph.D.(コーパス言語学)。東京都立航空工業高等専門学校英語科専任講師，東京学芸大学教育学部英語科教育講座専任講師等を経て，2001年より現職。

大量の言語データをコンピュータで分析して言語研究を行うコーパス言語学を専門とし，コーパス言語学に基づく第2言語習得(特に語彙習得)を応用したNHK教育テレビ「100語でスタート！英会話」の講師としても活躍。

Features: corpus mark-up

- Mark-up: a system of standard codes inserted into a document stored in electronic form to provide information about the text itself.

```
<header>
<textnum>0057</textnum>
<filedesc>
        <title>my family</title>
        <name>Hanako Yamada</name>
        <grade>10</grade>
        <date>1999-07-10</date>
</filedesc>
<textdesc>
        <medium>essay</medium>
        <domain>informative</imaginative>
        <genre>student writing</genre>
        <region>Japanese EFL</region>
</textdesc>
```

スライド1

Corpus annotation: POS tagging

- POS information (British National Corpus)

```
<s n="26">
<w PNP>It <w VBZ>is <w AV0>now <w DT0>more <w CJS>than
<w AV0>ever <w VVB>clear <w CJT>that <w AT0>every <w NN1>section
<w PRF>of <w NN1>society <w VVZ>needs <w TO0>to <w VBI>be
<w AJ0>involved <w PRP>in <w VVG>responding <w PRP>to <w NN1>AIDS
<c PUN>, <w PRP>including <w AT0>the <w NN2>churches<c PUN>.
<s n="27">
<w NP0>ACET <w VBZ>is <w AT0>a <w AJ0>Christian <w NN1>initiative
<w VVN>supported <w PRP>by <w DT0>all <w NN2>denominations<c PUN>.
<s n="28">
<w PNP>It <w VBZ>is <w AT0>the <w NN2>churches <w CJT>that
<w VVB>provide <w DPS>our <w NN2>volunteers<c PUN>;
<w PRP>without <w DPS>their <w NN1>support <w PNP>we <w VM0>would
<w XX0>not <w VBI>be <w AJ0>able <w TO0>to <w VVI>provide <w AT0>a
<w NN1>service <w AV0>at all<c PUN>.
```

スライド2

Corpus annotation: lemmatization

- POS information (CLAWS vertical format)

```
0000001 002  ------      ------      PUNC
0000002 001  NULL        <text>      PUNC
0000003 001  NULL        <s>         PUNC
0000003 010  PPY         You         you
0000003 020  VM          must        must
0000003 030  VVI         leave       leave
0000003 040  RT          now         now
0000003 041  ,           ,           PUNC
0000003 050  RR          otherwise   otherwise
0000003 051  ,                       PUNC
0000003 060  PPY         you         you
0000003 070  VM          will        will
0000003 080  VBI         be          be
0000003 090  JJ          late        late
0000003 100  IF          for         for
0000003 110  APPGE       your        your
0000003 120  JJ          social      social
0000003 130  NN2         studies     study
0000003 140  NN1         class       class
0000003 141  .           .           PUNC
```

スライド3

これはだいたいヘッダセクションに入れるのですけれども，例えばそのtextがだれによって書かれたかとか，何年に出版されたかとか，どこのウェブから取ってきたとか，そのようなtextに関する情報です。information about textということです。

もう1つは，textのstructureがあります。例えば，ここからここまでがparagraphなのだとか，ここまでがsectionだとか，ここがAbstractだとか，論文だったらここがConclusionだとか，そのようなparagraph，あるいは論文なりtextの構造がありますが，そういうものを記号で入れるというようなことをします。

もう1つは“corpus annotation”です。annotationという用語は，言語的な情報を入れるときに使います。textの言語的特徴というと，例えば品詞のタグです。これは動詞だとか，名詞だとか，そのような情報です。

2つ目はラマタイゼーション(lemmatization)といいまして，これは見出し語です。例えばbring，brought，bringingとか，そのように活用するときに，全部これはbringなのだとか，そのような見出し語情報を入れたりします。

“semantic tagging”というのは，意味領域を入れたりします。それから“parsing”，これは構文解析データです。

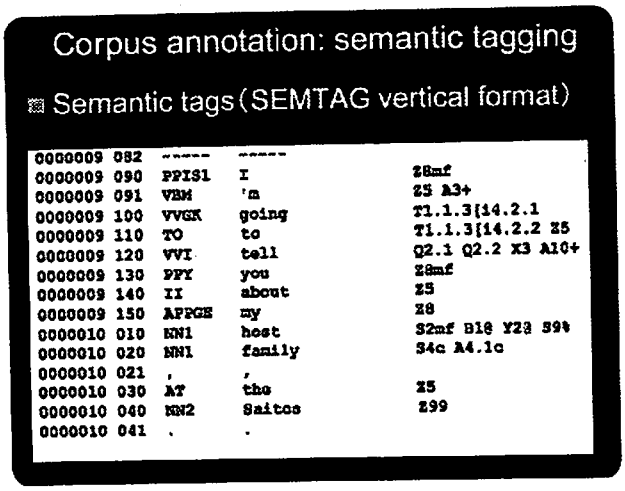
例えば，**スライド1**はcorpus mark-upの例です。これは私がつくっているlearner corpusの例ですけれども，text自体は出ていないけれども，textのdescriptionとかfile descriptionの部分をヘッダに格納しているという，このようなフォーマットです。ここを検索して，コンピュータが，例えば皆さんが目的とするようなtextをコンピュータが探すときの手がかりなわけです。例えば，医学英語でこういう領域のtextだけ取ってこいとか指示できるようになっているわけです。

同じように，その論文の構造みたいなことに関する情報をmark-upしておけば，abstractだけ取ってくるとか，あるいはreviewのところの表現だけ取ってくるとか，そんなこともできるかもしれません。

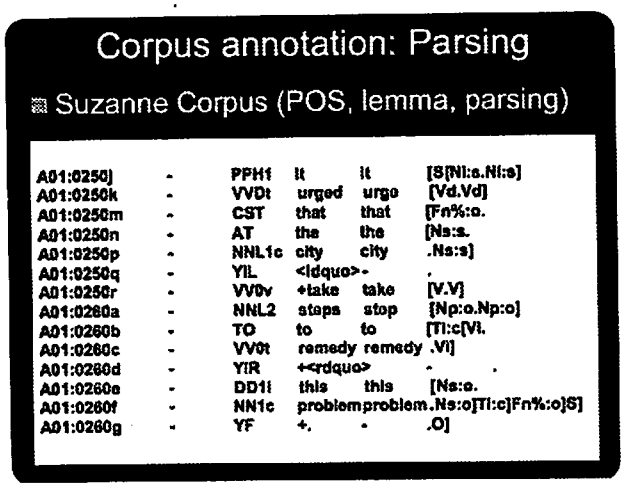
今度は言語的なannotationというのはどんなものがあるかというと，品詞のtaggingというのがあります。品詞タグ付与ですね。これは，いま，世界で最も大きいバランスのとれた，「均衡corpus」といいますけれども，その1億語のcorpusがイギリス英語であるのですが，そのBritish National Corpus (BNC)というものの内容です。

スライド2を見ていただくと，1個1個の単語の前の部分にangle blanketといいますが，<w PNP>とか書いてありますね。こういうものがタグなわけです。この品詞のタグを手がかりに，一括で，例えば動詞の形を全部取ってくるとか，その連鎖を取ってきたりということもできるのです。

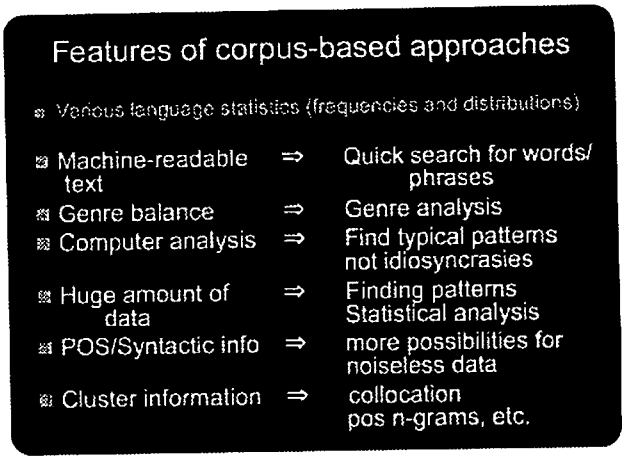
続けて，同じようなことにlemmaの情報をつけることができます。例えば，**スライド3**では1行に1単語ベースのフォーマットになっています。You/must/leave/now/otherwise/you/will/be/late/for/your/social/studies/class.とか。そうすると，1単語ごとに品詞がついて，これがlemmaになっているわけです。studiesというところがstudyになったりしています。そうすると，見出し語検索をすれば，複数形でも単数形でも一括で取ってきたり，そのようなことが全部できるわけです。



スライド4



スライド5



スライド6

もう少し進んだ研究をしている人たちもいます。彼らはsemantic fieldというタグをつけているのです(スライド4)。例えば，ランカスターにはUSAS(UCREL Semantic Analysis System)というタグセットがあって，これはLongmanのLexiconというthesaurusのdictionaryがあるのですけれども，その類義語辞書のような意味領域を単語に振っているのです。このようなことをすることによって，例えば，一般の人が読んでいるhealthと，専門領域のhealthということとを仕分けるようなことが技術的にできるようになる可能性があります。

また，context wordのタグの意味領域を見ることで，専門的なことを言っているのか，一般的なことを言っているのかをtextで分別するような技術が，ここから出てくる可能性があります。

スライド5はparsingのデータです。"It" 以下がずっと元の文なのですけれども，これがlemmaです。その隣の "it" 以下が品詞です。ここは全部構文解析した頭語情報が入っているのです。これはSuzanne Corpusというコーパスの一部分なのですけれども，このようにセンテンスのNODEから，NP，VPのかたまりなどということが入ってきたりすると，構造的な情報を検索したりすることもできます。

このような情報が入ってくると，いろいろなことの検索が可能になってくるわけです。皆さんはもうわかると思いますが，こういうデータを入れたものから，コーパス言語学というのはさまざまな頻度，そして分布の情報を取ってきます(スライド6)。

まず，コンピュータに入っていますから，非常に高速に単語やフレーズのサーチができます。それだけではなくて，ジャンルのバランスなどをとってあれば，こういう領域ではこんな使い方，みたいなこともできます。

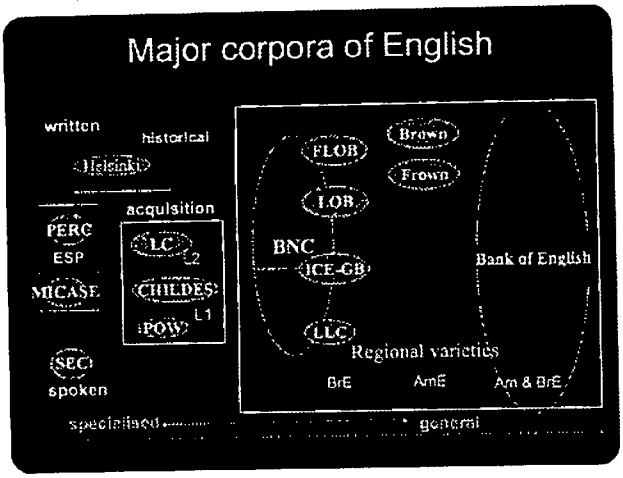
次に，コンピュータ処理しますから，人間が気づかないようなパターンを取ってきたりすることができます。

それから，ものすごく量が多くなってくると，いろいろな統計処理などをして，さまざまなデータのsummaryの仕方ができます。

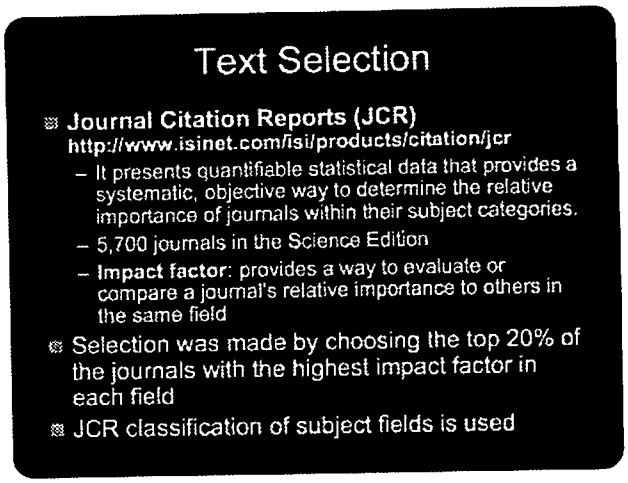
それから，単純に単語レベルで検索していたり，あるいは文字列処理しているだけだと正確に取ってこられないのですが，それに品詞や構造，構文解析のデータがついていれば，もっとノイズの少ない，正確なサーチができるようになります。

また，1単語だけではなくて，クラスタの情報も取ってこられます。ですから，どういう単語とどういう単語が連結しているかとか，どういう品詞とどういう品詞がよくくっついてくるかとか，品詞と単語の連鎖なども取ってくることができます。

以上がコーパスのおおよその特徴なのですけれども，英語のコーパスというのは世界中にどのぐらいあるかというと，いまはものすごくたくさんあるのです。前は自分でリストをつくっていたのですけれども，いまはあまりにもたくさんいろいろなコーパスがあるので，自分でリストをつくっているのが面倒くさくなってしまったぐらい，たくさんあります。



スライド7



スライド8

コーパスの分類

大きく分けると，書き言葉のコーパス(written corpus)，それからspoken corpusとに分かれます(スライド7)。それから，どちらかというとgeneralな汎用のコーパスと，specialized corpusといいますけれども，特殊領域のコーパスに分かれるのです。汎用のほうは，大規模なものが多いです。BNCは1億語，Bank of Englishがいまは4億数千万語と言われていますが，つい最近，Oxfordが Oxford English Corpusというのをつくっていまして，来年ぐらいに完成する予定なのです。そのコーパスのサイズは10億語です。ものすごい規模のコーパスの構築がいま盛んになってきています。

特殊コーパスのほうは，集めにくかったり，技術的にspokenだと書き起こしに手間がかかったりで，小さいものが多いです。ESP関係は，このあたりの領域に位置づけられるわけです。EAPなども含めれば，いくつか公開されているものがありますけれども。

医学英語というのは，実は私は門外漢で，今回も発表の中にたくさん難しい単語が出てきて，私の知らない単語がたくさんあったので，先生方に「これは何という意味？」と聞かれると恥ずかしいな，と思ったりしたのですが，Medical English Corporaというのは，私の知るかぎりではあまりたくさん公開はされていないのです。

1つは，歴史的な研究をしている人たちが公開しているものがあります。The Corpus of Early English Medical Writingというのがあります。これはCD-ROMでJohn Benjaminから出ているのですけれども，ご存じの方もいるかもしれませんが，昔どのように英語を使っていたかという，歴史的なものです。

あとは，コーパス言語学の学会に行っても，ESP関係でMedical Englishという発表は多いですが，ほとんどが自分でインターネット等で取ってきているものばかりです。ですから，コピーライトをクリアーしているような試みはあまりないのです。私的なinstituteの中でキープしていたりして，だれも使えないというデータがたくさんあります。

PERCとCPE

そこで，われわれはそのような壁を1つ打破しようということで，いま，大きなプロジェクトを立ち上げています。それがPERCという団体と，CPEというプロジェクトなのです。これをザッとご紹介したいと思います。

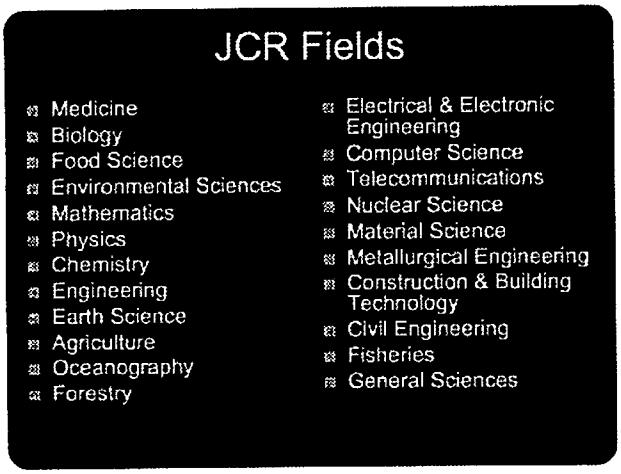
PERC(Professional English Research Consortium)は2002年に，劇と広くProfessional Englishというものに関するさまざまな研究をするようなassociationとして発足しました。コーパスだけではなくて，例えば教材作成やテスト作成，広くProfessional Englishの研究というものをする団体です。〈http://www.perc21.org〉

一応，日本でできたのですけれども，いまは会津大学のThomas Orr先生がヘッドになっています。世界的には，例えばミシガンのJohn Swalesとか，Ulla Connorとか，そのようなESP関係の大御所もPERCのメンバーとして活動してくださる予定です。

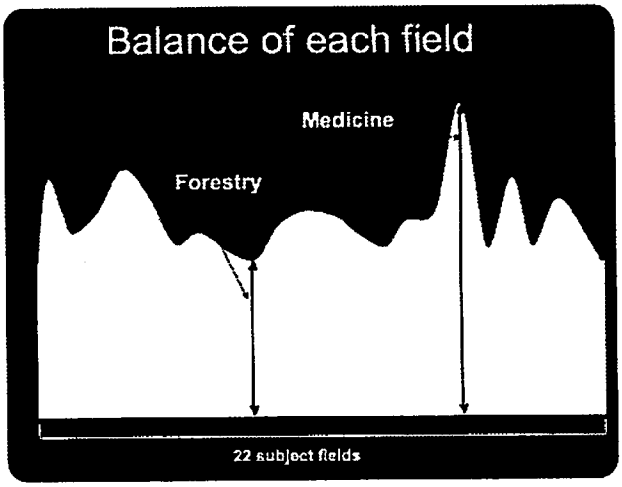
いま，PERCの活動の中心になっているのが，このCPE(Corpus of Professional English)ということで，科学技術系の英語に関してコーパスづくりをしようということになって，私がコーパス言語学のほうで，Professional Englishの専門家ではないのですけれども，コーパスづくりという意味で参画しています。

そして，さまざまな形のリサーチを，このデータを使ってやっていこうではないかということで始まっているのです。ただ，なかなか時間がかかって大変な作業なので，まだ完成には程遠いのですが，第1段階はほぼ終わってきたのでご紹介しています。CPEは，もともとはさまざまなコーパスのtextのタイプをバランスよく集めようということを考えています。例えば，医学なら医学なのだけれども，journalだけではなくて，この分野のさまざまなことをサンプルとして1つのコーパス・データとして持っているようなことをイメージしています。ただ，最初はacademic journalsを中心に収集しています(スライド8)。

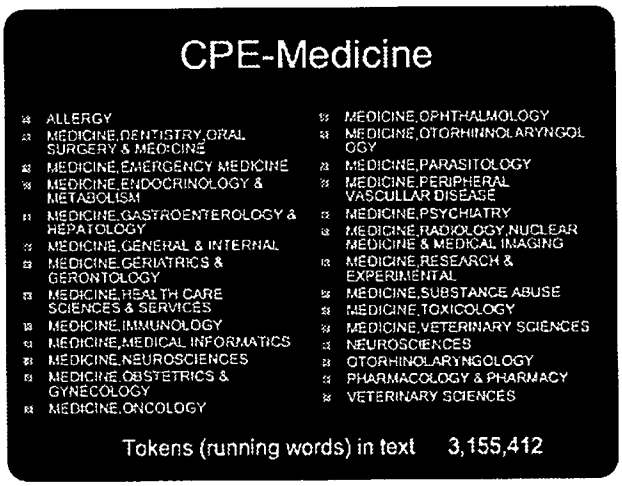
textセレクションは，Journal Citation Reportsというもの



スライド9



スライド10



スライド11

を元にして，ここのインパクトファクターが非常に高いjournalを選定して，サイエンスのエディションだと5,700誌ぐらいがこのレポートに載っているのですけれども，そのうち上位20％のjournalを選んでいます。そしてその出版元にConsortiumとして手紙を書いて，○○年から○○年までの全部のarticle（あるいは論文）を研究用に提供してほしい，というようなことをしています。

JCRのfieldsは，大体22 fields（**スライド9**）になります。medicineは，そのうちの1つなのです。科学技術関係の割と総合的なコーパスというようなデザインになっています。

バランスに関しては，実は22 fieldsをやってみると，JCRでも随分違うのです。例えば，medicineは非常にJCRに出てくる率が高いとか，あるいはforestryは低いとか，そのようなことがあります（**スライド10**）。

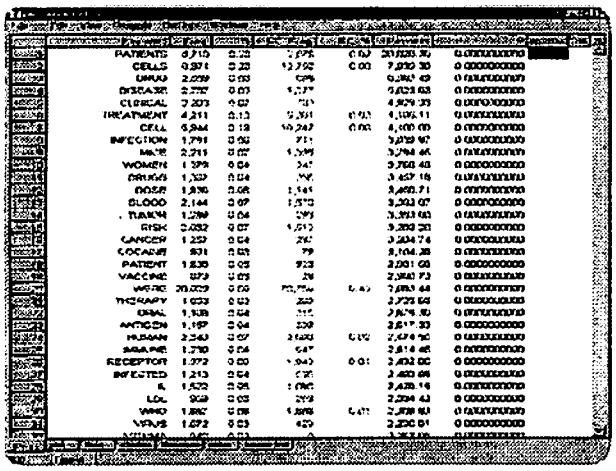
そこで，コーパスのデザイン的には，まずacademic journal textを，例えば30万語ぐらい必ずどの領域にも集めておく。そのあと，70万語ぐらいは他のtext typeを集めて，少なくとも100万語規模のものを22 fields集めると。そうすると，2,200万語になりますから。このようなもので，100万語単位の比較が科学技術分野に関してできるようなものを，まずベースに持っています。そして，それに集められるだけ上乗せしておいて，使える方は全部medicineを使いたいとか，そのようなことがあってもよいだろうというようなデザインになっています。

さまざまな出版社にコンタクトをとって，そして非常に大きな反響がありました。例えば，いま大御所だったらElsevierとか，IEEEとか，そのようなところからもコピーライトの許諾を得ています。ただ，彼らはテキストファイルで提供してはくれないのです。勝手にウェブページにアクセスして，取っていっていいよということは言うのですけれども。

われわれがずっとこの2年ぐらいやっているのは，PDFファイルをe-journalのサイトから取って，そしてそれを全部テキスト処理でダウンロードしたものをコンバートして，そしてクリーニングするという作業をしています。この部分の作業がものすごく大変で，これで何本も論文を書いているぐらいのものなのですけれども，PDFからテキストへの自動変換というのは非常に難しいのです。簡単だ，ソフトもあるよ，と言う人がよくいるのですけれども，そんな生易しいものではなくて，さまざまなゴミが出るのです。そのようなものをクリーニングしながら，皆さんがある程度使えるようなフォーマットにまでするのが一苦労という作業を，ずっとしてきました。

今回，CPEのfirst releaseというのがそろそろ出せるような状態になってきました。これはacademic journalだけです。ですから，先ほどのバランスはとれていません。ただ，すばらしいことに，すべてコピーライトは許諾を得ているのです。そして，さまざまな出版社から，いまのところ3,700編の学術論文を取れています。いまの規模ですと，1,700万語ぐらい，これは科学技術英語全体ですけれども，それが22 fieldsに分かれています（**スライド11**）。

medicineはどんな内容になっているかというと，GSERのブレークダウンなのですけれども，スライド9のような分野のものが取れていて，journalはまたこの下に何種類か



スライド12

blood: grammatical relations

- Triplets:
  - <subject, flow, blood>:
    - "*blood*" is the subject of "*flow*"
  - <object, suck, blood>:
    - "*blood*" is the object of "*suck*"
  - <modifies, transfusion, blood>:
    - "*blood*" modifies "*transfusion*"
  - <a_modifier, blood, cold>:
    - "*cold*" is an adjective modifier of "*blood*"

スライド13

入っているのだと思います。

現在，われわれがつくっている全体のコーパスの中で，medicineのセクションは300万語ぐらいです。ですから，それを大きいと思うか，小さいと思うかは皆さん次第だと思いますけれども，少なくともこのぐらいの規模のものがいま集まっています。

CPEのavailabilityということですけれども，このfirst releaseというのは，まず最初はPERC researchersという人たちが使って初期の研究をします。public releaseというのは，web-basedなインターフェースで，将来，来年の初めぐらいにオープンにしたいと思って，いま作業中です。小学館コーパスネットワーク〈http://www.corpora.jp〉という，小学館が企業としてここの科学技術関係の部分のサポートしてくれているので，そのようなサイトをオープンする予定です。

もう皆さんも知っているかもしれませんけれども，BNCとか，COBUILDのデータとか，大きなコーパス・データが，この小学館のサイトではweb-basedで利用できるようになっています。ですから，web-basedな，コーパスのポータルみたいな感じに使っていこうということです。

コーパスとEMP

では，残りの20分ぐらい，コーパスとEMPという話をしたいと思います。先ほどのようなデータができてきて，われわれがそれにアクセスできるようになってくると，どのような分析が可能で，そしてそれを元にすると，例えばどんな可能性があるのだろうかというようなことを，ちょっと考えてみたいと思います。

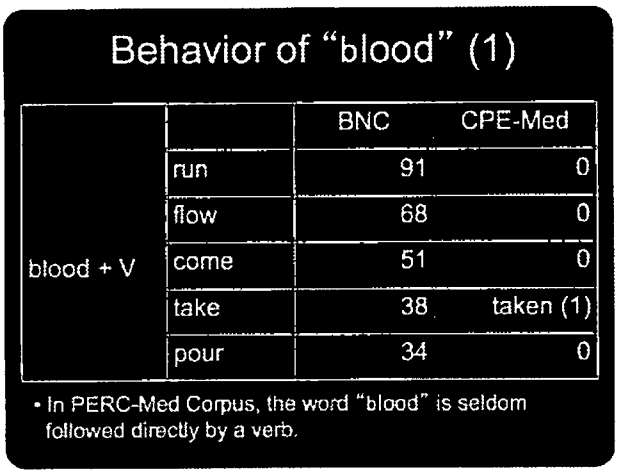
実は私の分野というのは，コーパス言語学といってもバリバリにコーパス処理の言語学的な部分ではないのです。私はどちらかというと，言語教育とのリンクを考えるようなことが専門なので，医学英語教育という意味では，私のやっているようなことを医学英語に適用するようなことをやってごらんになったらよいかな，と思うことがいくつかあります。そういう研究分野では，実はコーパスというのは，皆さんもおわかりだと思いますが，直接的に利用する場合と，間接的に利用する場合があるのです。そういうことについて，ちょっとお話ししたいと思います。

それから後半のほうでは，available corporaということで，すでにつくられているコーパスを使うというようなことと，あとはDIY (do it yourself) corporaですね，自分でコーパスをつくるような技術が，最近発達してきているのです。これはたぶん，インターネット等に興味のある方ならば，必ず役に立つ技術だと思いますから，ぜひメモでもしていってください。

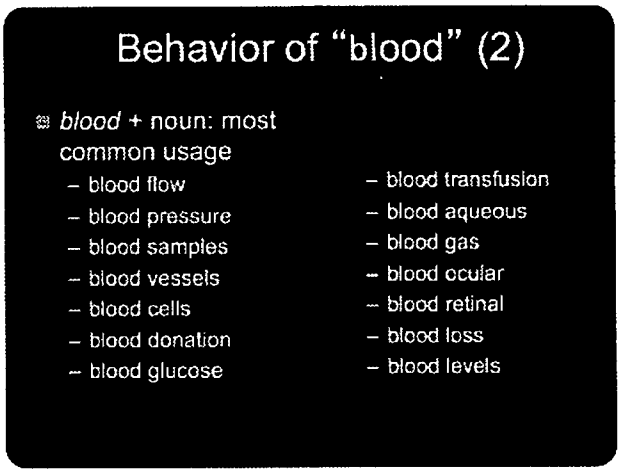
まず，間接的にはどんなことが可能かということの例をいくつかお見せします。まずはEMPのmaterials developmentということで，たぶん語彙リストみたいなものをつくっている方がいますね。コーパスをうまく使えば，非常に有効に語彙リストがつくれます。それから，そこから発展してlexicon database，単なるword listではなくてlexiconとしてそういうもののデータをつくっていくというようなことも可能になります。それから，reference toolsだったら辞書や文法書，それから教科書，テキストですね，それから言語テスト。先ほど，言語テストの話をされていたみたいですけれども。

一般の英語教育のほうでされているようなことをちょっと見て，それを元に，医学英語だったらどうかというようなことを考えてみたいと思います。

コーパスを使った語彙リストにはいろいろな技術があるのですけれども，いまは非常に目的特化したtextを集めます。そして，そのtextでどんな単語が使われているかを自動抽出するような技術があります。それはreference corpusといって，大量の一般的な英語の文章と比較することによって，その専門分野で特によく使われている特徴語を抽出するような，そういう手法があるのです。それぞれのコーパスから語彙リストをつくって，医学英語のコーパスから



スライド14



スライド15

頻度表をつくる。そして，そのreferenceになる大型コーパスから，無色の，割とニュートラルなリストと比べてみるわけです。そういうことによって，medical Englishに特徴的な語彙を取り出すということです。このようなことを，いまはコーパスではソフトウェアで簡単に行うことが，技術的には可能になっているのです。

これには，いろいろな統計の手法を使って抽出する，例えば代表的なものはlog likelihoodとか，そういう相対頻度のずれの度合いの大きいものを抽出するものを使います。

そうしますと，例えば先ほどのmedicineのセクションをCPE(科学技術英語)全体と比べてみたりすると，medicineの300万語のセクションのキーワードは，**スライド12**のような感じで出てくるのです。medicineのtext properに出てくる特徴語が，特に特徴の強いものから順番に並んだりします。なかには専門用語，専門用語との中間，機能語的なものももう少し下へ行くと出てきますけれども，医学英語の特徴をこういう形で機械的に抽出してみて，そしてわれわれが気がつかなかったようなポイントをコンピュータ的に取り出すようなことが可能になります。

これだけですと，単にword listを分野ごとに細かくつくる，というような発想になりますね。それだと単語単体のことになってしまうので，どのようにほかの単語と一緒に使うかというようなことがわかりません。

そこで，もう少し細かく文法的な関係，これをgrammatical relationsというのですけれども，そのようなものをしっかり見たり，それからcollocation分析みたいなことをすると，もっと単語の使い方が詳しくわかるようになります。

例えば，"blood"という単語を例にとると，一般の英英辞典であるMacmillanの辞書では，こんなことが書いてありました。1つめは普通の「血液」です。2つめは「血統」とか，町，地域，グループとしての，仲間意識としての「血」みたいなことでしょうか。そして3つめがviolenceやdeathを意味するという一般の英語で使われる"blood"の定義があります。

そして，一般英語のコーパスを見てみますと，"blood"というのは割と"blood"単体で出てきて，そして動詞と結合するようなものがたくさん類文で出てきます。ですから，一般の英語で"blood"と使う使い方というのは，動詞とのcollocationなどがとても多いのです。

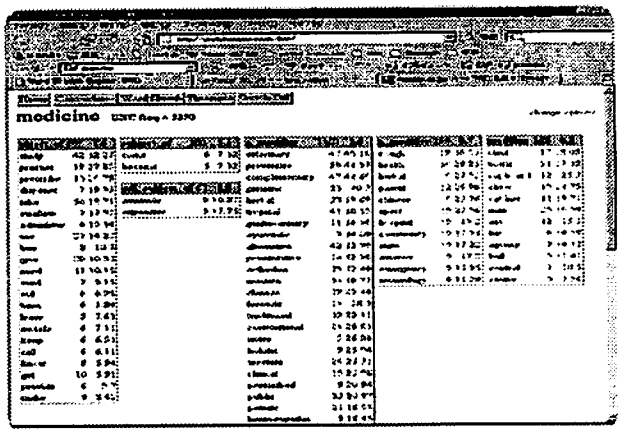
ところが，これをもっと詳細に分析してみるわけです。例えば"blood"というのがどのような関係をしているかというのを，tripletという形で取り出すわけです(**スライド13**)。「bloodはflowのsubject(主語)になる位置関係にある」というのを<subject / flow / blood>というtripletとして取り出します。このようなことを，機械的にやるわけです。そうすると，"blood"という単語の使われる場面や状況を，grammatical relationsの束で表そうとするわけです。こういうことをたくさんやると，"blood"という単語がどのように使われるかということを，もっとtextを分析した結果としてデータベース的に取り出すことができるのです。こういうものを取り出すと，かなりメリハリがわかってきます。動詞との連結の例を見るだけでもかなり違うのです。

BNCではこういう動詞との結合がたくさん出てくるのですけれども，おもしろいことに，medicineのコーパスではこういう例はほとんど出てこないのです(**スライド14**)。つまり，"blood"単体で結びつく動詞との連鎖は，medicalなtextではほとんど出てこないのです。逆にどういうものが出てくるかというと，"blood"というのはほとんどがcompoundで出てくるのです。

スライド15は名詞との連鎖で，大量に先ほどのmedicineのテキストから抽出したものなのですけれども，この"blood flow"のflowは動詞ではなく，名詞のflowなのです。このようなmedicalなテキストでは，"blood + noun"という形での結合が圧倒的に多いのです。単純に動詞とbloodが結びついている例というのは，ほとんどありませんでした。

これを見ると，われわれが知っている"blood"などという単語でも，医学では全く違う使い方をするということがわかります。こういう部分というのは，案外われわれにとっ

スライド16



スライド17

て，というか普通の英語を勉強してきた者にとっては難しい点というか，あまり気がつかない点ではないかと思います。

実は，これを大量にデータベースとして皆が使えるような研究をしている人がいるのです。Brighton大学のAdam Kilgarriffという人ですけれども，この人はLongmanの辞書，Macmillanの辞書，そしていまはOxfordなどを手伝っている自然言語処理(natural language processing)の研究者で，lexical databaseなどに非常に詳しい人です。この人はSketch Engine(http://sketchengine.co.uk)という先ほどのgrammatical relationsみたいなものをウェブ上で検索できるようなシステムをつくっているのです（**スライド16**)。

これは非常にパワフルなもので，例えば“medicine”という単語を検索すると，これはBNCなので，ESPのテキストではないのですが，それでも，例えば“medicine”というのは，“object of”ですから，こういう動詞の目的語になって現れる。“study medicine”，“practice medicine”，“prescribe medicine”とか，そのような目的語の関係。こちらはどういう形容詞がくっつくか。それから，どんなcompoundをつくるかなどということが，データベースの形でガーッと出るのです。

そうすると，この“medicine”という単語の使われ方がチャートになっていて，そしてこういうところのstudyの42番というところをクリックすると，具体的に例文に飛ぶわけです(**スライド17**)。そうすると，どのような使われ方をするかが一度にわかるわけです。これをESPのコーパスでやったら，とてもおもしろいのではないでしょうか。そうすれば，一般のgeneral Englishではなくて，ESPのtext，EMPのtextだったらどうなるかというようなことがわかれば，その差がかなりあると思うのです。そういう差を研究するようなことも，必要だと思います。

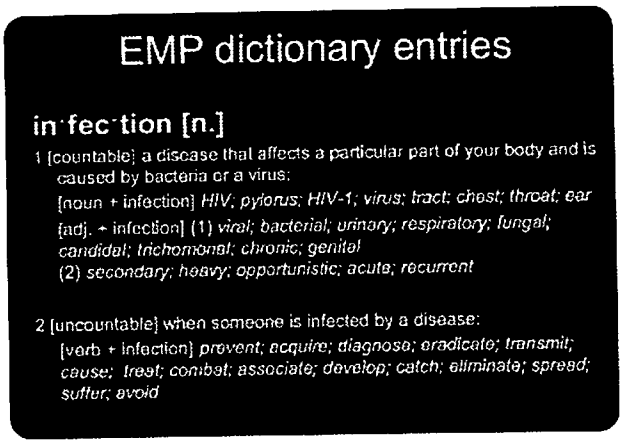
いま，コーパスを使った辞書というのは国内外でたくさん出ているのです。ですから，たぶん先生方も先ほどのようなことを元にして研究すれば，もっと使いやすいEMP用の辞書というのをつくれるはずなのです。いまのEMP用の辞書は，私はよくわからないですけれども，専門用語の辞書は単にterminologyの辞書ですよね。そういうものが圧倒的に多いと思うのです。ですから，本当の使い方がよくわからないものが多いと思うのです。

例えば“infection”という単語は，一般の英英辞典ではan ear infectionという用例が1つ，それからあとはmild/slight/severeなどの程度を表す形容詞と一緒に使われているとか，この程度の情報量です。これを医学生が聞いても，とてもではないけれども，彼らは“infection”の医学における全体像はわからないでしょう。

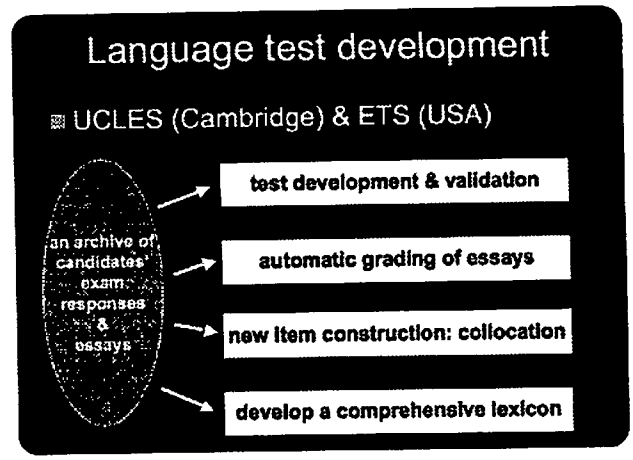
ところが，先ほどのようなデータベースを駆使して研究した結果，何かしらそういうものを元にしたテキストをつくれば，例えばですけれども，“infection”の2つの意味で，特にこちらのほうの意味では，どのような名詞とcompoundをつくるかとか，adjectiveだったらどんな種類のものと結びつくかなどということを整理したりできるかもしれません(**スライド18**)。これは，ちょっと不正確なものもあるかもしれません。ですから，これは1つの例だと思って見てください。こちらでは，動詞との結びつきです。そういうものも，“infection”にくっつく動詞などがあって，これがもしも訳語と簡単な例文つきなどで出てくれば，相当活用できるresourceになるのではないでしょうか。

コーパスを用いた辞書だけではなくて，いまは文法書もどんどん出始めてきているのです。先生方はどうお考えになるかわかりませんけれども，医学英語には医学英語なりの文法があるのでしょうか。そういうことは，ESPなどでもいろいろな議論があると思います。もしも一般の文法とは少し違う文法の特徴があるとすれば，そのようなものをコーパスで調べて，コーパスに基づくような医学英語の文法書があってもよいかもしれません。maybe ENP grammarということですね。

私は，コーパスを使ったテキストをたくさん出しています。NHK関係もそうですし，ほかにもいろいろなところから出ています。こういうものは，先ほどのiPodとか，iTunesとか，ああいうところにも私の教材が載っているの



スライド18



スライド19

ですけれども，そういうコーパスに基づくテキストという発想があれば，ひょっとしたらコーパスに基づくEMPのテキストもあってもよいかもしれません。私のような中高生モード，一般モードでなくても，もっと医学生向きの何かがつくれるかもしれません。

そして，Language Test Developmentということですが，実はいまは世界的な規模でいくと，CambridgeのUCLES，USAではETS，この辺が大量にコーパス・データを集めています(スライド19)。5年ぐらい前には，そういうことはしていませんでしたが，最近は彼らは非常にそういうことについての重要性を感じ始めていて，テストを受けた人たちのexam responsesとかエッセーなどを大量にアーカイブしているのです。そして，それを元に，いろいろな形のことをしています。例えば，自分たちのつくったテストの開発，そしてvalidationのためにそのようなものを参考にしたり，それからエッセーを自動gradingするということですね。これはETSなどがいま盛んにレポートを発表していますけれども，こんなこともやっています。

それから，新しいitem constructionというのをcollocationでやるようなことを，UCLESでやっています。つまり，collocationの知識で測ったほうが，単語単体の知識よりも弁別性があるというようなことを，彼らはレポートで言っているのです。

それから，comprehensive lexiconということで，このようなものを元にして，どのような単語の知識のデータベースがつくれると，それを元にテストが開発できるか，そのようなものを総合的につくろうというような発想も，テスト分野の人たちの間には出てきています。

いまの繰り返しになってしまいますが，general Englishとmedical関係のEnglishとの間の違いも，当然あるでしょうね。そして，それを習っている医学生が，あるいはお医者さんになる人たちが習っている過程で，どんなつまずきや間違いをするか。きょうも学習者コーパス(learner corpus)の発表がいくつかありましたけれども，そのような発想で，こういう3つのデータを組み合わせて研究するような可能性が出てきます。つまり，目標となるESPのターゲットの様子をしっかり研究するだけではなくて，そこまでに行くまでにどのようなプロセスを経ていて，どこがつまずいているかというようなことを，learner corpusで調べるわけです。これは，spokenだったり，writtenだったりしますけれども。こんなことを組み合わせていくと，EMPのさまざまなmaterials developmentに非常にプラスになるような素材をたくさん集めることができます。

Web as Corpus

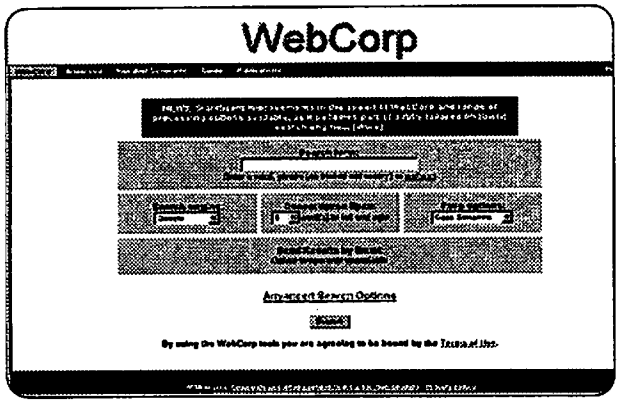
すでにあるコーパスを使うというのは，先ほどのPERCのコーパスなどを使えばよいのですが，そうではなくて，自分でつくりたいという方がいると思うのです。実は，こういう技術が最近ものすごく発達してきているので，ちょっとそのことをご紹介したいと思います。

実は，web as corpora，あるいはweb as corpusというような動きが，コーパス言語学者の中にあるのです。2006年度のAdam Kilgarriffによれば，Googleでインデックスされているduplicate-free(データベースの形になっていない，つまりstaticな形でインターネットに載っている)なtextは，英語の場合は10 thousand billion wordsです。どのぐらいか，日本語だと何兆と言ってよいのかわからなくなってしまうぐらい多い量ですけれども，このぐらいがestimateであるので，本当はどうかわからないぐらいたくさんの量があります。

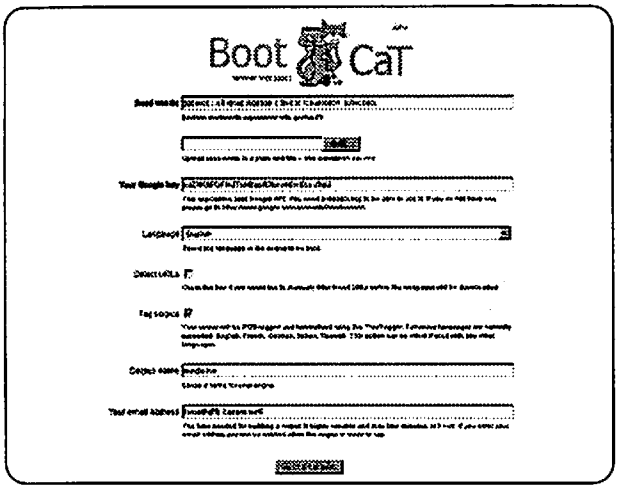
これを何とかして利用しなければ，というようなムーブメントがだんだん強くなってきて，実はGoogleに勤めているFranz Ochさんという人が，250 billion wordのトレーニング・コーパスというのをつくったと発表しました。すごい量ですね。ですので，こういうことが実際に研究としては可能になってきています。

そこで，いくつか既存の技術で皆さんが検索できるようなcorpusの使い方があるかというようなことを，ちょっと説明します。

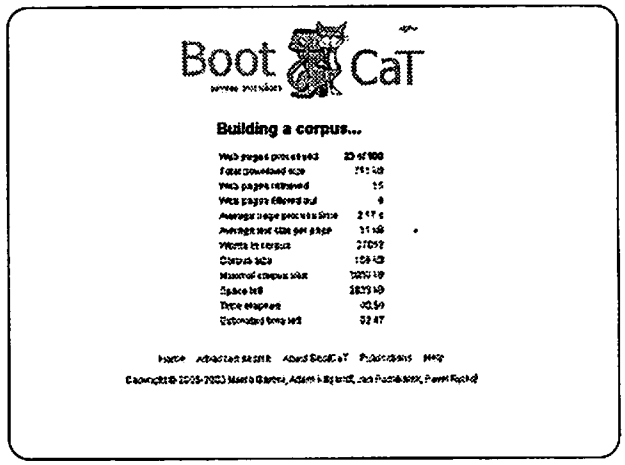
1つはWebCorp(http://webcorp.org.uk)という，これは



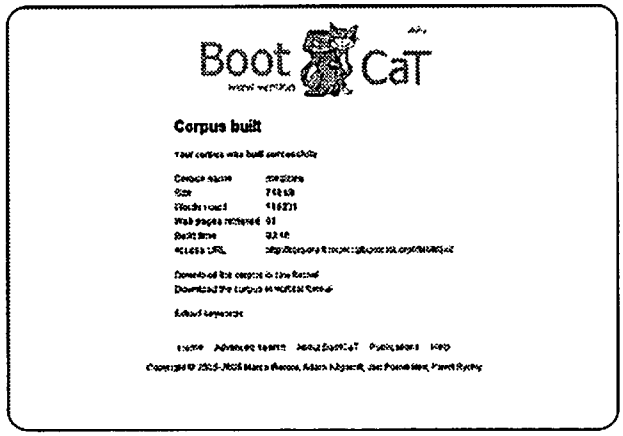
スライド20



スライド21



スライド22



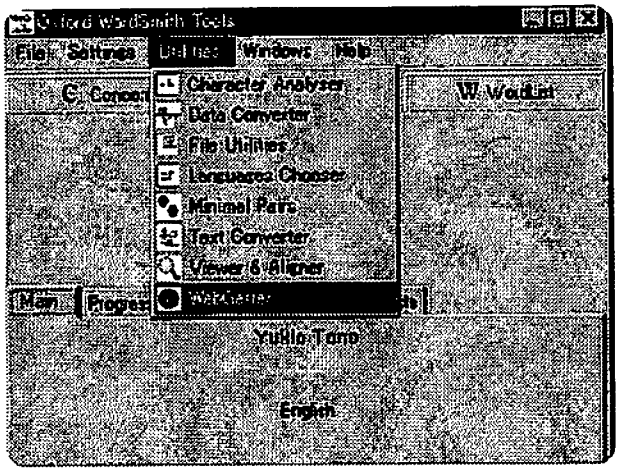
スライド23

Antoinette Renoufという人が中心でやっています（**スライド20**）。Liverpool大学から最近Central England大学へ移りましたけれども，これはウェブ上のGoogleなどのresourceを元にキーワードを検索して，それを表示するようなサービスです。それも，ウェブページをコーパス的に使おうというような発想のインターフェースなのです。

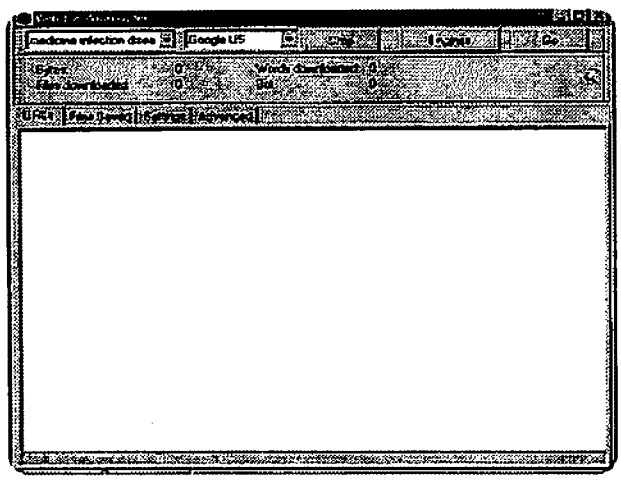
WebCorpを立ち上げてsearch termというところにご自分でキーワードを入れるわけです。これを組み合わせたりすることもできますし，自分なりにパターンを書いたりすれば，かなり医学的なtextだけを絞って取ってくるようなことができます。サーチエンジンを選んだり，いろいろなことをやって，submitを押すのです。ただ，ウェブをクロールして集めてくるので，ちょっと時間がかかります。ですから，WebCorpの1つの問題は，遅いということなのです。ただし，チョンとやっておいて，少しどこかへ行ってから戻ってきたりすると，このようになっているわけです。中には「何だこりゃ」と思うようなものも出てくるかもしれません。そういう場合には，もともとのURLへ行って，そこが信頼できるところかどうかということをチェックしてくればよいわけです。しかし，少なくともこういう形でたくさんのフレーズを一度に集めたりするようなことは，インターネット上でもできます。こういうインターフェイスのほうが，Googleよりもごみがなくて，比較的よいかもしれません。

次に，まだ開発中なのですが，すごいものが出てきました。BootCatというシステムです（**スライド21**）。これはBootstrapping Corpora and Termsというのですけれども，これは先ほどのような分野ごとの特徴語みたいなもの，「このキーワードを10個ぐらい入れれば，絶対に関連するtextがばっちり取れる」と思うようなキーワードのリストをつくりますね。そういうものをseedwordsといいます。そのseedwordsを元に，tuplesという，いくつかペアをつくるのです。例えば，3つずつぐらいのペアを，10個のうち最初の3つ，1，2，3。次には2，3，4。次には3，4，5というように，組み合わせた単語のリストをどんどんGoogleにほうり込みます。そうすると，その組み合わせでGoogleが取ってきたものの中の集合体みたいなものができますね。それをコーパスとして考えようという発想なのです。そして，Googleのヒットを，これはAPIでヒットしたヒット結果を元に，自動でダウンロードするようなことをするわけです。

このBootCatは現在開発途中なのですが，彼らは何とすべてをウェブインターフェースでやろうと考えています。



スライド24



スライド25

ですから，われわれ利用したい人がそのseedwordを入れれば，クロールしてきて，コンピュータでサイトから集めてきたウェブページを保存して，クリーニングして，そしてすべてコーパスとしてそれをフォーマットして見せてあげるというところまで，全部やるのです。すごい技術でしょう。このようなことを，いまWACCI(web as corpus cool initiative)という団体が中心でやっているのです。

どのようになっているのか，お見せしますと，私がmedicineのセクションから取ってきた“patient”“cell”“drug”“disease”“clinical”“treatment”“infection”などというキーワードトップの単語を入れたとします。これにはGoogleのAPIを使う，Googleに登録したキーが必要ですけれども，英語のデータを取ってきてくれ，それを“medicine”という名前にするよ，私のメールアドレスはこれですといって，submitするわけです。

そうすると，コーパスをつくっていますということで，ウェブページをプロセスしたのがこれだけで，というようなページが出て，自動的にこの数字がどんどんふえていくわけです（スライド22）。そして，corpus built(コーパスができました)というメッセージが表示されれば完了です(スライド23)。実は先ほどありましたけれども，いまのαバ

スライド26

ージョンはテスト版で出ているので3メガバイトまでしかつくるパワーがないのです。だけれども，この範囲でここのURLにアクセスすると，何と，単にtextをダウンロードできるだけではなくて，すぐに検索できるインターフェースに飛んでいけるのです。つくってくれたものを，すぐにコーパスの検索ソフトとして使用できます。そうすれば，“infection”というようなものの動き方などが，全部コーパス上でさまざまに分析することができます。

それだけではなくて，collocationの処理等もできます。例えば「動詞＋cancerという形を取ってこい」というと，動詞のところがハイライトされて，cancerとどんな動詞がくっつくかとか，そのようなこともわかったりします。

このようなインターフェースは，実はSketch Engineとも連動しています。将来的には，何と，先ほど私が紹介したgrammatical relationsみたいなものも瞬間的につくって，そしてデータベースとして同時に使えるような技術をいま開発中なのです。大量のデータにこういうことができるようになるには，まだちょっと時間がかかると思いますが，少なくとも先生方が教室内で非常に特定の文脈の医学トピックのようなものをseedwordにして，このtextを世界中から取ってきて，ミニ・コーパスを使って授業をしたりということは，すでに可能な時代になってきているということです。

似たようなことは，実はいちばん新しいWordSmithのバージョン4という，Windowsのソフトでもできるのです。WebGetterという機能を使うのですが，WebGetterというのを選びますと(スライド24)，こんな画面になって(スライド25)，そこに先ほどの“medicine”とか“infection”というキーワードを入れるわけです。そして「Googleで取ってきて」というようにやると，どんどん世界中のこういうページがあるところをクロールしていきます。そうすると，先ほどのWordSmithですぐに検索できます。ただしこれは，先ほどのように文法のタグ等はつきません。先ほどのBootCatのほうは，文法のタグまでつけるので，それがすごいの

ですけれども。こんな感じで手軽に，このWordSmithなどは1万数千円のソフトですけれども，こういうものを使うだけでも，すぐにかなりのことはできます。

こういうことを元にすれば，直接医学生がソフトを使って，そしてコーパスにアクセスして，そのコーパスでいろいろ調べるようなことが現実味を帯びてきます。インターネットさえあれば，自分でキーワードを入れて，瞬時にコーパスをつくったりすることもできるかもしれません。また，そういうものを元にしてエッセーを書いたり，いろいろなことをする指導をうまく組み合わせることもできます。こういうものをdata-driven learningというのですけれども，そのようなことを実際にやっている人たちもふえてきています。

ということで，皆さんはコーパスをどう思われたでしょうか。EMPに非常に役に立つ部分があると思うのです。まだあまりたくさん行われてはいないかもしれませんけれども。手間暇をどのぐらいかけるかということでは，ある程度テクニックさえ習えば，1人でもできます。それほど大変なことではないです。大変だと思ったら，われわれのようなコーパス言語学者と組んでやればよいわけです。そして，international committee，先ほどのBootCatみたいなものは，世界中の人たちが「おれもやりたい」といって参加して，オープンソースでやっているのです。あの辺の技術は，全部無料です。ですから，そのようなことを皆さんで一緒にやれば，resourceを共有したりできますね。Medical Englishというのは，たぶん世界中共通なところが何かあるのでしょうから，そのようなことをやってみるのもよいのではないでしょうか。どうもありがとうございました。

(2006年7月16日，ウェルシティ金沢にて収録)