英語コーパス学会第14回大会

日　　時　1999年10月9日(土)

会　　場　日本大学生産工学部実籾校舎（千葉県習志野市新栄2-11-1）

（京成本線：京成上野駅より急行で52分、実籾駅下車、徒歩7分。http://www.mmm.cit.nihon-u.ac.jp 参照）

ワークショップ　　10:30－12:00

《BNCのデータ構造とSARAによる検索》　　講　師　　徳島大学 中村　純作　梗概

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。できればBNC Samplerをご持参下さい。）

受付開始　　12:30

開　　会　　13:00

１．会長挨拶　　大東文化大学　　齊藤　俊雄

２．日本大学生産工学部長挨拶

３．その他

研究発表　第1セッション　13:15－14:15

司会　　神戸大学　　西村　　秀夫　　北海道大学　園田　勝英

１．電子版脚韻インデックスの開発と音韻変化を電子検索する試み：
AWKを用いた文字列処理課程を中心に　　大阪大学大学院生　　遠藤　裕昭　　梗概

２．　初期中英語における指示代名詞‘e(that)’と‘es(this)’の形態
タイプ別、意味別頻度数を中心として　　学習院大学院生　　新井　恭子　　梗概

〈休　憩　　14:15-14:30〉

研究発表　第 2 セッション　14:30－15:30

司会　　山形大学　　岡田　毅　　中央大学　　新井　洋一

１．　コーパス資料に基づく現代ビジネス英文の語彙的特性の研究　　梗概
東京大学院生　　染谷　泰正

２．　ウェブを用いた学習者用用例検索システムの作成　　梗概
群馬大学　　大名　　力

〈休　憩　　15:30--15:45〉

特別講演　　15:45--17:15

司会　　大東文化大学　　山崎　俊次

“Corpus Linguistics and the BNC”

講師　　ランカスター大学　　Geoffrey N. Leech

梗概

閉会の辞　　日本大学　　塚本　　聡

《懇親会　　17:45?19:30　　会費 4,000 円》

司会　　慶応大学　　吉村　由佳

英語コーパス学会　　(Japan Association for English Corpus Studies)

会長　斎藤俊雄

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　0886-56-7129　　郵便振替口座　00940-5-250586（英語コーパス学会）

E-mail: (E-mail address deleted)

URL ../../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 4,000 円　学生 3,000 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 14 回大会レジュメ

◆ ワークショップ《BNC のデータ構造と SARA による検索》

（講師　中村　純作）

JAECS Newsletter No. 25 でもお知らせしましたように，3 月下旬に BNC Sampler がリリースされました．すでに購入した会員諸氏も多いことと思います．この CD-ROM は，できるだけもとのコーパスのバランスを崩さないように BNC から抽出された 200 万語のテキスト（完全な書誌情報付きの書き言葉とやはり完全な documentation を備えた話し言葉，各々100 万語からなる 184 のテキスト）を収めたものです．付属の検索ソフトはすべて SGML 形式に対応したもので，そのうち Birmingham 大学の Oliver Mason の開発した Qwick，Oxford 大学で開発された SARA 及び Liverpool 大学の Mike Scott の WordSmith Tools の 3 つは Windows の環境（Windows 95, 98 あるいは NT)で動きます．これらの検索ソフトはこの CD-ROM 上のテキスト検索にしか使うことが出来ませんが，すべての機能を利用できる完全版です．それぞれ固有の特徴を持っているこれらの検索ソフトを用いることにより，研究目的に合わせた色々な作業をこなすことが可能だと言われています．今回は，複雑な SGML 形式の情報が付加された BNC のデータ構造そのものを理解することと，上記検索ソフトのうち SARA を利用した検索のデモンストレイションを行いたいと思っています．

ただ，今大会の開催校では BNC Sampler の CD-ROM を参加者全員に準備することが出来ませんので，お手数ですが CD-ROM そのものを持参して戴くことになります．コーパスそのもののサイズは 200 万語と限られてはいますが，学生を対象としたコーパス検索の実習には持ってこいの CD-ROM だと思われますので，この際，お持ちで無い方は，購入されることをお勧めします．価格は 10 枚までが 30 ポンド，それ以上は 20%引きになり，Web サイトから注文書をダウンロードして，郵送する形で手に入ります．賛助会員の桐原ユニからも 7,000 円で購入可能です．

◆研究発表　第 1 セッション

● 電子版脚韻インデックスの開発と音韻変化を電子検索する試み： AWK を用いた文字列処理課程を中心に

（遠藤　裕昭）

電子コーパスによる大規模な用例収集が可能になり，特に統語論，語用論の分野では研究効率，そして観察の妥当性ともに有効性が認められてきたが，残念ながら通常の書記英語には反映されにくい音韻情報を電子的に検索することはかなり困難である．14 世紀以前の中英語ではどうかというと，正書法が決まっていない時代であり，書いた本人の発音癖を綴字

が反映している可能性を考慮すると特定の文字列（綴字）を検索しての研究が成立しそうなものである．

しかし実際には，流通させる際に手書きで複写している写字生が綴字を改変しているのが普通であり，かつ作者自筆の原稿が残っていないという毎度の状況では，その発音癖がいつ，誰によって加えられたのかを特定することはできず，複数名の発音癖が混在しているテキストを分析しても発音変化の実証研究にはならないのである．これまでの英語音韻史研究は，その当てにならない綴字ではなく，当時多量に生産された頭韻詩や脚韻詩を用いて行なわれてきた（The Canterbury Tales 等が有名である）．

発表者が研究対象とする脚韻詩の場合，行の最後に位置する脚韻語が 2 行ないしそれ以上にわたって強勢位置の音素を合わせてある一方で，それらの語源情報を比較し，異なる語源幹母音を持った単語が脚韻している場合に，発音変化の一般傾向を考慮しつつ，発音変化を表す脚韻例として採取することができる．発表者はその方法論を踏襲しつつ，電子テキスト版脚韻詩において語源情報を記号として脚韻語に添付し，語源音素記号を検索キーとして高速な用例収集が行なえるよう，特殊なタグ付きコーパスを開発した．また，複数の脚韻詩コーパスを横断検索し，脚韻語および語源音韻記号を回収・表示するためのＡＷＫスクリプトも作成した．本発表においてはこの音韻史研究用電子コーパスの構造と構築過程，検索用ＡＷＫスクリプトの構造について詳細な説明を行ない，できれば発表者が現在までに行なってきた長母音変化に関する研究の中から 1 つ紹介して拙作脚韻コーパスの実用性を問うことにしたい．

● 初期中英語における指示代名詞 ‘e (that)'と‘es (this)'の形態

（新井　恭子）

今回の発表の主旨は，South-Western 方言に属し，現存する写本が 13 世紀に書かれた Laamon's Brut を主な資料として，初期中英語における指示代名詞の屈折の変化を明らかにすることにある．

Laamon's Brut の写本は Cotton Caligula AI X と Cotton Otho C XIII の 2 つがあり，後者は前者より 50 年ほど後に書かれたことから，両者を，またそれらを古英語と比較することにより，古英語から中英語への指示代名詞の形態的発達の経過をたどることができると考えた．今回は Caligula の写字生がそこで変ったとされる 1468 行目まで（C1）とそれに対応する Otho の部分（O1）とを対象に調査した．

資料は米国ヴァージニア大学がインターネット上に公開している電子テキストをダウンロードし，KWIC コンコーダンスを使って指示代名詞を含む行を抽出，更に MS

ACCESSTM を使いデータベース化した．主に指示代名詞が修飾している名詞の性数格，前置詞句における格支配，動詞の格支配などの必要事項を入力した後，MS EXCELTM にデータを移し，統計分析を行った．

その結果，West Saxon 地方における古英語の指示代名詞の屈折から C1,さらに O1 への屈折の変化には，先行研究において指摘されてきた，主に以下の 3 つの文法的，または音韻的変化の影響の現れが確認された．

1. ‘s’ → ‘’, ‘’ → ‘a’ などの変化，語尾の ‘um’ の ‘en’ への変化，語尾 ‘n’ や ‘e’ の脱落など(Moore 1928)

2. 主格・対格における，女性形と男性・中性複数形の不変化詞（undeclined）‘e’ による台頭

3. 全性数における対格・与格形の格融合の影響(Allen 1995)

さらに，上記の 3 の与格・対格の核融合の指示代名詞への影響を他の方言や Laamon's Brut より半?1 世紀前に書かれたテキストで確認する目的で，The Penn-Helsinki Parsed Corpus of Middle English: Helsinki Group 1（1150-1250）の下記の 11 のテキストのコーパスを利用して調査した．

The Katherine Group 1200-1250 WMidland
S. Katherine, S. Margarete, S. Juliene,
Sawles Warde, Hali Meidhad
Ancrene Riwle post - 1225 WMidland
The Holy Rood Tree 1150 - 1175 West Saxon
Lambeth Homilies 1150 - 1200 WMidland
Peterborough Chronicle 1132 - 1154 (N)EMidland
Trinity Homilies pre - 1200 EMidland
rewritten
Vice and Virtues - 1200 EMidland

The Penn-Helsinki Parsed Corpus のテキスト中で直接目的語，間接目的語のタグが付与されている指示代名詞をふくむ名詞句を抜き出し，指示代名詞の形態を確認したところ，Laamon's Brut より半世紀以上前に書かれたにもかかわらず，Lambeth Homilies 以外の

９つのテキストにおいては，指示代名詞に関しては対格・与格の区別がほとんどなくなっていることと，不変化詞（undeclined）‘e’ の台頭が著しいことがわかった．

今後は Laamon's Brut の写字生が変った 1469 行以降のテキスト（C2, O2）と The Penn-Helsinki Parsed Corpus of Middle English: Helsinki Group2（1250-1350）と Group3（1350-1420）のテキストの指示代名詞の形態の調査を続けていこうと考えている．

◆研究発表　第 2 セッション

●コーパス資料に基づく現代ビジネス英文の語彙的特性の研究

（染谷　泰正）

ESP の主要な分野のひとつに，いわゆるビジネス英語（BE)がある．現在，BE に関する教材は数多く市販され，この分野への関心の高さを示している．しかし，これらの教材の多くは執筆者の個人的経験と直感に基づいて記述されたものであり，実証的なデータによる裏付けや理論的支えを持たないのが普通である．ESP の観点からすれば，英語の使用人口が最も多いのは広い意味でのビジネス分野であることは疑う余地がない．それにもかかわらず，現実にはこの分野での学問的な研究は，他の ESP 分野に比較して大幅に立ち遅れており（Dudley-Evans & St John, 1996），その結果「われわれは，ビジネス英語の実態についてほとんど何も知っていないに等しい」(Holden, 1989)という現状にある．

本研究は，このような問題意識に立ち，独自に構築した約 130 万語からなる「ビジネス英語コーパス」の分析に基づいて，現代ビジネス英語の語彙的特性を明らかにしようとするものである．本研究を通じて，さまざまな事実が明らかになった．例えば，BE の語彙成長曲線はおよそ 1 万 5000 語から 2 万語（token）ほどで頭打ちとなることや，これらの語彙を基底語化した場合，およそ 1500 語（type)で全語彙の 92%を占めること，などが挙げられる．品詞別に見ると，動詞については頻度および使用度の観点から抽出した上位 350 語(type)で全使用動詞の約 91%をカバーし，副詞ではわずか 100 語で全副詞のおよそ 90%を占めることが明らかになった．これらの事実は，BE においては特定の語彙が集中的に使用される傾向が強いことを示している．

これらの分析に基づき，各品詞別に BE の「コア語彙リスト」を作成した．このリストには，頻度や難易度などの基礎的な情報のほかに，誤用率についての情報も加えた．これは，筆者が別途作成した「学習者コーパス」（日本人ビジネスマンの書いたビジネス文書データおよそ 22 万語を収録）の分析から得られたデータである．さらに，動詞を中心に，特に誤用率の高い語彙項目について「学習者コーパス」を使って詳細なエラー分析を行った．その結果，日本人学習者のエラーには一定のパターンがあることが明らかになった．このうち，

とりわけ重要と思われるのは，日英動詞の項構造の違いに由来するエラーと，各動詞に内在する意味論的･語用論的な制約にかかわるエラーである．前者については，日本人学習者は，英語の動詞を使うに当たって，意味的に対応する日本語動詞の項構造をそのまま適用する傾向があり，これが，両者の項構造が異なる場合に誤用となって現われるのである．これら一連の知見は，今後の教室における BE の指導，あるいはテキスト執筆に際して，従来にはなかった貴重な基礎資料を提供するものと考える．

なお，発表者は，本研究を進めるに当たって，大量のデータを効率的に処理するための各種コンピュータプログラムを AWK を使って自作したが，そのうちのいくつかはパッケージ化し，ごく簡単な手順で使用できるようにした．本研究発表ではこの点についても言及する予定である．

●ウェブを用いた学習者用用例検索システムの作成

（大名　力）

コンピューターの検索機能が，言語研究だけでなく言語学習や文章作成の役に立つことは，既に多くの人が指摘している通りであり，そのような目的で利用可能なコーパスや検索ツールも多く存在する．また，大学等の機関ではインターネットを利用できる環境も整ってきており，オンラインで大規模コーパスを利用したり，検索エンジンを利用し関連する英語のページを探したりして，英語学習や英文作成に役立てることも増えてきている．これらは，インターネットを利用できる環境であれば，特別なソフトやデータを自分で用意しなくとも利用できるという点で便利なものではあるが，実際に使いこなすには，それなりの知識と慣れ（そして場合によっては，検索結果を二次加工する手間）が必要であり，大学生レベルの学習者が日常的に利用するものとしては必ずしも使いやすいものであるとは言い難い．

本発表では，大学生レベルの学習者が手軽に使える小規模な検索システムの例として，現在，大名が作成している，ウェブと CGI プログラム（Perl）を利用した検索システムについて紹介する．既に，杉浦正利氏の WebGrep など，同種のシステムがいくつか発表されており，このようなシステムを構築すること自体は特に目新しいことではないが，1 つのシステムが全ての利用者にとって使いやすいものであることはまずなく，環境や利用者のレベル，利用目的に合わせて工夫する余地は残っている．本システムでは，学習者レベルの利用者が使いやすいシステムになるよう，検索方法や結果の表示の仕方などで工夫を凝らした．発表では，それらの工夫について，仕組みや具体的な利用方法などを，実際に検索を行ないながら説明する．また，いろいろな用例検索ツールの長所・短所を比較し，本システムのような検索システムを自作することの意義についても触れてみたい．

◆特別講演

● “Corpus Linguistics and the BNC”

（講師　Geoffrey N. Leech）

I will present a historical survey of corpus-based English linguistics, focussing on the present situation, and on new developments. I will particularly concentrate on the British National Corpus (BNC) as a “landmark corpus” which illustrates both the strengths and weaknesses of large-scale corpus resources in the 1990s. One need illustrated by the BNC is the requirement of higher-quality corpus annotation. I will also give attention to recent advances in corpus-based research on spoken language and dialogue, and on the corpus-based study of language change.