英 語 コ ー パ ス 学 会 第 １９ 回 大 会

日　　時　　**2002 年 4 月 20 日(土)**

会　　場　　**大阪大学言語文化研究科**（〒560-0043　豊中市待兼山町 1-8　TEL 06-6850-6111）

（阪急宝塚線石橋駅または大阪モノレール柴原駅より徒歩 15 分　詳細は http://www.lang.osaka-u.ac.jp/index.html 参照）

ワークショップ　10:00－12:00

《BNC 検索入門》　　講 師　徳島大学　**中村　純作**

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)

なお、ワークショップ会場は大会会場とは別棟のA棟３階３１５号室です。当日の案内に従って下さい。

受付開始　12:30

開　　会　13:00

１．会長挨拶　　摂南大学　**今井　光規**

２．総会

３．その他

研究発表　第 1 セッション　13:30－14:30

司 会　北海道大学　**園田　勝英**　東海大学　**朝尾幸次郎**

１．The *Paston Letters* **の** XML **版コーパスの作成とその課題・問題点**

東京慈恵会医科大学　**小原　　平**

２．応用言語学的視点による英語コーパスの解析

豊田工業高等専門学校　**高橋　　薫**

名古屋大学　**山下　淳子**

愛知学院大学　**伊藤　彰浩**

〈休　　憩　14:30－14:40〉

研究発表　第 2 セッション　14:40－15:40

司 会　大手前大学　**西村　道信**　元龍谷大学　**正保　富三**

３．D.H. ロレンス詩におけるイメージの展開　－作品コーパスの構成語彙分析に基づく研究－

広島国際大学　**石川慎一郎**

４．押韻俗語表現の特徴

大阪大学　**渡部眞一郎**

〈休　　憩　14:40－16:00〉

特別講演　16:00－17:30

司　会　大東文化大学　**齊藤　俊雄**

Corpus Linguistics: Past, Present and Future

講　師 Prof. Emeritus Nijmegen University　Jan　Aarts

閉会の辞　大阪大学　**田畑　智司**

《**懇親会**　17:45－19:30　会費 4,000 円》

司　会　神戸市外国語大学　**家入　葉子**

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

会長　今井光規　　事務局　770????2??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 19 回大会レジュメ

◆ワークショップ《BNC 検索入門》

（講師　中村　純作）

1 億語のイギリス英語を収録した BNC は 1994 年に完成, 翌年 2 月にヨーロッパの研究者を対象にリリースされるが，我々日本の研究者はインターネット上の BNC ホームページで簡単な検索を行うか，1999 年 3 月にリリースされた 200 万語の BNC Sampler を利用した研究しか行うことができなかった．2000 年 12 月になってようやく BNC World Edition が日本でも入手できるようになり，BNC を利用した本格的な研究が可能となる．この間，JAECS では，1999 年秋の第 14 回大会でワークショップ「BNC のデータ構造と SARA による検索」を開催，BNC 開発に重要な役割を果たした Leech 先生にも“ Corpus Linguistics and the BNC ”と題した講演をお願いした．また，昨年秋の第 18 回大会では，「BNC World Edition を使いこなす」と題して投野由紀夫先生（明海大学）と小学館のインターネット上での検索ソフト開発担当者にワークショップを依頼するなど，BNC に関する情報提供を行ってきた．ただ，複雑な付加情報を持った 1 億語のコーパスは，なかなか簡単には利用できないとの声も聞かれるので，今回は，もう一度その利用法をワークショップとして取り上げることにした．

BNC には，幅広い研究に対応できるよう全てのテキストにきめ細かい情報が付与されている．例えば，書き言葉には，そのテキストがどのようなドメイン（創作，純粋科学など 9 ジャンル）の，どのような媒体（書籍，定期刊行物など）から抽出されたものか，著者の性別，年齢，地域など，対象となる読者の年齢，性別など，刊行地，抽出法に関する情報なども含まれている．これらは全て SGML の Tag 情報として付加されているので，BNC 全体がどのような構造になっているかを知ることが BNC を使いこなすための大前提となる．さらに，これらの付加情報と組合わせた特定のキーワードによるコンコーダンス作成には，複合検索を行う必要が生じる．今回は，罵り言葉（例えば shit）の検索を主なトピックとして，書き言葉と話し言葉での対比，書き言葉でのジャンル別検索，話し言葉での性別，年齢別，社会階層別の検索を SGML 対応の検索ソフト SARA を利用して実際に体験して頂く予定である．

ただし，今回のワークショップでは全ての参加者に BNC World Edition を準備することは不可能であり，たとえ参加者に BNC World Edition そのものを持参してもらっても，インストールしている時間的な余裕は無いということもあり，多少 Version は違うものの BNC Sampler をご持参頂き，使用することにした．こちらの方は，インストールにそれほど時間もかからず，検索方法は基本的には全く同じなので，BNC 検索入門としては支障がないものと考えられる．コーパスそのもののサイズは 200 万語と限られてはいるが，学生を対象としたコーパス検索の実習には最適の CD-ROM だと思われるので，この際，お持ちで無い方には購入されることをお勧めしたい．価格は 10 枚までが 30 ポンド，それ以上は 20%引き，ただ BNC World Edition が 50 ポンドなので，多少高めの感じです．Web サイトから注文書をダウンロードし，郵送する形で入手可能．BNC World Edition をお持ちで，Sampler をお持ちでない方には，ある程度は準備可能です．事務局までその旨ご相談下さい．

◆研究発表

●The *Paston Letters* の XML 版コーパスの作成とその課題・問題点

（小原　平）

本発表では，15 世紀英国の The *Paston Letter*s の XML 形式によるコーパスの作成における課題とその問題点について報告する．The *Paston Letters* は，Norman Davis によるテキストの第 1 巻（1971）の部分が，早くから電子テキストとして OTA から公開されているが，これは単に文字テキストの部分の電子化であって，Davis のテキストの，それぞれの書簡の始めに書かれている付加的情報（例えば，書かれた年月日，紙のサイズ，実際に手紙を書いた人の名前等，書誌学的な情報）については一切触れられていない．さらに本来のマニュスクリプトから読み取れる情報（例えば，省略文字，上付き文字，文字の異型等に関する言語学的情報）もまったく無視されている．後者の情報に関していえば，電子テキストの元となった Davis のテキストにおいても，省略文字に関しての記述はなされているが，上付き文字や文字の異型に

関しては，注で若干述べられているに過ぎない.

コーパスの使用目的によっては, これらの付加情報の全てが必ずしも必要ではない. ただ筆者のようにこの書簡集を特化して研究の対象としている者にとっては, それぞれのマニュスクリプトから読み取れる情報が, 出来るだけ多くコーパスの中に凝縮されれば, それだけそのコーパスは，利用する価値の高いものとなる.

The *Paston Letters* は，Davis のテキストから判断しても，1000 通近くの書簡から構成されており，含まれているデータを整理してとらえようとするためには, 使いやすいコーパスの存在は不可欠である. 今回は最近注目され始めた XML の形式で，The *Paston Letters* のコーパスの再構築を試みるが，HTML と違い，デザインと内容を分離させた XML の形式でどのような新しいコーパスが作成できるのか, PPCME2 のような機械に読み込ませる専用のコーパスではなくて, 人間の目で見ても分かりやすいコーパスを設計するためには, どうすればよいのかを考えて, 筆者なりの結論を出し，出席者のご判断を仰ぎたいと思う.

なおこの報告は, 平成 13 年度日本学術振興会科学研究費補助金基盤研究 C「パストンレターズにおける完全版電子コーパス化の製作に基づく書記素の研究」（課題番号 13610592）の一環としてなされる研究に基づいている.

●応用言語学的視点による英語コーパスの解析

（高橋　薫・山下　淳子・伊藤　彰浩）

大規模コーパスにおける各ジャンルカテゴリーの linguistic features の生起率に注目して，多変量統計解析を行うことで, それぞれのジャンルに言語学的に解釈可能な複数の尺度上での数値が与えられ, このような手法 (Multi-feature and multi-dimensional approach) がテキストの類型化に役立つことを Biber (1988)が示した.

コーパス言語学では，2000 年 10 月より新たに公開された British National Corpus (BNC) により研究のさらなる発展が期待できるようになった．BNC はテキストに descriptive feature に関する情報を含むという注目すべき特徴を持つ. 具体的には，target age group (child, teenager, adult), target sex, social class, region (south, midlands, north) 等が挙げられる．これらの指標は，応用言語学，社会言語学への橋渡しとなることが期待できる．本発表では，コーパス言語学と応用言語学の接点を探り, コーパス言語学の新たな方向性を示唆するという目的のため, BNC に統計的解析を適用し以下の２点から分析結果を報告する.

(a) Readability（テキストの読みやすさ）の観点から

テキストの読みやすさの研究は, 書き言葉の情報伝達媒体としての効率性への興味から長い歴史を持つ. 言語（母語）教育の中では 1920 年代くらいから初等中等教育の教科書作成のために readability の研究が応用されてきた. 様々な研究の中で共通して指摘される要因は単語の難易度と文の複雑さであった．Takahashi (1996)は LOB コーパスの分析において，テキストの難易度を決定する要因と解釈できる３つの尺度を同定した（1. 統語構造　2. テキストの内容　3. 時制）．ここではその手法を生かしながら, 子供向けと大人向けのテキストの違いがどういう点から現れてくるのかを，BNC の target age group という下位範疇を利用することによって検討し，その結果を readability の観点から考察する.

(b) 関係節の難易度決定の観点から

次に，言語産出の観点から，関係節の難度決定要因としての枝型と名詞句の種類に着目する. 応用言語学では, 枝型および名詞の役割の違いが英語関係節の産出における難度に影響を与える傾向があるといわれている. その妥当性を明らかにするためコーパスの単語に付加された linguistic feature の情報をもとに，目的とする文構造を判断するプログラムにより分析を行い, すでに仮定された「テキスト難易度」の尺度により理論の検証を行う.

●D.H.ロレンス詩におけるイメージの展開 ―作品コーパスの構成語彙分析に基づく研究―

（石川　慎一郎）

文学作品の研究には，思想やイメージという内在的な要素を重視する伝統的なアプローチと, テクストの表面に顕現した言語上の諸特徴に着目する言語学的なアプローチとがある．今日，文学研究へのコンピュータの援用はもはや目新しいことではないが, それはどちらかといえば後者のアプローチに限られているように思われる．しかしながら，テクストの電子化とその機械的な解析は, 作品の内部世界を探る上でもきわめて有効な手段でありうる.

本発表では，20 世紀初頭に活躍した文学者 D.H.ロレンスの詩作品を取り上げる．ロレンスの詩には，基本的

なテーマとして，常に，生と死という二項対立がつきまとっているわけであるが，両イメージの相関性は先行研究においてもいまだ十分には明らかにされていない．そこで本発表では，ロレンス詩の一部を電子化し，その構成語彙を調査することで，生と死というイメージが具体的な詩作においてどのような位置付けにあるかを概観することとしたい．

発表ではまず，ロレンスの『最後詩集』について，それを「一面的な死の世界」とみなす立場と「生死のイメージが混在した世界」とみなす二つの矛盾した立場が先行研究に存在していることを指摘する．ついで，この矛盾を語彙の観点から解明するため，OCR を用いて『最後詩集』全体の電子化を行う．こうして得られた電子データを構成語彙に分解し，レマ化リストを作成したうえで，生と死を直接的・間接的に含意する語彙群の出現状況，および高頻度語彙の内容的特徴を管見する．これにより『最後詩集』の概括的な見取り図を得た後，今度はデータをコンコーダンサにかけ，生と死を含意する語彙の実際の出現状況を KWIC 画面で詳しく分析してゆく．以上の検証をふまえて，ロレンス詩における生と死のイメージの出現の特徴，さらには冒頭の解釈上の矛盾が発生した原因などについて，いくらかの考察を加えてゆく．

●押韻俗語表現の特徴

（渡部　眞一郎）

本発表は押韻俗語（rhyming slang）と呼ばれる言語表現の種々の特徴に関して，*OED*² on CD-ROM 等の辞書から抽出して作成した同表現のコーパスに基づいて論じる．また，その作成法についても触れたい．押韻俗語とは 19 世紀頃，ロンドンで生まれたとされる俗語表現で，現在でも，ロンドンのコックニーやオーストラリア，ニュージーランドの方言にみられる．この表現はもはや地域限定的表現ではなく，各種メディアでも一般的に用いられるようになってきているが，ほとんど研究の対象になることはなかったようである．

押韻俗語表現は，一般的に言えば，２つの内容語からなる表現形式をもち，その最後の語が意図される語と押韻する．たとえば，“Use your loaf of bread, get on t he dog and bone and call your trouble and strife.”という押韻俗語表現を含む文の意図するところは，“Use your head (loaf of bread) and call your wife (trouble and strife) on the phone (dog and bone).”である．これらの押韻俗語表現の最も興味深い点は，*the dog and bone* と *phone* の例のように，両者の意味はあまりにもかけ離れ，全く意味的なつながりを見いだせないものであるにもかかわらず，前者が後者を意図する関係が成立していて，この関係が同じ脚韻をもつという点だけに依存していることである．一方，押韻俗語表現 A *and/of* B の A の位置に起こる語は意図される語とは直接関わらないが，この位置に起こる語はどのような語でもよいわけではなく，B の位置で起こる語と意味的な連想関係にある．このように，押韻俗語表現の成立には一方で脚韻連想，他方で意味連想が関与している．

この押韻俗語表現は散発的に起こるようなものでなく，ひとつの語彙表現体系あるいは修辞法をなしていると言えるものである．本発表では，この押韻俗語表現の種々の特徴，意図される語の品詞，連語の意味的関係，脚韻語と意図される語の脚韻と音節数，意図される語の意味分類等について述べたい．

◆特別講演

●Corpus Linguistics: Past, Present and Future

（Jan Aarts）

It is only a few years ago that Chomsky in an interview maintained that corpus linguistics, as a linguistic discipline, does not exist, thus showing that where corpora are concerned his opinion has not changed over the last forty years. At a time when the use of corpora has become accepted practice in mainstream linguistics, such a statement might be shrugged off as the opinion of an old diehard, but it is perhaps wiser to see it as an occasion for some reflection on a discipline that has gone through such a rapid development during the last few decades that there has hardly been time to think very deeply about the nature of the ever-growing cluster of research activities that we have come to call ‘corpus linguistics’. In this paper I want to make a critical appraisal of the various aspects of the recent history and current practice of corpus linguistics and, on the basis of this, formulate some hopeful expectations with regard to its future development. My overall conclusion will be that corpus linguistics *does* exist and can look forward to an interesting future.

I shall start with some reflections on the nature of corpus data and their relation to other kinds of linguistic data:

introspective, elicitation and 'anecdotal' data. This will be followed by a bird's eye view of the developments that we have seen in the compilation of English language corpora in the last forty years. In many cases corpus building has also included linguistic annotation of some kind; I shall therefore also deal with the pros and cons of annotation in its various kinds, a topic that partly reflects my personal history in corpus linguistics. Having dealt with the various databases that have become available, we must then look at the ways they have been and are being put to use in English linguistics. This naturally leads to questions of corpus linguistic methodology, an issue that quite recently is receiving the renewed attention it deserves. After that, we shall have dealt with a sufficient number of aspects of corpus linguistics to have a basis for an informed look at its future.