生命科学学術用語データベースと利用ツールの制作

金子周司（京都大学・薬学部・薬理学講座）
大武　博（京都府立医科大学・医学部・第一外国語教室）
鵜川義弘（農水省農業生物資源研究所・DNA情報管理科）
河本　健（広島大学・歯学部・口腔生化学講座）
竹内浩昭（静岡大学・理学部・生物地球環境科学科）
竹腰正隆（東海大学・医学部・分子生物学講座）
藤田信之（国立遺伝学研究所・分子遺伝学部門）

生命科学（Life Science）はこれまで基礎医学，臨床医学，化学，生物学，物理学などに分けられてきた自然科学の学問分野のうち，生命現象の解明に関わる部分を包括する学際領域として捉えられる。生命科学に関連する学術用語のほとんどは英語として発生したものであるが，我が国においては学会研究組織等においてそれぞれ独自に訳語が規定されてきた。これまで，電子メディアで活用できる電子辞書は極めて少なく，進展の著しいこの領域にあって，従来の辞書編纂の手法で作られてきた用語集では内容的に不足が多い。

そこで我々は，電子化された論文抄録等の統計解析に基づいた客観的な語句の選択を行い，その訳語と共に品詞，出現頻度，用法，用例，分野，注釈などの情報を収録することによって生命科学の分野を網羅する学術用語のリレーショナルデータベースを制作した。さらに，生命科学の教育研究に際してこのデータベースを実用的システムとして誰もが使用できるよう，インターネットあるいはパソコンで使う各種の電子辞書と辞書ツールを開発した。

1. 英単語の解析

生命科学論文の抄録情報データベースCurrent Contents on diskette Life Science version (CCOD-LS) の最近3年間分（1994-1996）に収録された論文のタイトルおよび抄録（約1 GByte）を材料にして，perlスクリプトにて解析を行い，全単語の出現頻度を登場する論文数として求めた。この単語ごとの出現頻度をレベル分けし，頻度レベルに対して該当する単語数と，その語句の内容を検討した（図1左）。出現した単語は22万種類であり，単語の頻度と単語数との間には，両対数プロットにおいて反比例する傾向が認められた。3年間で1回しか使われない単語は，ほとんどがスペルミスと略語で占められていた。また，それぞれの頻度レベル以上で，CCOD-LSに出現する英単語の何%をカバーできるか調べたところ（図1右），3年間に10回以上出現する（これは全体の21％＝46,000語に相当する）単語で全抄録にある97%の単語をカバーすることが明らかになった。

以上の結果を参考にしつつ，さらにCCOD-LSの他，農学・地球環境学，物理化学，臨床医学のためのCCODや医学データベースMEDLINE，生物学データベースBIOSISなどについても同様の解析を行い，これらの1年分に2回以上出現する単語129,200語を正規化ファイルで構成されるリレーショナルデータベースに収録した。

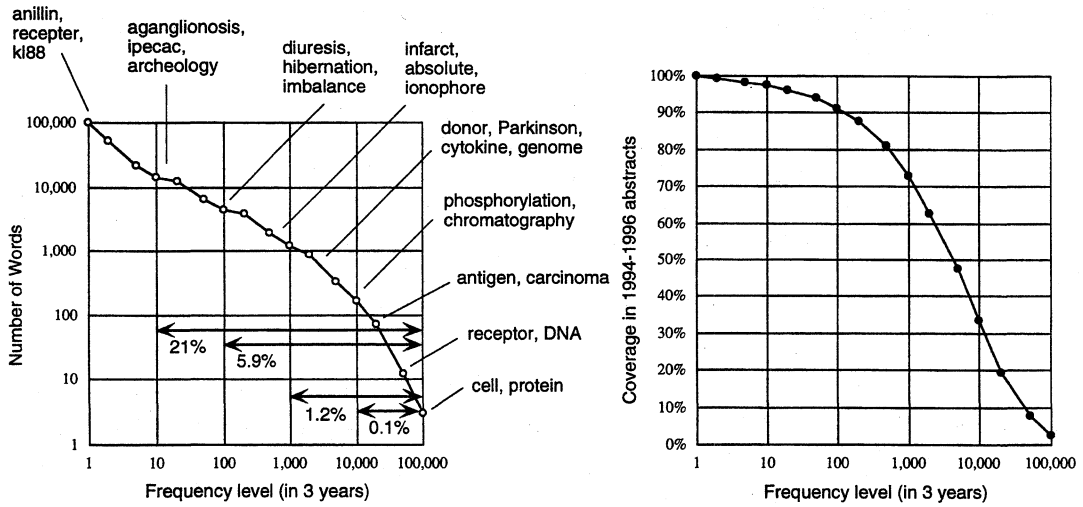

図1　（左）CCOD-LS 3年間の単語出現レベルと単語数の関係
（右）単語出現レベルと CCOD-LS 抄録中の全単語に占める割合

2. 複合語，訳語，用法，用例の選択

コンコーダンス作成プログラム Conc 1.71 を用い，CCOD-LS や MEDLINE 中に頻出する複数の単語から成る複合語を抽出した。さらに，既存の学術用語集を参考にして教育研究に必要な用語を補充し，合わせて 12,800 語の複合語をデータベースに補足した。以上の合計 142,000 語について，頻度の高いものや教育上必須の語句を中心にし，48,000 語について英日対訳を規定するとともに，日本語のよみ，日本語品詞，英語品詞，分野，注釈，意味情報などを収録した。語句の収録にあたっては，プロジェクトメンバー以外にのべ 21 名の研究者によるモニターを 2 回に分けて行い，誤りの最少化に努めた。

動詞，形容詞，副詞を中心とした用法については，別個に共起表現に基づく収集を行い（図 2），5,400 用法を選択して，英語の見出しに関連づ

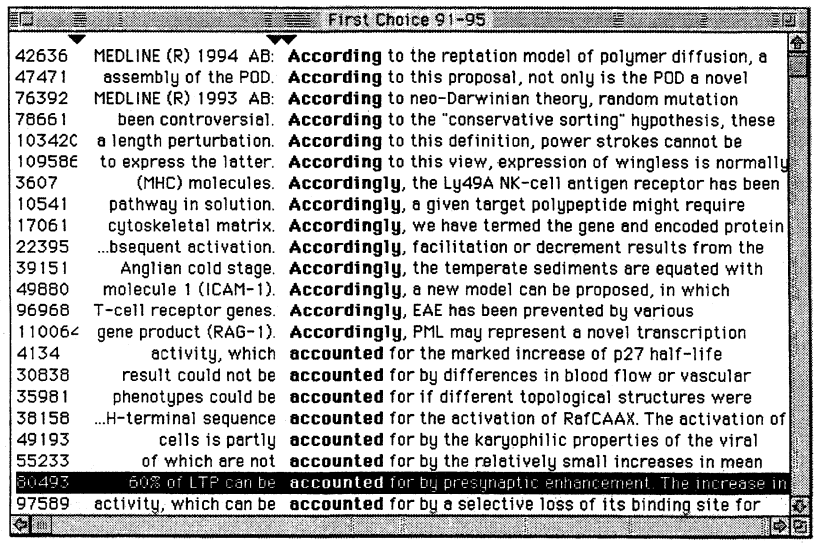

図2　共起表現の分析例（according, accordingly, account for）

けてデータベースに収録した。

次に，この用法を含む文章を実際の論文中から抽出した。材料には著名な学術誌に掲載された過去5年間の論文抄録（約60 Mbyte）のうちEnglish-nativeな著者の論文を用い，これまでに15,000文を発表年，研究が行われた所属機関の国籍，MEDLINEのaccession numberとともにデータベースに収録した。

3. オンライン辞書の制作

学生や研究者が直ちに利用できるよう，データベースから電子辞書を制作した。WebLsdはWWWブラウザでの英和・和英・用例の検索を可能にしたインターネットサーバであり，データベースに基づいて頻繁にデータを更新できる。利用者からの意見をフィードバックするページや用例からMEDLINEへのリンクなどに特徴がある（図3）。

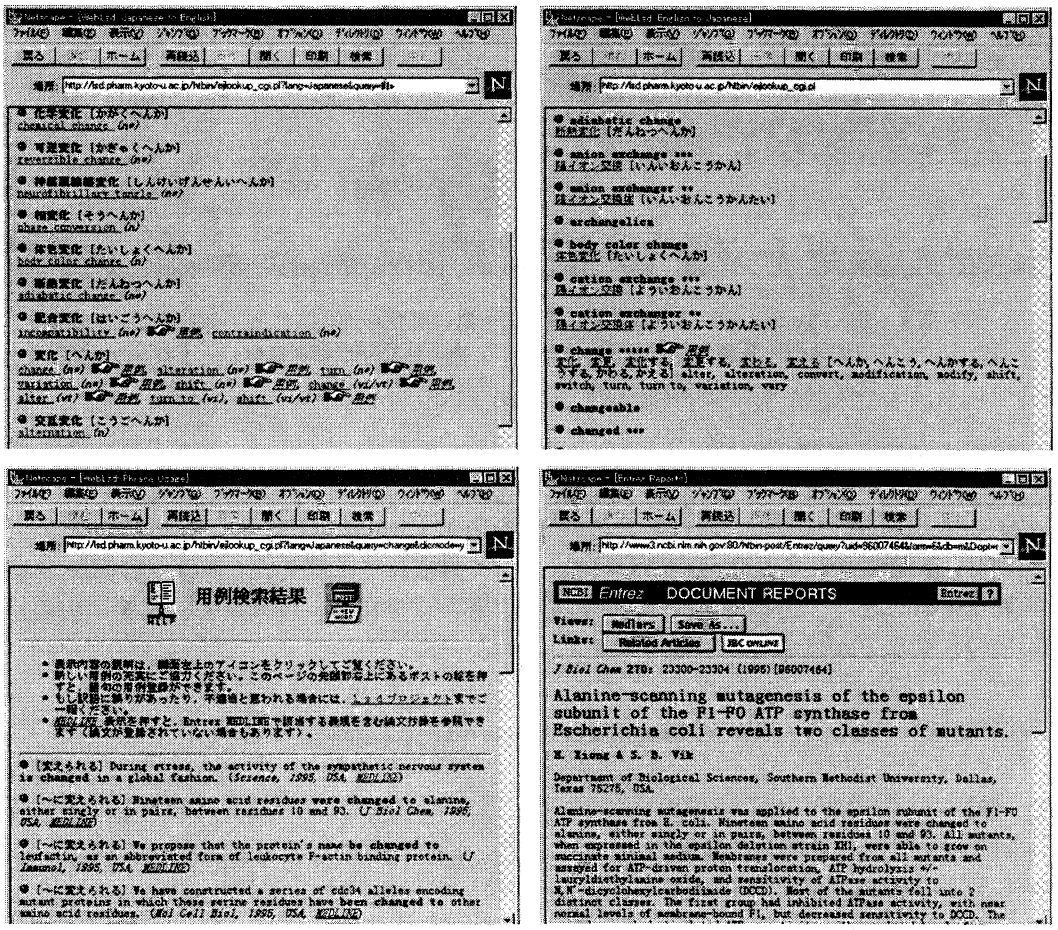

図3　WebLsdの画面表示例。英和表示（左上），和英表示（右上）は互いにクロス参照でき，用例表示（左下）にリンクしている。用例表示からはMEDLINE（右下）にリンクしている。

4. 英和逐次変換システムの制作

英文中の専門用語を簡便かつ高速に英和変換するためのツールEtoJを考案し，perl スクリプトにて制作した。さらに，この結果を WWW ブラウザを使って表示し，英文中の調べたい単語をクリックするだけで意味を表示するツール EtoJ Vocabulary を考案し，これをインターネット上のサーバとして公開した（図４）。

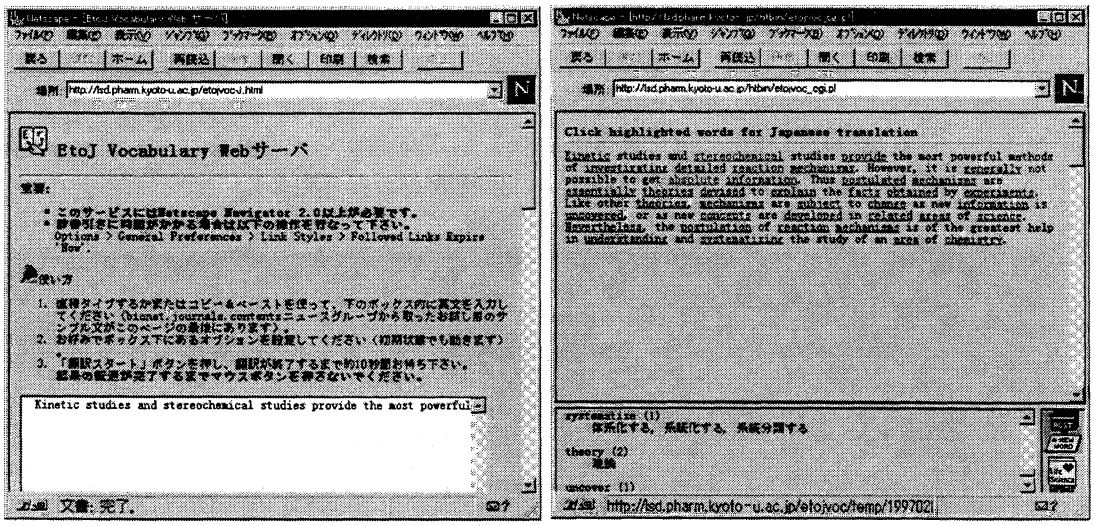

図４　EtoJ Vocabulary サーバ。左画面の入力窓に英文を流し込むと，右画面の上フレームのように辞書とマッチする単語ないし複合語にリンクを発生し，クリックすると下のフレームに意味が表示される。

5. スタンドアロン辞書の制作

データベースを元にして，これまでに30数種類のパソコン用辞書および自動翻訳ツールを制作し，下記アドレスのサーバにおいて，すべて無償にて公開した。学生や研究者がこれらを利用することで，かな漢字変換，英和および和英辞書検索，スペルチェック，用例および用法の検索などが効率化できると考えている。

公開サーバ　http://lsd.pharm.kyoto-u.ac.jp/index-J.html

【謝　辞】
本研究は，文部省科研費基盤研究(B)「インターネットにおける電子辞書利用システムの開発」および研究成果公開促進費（データベース）の助成によって行われた。