英 語 コ ー パ ス 学 会 第 18 回 大 会

日　時　　**2001 年 10 月 6 日(土)**

会　場　　**中央大学多摩キャンパス**（〒192-0393　**東京都八王子市東中野** 742-1　TEL 0426-74-2144）

（**多摩モノレール中央大学・明星大学駅下車**　詳細は http://www.chuo-u.ac.jp/chuo-u/indexM.html 参照）

ワークショップ　10:00－12:00

《BNC World Edition を使いこなす》　**講 師**　**明海大学**　**投野由紀夫**

小学館マルチメディア局　**森田　康夫**

同上　**中村　隆弘**

小学館外国語編集部　**井面　雄次**

同上　**星野　　守**

定員　先着 30 名（予定）　　**参加費**　会員無料・非会員 1,000 円　(申し込みは電子メール・郵便で事務局まで)

受付開始　12:30

開　　会　13:00

１．会長挨拶　**大阪大学**　**今井　光規**

２．臨時総会

３．その他

研究発表　13:30－14:40

司 会　**香川大学**　**永尾　　智**　**明治大学**　**久保田俊彦**

１．タグを使った分析手法による文法研究の精密化：英語における完了相の考察を通して

北海道大学大学院生　**小野　真嗣**

２．科学技術，経済，報道分野における日英語の再帰形の分布について

東京理科大学　**清水　　眞**

独立行政法人通信総合研究所　**村田　真樹**

〈**休　　憩**　14:40－15:00〉

シンポジウム　15:00－17:20

コーパス検索ツール解析：開発者自ら語る検索ツール

司 会　**大東文化大学**　**山崎　俊次**

TXTANA Standard Edition

講 師　**赤瀬川翻訳事務所**　**赤瀬川史朗**

KWIC Concordance for Windows

講 師　**日本大学**　**塚本　　聡**

ビジネスレターコーパス・オンライン KWIC **コンコーダンサ**（BLC KWIC Concordancer）

講 師　**ロゴス語学システム研究所**　**染谷　泰正**

Unix tools **：** perl **を中心に**

講 師　**北海道大学**　**園田　勝英**

閉会の辞　**中央大学**　**新井　洋一**

《**懇親会**　17:45－19:30　　**会費** 4,000 **円**》

英 語 コ ー パ ス 学 会　(Japan Association for English Corpus Studies)

会長　今井光規　　**事務局**　770–8502　**徳島市南常三島町** 1-1　**徳島大学総合科学部　中村純作研究室**

TEL:　088-656-7129　　E-mail: jun@ias.tokushima-u.ac.jp　　**郵便振替口座**　00940-5-250586

URL　http://muse.doshisha.ac.jp/JAECS/index.html

◆**大会当日、入会受付もいたしますので、お誘い合わせの上ご参加下さい（年会費　一般** 5,000 **円　学生** 4,000 **円）．また「当日会員」としての参加も受け付けております（**1,000 **円）**

英語コーパス学会第 18 回大会レジュメ

◆ワークショップ《BNC World Edition を使いこなす》
（講師　投野　由紀夫・森田　康夫・中村　隆弘・井面　雄次・星野　守）

本ワークショップでは先ごろ EU 圏外でも利用可能になった British National Corpus World Edition の基礎的な利用方法の演習を行う．BNC は１億語の巨大なコーパスなので，大別すると BNC 全体での検索をする場合と，特定のジャンルのサブコーパスを利用して比較する場合とに分けられる．前者の場合には，出来るだけ高速検索の可能なシステムを利用する方が得策である．そのため，今回は BNC の専用ソフトの Sara と小学館が開発中の Corpus Query System の２つを紹介する（注：Sara は投野の PC によるデモのみ．小学館 CQS は web-interface で全員で触れる）．また，サブコーパスの利用の際には必要なファイルが効率的に選択出来るように，XML format の概要を説明し，ファイル選択の手法などを，次のアウトラインに従って説明する．

(1) BNC World Edition の概要説明
　－corpus 全体の構成
　－XML format の説明
　－Subcorpus 選択の方法
　－WordSmith などでの subcorpus 読み込み方法
(2) Sara32 の概要と長所・短所（投野のデモ）
(3) Shougakukan Corpus Query System による BNC の演習
　－基本操作
　－サブコーパス検索
　－フレーズ，lemma，POS 検索
　－Collocation/colligation 統計
　　(T/MI/log-likelihood/log-log)
　－その他

◆研究発表

●タグを使った分析手法による文法研究の精密化：英語における完了相の考察を通して

（小野　真嗣）

本発表では，英語における完了相に焦点をあて，タグ付きコーパスを使うことによって，従来の研究からさらに精密化した観察が可能であることを示したい．

コーパスに基づく完了相の先行研究の一例として，Biber *et al*. (1999) と Ota (1963) が挙げられる．Biber *et al*. (1999) では，単純時制，完了相，進行相における言語使用域別の頻度；現在完了相，過去完了相の言語使用域別の頻度；完了相，進行相の方言別の頻度；現在完了相と動詞語彙の関係；get と have の現在完了形の方言別頻度；過去完了相と動詞語彙の関係；過去時制と過去完了相の相対的使用度に関するコーパス分析結果を与えている．

一方，Ota (1963) では，単純時制，完了相，進行相の本質的意味を，動詞や副詞句との共起関係に基づいた分析によって示しており，完了相では「ある期間における行動の生起や状態の存在」を，進行相は「行動の過程」をそれぞれの文法的意味としている．さらに多くの学者によって示されてきた “completion” や “continuation” などの完了相の意味は，完了相の文法的意味ではなく，その文法的意味と共起する動詞の語彙的意味の相互作用に起因するものであると分析している．しかしながら，Ota (1963) で行われた英語におけるコーパスに基づいた時制研究は先駆的であるものの，その種のコーパス研究はその後はあまり行われておらず，Biber *et al*. (1999) でも，言語使用域間の変異に焦点が置かれ，時制あるいは相そのものについての分析は行われていない．

Ota (1963) では，次の２点が分析できなかった問題として挙げられている．一つは単純現在形，現在完了形，現在進行形における副詞 “already,” “yet,” “still” の共起に関して，「状態動詞」「動作動詞」別の数的分布を示したが，分析の際使用されたコーパスの規模による制約で十分な結果が得られなかったことが問題として残された．もう一つは進行相と共起する動詞に関して，頻度を基準に 5 つの分類をし，分析を行ったが，最も頻度の低い動詞分類では「状態動詞」「動作動詞」という概念を使用した分析では例外が多く十分な特徴付けを行うことができなかった点が問題として残された．

本発表では，Ota (1963) で指摘された 2 つの問題

点を中心にタグ付きコーパスを分析する．また，完了相の派生的意味として下位区分され所謂学校文法で言われている継続用法，経験用法，完了用法について，動詞や副詞との共起関係を，コーパスのタグを利用した分析によって明示的に示す．さらに，その動詞や時の表現との共起関係に基づいて，さらに精密化した完了相のコーパス研究の可能性を提示したい．

本研究で使用したコーパスは，Brill の Rule Based Tagger 1.14 によって処理されたテキストであり，このテキストに付けられたタグをもとに適切なテキスト処理を行うこととした．本発表のために行った一連の研究が，従来の pattern matching と sorting による文法研究を発展させるものであることを示し，タグ付きコーパスに基づいたさらに精密化した文法研究が可能であることを示したい．

●科学技術，経済，報道分野における日英語の再帰形の分布について

（清水　眞・村田　真樹）

日英語の指示表現は，代名詞，再帰形，名詞句に分類することができる．その対応関係を考える時，従来，それらは対象言語のそれぞれの代名詞，再帰形，名詞句に対応すると考えられてきた．しかし，コンピュータコーパスには多くの反証を見いだすことができる．

清水他(1997a，b)，清水(1998，2000)は，和英辞典，文学作品等のパラレルコーパスを用い，再帰形を中心に日本語と英語の指示表現を研究した．英語の再帰形が日本語においてどのような表現に翻訳されているか，また，どのような英語の表現が日本語の再帰形に翻訳されているか，それぞれの翻訳パターンを分析し，統計をとったのである．その結果，翻訳の方向の違いにはあまり関係なく，日本語の再帰形に対応する英語の指示表現は，頻度の順に，代名詞 (65-70%)，再帰形 (13-38%)，空表現 (4-14%)，名詞句 (1-7%)，self+X (self-disgust 等，0-3%) であり，英語の再帰形に対応する日本語の指示表現は，頻度の順に，空表現 (50-70%)，再帰形 (11-30%)，名詞句 (5-24%)，代名詞 (3-7%)，自+X (「自活」等，1-4%) であるということがわかった．

この発表では，別のデータ，すなわち，科学技術，経済，報道の各分野のパラレル コーパスより例を検索し，同様の分析を行う．清水他(1997a，b)，清水(1998，2000)の分析結果と比較し，科学技術，経済，報道の各分野の独自性，あるいは，辞書，文学作品との類似性を，統語的，意味的に考察する．特に，日本語の空表現に対応する英語の再帰形は，いくつかのサブタイプに分類できるようである．その分析の仕方，取扱いを考えてみたい．

本発表の発表者のうち，清水はデータの言語学的分析，議論を，村田はデータの自然言語処理的分析，議論を担当した．

◆シンポジウム

●《コーパス検索ツール解析：　開発者自ら語る検索ツール》

（司会　山崎　俊次）

コンピュータコーパスを用いた言語研究には，検索ソフトウェアが不可欠である．一般に使用されている検索ツールには，WordSmith，TXTANA，KWIC Concordance，Conc などがあり，これらは，その機能においてそれぞれ特徴を持っている．本シンポジウムでは，TXTANA，KWIC Concordance，オンライン Web 検索システム，Unix 上のツール（perl，awk など）について議論を行うが、その目的は各ツール間に優劣をつけることではない．一般のソフトウェアでも，高機能，高価格のものが望ましいものとは限らず，ソフトウェアには適材適所があることは明らかである．同様に，これらのコーパス検索用のツールにも，長所，短所，処理量の制限や処理速度など，それぞれに特徴がある．

本シンポジウムでは，各ツールについて，開発の動機・趣旨，ツールの特徴（長所・短所），使用環境・開発言語，インストールの手順，利用可能なコーパスの種類などをキーポイントに，開発者自身が講師となり，実践しながら紹介し，議論を行う．それぞれのツールの長所を分かり易く論じてもらうことと、複数のツールを同時に取り上げるため，各々の特徴が整理されるとともに，各ツールの価値がよりいっそう理解しやすくなることを期待している．また，全ての目的に対応できるツールは残念ながら存在しないため，必要に応じて，既存のアプリケーションソフトや，複数のツールを組み合わせることにより，効率的にコーパスの検索，データ整理が可能になるような利用例も紹介する．

本シンポジウムの特徴の一つは，各ツールの開発者自身が講師として発表にあたり、これらのツールの要となる内部処理の仕組みなども紹介してもらうことである．

このことにより、日頃プログラミングそのものにはあまり関心を持たない一般のユーザーに対して，その重要性を認識してもらうと共に、プログラミングをもっと身近なものと感じてもらうこともねらいの一つである．

●TXTANA Standard Edition
（赤瀬川史朗）

TXTANA Standard Edition（以下，TXTANA SE）は Windows98/Me/NT/2000 で動作するコンコーダンサである．コンコーダンサは２種類に大別できる．１つはコーパスファイルから言語情報を抽出してファイルとして保存するコンコーダンサ（情報抽出型），もう一つはコーパスファイルからの検索結果をユーザがリアルタイムで検討できるコンコーダンサ（情報吟味型）である．TXTANA SE のモデルとなった WordSmith Concordancer や MonoConc は後者のタイプで，TXTANA SE ではさらにそうした機能を拡充するために Pascal に基づいたプログラミング言語 Delphi で開発した．

情報吟味型のコンコーダンサの基本機能としては，KWIC コンコーダンスラインのリアルタイムの並べ替え，コロケーション情報の表示，コーパス原文の参照などが挙げられるが，TXTANA SE では，検索結果を絞り込むためのクエリー機能（クエリーウィンドウ），共起語の頻度を表示しコンコーダンスラインと連動させた機能（フリークエンシビュー），品詞ごとの単語リストを使用したフィルタリング機能などを追加し，検索結果を精緻にしかも多角的に検討できるように配慮している．検索機能でも他に見られない特長を持つ．その一つがコンセプト辞書である．この辞書を利用すると，複数の語を１つのキーワードにまとめて検索することができるので，煩わしい検索式を入力せずに済む．この他，正規表現検索，センテンス単位の検索，排除語の指定，無作為抽出など，言語分析に必要な機能を搭載している．コロケーション統計では，WordSmith のコロケーション統計に相当するシートビューと，CobuildDirect の Picture 画面に相当するピクチャービューの２方式の表示が可能である．大規模コーパスへの対応も TXTANA SE の重要な課題であった．数万件に及ぶ検索結果を柔軟に処理できるように，内部にデータベースを使用している．今後の機能拡張としては，WordSmith のワードリストに相当するプログラムの開発，タグ付きコーパスへの対応，MI スコアなどの統計機能の充実などを考えている．

●KWIC Concordance for Windows
（塚本　聡）

本シンポジウムで取り上げる KWIC Concordance for Windows（以下 KWIC）は，Windows 上で作動するソフトウェアである．本ソフトウェアは，MS-DOS 上で作動する，COCOA 形式で作成された Helsinki コーパスを検索する際に便利であった Micro-OCP をモデルに，その主要な機能を網羅するよう C++を用いて開発された．開発当初は，Micro-OCP をモデルとしたが，以後，WordSmith なども参考に機能を拡張し，また，当初の目標コーパスであった Helsinki コーパスのみでなく，Brown や Frown など他の既存のコーパスを扱う際の利便性も考慮されている．

現在広く使用されている WordSmith は，高機能な検索ソフトとして一般的であるのに対し，KWIC は，統計情報や，ファイル上の分布度合いを示す機能などを持っていない．さらに，検索結果を絞り込んだり，コーパス内の原文を参照したり，参照範囲をユーザが検索結果を見ながら随時変更したりといったデータベース的な機能を持たない．しかしながら，KWIC は，SGML，固定長，COCOA などの各種のコーパス形式に対応しており，これらの形式で編集されたコーパスから，タグ情報を取り出すことが最大の特徴である．単に，用例を検索するのではなく，ある種の語彙・用法をもとに，コーパスの構成や，コーパス間の比較検討などには，このようなタグ情報を取り出すことは必須であり，有効な機能である．残念ながら，これらのタグ情報を利用する際にも，上記の短所であるデータベース的な機能を有していないため，別途データベースソフトなどを使用する必要がある．

以上のような長所・短所があるため，それらを認識した上で，必要に応じて，使用されることが望ましい．シンポジウムでは，FLOB コーパスを利用して，長所の一つである，タグ情報の取り出し，およびそのタグを利用してデータベース化を行う点を実例として紹介する．

●ビジネスレターコーパス・オンライン KWIC コンコーダンサ（BLC KWIC Concordancer）
（染谷　泰正）

"accuracy" の向上を主眼にした英文ライティングの指導は，通例，学習者が書いた英文を教師が添削するという方法で行われることが多い．ただし，その教育上の

効果は必ずしも明らかではなく，最近の研究の多くはむしろその効果を疑問視している．これは，いわゆる "after-the-fact corrective intervention" という従来の指導方法の限界を示している．これに対して「エラーが起こる前にその発生を防ぐ」ための方法というのは，これまでほとんど考慮されることがなかった. 本シンポジウムで紹介する BLC KWIC Concordancer は，そのひとつの具体例として開発されたもので，学習者（とりわけ「ビジネス英語」学習者）が，自分が書いた英文のうち疑問を持った箇所についてその適格性をデータに基づいて検討し，自ら判断するためのツールを提供することを意図したものである．

BLC KWIC Concordancer はオンライン CGI プログラムであり, インターネットへのアクセスが可能なユーザは, 何らの制限なくいつでも自由にこれを使用することができる．ユーザ・インターフェイスはきわめてシンプルなもので, 初めてのユーザでもすぐに使えるようになっている．現在のところおよそ 18 種類のコーパスが実装されており，ユーザは，英文を書いていて特定の語句や表現の適格性に疑問があった場合, 任意のキーワードを検索語としてこれらのコーパスから用例検索をすることで, それぞれの適格性について判断するためのデータを簡単に取得することができる．なお，これらのコーパスのうち，メインになっているのは Business Letter Corpus (BLC2000) と POS-tagged BLC の２つで，前者には 1970 年代以降に発行された英米その他の出版物から収集したデータがおよそ 100 万語収録されている. 後者は BLC2000 のデータに品詞タグを付与したものである．このほか，サブプログラムとして Bigram Plus がある. 本シンポジウムではこれについても簡単にご紹介するつもりである．

●Unix Tools：Perl を中心に

（園田　勝英）

コーパス研究の道具としての perl を紹介する. perl はプログラム言語であるが，「簡易インタプリタ言語」と言われる類に属し，習得が容易である．コーパス研究者から見たときの最大の長所は, 非常に強力な正規表現が使える上に, その正規表現を活用できるように多くの便利な仕組みが用意されていることである．KWIC コンコーダンスを作成したり，語の頻度を求めたり，タグ付きコーパスからタグを取り去るといった再編集をするのに，使うことが出来る．以上のことを実例とともに説明する．

perl は UNIX 流のテキスト処理の伝統を受け継ぐ道具である．このことは perl の利点でもあり欠点でもある．利点としては，他のテキスト処理用の道具との連携が容易であることがある．欠点であるというのは，perl を使いこなすためには，UNIX 流のテキスト処理に慣れる必要があることである. この欠点の乗り越え方の一つとして，Windows での perl の利用方法を紹介する．

perl の別の長所として, コンピュータ界の発展と歩調を合わせて perl も進歩を続けていることがある．（これは perl をフリーソフトウェアとして提供している開発者の Larry Wall 氏に負うところが大きい. ）例えば，日本語へ対応した perl として jperl があり，英語のテキスト処理で習得した正規表現をそのまま日本語に適用することができる．さらに，SGML や XML によって構造化されたテキストの扱いも容易にできるように拡張モジュールが開発されている. 今後の英語コーパス研究の発展にも十分対応できる道具であること力説したい．