2. LSDサービス

図1はLSDプロジェクトのホームページである[2]。京都大学薬学部にあるメインサーバのほかに，東京都内と国立情報学研究所にもミラーサーバを設けて運用している。サービス内容は次のように大別され，誰でも無料でWebブラウザから利用できるようにしている。

(1) オンライン対訳辞書であるWebLSD
(2) 専門英語の自習ができるオンデマンド英語教材
(3) 逐語訳EtoJエンジンによるオンライン変換サービス
(4) パソコンにダウンロードできる変換辞書等の提供

これらの辞書コンテンツおよび検索等に要するプログラムは，すべてプロジェクト内で独自に開発したものである。なお，無償公開サーバでは検索された語句や接続状況について監視と解析を行っているので，秘匿性の高い研究開発を行う会社内LAN等でこれらサービスを利用するためのローカルサーバも開発し，有償での提供を行っている。

以下では，これらのサービスの概要を紹介する。

2.1 オンライン辞書WebLSD

WebLSDは生命科学用語に特化したインターネット対訳辞書である（図2）。2006年1月現在で英和49,000語，和英48,000語の見出し語が収録されており，現在では1日平均で7万件以上の利用がある。「音声付き英和・和英検索」では入力された語句が漢字・かな・英語のいずれの文字種かが自動的に判断されて，その語句の出現頻度から算出した重要度に始まり，外国人による英語の発音例，対訳と簡単な解説，関連する語句，英語の用法や例文，共起表現が表示されるようになっている。

この辞書はWebリンク技術を活かし，見出し語でのPubMed（英語の場合）あるいはGoogle（日本語の場合）検索，訳語の逆引き，例文から出典元のPubMed抄録の閲覧などを可能にしている(図3)。したがって，スペルがあいまいなキーワードについて，和英辞書から英和辞書へと移行してPubMed検索を実行するという応用的な使い方もできる。

WebLSDの最大の特長は，共起表現(concordance)を3,000万語のPubMed抄録コーパスからオンデマンドで高速検索してKWIC（Key Word In Context）形式で表示する点である。KWIC形式の共起表現を用いると，任意の単語の前後にどういった別の単語が使われるかを直感的かつ定量的に理解できる。

図1　ライフサイエンス辞書プロジェクトのホームページ

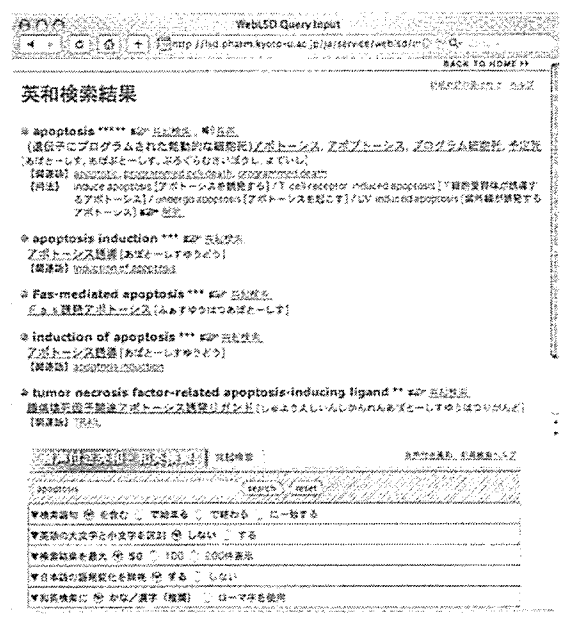


図2　オンライン辞書WebLSD

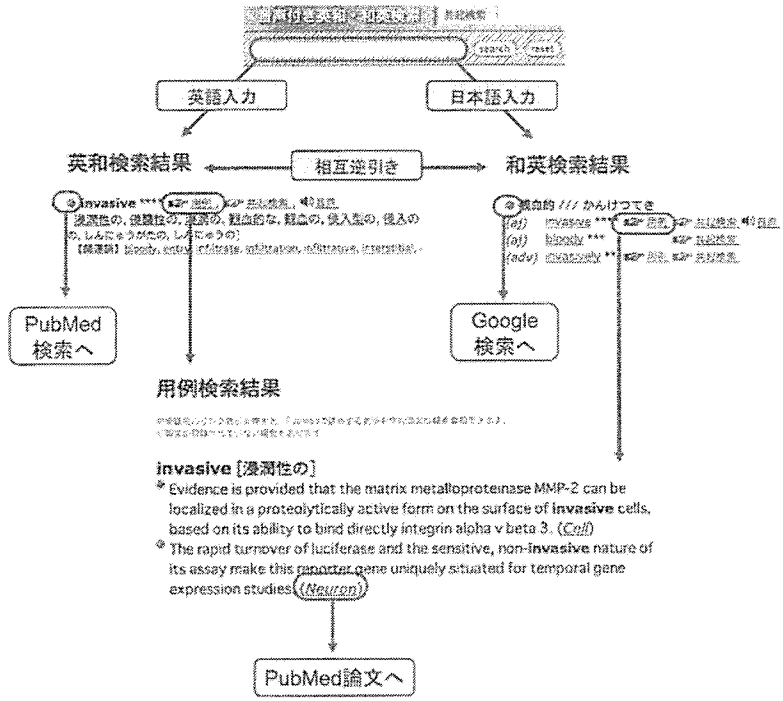


図3 WebLSDの基本動作

例えば「consequence」という単語は「as a consequence of」という構文で用いられる例が多数であることが理解できる（図4）。同様にして，名詞と動詞の親和性や，形容詞および副詞の用い方，あるいは名詞が複数形を取り得るか否かなど，さまざまな英語の正しい用法を知ることができる。ここで解析材料としている英文は，インパクトファクターの高い学術誌に欧米の研究機関から発表された論文抄録のみを厳選しているため，われわれ日本人が英文を書くに当たって非常に有用な資料を提供している。

またWebLSDでは辞書に収録されていない専門用語をユーザーに投稿していただいたり，さまざまなフィードバックを受け付けたりする入口を用意している（図5）。利用者からの新語追加の要望は年間3,000件以上寄せられており，それら候補については出現頻度などの解析を行った上で，改訂時に収録を検討することにしている。

2.2 オンデマンド英語教材

医歯薬学や生命科学を専攻する大学課程において，学部カリキュラムの早期から専門英語教育の必要性が増している。しかしながら，医学英語の学習教材はわが国では十分になく，初学者レベルであっても専門領域と関連した分野の英語を学ぶ教材が医学英語教育の現場で求められている。こうした状況を打開すべくLSDメンバーである京都府立医大の大武博教授を中心に行っているのが，オンデマンド英語教材プロジェクトである（図6）。このページでは，欧米の著名な雑誌等に掲載される医学関連のWeb記事を題材にして，その英文テキストをわれわれが開発した逐語英和訳ツールEtoJ vocabularyを通して読むことができる。このツールは英文テキストの中に日本語での注釈付けを行い，それをWebブラウザのハイパーリンク機能を利用して，注釈を見たいときにクリックするだけで表示するものである。従って，マスコミに掲載されるやや難解で格調高い英文を読もうという読者は，英語で読める範囲では英文を読み，注釈を参照したい場合にはクリックするだけでブラウザの下部に表示される訳語や解説を見て読み進めることができる。現在までに260余の英文を選んで連載を続けており，中には実際に大学院入試で出題された英文も収録している。

2.3 逐語訳EtoJ

日本語は，漢字，カタカナ，ひらがなという3種類の記号が適度に交じり合った「字面」をしており，視覚的にパターン認識がなされて直感的に内容を理解できる日本人にとって都合の良い言語で

図4　WebLSDでの共起表現リスト

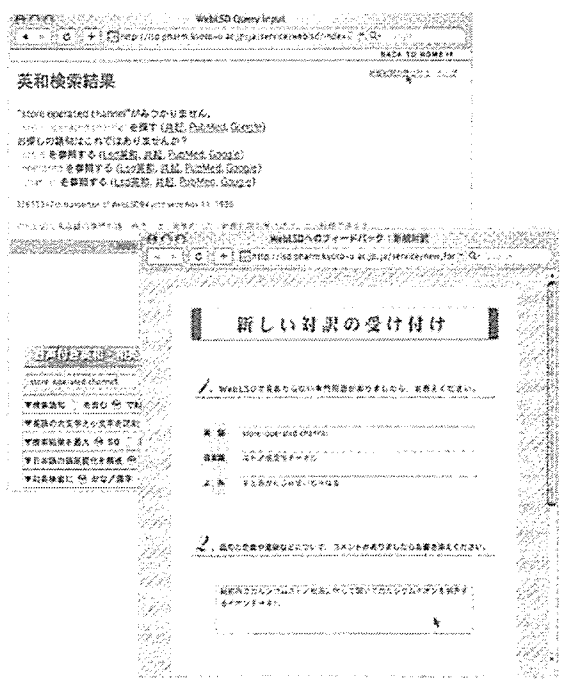


図5　ユーザーからのフィードバック

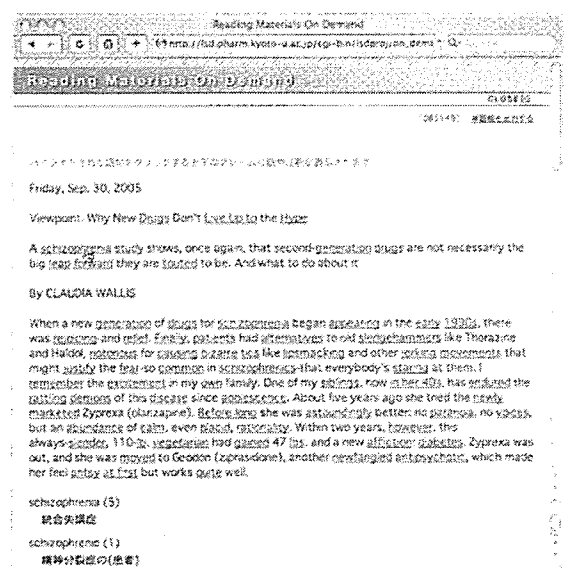


図6　オンデマンド英語教材

ある。一方，英語はアルファベットという記号で表された言語であり，日本人が一見して内容を把握することは困難である。そこで生命科学英語は専門用語だけ翻訳すれば内容のおおよその見当がつくだろう，という発想から開発したのがEtoJ逐語訳エンジン[3]である。これはWebだけでなく，自動応答メールエンジンでも利用できる。

このEtoJではWebLSDとは異なる独特のチューニングを施した辞書を用いており，適度な割合で専門用語が日本語に置換され，構文自体は原文の英語のままという文章が返される（図7）。論文抄録やタイトル情報などを斜め読みしたい場合には意外と重宝するため，あまり知られていないが根強いファンがいるサービスとなっている。このほかにも，先のオンデマンド英語教材でも使用しているEtoJ vocabularyを任意のテキストに対して行うためのサービスやスペルチェックを行うためのWebSpellも公開している。

2.4 パソコンで使える辞書

読み書きと同様に，あるいはそれ以上に日本人を悩ませるのが英語のヒアリングである。WebLSDでも音声例を提供して正しいアクセントや発音を参照できるようにしているが，学習者が最低限の生命科学ボキャブラリーについて，すべての発音を聞けるように工夫したのが「耳で覚えるライフサイエンス英語」である（図8）。このファイルをダウンロードして，アップル社が無料配布しているiTunesで開くと，2,000語の基本用語についてパソコンで英語の正しいスペル，日本語訳，発音を参照できる。さらにiPodにそれらを転送すると，電車の中での生命科学英語の勉強も可能になる。iTunesやiPodはランダムに再生する機能を有しているので，聞き取りテストを試してみるのも良いだろう。なお，初学者向けには6,000語の基本語彙をまとめた書籍も刊行している[4]。

ところで，筆者がLSDプロジェクトを開始した直接の引き金は，昔のパソコンのかな漢字変換辞書であまりにも専門用語の変換ができず，自分でコツコツと変換辞書を作っていたことにある。それに比べると最近のWindows XPやMac OS Xに装備されている変換辞書はかなり優秀になってきたが，生命科学に特化した変換辞書はまだまだ必要とされている。LSDプロジェクトでは，これらのパソ

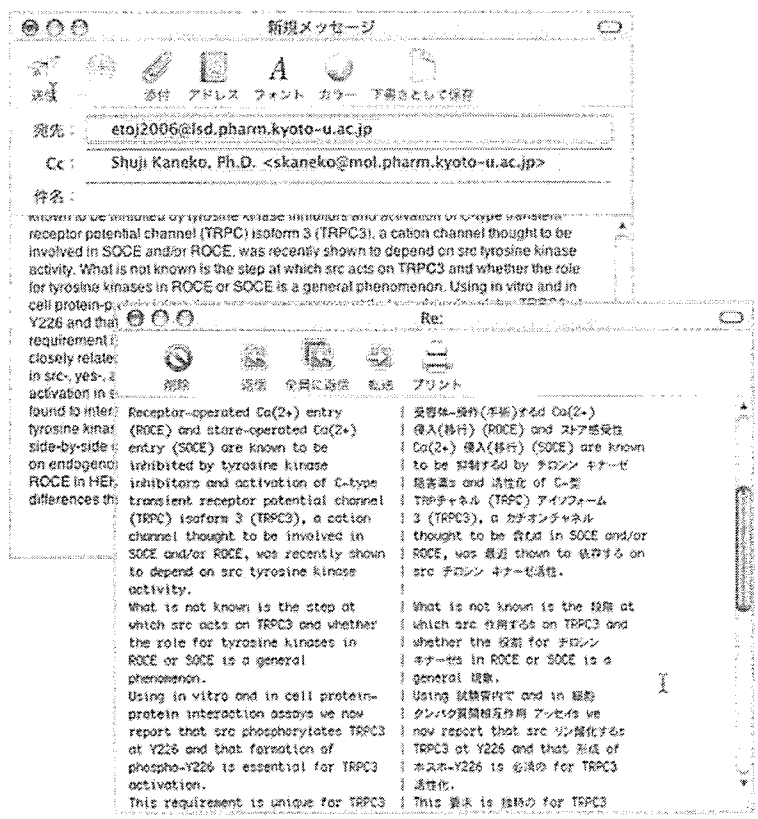


図7 EtoJメール逐語訳サービス

図8　耳で覚えるライフサイエンス英語

コンOSで使える2万語規模の変換辞書を無料で配布しており，日本語の教科書や総説で高頻度に出現する用語を網羅している。

また近年は漢字変換プログラムの高性能化が進み，OSやアプリケーションと連携して辞書ツールとしての機能を有する例が出てきた。LSDプロジェクトでは（株）ジャストシステムのATOK2006に最適化した辞書を開発し，かな漢字変換だけでなく，かな英語変換，英和辞典，和英辞典の4種類の機能をATOKから利用できるようにした（この辞書についてはジャストシステムから販売されている）。これを使うとInternet Explorerに表示されるPubMedページで不明な単語があった場合に，右クリックだけで辞書検索を行うことが可能になる（図9）。今後もさらに高機能で便利な辞書ツールを開発し，わが国の生命科学が発展していくための一助になればと考えている。

3. LSDの構築

以上紹介した電子辞書群は，すべて単一のデータベースを元にして定期的に更新している。その元となっているLSDデータベースは，独自に収集した英語および日本語コーパスを自作ツール等で解析し，得られた用語をリレーショナルデータベースに収録し，最終的には手作業で対訳等を定義している。それらのプロセスについて，概要を紹介する。

3.1 計量的解析に基づく収録語の選択

いくつかの用語集の編纂（へんさん）に関わった経験では，専門家が必要と思われる用語をリストアップし，そうしてできた複数のリストを摺り合わせることで見出し語が決められていた。つまり，日本人が用語集を作る場合は，最初に日本語で考える場合が多い。しかしながら生命科学の論文はほとんどが英語で記述されるので，用語は英語で抽出されるべきである。また，従来の用語集には実際に使われていない英語が収録されている場合も多いことが予備調査から分かっていた。そこで，LSDにおいては英語の論文テキストを計量的に解析した結果に基づいて，用語集に収録すべき見出し語を選定し，死語を可能な限り排除することにした。

3.1.1 英語コーパスの解析

LSDプロジェクトでは毎年，影響力の大きな学術誌からサンプリングした23,000〜24,000論文の抄録テキストを解析して，その中での単語の出現頻度や共起関係を蓄積している。抄録を選択するに

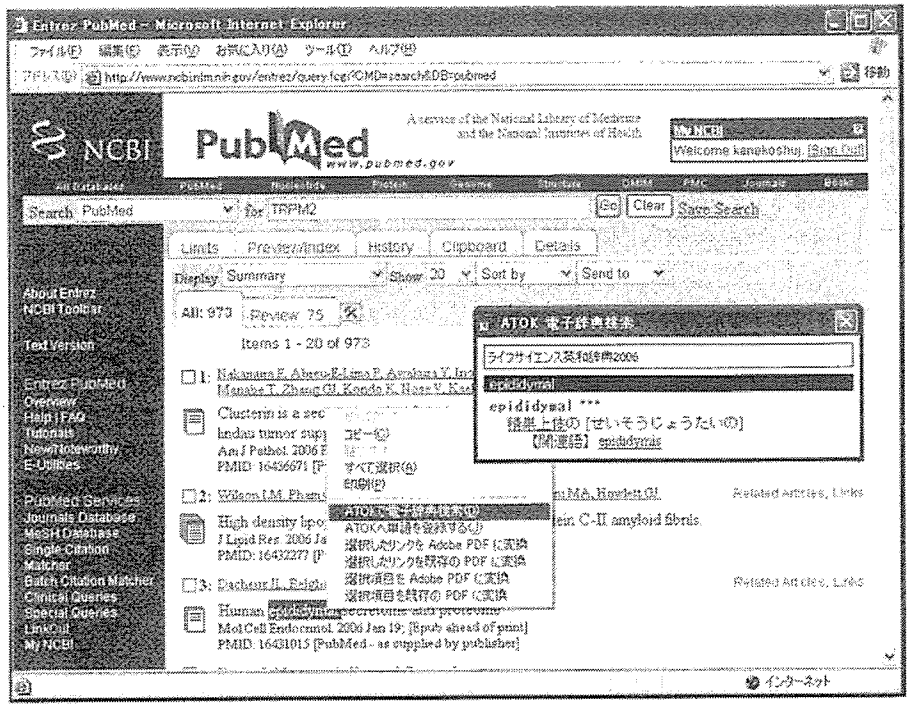


図9　ATOK2006でのLSD検索

当たっては，あらかじめインパクトファクターなどを考慮して生命科学の各分野から選んだ代表的な学術誌（89種類）において，アメリカおよびイギリス国内の研究機関から報告された論文だけを選択することによって，質の高い英文を収集するように努めている。また，原著論文だけでは教科書に記載されるような基本的理解に必要な語彙が不足するため，NCBI Bookshelfで公開されている教科書や，協力を得られた出版社から提供された電子テキストなども使用している。現在，解析に使用している1995年以降の英文コーパスは368 Mbyte（約6,000万単語）である。

出現頻度や連接頻度の解析は独自に制作しているPerlスクリプトによって行っている。上記の英文コーパスの単語頻度解析では62万種類の単語が抽出されたが，半数を占める30万語はコーパスで1回しか出現しない記号および数字やスペルミス等の収録の必要がない単語であった（図10上）。用語数は年々，増加しており，必要とされる辞書の規模は増大し続けている。現在のLSDでは頻度の高い約3万語の単語と2万語の複合語（複数の単語から形成されている語句）を収録しており，年間1万語を目標に収録語の増加を計画している。なお，単語の出現頻度と英文コーパスに占める割合を計算すると，10年間に20回以上出現する56,000語で全

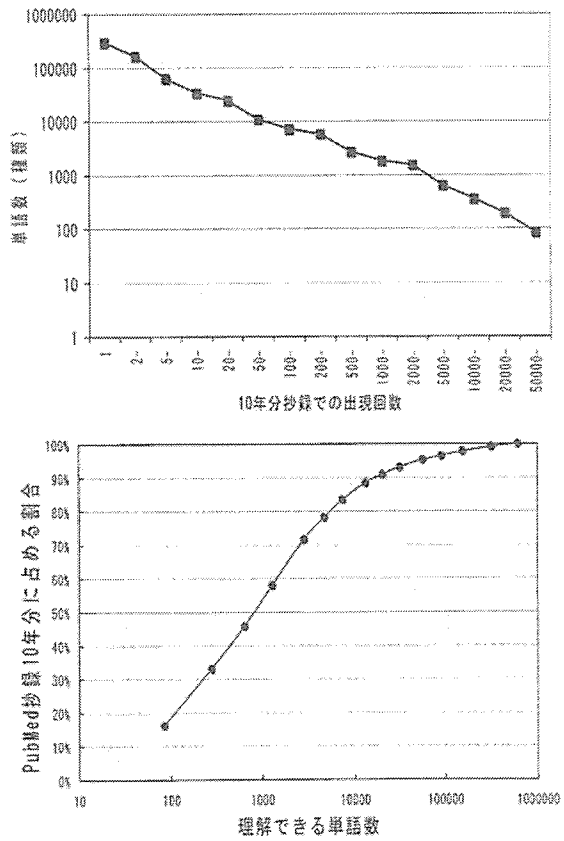


図10（上）PubMed 1995-2004 に出現する62万語の頻度
（下）PubMed を理解するために必要な単語数

テキストの95%がカバーできるが，現状のLSDはこの規模にようやく達しようとしている段階といえる（図10下）。

3.1.2 日本語コーパスの解析

日本語を系統的に解析し始めたのはごく最近のことである。解析が遅れた原因としては，専門的な日本語のコーパスが著作権の問題から得難かったことが挙げられる。日本語解析は学会抄録を利用することによって行うことが可能であるが，分野の網羅性に乏しいことと，学会抄録の場合には校閲が入っていないために間違いが数多く存在するので，これまでは積極的に利用してこなかった。さまざまな出版社への協力依頼を行った結果，少数から前向きな回答が得られ，これまでに基礎医学・ゲノム科学の最新研究成果に関する総説誌を発行している出版社の協力を得て，1996年から2002年にかけて出版された総説誌原版（QuarkXpressファイル）を入手した。ここからタイトルを含む本文テキストを抽出して，2005年9月現在で26Mbyteの基礎医学コーパスを作成した。また，そのほかに入力した臨床医学関連の文書を合わせると現在，解析に用いているのは29Mbyte（1,500万文字）の日本語コーパスである。

日本語専門用語の抽出は，ごく単純に漢字やカタカナとひらがな，記号，アルファベットの境目で切断して最長連続の語句を抽出する方法と，公開されている形態素解析システム（茶筌（ちゃせん））や専門用語抽出プログラム（termex）を組み合わせて行っている。同様に日本語の頻度解析は，茶筌にLSDを辞書として登載して語分割と計数を行う方法と，LSD収録語の個々について出現頻度をスキャニングする方法を併用し，その効果を比較しているところである。

現在までに21万語の日本語専門用語を収集したが，驚いたことに37%に相当する7.8万語は英語のままであり，カタカナのみから表記されるものが3.4万語もあった。つまり，漢字を含む用語は10万語程度しかなく，英語と比べてコーパスサイズの差はあるものの語彙数は明らかに少なく，英語の専門用語が日本語に翻訳しきれていない現状が数字からもうかがい知ることができる。

3.2 データベースの設計

従来の用語集が書籍として公開されるものという発想であったのに対して，LSDは最初からパソコンでの利用のみを考え，汎用性のあるデータベースとして設計した。電子辞書は更新にかかる時間とコストが節約でき，常にアップツーデートな辞書を供給できるというメリットがある。

パソコンで活用するためには英和・和英辞書とともに，かな漢字変換辞書やスペルチェック辞書も必要となる。これらの制作を1つのデータベースから可能にするため，LSDは英語と日本語をユニークなレコードから構成されるテーブルとして，対訳を中間テーブルとして設計した（図11）。さらにかな漢字変換辞書やスペル辞書も，それぞれ漢字変換テーブルや英語の頻度テーブルから制作できるように配置している。また，対訳において事物の意味分類情報を定義しているため，機械翻訳辞書の制作も容易になっている（後述）。

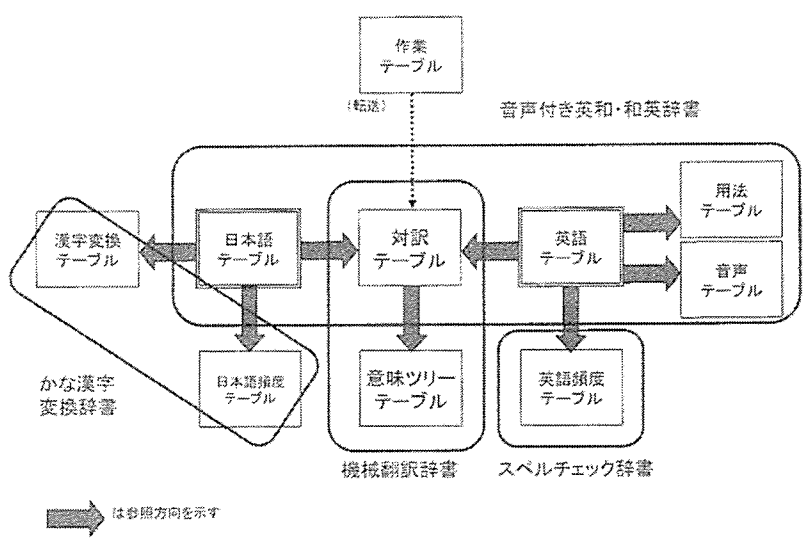


図11　LSDデータベース設計と応用

対訳を中間テーブルとすることにより，多対多の訳語関係や英和・和英の方向によって異なる内容を記述することが可能になった。実際に，多くの用語は英和，和英のいずれの方向でも1対多の関係で複数の訳語が用いられるが，生命科学の場合でも3〜4割の用語に表記の多様性も含めた複数の訳語が存在している。

辞書データベースは現在，汎用ソフトウエアであるFile Maker Server + File Maker Proをエンジンに使用している。SQLへの移行も検討したが，WebLSDなどの公開サービスはデータベースへの直接アクセスではないことや，データ構造の設計変更の自由度と使いやすさから考えて，この形に落ち着いている。

3.3 対訳の定義

以上のようにして収集した英語および日本語を関連付けて対訳テーブルを作成する作業は，今のところすべて人間が行っている。実際には大学院生等からなる協力員がインターネットを介して上記のFile Maker Serverにログインして作業テーブルにおいて対訳を定義し，最終的には筆者が頻度調査などを行って確認した後，正式にFile Maker Proデータベースに登録している。この対訳調査は複数の出典において確認することを義務付けているが，最近ではGoogleなどの検索エンジンを用いてインターネット上のドキュメントで使用頻度を確認する手法も有用性が増しており，採用対象としている。

ところで，専門用語は表記法の多様性が大きいが，LSDでは表記を統一していない。英語では米英の相違はもちろん，解剖学用語や植物名で一般的な表現とラテン語由来の学術的な表現が混在することが多く，また複合語ともなると類義語の掛け合わせだけ，表現が存在する場合がある。日本語はさらに複雑で，漢字の異字体，カタカナ外来語の表記の不統一がそれに拍車をかけている。典型的な例として「protein」に対しては，「タンパク質」「タンパク」「蛋白質」「蛋白」「たんぱく質」「たん白質」「プロテイン」等々の訳語があり，異なる学問領域では異なる表記が推奨されている。そして過去の用語集ではそれらを統一しようという意図で編纂がなされてきた。

しかしながら，LSDにおいては実際に使われている表記をそのまますべて収録しており，統一は行わない。その代わりに頻度を重要度という形で提示することによってユーザーがその妥当性を判断できるような工夫を施している。

3.4 オントロジーへの展開[5)]

前述したように，LSDではすべての対訳の規定時に必ず意味情報を付与することによって機械翻訳辞書やオントロジーの構築基盤を築いてきた。一般的なシソーラスでは英語または日本語のいずれかにおいてのみ同義語関係や概念の上下関係を整理している例が多い。しかしLSDでは英語と日本語の関係が1対1で定義されている対訳テーブルにおいてシソーラスを規定している。これは特に英語と日本語が1対1で対応していない用語において重要である。例えば，行為や方法を表しているstimulationと，刺激そのものを表すstimulusはいずれも日本語で「刺激」と訳されるため，和英翻訳ではいずれかを判別する必要がある。同様な例は，「基質」という言葉において対象が酵素ならsubstrate，細胞ならmatrixと別のものを意味したり，あるいはdischargeが生理学では「発火」，内分泌では「放出」「分泌」，臨床では「退院」を意味したりというように，出現頻度の高い多くの語句で見られる。さらに，形容詞が名詞としての派生的意味を持つantibioticのような語においても，対訳で形容詞と名詞を別々に収録できる。

シソーラスを構築する場合には他の体系との整合性に配慮が必要である。LSDはPubMedや論文情報と組み合わせて使われる場合が多いと考えられるため，NLMがPubMedの統制語として用いているMeSHへのリンクを設け，標準的なシソーラスとして参照できるようにした。しかしながら，MeSHには通常の用語集と同じく名詞のみが収録されており，一方，LSDは動詞，形容詞，副詞から構文までも収録しているため，ライフサイエンス辞書に収録された7万語の全対訳の英語表記のうち，MeSHツリー語と一致するレコードは2万語弱である（表1）。また反対にLSDは元々頻度解析から最低必要限の語彙を収録したため，MeSHに多い複合語が収録されていない場合も多く，現在，それらの補充を行い，MeSHとの整合性を高めようとしている。

なおMeSHツリー7万語のうち，LSDの英語コーパスに1度も出現しない語句は生物名や物質名を中心に2万語近くにも上っている。これは実際の論文で話題にされない用語も博物学的に収録されているためであり，網羅的な辞書としては仕方ないことであろう。反対に，新しいタンパク質名や新しい現象や機能についてはMeSHに未収録のことが多く，特に発生学や物理化学など生命科学の周辺

表1　LSDデータベース対訳とMeSH2006との分類別収録数

分類（MeSH カテゴリー）	LSD 対訳	MeSH2006	共通
Anatomy (A)	4,979	2,867	2,187
Organisms (B)	3,130	5,983	1,615
Diseases (C)	7,398	15,111	5,419
Chemicals and Drugs (D)	9,734	35,098	5,483
その他の名詞（現象等）	23,383	17,777	4,606
形容詞	12,053		
動詞	3,974		
副詞	1,686		
その他の品詞（略語等）	4,285		
合計	70,622	76,836	19,310

領域においてMeSHは収録語彙が乏しい。これらの事情を考慮して，他のシソーラスを参考にしつつLSDを新しい機能的な類義語辞書として体系化する試みを現在，行っているところである。

4. 英語と日本語のズレ

LSDデータベース構築に当たって英語と日本語の専門用語について出現頻度を比較してみると，2言語間における事物や概念は必ずしも1対1の対応関係になく，背後にある理解や思想の違いや独自の取捨選択過程によって，ズレを生じている場合がかなり多いことが明らかになってきた。これらを理解することは情報検索や翻訳において重要な点であると思われるため，本稿の最後にいくつかの事例を整理して紹介したい。

(1)類義語は翻訳時に混同されることがある(表2)

典型的なのは，薬物を表す用語である英語の「agent」と「drug」の関係が日本語における「剤」と「薬」に対応していない例である。英語においてagentが作用する機能を有する物質の総称として，drugはより狭く人体に対して主として治療効果をもたらす物質として定義されるのに対して，日本語での「剤」と「薬」の関係は，もともと「接着剤」や「殺虫剤」などの名称から機能単位としての「剤＝agent」，また，人間に対する「薬＝drug」とそれぞれ対応していたと考えられる。しかし実際には「免疫抑制剤」「抗菌薬」「抗癌剤」「抗炎症薬」など，執筆者によってどちらか一方が好んで（しかし混同して）用いられる。同様な混同は「dysfunction＝障害」と「injury＝傷害」のようにたまたま訳語が同音異義語であった場合に，「injury」であっても日本語で「(機能の）障害」が用いられる例が多く見られる。

(2)日本語が適切でない場合は置き換わる(表3)

ほとんどの学術用語集や辞書で「cancer＝癌」「tumor＝腫瘍」「sarcoma＝肉腫」といった対応関係が整理されているが，「carcinoma」や「neoplasm」に関しては学術用語としての「癌腫」や「新生物」はほとんど廃れており，ほとんど「癌」にとって代わられている。これは日本語訳の音や表記が直ちに認識されづらく嫌われたケースのように思われる。また，「腫瘍プロモータ」が一般的に「発癌プロモータ」と表記されるような例も，より分かりやすい言葉へと変遷していった例であろう。これらは学術的な定義よりも，表記や音の響きによって日本人が好む用語があることを意味しているように思われる。

(3) 日本語と英語は直訳関係にならない（表4）

一般的に英語は唯物論的に事物を命名するのに対して，日本語では機能単位として事物をとらえて名付ける場合が多い。一例として「消化器系」を考えた場合，英語においては「消化」に相当する形容詞「digestive」や「alimentary」はあまり用いられず，形態学的に「gastrointestinal」を用いる例が圧倒的に多い。しかしながら日本語においては「胃腸症状」は用いられるが，「消化器系」「消化性潰瘍」「消化器症状」「消化管出血」といった機能的記述が圧倒的に多い。同様な例は「cardiovascular＝心血管」が「循環」となるように数多く挙げられる。

以上，例示してきたような差異は従来の辞書やシソーラスでは十分に配慮されておらず，直訳的で使われない表現が訳語として採用されている場合が多い。表記の多様性も含めて，等価な意味を持つ日本語と英語を頻度情報とともに網羅的に収録するLSDデータベースは，新しい対訳シソーラ

表2 類似している複合語の日本語訳で見られる偏り

(数値はコーパス中での頻度を表す。いずれも編者が手を加えている場合があることに注意)

英語 (* MeSH term)	agent	drug	日本語	剤	薬
immunosuppressive	263*	229	免疫抑制	102	6
antihypertensive	107*	82*	降圧	2	71
antibacterial	131*	55	抗菌	3	313
anti-inflammatory	197*	411	抗炎症	6	102
anticancer	125	121	抗癌	633	6

英語 (* MeSH term)	dysfunction	injury	日本語	障害	傷害
hepatic (liver)	107*	540	肝	114	29
cellular (cell)	188	491	細胞	180	201
endothelial	280	76	内皮	13	14
balloon	0	107	バルーン	14	57
cardiovascular (1)	13	2	脳血管	82	0

■ は80%以上の偏り, □ は60～80%の偏りを表す。
(1) cardiovascular event = 305

表3 日本語で好まれる表記と好まれない表記

英語	PubMed	Textbook	日本語	基礎系	臨床系
cancer	28,302 (43%)	5,158 (50%)	癌	13,630 (72%)	1,154 (52%)
carcinoma	5,846 (9%)	1,368 (13%)	癌腫	29 (0%)	5 (0%)
sarcoma	1,235 (2%)	340 (3%)	肉腫	289 (2%)	57 (3%)
tumor	31,207 (47%)	3,242 (31%)	腫瘍	4,192 (22%)	989 (44%)
neoplasm	504 (<1%)	207 (2%)	新生物	4 (0%)	6 (0%)
oncogenesis	373 (<1%)	6 (0%)	発癌	763 (4%)	39 (2%)
計	66,357 (100%)	10,321 (100%)		18,916 (100%)	2,250 (100%)

□=英語に比べて割合が高い
■=英語に比べて割合が低い

表4 直訳的な対訳関係にならない例

英語 (PubMed)	digestive alimentary	gastrointestinal
(total)	440	3,372
ulcer ulceration	0	24
symptom	3	112
bleeding, hemorrhage	0	283
disorder, disease	23	116
organ, system	35	24
tract	170	1,077

日本語 (臨床)	消化性	消化器	消化管	胃腸
(総数)	32	152	279	72
～潰瘍	32	0	2	1
～症状	0	68	1	7
～出血	0	1	41	0
～疾患	0	10	7	0
～系	0	9	0	0
～管 (路)	0	0		0

スとして価値あるものと思われる。

5. おわりに

生命科学領域における対訳シソーラスとしてはUMLS（Unified Medical Language System）などの先例があるが，実際に英語および日本語テキストの解析を行った結果に基づき，2言語の特徴や多義ないし相違にまで踏み込んで対訳シソーラスを構築した例は他にない。LSDデータベースから制作される電子辞書は英語ないし日本語で記述される専門文書の機械翻訳やテキストマイニングを可能にする発展性を有しており，今後も内容の充実を図っていきたいと考えている。また，生命科学を志す学生や研究者が広く使える電子辞書ツールを提供することによって，日本の生命科学がますます発展していくことに少しでも寄与できれば幸いである。

謝辞

本研究は，文科省科研費研究成果公開促進費，同特定研究「応用ゲノム」，カシオ科学研究振興財団，21世紀COEプログラム「ゲノム科学の知的情報基盤・研究拠点形成」の助成を得て行われた。LSDプロジェクトのメンバー（藤田信之，大武博，河本健，鵜川義弘，竹腰正隆，竹内浩昭）の多大なる協力に感謝する。

参考文献

1) 金子周司ほか．フリーウェアのライフサイエンス学術用語辞書を作った理由．コンピュータサイエンス．Vol.1, No.1, 1994, p.25-32.
2) 金子周司．ライフサイエンス辞書．医学のあゆみ．Vol.210, No.13, 2004, p.1062-1063.
3) 藤田信之；金子周司．ライフサイエンスのための英和変換ツール．コンピュータサイエンス．Vol.2, No.1, 1995, p.41-45.
4) ライフサイエンス辞書プロジェクト．ライフサイエンス必須英和辞典．東京，羊土社，2005, 403p.
5) 金子周司ほか．ライフサイエンス辞書から生命科学オントロジーへ．情報知識学会誌．Vol.14, No. 5, 2005, p.1-10.

話題
WADAI

無料ライフサイエンス辞書の活用と効能

金子周司
Shuji KANEKO
京都大学大学院薬学研究科教授

1. はじめに

ライフサイエンス辞書(LSD)とは，筆者が主宰するライフサイエンス辞書プロジェクト(LSDプロジェクト)が1993年以来作り上げてきた薬学を含む生命科学領域の電子辞書である．この辞書の最大の特徴は，英語及び日本語のいずれも大量の学術テキストを計量的に解析した独自のデータに基づいていることである．本稿では現在，無償で公開している辞書サービスの活用法を紹介し，併せて将来の展望も述べてみたい．

2. LSDプロジェクトとは

ライフサイエンス(生命科学)とは薬学の全領域はもちろん，様々な生命の理解に資する学術領域を包含する巨大かつ境界のない知識体系である．急速に増加する生命科学の知識はほとんど論文，すなわちテキスト情報として記述され電子的に蓄積されており，研究者がみいだした新しい事物や概念は，膨大な数の新しい専門用語を生み出している．

しかし歴史的に他の科学もそうであったように，欧米から次々に「輸入」される新しい学問を記述する「ことば」は，我が国において十分に定義あるいは翻訳される間もなく，研究者社会において流通している．LSDプロジェクトは，そういった状況下で生命科学の教育や研究現場で使いやすい辞書を自ら作ろうという目標を掲げて集まったボランティア研究グループであり，各種の辞書と検索エンジンをすべて独自に開発してきた．ここでは現在LSDプロジェクトがホームページ(http://lsd.pharm.kyoto-u.ac.jp/)で提供している辞書サービスを概説する．

3. オンライン辞書WebLSD

WebLSDは生命科学用語に特化したインターネット対訳辞書である(図1)．2006年1月現在で英和49,034語，和英48,024語の見出し語が収録されており，現在では1日平均で7万件以上の利用がある．「音声付き英和・和英検索」では入力された語句が漢字・かな・英語のいずれの文字種かが自動的に判断されて，その語句の出現頻度から算出した重要度にはじまり，外国人による英語の発音例，対訳と簡単な解説，関連する語句，英語の用法や例文，共起表現が表示される．

この辞書はWebリンク技術を活かし，見出し語でのPubMed(英語の場合)あるいはGoogle(日本語の場合)検索，訳語の逆引き，例文から出典元のPubMed抄録の閲覧などを可能にしている(図2)．したがって，スペルが曖昧なキーワードについて和英辞書から英和辞書へと移行して，PubMed検索を実行するという応用的な使い方もできる．

WebLSDの最大の特長は，共起表現(concordance)を3,000万語のPubMed抄録コーパス

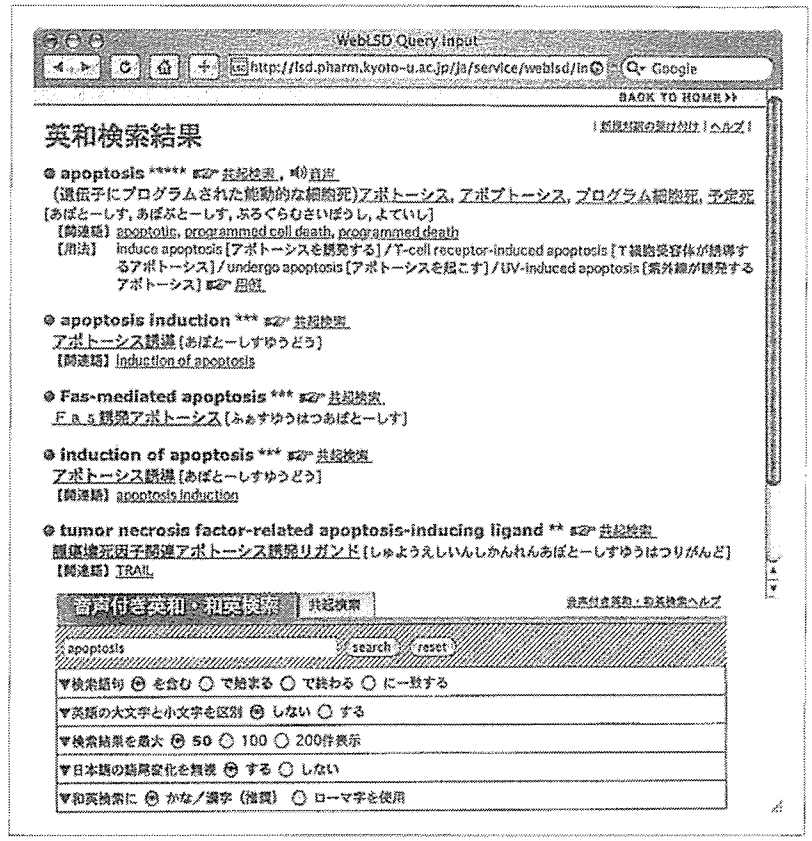


図1　オンライン辞書 WebLSD

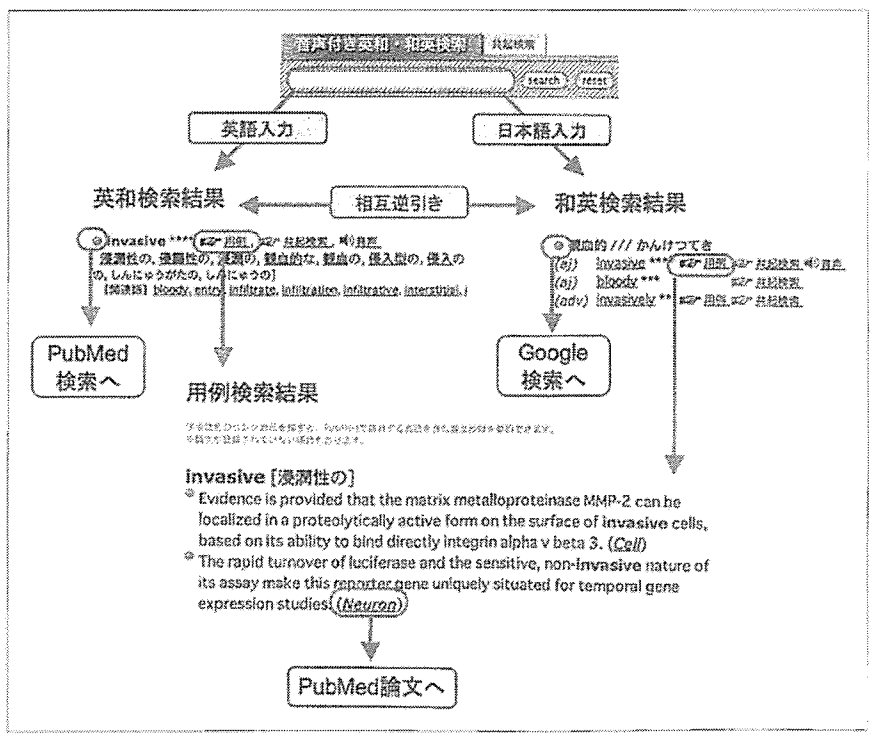


図2　WebLSD の基本動作

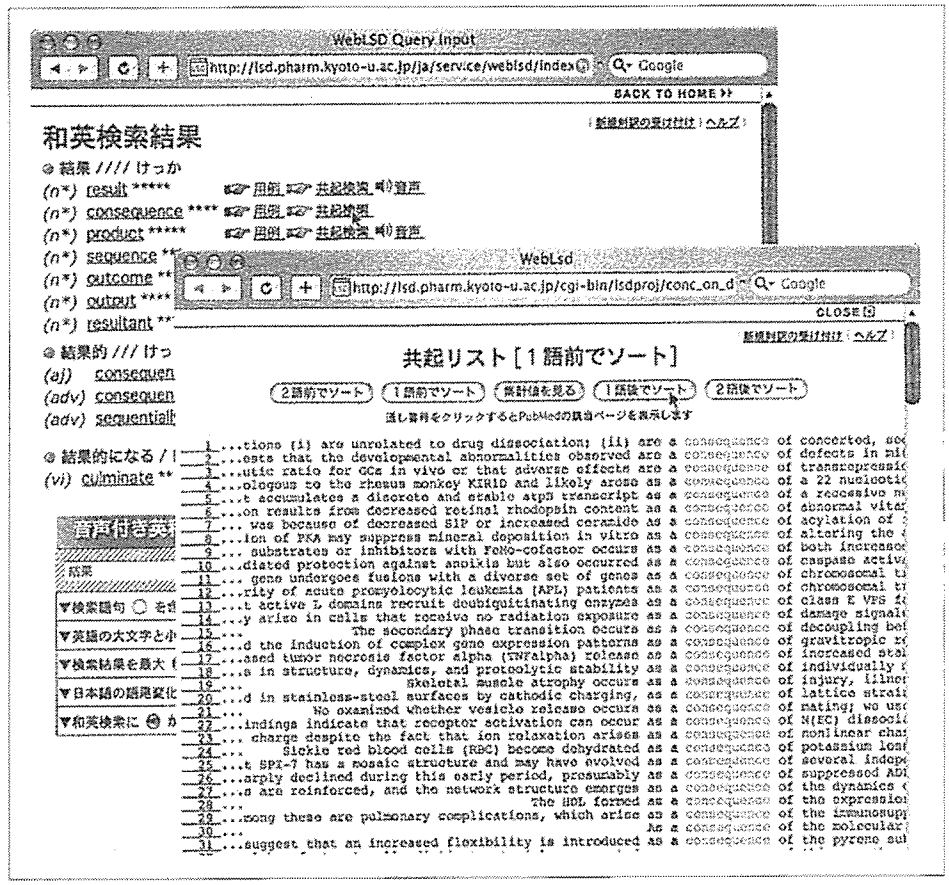


図3　WebLSDでの共起表現リスト

からオンデマンドで高速に検索してKWIC(Key Word In Context)形式で表示する点である．KWIC形式の共起表現を用いると，任意の単語の前後にどういった別の単語が使われるかを直感的かつ定量的に確認でき，例えば「consequence」という単語は「as a consequence of」という構文で用いられる例が多数であることが理解できる(図3)．同様にして，名詞と動詞の親和性や，形容詞及び副詞の用い方，あるいは名詞が複数形を取り得るか否かなど様々な英語の正しい用法を知ることができる．ここで解析材料としている英文は，インパクトファクターの高い学術誌に欧米の研究機関から発表された論文抄録のみを厳選しているため，我々日本人が英文を書くにあたって非常に有用な資料を提供している．

なおWebLSDでは辞書に収録されていない専門用語をユーザーに投稿していただいたり，様々なフィードバックを受け付けたりする入口を用意している．利用者からの新語追加の要望は年間3千件以上寄せられているが，化学系や物理系などの専門用語はまだ不十分と自覚しているので，ファルマシア読者のご協力を賜れば幸いである．

4. オンデマンド英語教材

医歯薬学を専攻する学部カリキュラムで早期から専門英語教育の必要性が増している．しかしながら，薬学英語の学習教材は我が国では十分になく，初学者レベルであっても専門領域と関連した分野の英語を学べる教材が現場で強く求められている．こうした状況を打開すべくLSDメンバーである京都府立医大の大武博教授を中心に行っているのが，オンデマンド英語

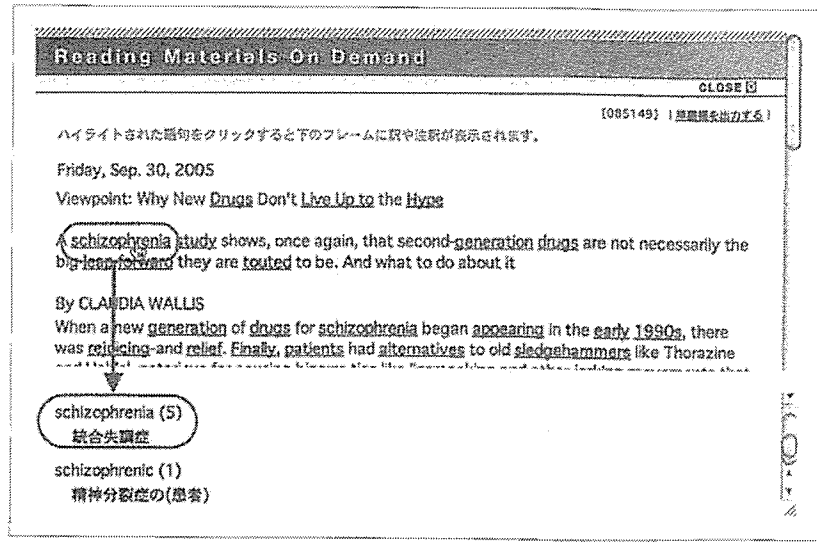


図4　オンデマンド英語教材

教材プロジェクトである．このページでは，欧米の著名な雑誌等に掲載される医療関連の web 記事を題材にして，その英文テキストを我々が開発した逐語訳ツール EtoJ vocabulary を通して読むことができる．このツールは英文テキストの中に日本語での注釈付けを行い，それを web ブラウザのハイパーリンク機能を利用して，注釈を見たいときにクリックするだけで表示するものである(図 4)．したがって，マスコミに掲載されるやや難解で格調高い英文を読もうという読者は，英語で読める範囲では英文を読み，注釈を参照したい場合にはクリックするだけでブラウザの下部に表示される訳語や解説を見て読み進めることができる．現在までに 260 余の英文を選んで連載を続けており，なかには実際に大学院入試で出題された英文も収録している．

5. 逐語訳 EtoJ

日本語は，漢字，カタカナ，ひらがなという 3 種類の記号が適度に混じり合った「字面」をしており，視覚的にパターン認識がなされて直感的に内容を理解できる日本人にとって都合の良い言語である．一方，英語はアルファベットという記号で表された言語であり，日本人が一見して内容を把握することは困難である．そこで生命科学英語は専門用語だけ翻訳すれば内容のおおよその見当がつくだろう，という発想から開発したのが EtoJ 逐語訳エンジンである．これは web だけでなく，自動応答メールエンジンでも利用できる．

EtoJ では WebLSD とは異なる独特のチューニングを施した辞書を用いており，適度な割合で専門用語が日本語に置換され，構文自体は原文の英語のままという奇妙な文章が返される(図 5)．論文抄録やタイトル情報などを斜め読みしたい場合には意外と重宝するため，あまり知られていないが根強いファンがいるサービスとなっている．この他，先のオンデマンド英語教材でも使用している EtoJ vocabulary を任意のテキストに対して行うためのサービスやスペルチェックを行うための WebSpell も公開している．

6. パソコンで使える辞書

読み書きと同様に，あるいはそれ以上に日本人を悩ませるのが英語のヒアリングである．

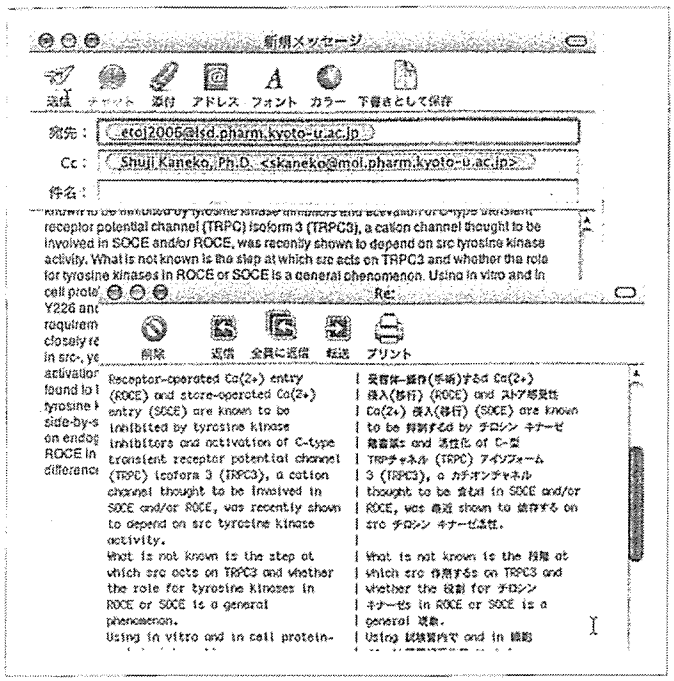


図5　EtoJ メール逐語訳サービス

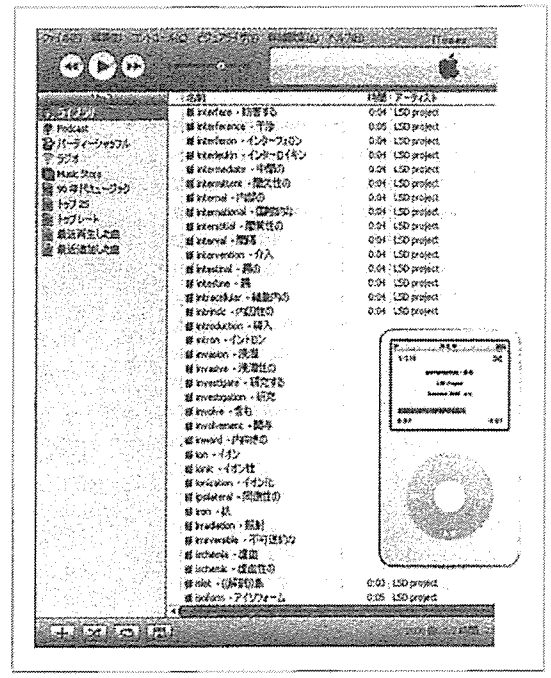


図6　耳で覚えるライフサイエンス英語

WebLSD でも音声例を提供して正しいアクセントや発音を参照できるようにしているが，学習者が最低限の生命科学ボキャブラリーについて，すべての発音を聞けるように工夫したのが「耳で覚えるライフサイエンス英語」である(図6)．このファイルをダウンロードしてアップル社が無料配布している iTunes で開くと，2,000 語の基本用語についてパソコンで英語の正しいスペル，日本語訳，発音を参照できる．さらに iPod にそれらを転送すると，電車の中で生命科学英語の勉強も可能になる．iTunes や iPod はランダムに再生する機能を有しているので，聞き取りテストを試してみるのもよいだろう．なお，初学者向けに 6,000 語の基本語彙をまとめた書籍も刊行している．[1]

ところで筆者が LSD プロジェクトを開始した直接の引き金は，昔のパソコンのかな漢字変換であまりにも専門用語の変換が正しくできず，自分でコツコツと変換辞書を作っていたことにある．それに比べると最近の Windows XP や Mac OS X に装備されている変換辞書はかなり優秀になってきたが，生命科学に特化した変換辞書はまだまだ必要である．LSD プロジェクトでは，これらのパソコン OS で使える 2 万語規模の変換辞書を無料で配布しており，日本語の教科書や総説で高頻度に出現する用語を網羅している．

また近年は漢字変換プログラムの高性能化が進み，OS やアプリケーションと連携して辞書ツールとしての機能を有する例が出てきた．LSD プロジェクトではジャストシステム社の ATOK 2006 に最適化した辞書を開発し，かな漢字変換だけでなく，かな英語変換，英和辞典，和英辞典の 4 種類の機能を ATOK から利用できるようにした(この辞書に限っては有償である)．これを使うと Internet Explorer に表示される PubMed ページで不明な単語があった場合に，マウス操作だけで辞書検索を行うことが可能になる(図7)．今後もさらに高機能で便利な辞書ツールを開発し，我が国の生命科学が発展していくための一助になればと考えている．

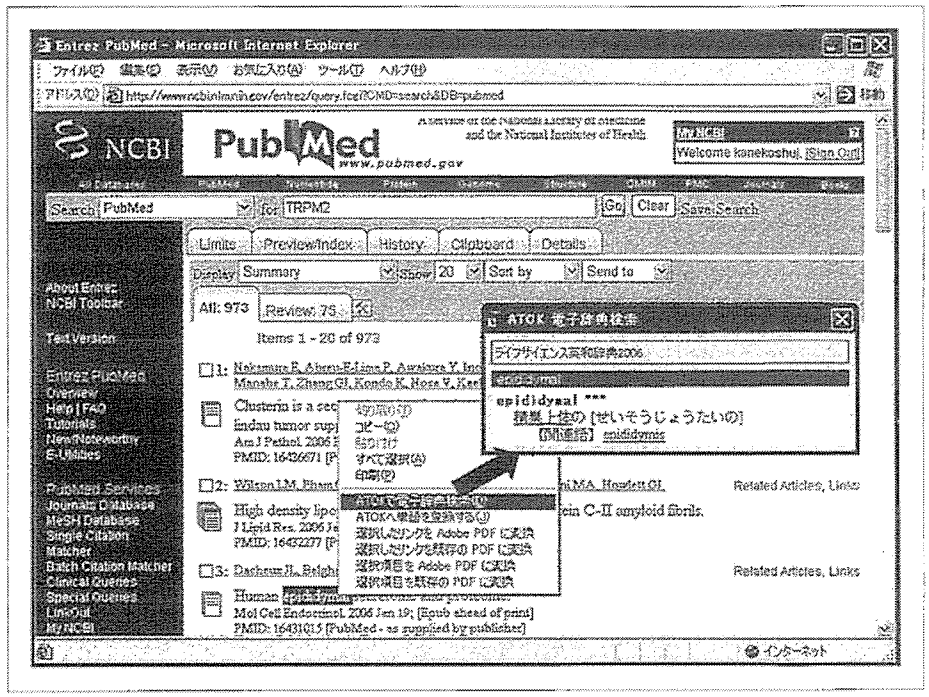


図7　ATOK 2006での電子辞典検索

7. 情報科学への展望

LSDは論文，抄録，総説など大量のテキストを解析する手法[2,3]によって構築しているが，その中で生命科学の新しい事象や概念に対して視認性に優れた日本語という文字での訳語を付与するという我が国の優れた「翻訳文化」が既に破綻しつつあることを指摘している．一部の研究者からは「生命科学はすべて英語で行うべし」との極論も耳にするが，いかに専門家が英語ですべてを理解して論文を読み書きしたとしても，教育や医療を日本語抜きで行うことは不可能であり，どこかの段階で「翻訳」は必要となる．しかし翻訳すべき情報が大量である現在，情報の解読を人間の英知に頼るのみならず，コンピュータの力を借りた情報自動抽出や機械翻訳などで前処理する必要性が感じられる．また，人間がポータルサイトで検索を行った場合でも，おびただしい量で混入する不必要な情報を排除し，真に必要な科学的情報を探し当てることはなかなか難しい．それを解決するためのsemantic　web(文脈を理解した検索)など，情報科学において開発される新技術を生命科学に応用することが求められる．

このような将来の発展を考えた場合に最も重要な点は，語彙の概念構造(オントロジー)を整理した用語シソーラスが英語と日本語の対訳関係において構築されていることである．文献の解析からスタートしたLSDプロジェクトは，そうした新技術にも役に立つ情報資源となるよう，現在はオントロジーの構築と情報抽出の研究を展開しつつある．[4] ひいては，LSDという資源を生命情報科学に活用し，まったく新しい薬学の研究領域を開拓したいと考えている．

参考文献

1)　ライフサイエンス辞書プロジェクト編著，ライフサイエンス必須英和辞典，羊土社，東京，(2005)．
2)　金子周司，医学のあゆみ，210，1062(2004)．
3)　金子周司ほか，情報知識学会誌，14(5)，1(2005)．
4)　金子周司，情報管理，49(1)，印刷中(2006)．

研究速報

文献情報の解析に基づく対訳シソーラスの評価

金子周司[*1]　藤田信之[*2]

生命科学および医学の教育研究を支援する目的で，筆者らは広範な生命科学の諸領域で使われる英語および日本語の専門用語を文献コーパスの定量的解析から抽出し，独自の対訳辞書をライフサイエンス辞書（LSD）として公開してきた．今回，テキストマイニング等に応用できるシソーラスへの発展を目標に，LSDで対訳と意味情報を定義した5万語の英語と5万語の日本語について文献情報による評価を行い，続いて，MeSHツリーとのマッピングによる体系化を試みた．主にPubMed抄録からなる英語コーパスをLSD収録語で解読できる割合は80%であった．MeSHツリーから得られた6.5万語のうち，LSDと一致したのは20%であったが，PubMed中に出現するMeSH termについては40%がカバーされていた．MeSHにないLSD収録語としては略語，名詞以外の品詞，未分類の学問分野の用語などが浮かび上がった．今回の解析から，我が国で今後，医療現場等で発生する大量のテキストをコンピュータで解析するためには新たな対訳シソーラスの必要性が示唆された．
■キーワード：シソーラス，オントロジー，電子辞書，ライフサイエンス，用語法

Evaluation of an English-Japanese Thesaurus Based on the Analysis of Biomedical Corpora：Kaneko S[*1], Fujita N[*2]

Life Science Dictionary (LSD) is a versatile database of English and Japanese terms based on the quantitative analyses of biomedical corpora. To develop a thesaurus of LSD terms for future application to computer-assisted text mining, we have evaluated the frequency of LSD terms in the literature-based corpora, and mapped the LSD terms to the MeSH tree. Coverage of LSD English terms in a PubMed-based corpus was 80%. In 65,000 MeSH tree terms, LSD-matched terms were 20%, which was increased to 40% in a subpopulation of terms occurred in the English corpus. The MeSH-unmatched LSD terms included abbreviations, verbs, adjectives, adverbs and MeSH-unclassified terms. These results indicate the requirement of new comprehensive thesaurus tree covering complex English-Japanese translations.
Key words：Thesaurus, Ontology, Electric dictionary, Life sciences, Terminology

1．はじめに

ライフサイエンス辞書(Life Science Dictionary, 以下LSD）は医学を包含する生命科学の教育研究を支援する目的で，様々な学問領域の研究者が協力しあい，1993年以来，制作されてきた電子対訳辞書である[1]．LSDの特長は，学術論文の計量的な解析を行って作成したデータに基づき，頻度の高い用語を網羅的に収録している点や，音声や共起表現の提示など，学習者にとって有用な機能を実装している点などが挙げられる[2,3]．

一方，生命科学の研究が加速し，テキスト情報

[*1] 京都大学大学院薬学研究科生体機能解析学分野
〒606-8501　京都市左京区吉田下阿達町
[*2] 製品評価技術基盤機構ゲノム解析部門
〒151-0066　渋谷区西原2-49-10
E-mail：skaneko@pharm.kyoto-u.ac.jp
受付日：平成18年2月10日

[*1]Department of Molecular Pharmacology, Graduate School of Pharmaceutical Science Kyoto University Shimoadachi-cho, Sakyo-ku, Kyoto, 606-8501, Japan
[*2]Department of Biotechnology, NITE 2-49-10 Nishihara, Shibuya-ku, Tokyo, 151-0066, Japan

量が指数関数的に増大している現状において，すべての情報を人間が解読する努力はすでに限界を超えている．また，ゲノムやタンパク質の配列情報が正規化データベースとして蓄積され，コンピュータによって情報の解読や推測が可能となっている現状とはきわめて対照的に，医学研究報告に記述された人智は有効利用されないまま蓄積され続けている．さらに，インターネットの普及によって一般化した検索エンジンによる情報検索において，無駄な情報を排除して求める情報を提示するためには，分野に特化したセマンティックウェッブ辞書の開発が求められる．

我々は将来LSDを英語と日本語を包括したテキストマイニング，機械翻訳，セマンティックウェッブ等に応用できる「コンピュータのための辞書」に発展させることを新たな研究目標としている[4,5]．生命科学領域における対訳シソーラスとしてはUMLSの日本語化[6,7]などに先例があるが，実際に英語および日本語テキストの解析を行った結果に基づき，対訳における多様性と各々の言語の特徴にまで踏み込んで対訳シソーラスを構築した例はまだない．そこで本研究では，まずLSDに収録された5万語の英語と5万語の日本語が，実際に最近の論文や総説で用いられている専門用語を網羅する割合を評価し，到達目標を明確にしようとした．次に，これら英語と日本語の関係を定義している約7万対の対訳を体系化するため，PubMed検索への応用性を考慮して，代表的な既存シソーラスとしてMeSH（Medical Subject Headings）との照合を行い，その結果から実用性の高い対訳シソーラスを構築するにあたっての問題点と方向性を考察した．

2. 方　法

1）LSD用語

2006年1月にWebLSDとして公開した辞書[1]の元となったFileMaker Proのリレーショナルデータベースより，英語49,034語（対訳定義ずみ），日本語73,103語（そのうち，対訳の定義されたもの48,024語），およびそれらを結合する中間テーブルから70,622対の英語および日本語を出力して用いた．なお，LSDにおいて英語とは英単語と語句を含んでおり，規則変化を伴う名詞，形容詞，動詞ではそれぞれ単数形，原級，現在形がラテン語由来の名詞は単数形と複数形が別個のレコードとして収録されており，不規則変化を伴う動詞は現在形のみを収録している．また，LSD対訳には従来より意味情報および品詞が付与されているので，それらをMeSHに合わせて「A: 解剖」「B: 生物」「C: 疾患」「D: 薬物」「E: 技術」等に再分類した．

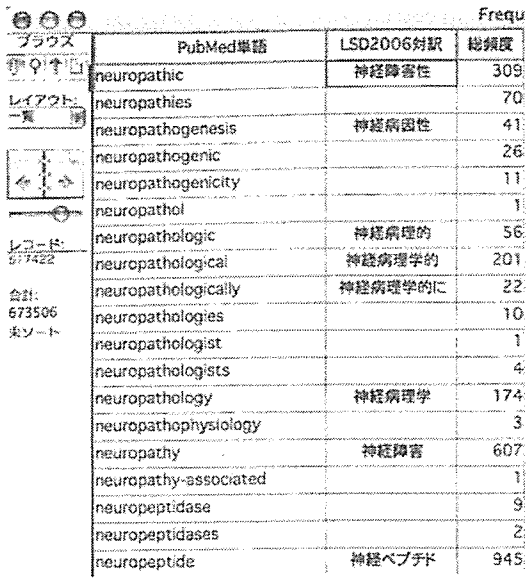
2）コーパスの作成

英語については，PubMedに抄録が収録されている学術誌のうち，インパクトファクターなどを考慮して生命科学の各分野から選んだ代表的な学術誌（89種類）にアメリカおよびイギリス国内の研究機関から1995年から2004年に報告された論文抄録テキスト（年間23,000～24,000抄録，368 MB）を蓄積した．原著論文だけでは教科書に記載されるような基本的理解に必要な語彙が不足するため，NCBI Bookshelfで公開されている教科書や，協力を得られた出版社から提供された電子テキストなどを加えて，463 MB（約6千万単語）の英語コーパスを作成した．

一方，日本語については，基礎医学・ゲノム科学の最新研究成果に関する総説誌を発行している出版社の協力を得て，1996年から2002年にかけて出版された総説誌原版からタイトルを含む本文テキスト26 MBを抽出した．これに臨床医学の教科書テキストを合わせて34 MB（約2千万文字）の日本語コーパスを作成した．

3）解析プログラム

英語コーパスからの英単語の抽出と出現頻度解析には，単語間のスペースを認識して切断し，単語ごとに計数するPerlスクリプトを作成した．この結果をFileMaker Proに読み込み，LSDデータベースに対して同一見出し語間でリレーションを設けることによって語尾変化を考慮しないLSD収録語とのマッチングを行った（図1）．続いて，ラテン語由来の名詞の不規則変化には対応しないが名詞複数形の規則変化や規則動詞および不規則動詞の変化形に対応する公開逐語訳エンジンEtoJ[8]を改変し，LSDに収録された英単語および英語句の英語コーパス中での出現頻度解析を語尾



Frequ

PubMed単語	LSD2006対訳	総頻度
neuropathic	神経障害性	309
neuropathies		70
neuropathogenesis	神経病因性	41
neuropathogenic		26
neuropathogenicity		11
neuropathol		1
neuropathologic	神経病理的	56
neuropathological	神経病理学的	201
neuropathologically	神経病理学的に	22
neuropathologies		10
neuropathologist		1
neuropathologists		4
neuropathology	神経病理学	174
neuropathophysiology		3
neuropathy	神経障害	607
neuropathy-associated		1
neuropeptidase		9
neuropeptidases		2
neuropeptide	神経ペプチド	945

図1　英語パスの単語頻度解析とLSDとのマッチング

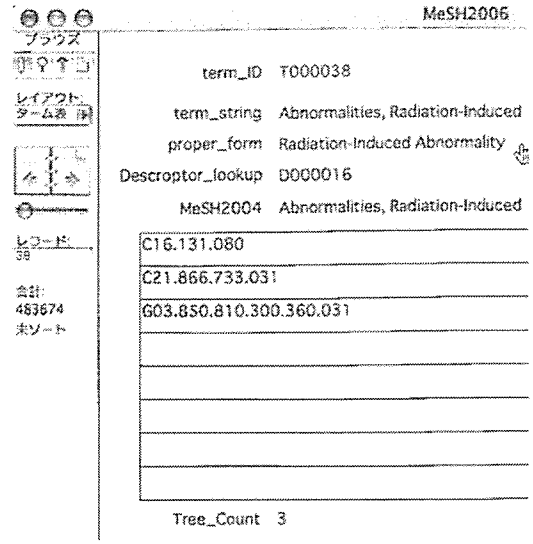


図2　変換されたMeSH2006

変化を含めて行った．

日本語の抽出および出現頻度解析には，漢字，カタカナ，ひらがな，アルファベットおよび数字の境目を認識して最長連続する要素（仮に単語と呼ぶ）を抽出するPerlスクリプトを作成し，コーパスにおけるそれぞれの文字種の割合とユニークな単語の種類を計数した．また，日本語コーパス中で日本語の単語が出現する頻度の計数もPerlスクリプトを作成して行った．

4）MeSH加工と対訳リンク

MeSH 2006（XML形式）からTree，Descriptor，Concept，Termテキストを抽出し，FileMaker Pro上でリレーショナルデータベースとして再構築した．MeSH Term表記の語順を自然な形に置き換え，規則的な名詞の語尾変化を吸収してLSDで用いている標準的な辞書表記と一致させるスクリプトを作成し，規則変化を伴うMeSH Term表記の変換を行った（図2）．MeSH Term 483,674語のうち，ツリーに帰属できたTermは76,836語であったが，語順変換などによって同一の文字列となった重複レコードを除外すると65,733語となった．一方，LSDに収録されている対訳レコード中の英語について，FileMaker Pro上で表記の一致するMeSH Termへのリレーションを設定することによってLSDとMeSHツリーの照合を行った．なお，FileMaker Proのリレーションでは大文字/小文字は同一視される．対訳テーブルあるいはMeSHテーブルでは重複する見出しが存在するため，それらの重複は後で除外した．

以上の関係を整理すると図3のようになる．7万対のLSD対訳は5万語の日本語と5万語の英語から参照されており，それらはコーパス解析結果である頻度情報へリンクしている．日本語の場合は対訳を有しないレコードを含む7万語の漢字変換テーブルも頻度解析に用いた

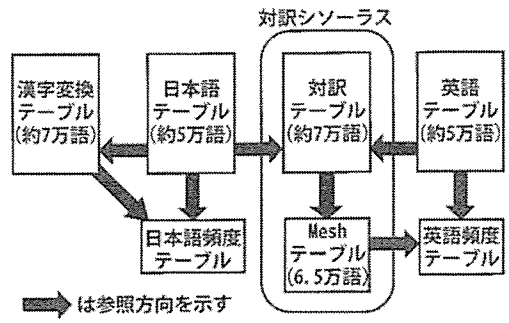


図3　LSDとMeshのリレーション概略