英語コーパス学会
第 **44** 回大会資料

日時：2018 年 10 月 6 日（土）–10 月 7 日（日）
会場：東京理科大学（神楽坂キャンパス）
〒162-8601 東京都新宿区神楽坂 1-3

英語コーパス学会 2018

英語コーパス学会　第 44 回大会　プログラム

■第 1 日目

ワークショップ受付：2 号館 1 階　211 教室前にて 9:30 受付開始

ワークショップ参加費：会員無料。非会員 2,000 円（当日会員としての大会参加費、2 日間共通）。

ワークショップ 1【言語研究のための Word2Vec 入門】
会場：東京理科大学（神楽坂キャンパス）2 号館 2 階　223 教室
日時：10 月 6 日（土）10:00–11:00
講師：内田　諭（九州大学）

ワークショップ 2【コーパス研究の作法】
会場：東京理科大学（神楽坂キャンパス）2 号館 2 階　223 教室
日時：10 月 6 日（土）11:00–12:00
講師：井上　永幸（広島大学）

日時　2018 年 10 月 6 日（土）
受付開始　12:00（2 号館 1 階　211 教室前）
開会式　13:00（2 号館 1 階　212 教室）

司会：石井　康毅（成城大学）

1. 会長挨拶　投野　由紀夫（東京外国語大学）
2. 開催校挨拶　渡辺　一之（東京理科大学　副学長）
3. 総会
4. 学会賞審査報告　西村　秀夫（三重大学）
5. 事務局からの連絡

●研究発表第 1 セッション（場所：2 号館 1 階　212 教室）　司会：柴﨑　礼士郎（明治大学）

研究発表 1：14:00–14:30
英語史的コーパスによる言語変化速度の測定　塚本　聡（日本大学）

研究発表 2：14:35–15:05
アメリカ英語における分離不定詞使用に関する研究：“splitter”に焦点を当てた分析　福本　広光（大阪大学大学院生）

研究発表 3：15:10–15:40
コーパスを利用した 18・19 世紀オーストラリア文学作品における自動詞完了形助動詞に関する研究　守家　輝（京都大学大学院生）

●研究発表第 2 セッション（場所：2 号館 2 階　223 教室）　司会：水本　篤（関西大学）

研究発表 1：14:00–14:30
（キャンセル）

研究発表 2：14:35–15:05
Using Corpora to Examine Lecturing Styles in American and Japanese University Engineering Courses　Judy Noguchi（Kobe Gakuin University）
Kazuko Tojo（Osaka Jogakuin University）
Nilson Kunioshi（Waseda University）

研究発表 3：15:10–15:40
Collaborative Texts under a Stylometric Microscope: Investigating Texts of Mixed Authorship　Tomoji Tabata（Osaka University）

＜休憩　15:40–16:10＞

●シンポジウム：16:10–17:40（場所：2 号館 1 階　212 教室）

Issues on multi-word units (MWUs) and collocation	Chair：Laurence Anthony（Waseda University）
On the Creation of a Large-Scale Multi-Word Unit Resource for Learners of English for Academic Purposes	James Rogers（Meijo University）
Applying a bundle-move connection approach to the development of an online writing support tool for research articles	Atsushi Mizumoto（Kansai University）
Collocational patterns beyond word pairs	Stefan Evert（Friedrich-Alexander-University of Erlangen-Nürnberg, Germany）

＜懇親会　18:00–20:00＞

■第 2 日目

日時　　2018 年 10 月 7 日（日）
受付開始　　9:30（2 号館 1 階　211 教室前）

●研究発表第 3 セッション（場所：2 号館 1 階　212 教室）　司会：久保田　俊彦（明治大学）

研究発表 1：10:00–10:30
Charles Dickens の *The Mystery of Edwin Drood* の、“Thomas Power James” による続編の「著者推定」－テクスト全体・地の文・発話の 3 種類の語彙による推定の比較・評価　後藤　克己（中部大学大学院生）

研究発表 2：10:35–11:05
（キャンセル）

研究発表 3：11:10–11:40
トピックモデルを用いた Agatha Christie 作品へのアプローチ　土村　成美（大阪大学大学院生）

●研究発表第 4 セッション（場所：2 号館 2 階　223 教室）　司会：能登原　祥之（同志社大学）

研究発表 1：10:00–10:30
小学生のための英語 DDL 支援サイトの開発に向けた作例参照用コーパスの構築と検索ツールの開発　西垣　知佳子（千葉大学）
赤瀬川　史朗（Lago 言語研究所）
中條　清美（日本大学）

研究発表 2：10:35–11:05
英日バイリンガルエッセイコーパスに見るコロケーションの比喩的な意味拡張　鎌倉　義士（愛知大学）

研究発表 3：11:10–11:40
英語学習者コーパス構築のためのタスク設計：特定の文法項目抽出に向けて　工藤　洋路（玉川大学）
内田　諭（九州大学）

＜休憩　11:40–13:00＞

●研究発表第 5 セッション（場所：2 号館 1 階　212 教室）　司会：秋山　孝信（日本大学）

研究発表 1：13:00–13:30
アカデミックライティングにおけるヘッジの活用：研究論文における Discussion のコーパス分析　中谷　安男（法政大学）

研究発表 2：13:35–14:05
強意語的機能を持つ罵倒語の進化特性について　新井　洋一（中央大学）

●研究発表第 6 セッション（場所：2 号館 2 階　223 教室）　司会：阿部　真理子（中央大学）

研究発表 1：13:00–13:30
Reliability and Replicability of Annotation Schemes for Learner Corpora　Aika Miura（Tokyo University of Agriculture）

研究発表 2：13:35–14:05
Searching for grammatical items as criterial features of CEFR levels in spoken and written learner corpora: Using the CEFR-J Grammar Profile　Yukio Tono（Tokyo University of Foreign Studies）
Yasutake Ishii（Seijo University）

＜休憩　14:05–14:30＞

●講演　　14:30–16:00（2 号館 1 階　212 教室）

司会：投野　由紀夫（東京外国語大学）

Measures of Productivity and Lexical Diversity

Stefan Evert （Friedrich-Alexander-University of Erlangen-Nürnberg, Germany）

閉会式　　16:00（2 号館 1 階　212 教室）

閉会の辞

■10 月 6 日（土）
【ワークショップ 1】

言語研究のための Word2Vec 入門

内田　諭（九州大学）

近年の自然言語処理の技術の発展は目覚ましい。その裏にはコンピュータの飛躍的な進歩と言語情報の大規模な電子化がある。これはコーパスを使った言語研究の実施には恵まれた状況であるといえるだろう。その一方で、多くの技術や技法が溢れるが故に、特に文系の研究者にとって自身の研究テーマに沿ったものを選択することが難しく、日進月歩の進化についていくことは容易ではないという現状がある。

本ワークショップでは、多くの自然言語処理の研究で用いられるようになった word embedding（単語埋め込み）について基礎的な概念を説明し、その中でも特に広く利用されている Word2Vec について概観する。word embedding は単語の意味をベクトル（数字の集合）として扱う技法で、意味の似た単語は似た文脈で用いられるという分布仮説が背景にある。単語の意味をベクトルで表すことで、単語と単語の距離を測定することが可能となり、類義語の抽出やベクトルを用いた意味の演算が可能となる。これらの性質を用いることで品詞解析や機械翻訳、語義曖昧性解消などの精度が向上することが報告されており、自然言語処理において基幹的な技術の一つとなりつつある。Word2Vec は Google の研究チームが開発した word embedding のアプリーケーションの一つで（cf. Mikolov et al. 2013）、Python や TensorFlow などの広く使われているプラットフォームを通して比較的容易にかつ高速に実行することができるものである。本ワークショップではウェブのインタフェース（Python CGI）を用意し、word2vec を実際に動作するところを示す。例えば、約 1 億語のコーパスを用いて作成したモデルを用いて speak の類義語（ベクトル間の距離が近いもの）を検索すると、talk, respond, communicate などの意味が近いと考えられるものや、listen, hear などの知覚を表す動詞などがリストされる。このような結果が言語研究にはどのように応用できるかを議論し、word2vec を用いた新たなコーパス研究の可能性について探る。

【ワークショップ 2】

コーパス研究の作法

井上　永幸（広島大学）

比較的簡単に大規模コーパスにアクセスできるようになったり，膨大なインターネット上のデータをコーパス代わりに利用する機会が増えるなど，コーパス研究のための環境変化には目を見張るものがある。本ワークショップでは，特にコーパス研究初心者が遭遇しそうな状況を想定しながら，問題解決の際の糸口を提案してゆく。

大規模コーパスの利点は，それほど頻度の高くない語句について，小さなコーパスでは特徴として認識できなかった形が特徴として認識できるようになることであろう。たとえば，enough as it is は，Brown，LOB，Frown，FLOB を合わせた 400 万語のコーパスでは 1 例しか出てこないが, 4 億 6230 万語の Wordbanks*Online* では 100 例出てきて，その定型表現的性質を垣間見せてくれる。また，大規模コーパスを使った研究の醍醐味を見せてくれるのは各種統計値であろう。MI-score は，低頻度のものにも反応するため，その取っ付きにくさからか敬遠されがちであるが，逆にキーワードと連想関係の強い語を高頻度のものから低頻度のものに至るまで焦点を当ててくれるので，キーワードの特性を見事にあぶり出してくれるのである。無論，MI-score が高くて頻度も高い方が信頼度も高まるので，そういった語から優先的にアプローチするのが正攻法であろう。MI-score の「機微」にふれてみたい。

インターネット上にあふれるデータをコーパスとして利用する方法も注目を集めている。日々増大するデータを言語資料として利用しない手はないであろう。ただ，Mark Davies による BYU corpora で提供されている 140 億語から成る iWeb で複数形の “informations” を検索すると 8,000 を越える例が出てくることが分かる。英語を地球規模で研究するのであればともかく，少なくとも教育現場で使う英語を吟味するには適さないであろう。ウェブデータをコーパスとして用いる際の留意点などにもふれたい。

■10 月 6 日（土）
【研究発表第 1 セッション】
【研究発表 1】

英語史的コーパスによる言語変化速度の測定

塚本　聡（日本大学）

古英語以来，英語には多くの統語的変化が生じている。Denison (1993)が示す通り，動詞にかかわる変化は特に顕著である。Fries (1940)や Ellegård (1953)をはじめ，個別の言語変化についての研究は多数行われている。それらの個別の研究では，各言語変化の開始時期や終了時期，あるいは量的な変化は示されるものの，言語変化に要した時間を横断的に扱った研究は見られない。本研究では，以下に挙げる言語項目の変化速度について，コーパスを使用し，その変化に要した時間を観察し，その変化速度にある一定の規則性がみられるかを検討する。

異なる言語資料を使用することによる差異が生じることが無いよう，本研究では Penn-Helsinki Parsed Corpus（ME2, EME, MBE2）を中心的言語資料とし，YCOE などを追加的に使用し，下記の言語項目の変化を観察する。

(1) 非人称動詞 like の人称動詞化
(2) 変移自動詞の助動詞選択
(3) 進行相の確立
(4) V+-ing 補部の確立
(5) 属格形および of 迂言形の交替
(6) thereof または of it の選択
(7) 限定詞の義務化
(8) 二重限定詞
(9) 二重比較級・最上級

具体的には，(1) ME 後期から eModE にかけて非人称動詞 like が人称動詞化した変化，(2) ModE 後期に be+V-en であった変移動詞の助動詞が have+V-en に変化した現象，(3) eModE に be+-ing 構文の主要な動詞が dwell などの非有界動詞から有界動詞へと拡大した変化，(4) ModE 後期に，一部の動詞補部に-ing が生起しその頻度を高めた変化，(5) OE から ME にかけて屈折語尾の衰退に伴い前置詞の使用が増加した変化，(6) ほぼ同義となる thereof および of it は一部のレジスターを除き，後者の使用が優勢となる変化，(7) 可算名詞において，a, the, my などの限定詞が必須となる変化，(8) ModE まで生起していた these my words のような限定詞の連鎖が許容されなくなった変化，(9) most easiest のように迂言形を比較級・最上級に使用する現象，を調査対象とする。なお，(8)および(9)は定着しなかった変化である。

調査は Penn-Helsinki Parsed Corpus で解析された統語情報から条件に合致する構造を検索し，その生起数や対立する形態・構造との構成比率の変化を観察することにより行う。予備的調査では，定着した変化のいくつかは，ほぼ 200～250 年くらいの時間で変化が完了していることが観察されている。これらの観察から，言語変化は一定の変化速度を有していることを示す。

【研究発表 2】

アメリカ英語における分離不定詞使用に関する研究：“splitter”に焦点を当てた分析

福本　広光（大阪大学大学院生）

本研究の目的は，アメリカ英語における分離不定詞の使用の実態について，特に to と原形動詞を「分離」する 要素である‘splitter’を中心に通時的に考察することである。

分離不定詞は，Rohdenberg (2009), Mitrasca (2009), Mikulova (2011)などの研究により，イギリス英語よりもアメリカ英語において頻繁に確認される構造であることが分かっている。そのことから，アメリカ英語は分離不定詞の主要な使用域であると考えられる。また,分離不定詞に関する通時的研究は Calle-Martin and Miranda Garcia (2009)などを除いて多くはなく，計量的に見るとまだ知られていないことも多いと思われる。本研究では主に COHA, Time Magazine corpus の 2 種類のアメリカ英語コーパスを用いる。

本研究における RQ は以下の 3 つである。

1.分離不定詞は，アメリカ英語においてどのように拡大してきたのか
2.splitter そのものに，経年的な変化というものがあるのか，またどのような変化なのか
3.分離不定詞が，どのような context で生起しているか

この“splitter”を，例えば(1)での really のような 1 語 splitter と(2)での all of a sudden のような複数語 splitter に分け，それぞれについて通時的コーパス COHA を用いて量的・質的に経年的変化について分析する。

(1)He seemed *to really want* to find a way to serve. (Time Magazine Corpus)
(2)The states he’s won, the red states, is they’re not likely *to all of a sudden turn* blue in November. (COCA)

具体的には，量的分析として，分離不定詞の 200 年間の頻度変遷（Fiction などのジャンル別の頻度変遷を含む）を調査し，さらに各ジャンルでの頻度変化について見ることで頻度変化のメカニズムを探る。

1 語 splitter を伴う分離不定詞の使用頻度は、20 世紀前半まで増減を繰り返していたが，20 世紀後半，特に 1990 年以降大きく伸びている。その背景には Fiction と Magazine ジャンル，とりわけ Magazine ジャンルにおける増大がある。（Time Magazine corpus でも同じような結果が得られた）また，20 世紀初頭に分離不定詞は頻度を減少させているが，その背景に当時の分離不定詞のメイン使用域であった Fiction における減少があった。これらの時代の Magazine や Fiction をさらにサブジャンルに下位分類し、当時の使用の実態について考察する。

splitter の経年的変化については，おもに COHA の Fiction ジャンルから分離不定詞として使用された副詞をすべて抽出，集計し，時代ごとに偏って使用されている splitter の存在を観察する。そして長年にわたって使用されてきた副詞 (even, just など)もある一方で，時代ごとに特徴的な副詞 (thus, actually など)もまた存在することを明らかにする。

分離不定詞には,「使用すべき理由」もまた存在する。「曖昧性を解消し，文の意味を正確に伝えるとき」や,「リズムを整え，発音しやすく心地よい響きを与えるとき」など (Crystal(1984:29-30)) である。抽出した例文のうち，どれだけがこれらの理由に則って使用されているのか，コンコーダンスライン全体の意味を解釈しながら考察する。

複数語からなる splitter が使用される場合、そのパターンとして
(i)at once や at least などの、lexicalize されたもの
(ii) more strongly や at least passively などの、一方（一部）の副詞がもう一方の副詞を修飾するもの
のどちらかに区分されるということがほとんどであるということが確認された。

【研究発表 3】

コーパスを利用した 18・19 世紀オーストラリア文学作品における自動詞完了形助動詞に関する研究

守家　輝（京都大学大学院生）

本研究は，*go*，*come* 等動作主の移動・変化を表す自動詞の完了形構文において観察される，助動詞 *be/have* の交代現象と，言語使用者の出身や社会階層といった社会言語学的要因との関係解明を目的とする。元来自動詞の完了形には，次の例文のように *be* 動詞が主に使用されていた：“Mr. Sandford said, “Hannah what is become of your frock?” (COOEE 2-308)。他動詞の完了形構文で使用される助動詞 *have* が自動詞完了形に対して用いられることが増加し，後期近代英語期(18・19 世紀)には移動・変化を表す自動詞に対しても *be* 完了形の使用頻度を *have* 完了形が上回ったことが，Rydén and Brorström (1987)， Kytö (1997)等多くの先行研究で指摘されてきた。

Have 完了形の増加現象を研究する上で，各植民地社会で新たな言語変種が成立しつつあった後期近代の特性から鑑みて，地域・社会的要因を検証する余地が多分にある。本研究では特に，19 世紀以前のオーストラリア社会における，アイルランド英語(IrE)の影響という側面に焦点を当てる。アイルランド移民は 19 世紀以前オーストラリア入植者の約 20%を占め，イングランド移民に次いで多い (Fritz 2007: 22-27)。IrE 由来の統語的特徴は小説・演劇等の会話文でアイルランド人の発話内に頻出するものがある一方，一部はオーストラリア英語(AusE)特有の表現として定着した (Burridge and Musgrave 2014: 40-41)。自動詞 *have* 完了形の発展に関する IrE 等言語変種の影響を解明する上で，初期オーストラリア社会を対象とした社会言語学的研究は有意義だと考えられる。

本研究では 18-19 世紀のオーストラリアで執筆された文献を収録した Corpus of Oz Early English (COOEE)と，AustLit の 2 種類のコーパスから抜粋した，18・19 世紀の fiction ジャンルに分類されるテキスト約 300 万語を対象に分析を行う。Rydén and Brorström (1987)を参考に 19 世紀において *be/have* の交代が確認される自動詞 12 種を選び，各自動詞の完了形構文中で *have* が選択される比率を計算して，その結果を話者(著者あるいは発話者である登場人物)の出身地・社会階層という観点から分析する。また同様の分析を ARCHER コーパスに収録された同時代・同ジャンルのイギリス英語(BrE)，アメリカ英語(AmE)テキストにも適用し，AusE と比較する。

その結果，AusE, BrE, AmE の三変種間における *have* 完了形使用頻度の推移に差があることが判明し，その変化が確認される時期とオーストラリアへのアイルランド系移民の流入時期との間に相関がある可能性が示唆される。また，同じアイルランド系作家の文章でも，AusE のデータでは *have gone* の使用率が減少傾向にある期間において，BrE では逆に増加傾向である等, 移民と本国の作家との間で特徴の差異が認められる。さらに, “an’ I wouldn’t a’ come anigh you” (*AustLit*. *Dave’s Sweetheart*.)のような会話文中で助動詞の形態が特殊な場合(*a’come*)が確認されることなどから，初期のオーストラリア移民が用いた言語変種の特徴が作品の著者によって登場人物が用いる完了形構文においても表現されているという仮説を立て，*have* 完了形の使用率を発話者の出身地・社会階層といったオーストラリア植民地時代における社会学的側面との関係から分析し，解明していく。

■10 月 6 日（土）
【研究発表第 2 セッション】
【研究発表 1】

（研究発表第 2 セッションの研究発表 1 はキャンセルされました。）

【研究発表 2】

Using Corpora to Examine Lecturing Styles in American and Japanese University Engineering Courses

Judy Noguchi（Kobe Gakuin University）
Kazuko Tojo（Osaka Jogakuin University）
Nilson Kunioshi（Waseda University）

The rapid globalization of society today has accelerated the need for internationalization of higher education in Japan. In 2008, the Japanese government announced the “300000 Foreign Students Plan” Campaign (Ministry of Foreign Affairs, 2010) and as of May 1, 2017, the number of international students studying in Japan had reached 267,042, with 188,384 students being enrolled in institutions of higher education (Japan Student JASSO, 2017). This has led to an increase in the demand for university degree courses which are offered using English as the medium of instruction. With an interest in aiding Japanese instructors faced with delivering lectures in their disciplines in English, especially those in science and engineering, we built OnCAL, a corpus of university lectures for science and engineering courses. We aimed at identifying the pedagogical functions and useful expressions that could be used for delivering lectures in English. The OnCAL concordancing interface allows access to 430 science and engineering lectures given at MIT Opencourseware (MIT OCW, http://ocw.mit.edu/index.htm) and Stanford Engineering Everywhere (SEE, http://see.stanford.edu/). This work led us to wonder about how functions that we had uncovered, such as the asking of questions and the proposal of thought experiments to initiate consideration of the lecture contents, would be delivered in comparable lectures at Japanese universities. We therefore began construction of a corpus of Japanese lectures given at four major national and private universities in similar disciplines. At present, there are 104 Japanese lectures and comparing them with the American lectures revealed marked differences in the lecturing styles. In this paper, we focus on questions and their functions in the lectures. In general, the 430 lectures given at American universities included many question-word phrases to elicit student thinking with frequent uses of personal pronouns such as *you* and *we*, indicating more interaction during the class. Question-word phrases with *what* (1706 average per lecture), *how* (697), and *why* (244) were frequently used, in contrast to the Japanese lectures where *何*、*なん* (*what)* was used only 2.8 times on average per lecture, *なん（て）*、*どう*、*どのよう* (*how)* only 2.8 times and *何（で）*、*なん（で）*、*どう（して）*、*なぜ*（*why*） only 0.6 times. The lectures at Japanese universities were presented in a manner suggesting the presence of a top-down authority. Overall, the American lectures tended to be presented in a conversational style with a strong audience orientation while the Japanese lectures were given in a more formal style with an emphasis on content dissemination. Our findings led us to conclude that English-medium instruction may not be successful if a lecture which was originally intended for a Japanese university audience was just simply delivered in the English medium. What is essential for successful instruction in a “globalized” classroom with students from different educational backgrounds is an awareness of differences in lecturing styles and the structuring of lectures in order to reach students who may have different expectations with respect to classroom instruction.

【研究発表 3】

Collaborative Texts under a Stylometric Microscope: Investigating Texts of Mixed Authorship

Tomoji Tabata（Osaka University）

The Victorian author Charles Dickens was among the first publishing entrepreneurs to run mass-produced weekly magazines on a successful commercial basis. He employed many “salaried staff writers” (Nayder, 2002), who had to write under anonymity, including Elizabeth Gaskell, Adelaide Anne Proctor among others, in *Household Words* and *All the Year Round*, the journals “conducted by” Dickens (Stone, 1968; Thomas, 1982; Allingham, 2011). On the other hand, Dickens collaborated with his younger contemporary Wilkie Collins on a number of stories, typically for the Christmas Numbers of his journals. While some of their collaborative pieces were written with the assistance of other staff writers, four works are known to have been co-authored by Dickens and Collins alone (Nayder, 2002): *The Frozen Deep* (1857), *The Lazy Tour of Two Idle Apprentices* (1857), *The Perils of Certain English Prisoners*’ (1857), and *No Thoroughfare* (1867). The four collaborations can be seen as betokening what appears to be a firm presence of Collins, a foothold he had gained, in the Dickens circle by the time he and Dickens launched into the joint works beginning in 1857.

The present study draws on a corpus of Dickens set comprising 22 texts and a Collins set with the same number of texts as a training corpus, with which we compare the style of the collaborative texts. Just as Dickens was a prolific writer, so was Collins. His career spans 38 years including his ‘etudes’, *Antonina* (1850) and *Rambles Beyond Railways* (1851), which are not included in this study. The set of corpus texts can be transformed into a vector of figures, word-frequencies. Word-frequency profiles make it possible to compare between texts to see a certain set of texts have more frequent access to a particular set of words than others, while another group of texts may be characterized by consistent avoidance of a particular set of words among others. However, the larger number words the frequency profiles encompass, the greater difficulty they impose upon us when we try to grasp complex interrelationships between a large number of vocabulary items and relationships between texts. This is

exactly where we need a statistical technique for visualization. This study employs Support Vector Machine in an effort to attribute a chunk of text in question to the more likely author with rolling chunks of the collaborative text progressively compared with the training corpus.

A rolling chunk is designed to be sensitive to a stylistic shift in texts in order to pinpoint where one author takes over from the other in the text of mixed authorship. For this study, collaborative texts are segmented into equal-sized, partially overlapping chunks. If we specify a 'chunk size' of 3,000 and a 'step size' of 300, for example, the first chunk of a text contains 1st–3,000th words, the second has 301st–3,300th words, the third 601st–3,600th words, and so forth. The procedure uses the relative frequencies of *n* most frequent words in the reference collection. Emerging results from this analysis show that it is possible to assign a particular chunk to the more likely author with a 100% accuracy, thus allowing us to locate exactly where authorial takeovers happen in the texts of mixed authorship. The findings from this research show that it is always Dickens who starts joint chapters, setting the keynote of each of the collaborated chapters. A typical pattern is that Dickens runs about one-third to halfway into a chapter before passing over to Collins. Of remarkable interest with respect to the making of collaborations is that the pattern in authorial takeovers can be interpreted as reflecting an "unequal" partnership (Nayder, 2002) between Dickens and Collins, just like one between a master and his disciple.

【シンポジウム】

Issues on multi-word units (MWUs) and collocation

Corpus linguistics research is paying increasing attention to the identification and usage of multi-word units (MWUs). To date, research on MWUs has provided new insights in a wide range of areas, including vocabulary development, learner writing, discourse signaling, and disciplinary variation. However, the precise definition of MWUs has proved elusive, leading to widely different interpretations of what counts as a MWU and how they are used in spoken and written language. One common view has been to regard MWUs as contiguous word units that appear above a threshold frequency and dispersion value in a target corpus. Another common view has been to regard MWUs as contiguous and non-contiguous word units in which the members have a collocation strength above a certain statistical or effect-size threshold. As a result of the current confusion, corpus tools provided widely differing functions for identifying and analyzing MWUs. They also differ greatly in regard to the measurements of association between candidate members offered to researchers.

In this symposium, MWUs and their relation to the concept of collocation will be discussed from three different perspectives: the creation of high-frequency, pedagogic MWU lists using lemmatized collocates; applications of MWUs in the development of practical learning and teaching tools, and novel methods for automatically identifying non-idiomatic MWU combinations. Following presentations by three influential researchers in this field, the audience will be invited to join the panelists in a discussion of collocation measures and the challenges of identifying and utilizing multi-word units (MWUs).

Chair: Laurence Anthony（Waseda University）
Title: Issues on multi-word units (MWUs) and collocation

To open the symposium, I will briefly review some of the definitions of multi-word units (MWUs) proposed in the literature and summarize the various challenges that researchers face when identifying and utilizing MWUs in linguistic research as well as language learning and teaching resources. As part of this introduction, I will briefly discuss a new software tool designed specifically for MWU identification and analysis.

Panelist 1: James Rogers（Meijo University）
Title: On the Creation of a Large-Scale Multi-Word Unit Resource for Learners of English for Academic Purposes

Previous research indicates that there are gaps in the literature in regard to a methodology of identifying high-frequency multi-word units (MWUs) for general English purposes, and specifically, English for academic purposes (EAP). Thus, there is also a lack of large-scale resources. In this talk, I present a study in which a novel methodology used to identify high-frequency MWUs of general English is applied to create a similar large-scale resource for EAP. First, the most frequent 500 lemmas in an academic vocabulary list were utilized in the search for lemmatized collocates. Then, these lemmatized collocates were used to identify commonly occurring EAP MWUs, leading to the creation a large-scale EAP MWU list. This results of this study confirmed the importance of native speaker judgments when relying upon corpus data to create a list of MWUs for second language learners that is used to improve their EAP fluency. The results also shed light on the importance of manual checking of corpus data, and the type of low-value items that only manual checking can identify. Most importantly, the study has also resulted in a large-scale EAP MWU resource that not only fills a major gap in the literature, but also confirms previous findings and potentially leads to new discoveries in regard to MWU identification.

Panelist 2: Atsushi Mizumoto（Kansai University）
Title: Applying a bundle-move connection approach to the development of an online writing support tool for research articles

Achieving a high level of English proficiency requires a comprehensive English vocabulary of which multi-word units (MWUs) are a critical component. However, acquiring and using these MWUs poses a formidable challenge for second language users of English. In order to facilitate the learning of these units, various online reference resources based on different types of corpora have been developed in recent years. Also, there is a growing interest in resources that are specifically designed to help learners develop an understanding of MWUs above the level of the sentence. In this talk, I introduce some of the current resources available for accessing MWUs that can help to develop rhetorical competency. Specifically I will focus on a data-driven and theory-based practical writing support tool for research articles (RAs) called AWSuM. This innovative, web-based tool is

powered by a combination of rhetorical moves and lexical bundles. It also has an auto-complete feature that suggests the most frequent lexical bundles in a move within an RA section. AWSuM was developed as a proof-of-concept of the bundle-move connection approach. Preliminary user feedback was positive overall, and the writing support tool was found to bring about beneficial effects that genre writing pedagogy explicitly aims to achieve. In light of these findings, the pedagogical implications of the developed tool are discussed, with particular focus on the potential role that it can play in the teaching and learning of technology-enhanced genre writing.

Panelist 3: Stefan Evert（Friedrich-Alexander-University of Erlangen-Nürnberg, Germany）
Title: Collocational patterns beyond word pairs

While there is a substantial body of work on the identification and lexicographic description of collocational word pairs as well as idiomatic multiword expressions, only a few studies have addressed longer non-idiomatic word combinations (MWCs). Such MWCs can include collocational patterns involving three or more lexical items that form a series of semantically related MWCs, for example, “set a {dangerous | bad | unfortunate | damaging} precedent”). They also include grammatical constructions with marked lexical or semantic/morphosyntactic preferences, such as the ditransitive use of “earn” (“sth earns sbdy sth”), where the direct object is almost always selected from a narrow semantic field (“nickname, reputation, title, …”). In this talk, I present ongoing research towards a description of MWC phenomena and the automatic identification of MWC candidates. This approach builds on two premises: 1) Co-occurrence patterns between words cannot be reduced to a one-dimensional association score, but comprise multi-faceted aspects including frequency, salience, and the type-token distribution of each slot; and 2) The complex interrelations between different slots of a MWC can be modelled in terms of nested hypothesis tests, taking into account both significance and association strength. Such nested hypotheses may also involve semantic or morphosyntactic restrictions on the slots, or test whether a larger MWC is composed of overlapping smaller MWC (e.g. “earn good money” from “earn money” + “good money”).

■10 月 7 日（日）
【研究発表第 3 セッション】
【研究発表 1】

Charles Dickens の *The Mystery of Edwin Drood* の、“Thomas Power James”による続編の「著者推定」
ーテクスト全体・地の文・発話の 3 種類の語彙による推定の比較・評価

後藤　克己（中部大学大学院生）

米国人 Thomas Power James （以下，James）は，Dickens の遺作となった *The Mystery of Edwin Drood*（以下，*ED*）に続編を加えた「完全版」(Dickens, [James], 1873)を発表し，その続編を"By the Spirit Pen of Charles Dickens, through a Medium."とアピールしている。後藤(2017)は地の文の語彙頻度を用いた対応分析結果，およびコロケーションの生起頻度の比較から，*ED* と続編の文体の類似性は低いと結論づけ、このアピールを疑わしいものと評価している。本研究では，対応分析とともに文体比較／著者推定に多く用いられている多次元尺度法およびクラスター分析を使用し，あらためて続編の著者が「Dickens の霊」といえるかどうかの推定を試みる。この推定は，続編が Dickens の霊による自動書記(automatic writing)で創作されたならば，それは Dickens の文体特徴をもつはず，との考えによる。なお著者推定には，地の文に生起する語彙頻度が多く用いられる(Hoover, 2001, 2002)が，一般に発話テクストの除去には，１重引用符とアポストロフィの区別，引用符を欠いた自由直接発話の抽出などが必要となり大きな負担となる。そこでテクスト全体の語彙／地の文のみの語彙のそれぞれで分析し，分析語彙を地の文に限定することの効果を評価する。

コーパスには後藤(2017)と同様に，*ED* と続編に，参照用として Dickens の *Our Mutual Friend*（以下，*OMF*）を加えた 3 作品を用いた。これらを地の文と発話に分離し，それぞれ 4〜12 のサブコーパスに分け Lemma による語彙頻度数を抽出した後，Hoover(2004)を参考に 1 作品に 70%以上偏在する語，人称代名詞および固有名詞を除外した。このデータをもとにテクスト全体・地の文・発話の 3 種類のデータを構成した。

分析の結果，地の文の語彙はもとよりテクスト全体の語彙によっても，さらには発話の語彙によっても，続編のサブコーパスは *ED*・*OMF* のそれとは別クラスターを構成しており，続編を Dickens の「遺作」とする James のアピールは疑わしいとの，後藤(2017)と同じ結論を得た。

なお，多次元尺度法の散布およびクラスター分析の樹形図において，続編と *ED*・*OMF* との距離は，地の文語彙の場合が最も大きく，テクスト全体，発話の順に減少しており，地の文語彙によった場合の著者推定性能の高さが確認された。なお，本研究ではテクスト全体でも地の文と同様な結果が得られており，推定すべき著者が限定的な場合であれば，テクスト全体の語彙の使用も許容されることを示唆している。

【研究発表 2】

（研究発表第 3 セッションの研究発表 2 はキャンセルされました。）

【研究発表 3】

トピックモデルを用いた Agatha Christie 作品へのアプローチ

土村　成美（大阪大学大学院生）

本研究では，イギリスのミステリー作家 Agatha Christie の作品に関して，同時代作家と比較して統計的手法を用いた特徴の分析を行うことを目的とする。比較対象として，Christie と同時代に活躍した女性ミステリー作家，Dorothy Sayers, Margery Allingham, Ngaio Marsh の作品を用いる。これら 4 名の女性作家は第一次世界大戦と第二次世界大戦の戦間期を中心に活躍しており，「イギリスミステリーの 4 大女王」と呼ばれているためである(Joannou (Eds.), 2013)。

分析に用いるデータは以下の通りである。Christie 作品 66 作品(4,183,485 語)，Sayers 作品 11 作品(1,115,019 語)，Allingham 作品 20 作品(1,534,462 語)，Marsh 作品 33 作品(2,510,391 語)である。分析対象作品は長編作品に限定している。作品ごとに総語数が大きく異なるため，各作品のテクストファイルを 2000 語単位に分割した。この処理を経た合計 4661 ファイルを用いて分析を行った。

分析手法として，機械学習の一種であり，確率論的アルゴリズムに基づいてトピック(話題，テーマ)の抽出を行うトピックモデルを用い，分析を行う。本研究では Blei et al. (2003)によって提唱された潜在的ディリクレ配分法を用いたトピックモデルを行う。トピックモデルを実行するにあたり，マサチューセッツ大学で開発された自然言語処理ツールキット MALLET (Machine Learning for LanguagE Toolkit)バージョン 2.0.7 を使用した。この分析を通して，Christie 作品において同時代作家と比べて特徴的なトピックを抽出することが本研究の目的である。

Christie 作品と特に関連性が強いトピックを確認すると，第一に挙げられるのは，縮約表現を中心とした口語的表現から成るトピックである。このトピックには *I'm, that's, didn't, I've, you're* のような語が含まれている。Christie の作品は登場人物の会話を中心として物語が展開されるものが多く，その特徴が反映された結果であることが考えられる。また，このトピックと関連する作品は Christie の初期の作品よりも，晩年の作品の方が多くなっている。Christie

は 1952 年に腕を骨折し，従来までの手書きの原稿をタイプライターで打つという執筆方法から，ディクタホンを使用した執筆へと，執筆スタイルが変化している(Le et al., 2011)。この執筆方法の変化も Christie 作品がこのトピックと関連性が強い要因となっていると考えられる。

全作家に対してミステリー作品のみでコーパスを構築しているため，*case, murder, death, evidence, police* などの語が含まれる犯罪捜査に関するトピックも生成されており，このトピックに関しても Christie 作品との関連性が強く出ている。Christie が直接的に犯罪を表現する語を多く使用していることが反映されている可能性が高いと言える。また，トピック数を増やすと Christie 作品の探偵 Poirot の言動を表す語から成ると推察されるトピックも現れることが確認された。同ジャンル作品を対象としたトピックモデルで，どの程度作家間の特徴を識別することが可能であるかに関して可能性を検討する。

■10 月 7 日（日）
【研究発表第 4 セッション】
【研究発表 1】

小学生のための英語 DDL 支援サイトの開発に向けた作例参照用コーパスの構築と検索ツールの開発

西垣　知佳子（千葉大学）
赤瀬川　史朗（Lago 言語研究所）
中條　清美（日本大学）

本研究グループでは，小学校英語において，児童が英語の語彙・文法のルールを発見して学ぶ「データ駆動型学習」(Data-Driven Learning: 以下，DDL)を実践するための，英語 DDL 支援サイトを開発している。本発表では特に，その開発に向けた 1) 英文作例参照用コーパスの構築，2) 検索ツールの開発，および 3) 作業を進める中で遭遇した課題とその解決方法について報告するものである。

発表者らは，DDL を小学生から大学院生の授業に導入し，その普及に努めている。そのなかで，小学校ではペーパー版 DDL を実践し，言葉への気づきを引き出し，英語の語彙・文法の深い学びの促進に効果をあげてきた。小学生のペーパー版 DDL の実践にあたっては，入門期の学習者を，英語のルールやパターンの発見に導くように精選された英文の提示が不可欠で，英文の長さ，語彙・文法のレベル，内容やトピックなどの難易度に関して格段の配慮が要であり，その作業は容易ではない。そこで発表者らはそのような例文を集積した「英文作例参照用コーパス」を構築した。

実際の作業では，はじめに，英文用例作成の際に参照するための 2,500 万語規模の入門期英語を集めたコーパスを作成した(2016～2017 年度)。ソースとなるデータは，アメリカの教科書，日本を初めとするアジア諸国の検定教科書，入門期英語教材，Graded Reader，語彙を制限した子ども向けのニュースなどのインターネット上の入門期英語コンテンツなどから精選した。この参照用コーパスのデータは，センテンスに区切り，Stanford POS Tagger を用いて品詞タグ付けを行った。

次に，参照用コーパスを活用するための検索ツールを開発した(2018 年度前半)。効率的な用例作成を支援するという課題を解決するために、以下のような工夫を取り入れた。

(1) 検索結果はセンテンスの短いものから順に表示される
(2) 検索するセンテンスの長さを語数で指定できる
(3) 文法項目が検索できるように，表層形だけでなく，レマや品詞を指定した CQL(Corpus Query Language)による検索ができる

(1)，(2)については，参照用コーパスにセンテンス長の情報を付与し，センテンス長とセンテンスの文字列をソートキーとしてコーパスの全用例を並べ替え，重複するセンテンスを排除した。次に，センテンス長が 1 語から 100 語までの用例のサブセット(約 197 万センテンス，約 2,234 万語)を作成し，作例参照の用途により適した形に再編集した。

検索ツールの開発により，用例作成の効率が大幅に向上した。また，英語教育の専門家ではない小学校英語指導者が用例作成に参加できるようになり，授業で使いたいような英文を DDL 支援サイトに組み込むことができるようになった。

【研究発表 2】

英日バイリンガルエッセイコーパスに見るコロケーションの比喩的な意味拡張

鎌倉　義士（愛知大学）

第二言語学習者の習熟度を測る指標としてメタファーを含む比喩表現の使用が注目されている。metaphoric competence（Littlemore, 2001）の概念において，目標言語での比喩表現の理解や運用の度合が学習者の言語の熟達度と関連すると言われる（Littlemore, Krennmayr, Turner & Turner, 2014）。本研究では日本人英語学習者の比喩表現使用の分析に，英語と日本語の両言語でのエッセイをデータとするバイリンガルコーパスを使用した。なぜなら，学習

者が英語で表現できなかった比喩も対訳の日本語作文でその内容を確認することが可能であるからである。特にコロケーションに注目し日本人学習者の英作文に確認される基本的な意味から比喩的な意味への拡張について分析を行う。例えば，big, serious, significant, important など重大さを示す形容詞が problem(s)と共起する頻度が高い一方，complicated, complex という困難さを示す形容詞や basic, central など基礎的な意味を指す形容詞は日本人学習者の英作文ではネイティブよりも頻繁に使用されていない。このように日本人英語学習者のコロケーションの特徴を problem(s)との共起に限定して本研究では分析する。

研究課題は以下の３つとなる。１）problem(s)と共起する形容詞・名詞・前置詞の分析から日本人英語学習者に特徴的な problem(s)の意味拡張ネットワークが存在し，記述することは可能か。２）problem(s)の意味拡張ネットワークは単純に対訳語となる「問題」とその共起する日本語の語彙の組み合わせとは異なり，日本人学習者の英語独自の意味拡張が観察されるか。３）TOEIC スコアを基準にエッセイ作成者のグループを三つに分け，その上位と下位グループの間に意味拡張の使用頻度と英語の熟達度が関係するのか。

研究課題として述べた日本人英語学習者特有の意味拡張ネットワークは図１の「形容詞＋problem(s)」のネットワークに加え，表２「名詞＋problem(s)」と表５「problem(s)＋前置詞」からもその傾向が確認された。そのコロケーションから見る比喩的な意味拡張は，単に母語からの影響だけではなく，英語学習を通じて育まれた独自の発展である。

本研究の目的は英語母語話者が使用するコロケーションと比較し，第二言語学習者である日本人学生のコロケーション使用の狭小さや貧弱さを指摘することではない。むしろ，日本人に好まれる英語でのコロケーションを分析することで日本人英語学習者の比喩表現を記述し，日本人らしい英語を解釈可能なコロケーションと認めた上で更なる上達を促し，未発達の意味拡張の方向性を提示するのが目的である。日本人英語学習者は比喩表現の使用に積極的ではない傾向があるが，発信するための英語の上達には比喩表現の学習は不可欠である。

【研究発表 3】

英語学習者コーパス構築のためのタスク設計：特定の文法項目抽出に向けて

工藤　洋路（玉川大学）
内田　諭（九州大学）

英語学習者コーパスから得られる学習者の文法の使用状況は，言語テストの開発や教材作成，そして効果的な指導法の確立などに対して，有益な情報を提供することができるが，各文法項目と学習者のレベルの関係を調べるためには，コーパスに蓄積されている学習者の言語データがどのようなタスクを用いて抽出されたかという情報が必要になる。例えば，同じ学習者でも，あるタスクの下で行われたライティングでは，ある特定の文法項目を使用するが，別のタスクの下では，その文法項目を使用しないという現象が起きることは容易に想像がつく。つまり，学習者が特定の文法項目を習得していても，対象タスクではそれを使う必然性が生じず，その項目を使用しないということは起こり得るということであり，タスクと特定の文法項目の出現のしやすさについて検証が必要である。そこで，本研究では，「特定の文法項目を学習者から必然的に引き出すためのライティングタスクの要件は何か」をリサーチ・クエスチョンに設定した上で，文法項目抽出を目的とした複数のライティングタスクを設計し，それぞれのタスクデザインが学習者の文法使用にどの程度影響を与えるかというリサーチを行った。特定の文法項目が使用される確率が高いタスクを開発することができれば，学習者の文法習得状況を効果的に測定することが可能となり，また文法項目を軸とした学習者コーパスの構築が可能となる。

本研究では，高校の「英語表現I」の教科書のライティング活動を用いて，2 つのグループの参加者に，それぞれ異なる設計のライティングタスクを実施した。対象となる文法項目は受動態，現在完了，不定詞の名詞的用法（形式主語の it）で，1 つ目の学習者グループには，(1)教科書の活動からモデル文を取り除いたタスク（指示文のみ）と，(2)教科書の活動（指示文あり，モデル文あり）をそのまま利用したタスクの 2 つを実施した。2 つ目のグループには，(3)教科書の活動からモデル文を取り除いたタスク（指示文のみ=(1)）と，(4)教科書の活動からモデル文を取り除き，指示文を詳細に書き直したタスクの 2 つを実施した（指示文改，モデル文なし）。その結果，それぞれのタスクで使用することが想定される文法項目（受動態，現在完了，不定詞の名詞的用法）を使用した学習者の割合は，(1)よりも(2)のタスクにおいて，概して約 3 割程度高く，特定の文法項目抽出において，モデル文を提示することの有効性が示された。同様に，(3)よりも(4)のタスクにおいて，想定される文法項目（同上）を使用した学習者の割合が高くなり，特に現在完了については約 35%から 89％へと大きく上昇した。これらの結果から，モデル文の有無や指示文の詳細さなど，タスクの設計方法によって，テーマやトピックが同じライティングタスクであっても，学習者が特定の文法項目を使用できる確率が変わってくることが明らかとなった。

■10 月 7 日（日）
【研究発表第 5 セッション】
【研究発表 1】

アカデミックライティングにおけるヘッジの活用：研究論文における Discussion のコーパス分析

中谷　安男（法政大学）

1.はじめに

Wood(1982)が主張しているように，国際ジャーナルにおいては，単に結果を報告するだけでなく，その成果が該当分野にどのような貢献をもたらすのか明確な英語で示す必要がある。さらに，研究の価値を明確に読者に報告し，論文の貢献度を訴求する必要がある。そのためには，研究手法に基づき，どのような結果が得られたか正確に伝えなければならない(Cohen, 1994)。これらをまとめて伝えるのは，通常 Discussion と呼ばれる章である。しかし，Swales(2004)が指摘しているように，この箇所を量的に探究した研究は少なく，確固たる証拠よりも，研究者たちの経験や主観でこの章の書き方を示しているものが多い。一つの理由として，この箇所は様々な書き方があり，執筆者の裁量で比較的自由に記述できるという認識がある（例 Swales and Luebs, 2002)。この章では，研究の成果をアピールする必要がある。この際、研究分野の権威者である査読者に主張を伝える際に，自論に関する指摘を予め防御する必要がある。このような時に活用とされるのがヘッジ（Hedge）表現である。しかし、非母語話者にとって，英語論文の執筆の際に，いかなるヘッジ表現をどのように使用すればよいのか習得するのは容易でない (Charles, 2006)。より具体的なヘッジの考察の章における使用方法を提示すべきである。以上のような観点から本論は，社会科学，人文科学，自然科学分野の代表的な英語による学術論文の比較的規模の大きなコーパス・データを分析することにより，Discussion の章に関するこれまでの主張や概念を検証していく（例 Biber et al., 2002）。この際、既存では特定のヘッジ表現の頻度や使用傾向や目的があまり明確ではない。このため、コーパス分析により適確な言語データを抽出し，より説得力のある Discussion の望ましいヘッジ使用形式を提示したい。

2.検証方法

研究論文の一般的な傾向をみるため，自然科学，社会科学の経済・経営，人文科学の応用言語学から，それぞれインパクトファクターの高い代表的な学術誌を 2 つずつ選んだ。具体的には，*Science*，*Nature*，*International Economic Review*，*Journal of Management*，*Modern Language Journal*，*Language Learning* の 6 誌の 2006 年より 2011 年に掲載された研究論文の中から，第一著者が英語ネイティブと思われる 17 本をそれぞれ選んだ。これらを電子ジャーナルからダウンロードしテキストファイルに変換した。この合計 102 本の論文による総語数 105 万語の学術論文コーパスを作成した。

この中の Discussion として明記している章，または明記されていない場合は，それと同等の章の総計 79,876 語を抜き出した。この章のコーパスを，学術論文コーパスの他の部分を参照コーパスとして，*AntConc Windows 3.5.7* 版を活用し特徴語を抽出し，さらにクラスター表現を確認した。また Hyland(2005)の Hedge のリストを活用し，Discussion におけるこれらの表現の頻度と，他の章の参照コーパスにおける頻度と比較した。

3.結果

アカデミックライティングでは，編集者や査読者を納得させるに，巧みに成果を伝える必要があると考えられている。権威者である読者と文章で交渉を行う英語の記述が必要となる。本研究の成果では，このような目的を達成するために様々なヘッジ表現が使われていることが明らかになった。特に法助動詞の may, might, would などが、実際のテキストデータでは考察のムーヴを組み立てるのに有効であることが示された。本発表では，これらの分類に該当する事例を検証していく。

【研究発表 2】

強意語的機能を持つ罵倒語の進化特性について

新井　洋一（中央大学）

英語の罵倒語（expletives or swear words)には，例えば damn, fuck, bollocks, shit などのように，それぞれ宗教，性行為，身体器官，排泄物に関係するものが多い。今回の発表では，これらを含む罵倒語の中でも，おもに形容詞や副詞の機能を持つ罵倒語である bloody, fucking, damn などを取り上げ，これらの統語的，意味的特性と，ここ最近の際立った傾向について考察したい。なお，以下であげられる用例はすべて BNC からのものである。

これらの罵倒語は，たとえば形容詞として生起する場合，不快な（あるいは否定的）意味を持つ名詞（句）と共起し，それらを強調する強意語(intensifier)的機能を持つ。

(1) a. They must have been a **bloody** nuisance.
　　b. ... we'd have a **fucking** rain forest.
　　c. I knew later I had been a **damn** fool.

また，不快な意味を持たない名詞とも共起するが，その場合は，共起名詞に対して不快な意味合いを付帯する意味的韻律（semantic prosody）現象が感知される。

(2) a. That **bloody** bird has annoying me for days.
　　b. My God, we're all victims in this **fucking** city.
　　c. Let's get the **damn** thing finished.

一方，これらが副詞として機能する場合は，不快な意味や中立的な意味を持つ形容詞や副詞のみならず，次の用例の下線部で示されるように，むしろ快適な（あるいは肯定的）意味を持つ形容詞や副詞と共起する場合がある。以下では形容詞の例のみあげる。

(3) a. **Bloody** amazing it was!
b. I think he is **fucking** brilliant.
c. And Eleanor was **damn** lucky to have him as an escort once in a blue moon.

また次の用例で示されているように，形容詞のみならず，動詞（句）の強意語として生起している場合もある。

(4) a. I'll **bloody** come down and ask you then.
b. Gary and Margot just **fucking** kill me!
c. That's what they **damn** want, isn't it?

この発表では，(3)や(4)のような例について既存研究を概観し，これらについて深く掘り下げた研究がほとんどない点をまず指摘した上で，BNC とそれ以降の大規模英語コーパスからの用例を鳥瞰し分析を試みたい。その結果として，最近の約 30 年間のあいだに進行している，罵倒語と共起する[+PLEASANT]な意味素性を持つ形容詞や副詞の種類の増加傾向と，形容詞や副詞のみならず，動詞（句）とさえも共起する副詞機能のいっそうの進化傾向を明らかにし，その文法的仕組みについて考察する予定である。

■10 月 7 日（日）
【研究発表第 6 セッション】
【研究発表 1】

Reliability and Replicability of Annotation Schemes for Learner Corpora

Aika Miura（Tokyo University of Agriculture）

The aim of this study is to compare the inter-annotator agreements for three annotation schemes for a spoken learner corpus, the NICT JLE Corpus. The author developed the following multi-layered annotation schemes and conducted the initial annotations manually: (i) identification of learners' requestive speech acts, (ii) labeling of the functions of learners' utterances, and (iii) assignment of their degree of grammatical accuracy and acceptability in the utterances. In order to examine whether the annotation schemes were "reliable" and "replicable," and to conduct analyses in a "transparent" manner (Fuoli and Hommerberg, 2015, p. 316), the author reports the obtained agreement measure of Krippendorff's *alpha* for each annotation scheme (Artstein and Poesio, 2008; Krippendorff, 2004; Geertzen, 2012).

In the present study, the author examined learner productions during shopping role plays from the NICT JLE Corpus, which contain 68, 114, and 66 files of utterances at the Common European Framework of Reference for Languages (CEFR) levels of A1, A2, and B1, respectively. In the first scheme, the author identified pragmalinguistic features of requests, drawing on the Cross-Cultural Study of Speech Act Realization Patterns (CCSARP) coding scheme (Blum-Kulka, House, and Kasper, 1989): desire verbs (e.g., *want*) and imperatives were categorized as direct strategy, while ability/permission modals (e.g., *can*) and suggestory (e.g., *how about*) were grouped as conventionally indirect strategies. The second and third schemes were involved with classification of the function of every utterance. The utterances were divided into two major functions: "dealing with transaction" (including subcategories such as "expressing their intention of purchase" and "expressing or asking about the item") and "communication for transaction" (including "explaining the background," "requesting an action," and "confirming"). The degree of grammatical accuracy and discoursal acceptability was also assigned to each utterance: "high" indicated that the utterance was grammatically accurate and acceptable in terms of discourse; "low" was further categorized into (i) coherent (i.e., discoursally coherent but slightly ungrammatical), (ii) slightly incoherent (i.e., semantically inferable but grammatically unacceptable), (iii) incoherent (either discoursally unacceptable due to ungrammatical features or structurally and semantically acceptable but completely incoherent in terms of discourse), and (iv) featuring the use of Japanese. Table 1 shows the total numbers and ratios of the annotated segments.

Table 1
Total numbers and ratios of the annotated segments

The CEFR Level	A1	A2	B1
Total segments for the 1st scheme	597	1,170	412
Total segments for the 2nd and 3rd schemes	893	1,911	1,159
The ratios of the functions			
Dealing with transaction	55.03%	59.2%	0.65%
Communication for transaction	44.97%	40.8%	99.35%
The ratios of the high and low segments			
High	52.18%	54.98%	66.82%
Low			
Coherent	41.72%	41.74%	31.88%
Slightly incoherent	3.68%	2.13%	1.11%
Incoherent	2.18%	1.09%	0.19%
Japanese	0.23%	0.05%	0%

To attain the reliability and replicability of annotation schemes, the author: (i) documented manuals, (ii) provided annotation training to an external annotator (i.e., checker), (iii) conducted a random check of annotated segments following face-to-face discussions with the checker (i.e., 49.9% and 21.77% of the whole annotations were double-checked by the author and checker, respectively), and (iv) requested the checker to replicate the annotations of 12 files for each annotation scheme without

referencing the manuals, followed by some practice. Krippendorff's *alpha*, appropriate for "semantic and pragmatic features" involved with "different magnitudes of disagreement" (Arstein and Poesio, p. 564), were 0.842, 0.862, and 0.481 for the first, second, and third schemes, respectively. Even if it was highly revised after process (iii), the third scheme obtained only the lowest reliability. The discrepancy between the annotators may be attributed to ambiguous definitions in the manuals and the checker's insufficient training, as well as to the overly detailed and complicated classification schemes.

【研究発表 2】

Searching for grammatical items as criterial features of CEFR levels in spoken and written learner corpora: Using the CEFR-J Grammar Profile

Yukio Tono（Tokyo University of Foreign Studies）
Yasutake Ishii（Seijo University）

There is a growing interest in profiling L2 learners' proficiency levels based on the CEFR, and research projects such as *English Grammar Profile* (2015), *Global Scale of English* (2015), and the CEFR-J (Tono, 2013) seek to identify so-called "criterial features" for distinguishing one CEFR(-J) level from the others.

This study investigates what grammatical items can serve as criterial features of the CEFR(-J) levels to evaluate the English utterances and writings by Japanese EFL learners, and whether two different modes of production, i.e. spoken and written, need different criterial features. Two learner corpora, the NICT JLE Corpus (Izumi, Uchimoto, & Isahara, 2004) and the JEFLL Corpus (Tono, 2007) were used for the analysis of L2 spoken and written production respectively. All data were re-classified according to the CEFR(-J) level of each spoken/written production. Using the inventory of grammatical items developed for the CEFR-J Grammar Profile (Ishii & Minn, 2015), we obtained the relative frequency of grammatical items in each participant's data. The items used in this study were selected based on their frequency in the whole data in each corpus, whereby 124 items were chosen for the NICT JLE and 196 for the JEFLL.

Random Forest implemented in the "randomForest" package in R was used for binary classification of different levels or level groups (e.g. A1 vs A2, A1 vs non-A1, A2 vs B1.1, and A vs B) in order to search for discriminating features useful for classification. The reason why we tried to separate the data into two level groups, not into all four or five levels present in our data sets, is the rather low accuracy rates we got for multi-level classification in our pilot analysis. The data from odd line numbers were used to train our model while the remaining data was used to test the model's accuracy.

The results show that in the JEFLL, accuracy rates for telling A1 from other levels is lower than those for discriminating more advanced levels (A1-A2 79%; A1-nonA1 82%; A2-B 93%; A2-B1 93%), whereas in the NICT JLE the opposite relationship is discerned and we can more accurately distinguish A1 from other levels (A1-A2 92%; A1-nonA1 94%; A2-B 75%; A2-B1.1 71%). (All the percentages given above are approximate average figures.)

Relative weights of grammatical items as predictors were evaluated by the mean decrease in Gini index. Variables which are effective in distinguishing A1 and A2 levels in the JEFLL include prepositions, coordinating conjunctions, the definite article *the*, and to-infinitives, while in the NICT JLE, *have to* (affirmative), *can* (affirmative), the present tense of lexical verbs (affirmative), and personal pronouns *me*/*us*/*him*/*her*/*them* are measured to have great contribution to the distinction between A1 and A2.

There seems to be a big difference between the two different learner corpora in terms of what grammatical items count for discriminating the levels and how effective they are. The difference may be partly due to the different tasks in the two corpora, but it should be noted that apparently different modes (i.e. spoken vs. written) require different criterial features.

We will discuss in detail learners' use of grammatical items as possible candidates of criterial features for their CEFR(-J) levels, together with some methodological and pedagogical implications.

【講演】

Measures of Productivity and Lexical Diversity

Stefan Evert（Friedrich-Alexander-University of Erlangen-Nürnberg, Germany）

Quantitative measures of productivity and lexical diversity—such as the type-token ratio (TTR), Baayen's productivity index P or Yule's K—play an important role in many corpus studies. They have been used to assess the degree of morphological productivity, to estimate the size of an author's vocabulary, to investigate stylometric differences between writers and settle questions of disputed authorship, to study diachronic changes in grammar, to assess the readability and difficulty level of a text, to explore the linguistic correlates of dementia, and as a feature in the multivariate analysis of linguistic variation.

However, most of the approaches and quantitative analyses found in the literature suffer from serious methodological problems: (i) productivity measures often are sensitive to text size, the presence of lexicalized types and other confounding factors; (ii) there are no well-established methods for assessing the significance of observed differences in productivity, especially in the light of repetition effects due to the non-randomness of natural language; and most importantly, (iii) quantitative measures usually lack a clear linguistic interpretation that links them to intuitive notions of productivity.

In my presentation, I will show how these issues can be analyzed systematically with the help of simulation experiments based on statistical LNRE models. I will also suggest improved approaches and measures that overcome some of the problems and highlight open questions for future research.

≪会場案内図≫


10号館
5号館
10号館別館
神楽坂校舎
至 市ケ谷
11号館別館
11号館
12号館
大会会場
(2 号館)
1号館
1
6号館
13号館(森戸記念館)
8号館
7号館
2号館
3号館
9号館
3
2
東京メトロ・都営地下鉄
飯田橋駅 B3番出口
東京メトロ・都営地下鉄
飯田橋駅 B2a番出口
富士見校舎
4
JR 飯田橋駅
西口
九段中等学校前
至 水道橋
日本武道館
東京メトロ・都営地下鉄
九段下駅 1番出口

《大会参加者へのご案内》

- 会場では eduroam による無線 LAN 接続サービスが提供されています。(利用可能なのは、eduroam 参加機関のアカウントをお持ちの方に限ります。)
- 会場には駐車場の用意はございませんので、公共の交通機関をご利用ください。
- 大会（ワークショップを含む）への事前参加予約は不要です。ただし、懇親会（下記）への参加には予約が必要です。
- 第 1 日目のワークショップの受付は 2 号館 1 階 211 教室前で 9 時 30 分から行います。
- 大会受付は、第 1 日（10 月 6 日）は 2 号館 1 階 211 教室前で 12 時から行います。第 2 日（10 月 7 日）は 9 時 30 分から行います。
- 構内での喫煙は 1 号館横の喫煙室にてお願いいたします。
- 展示について：第 1 日（10 月 6 日）午後より、221 教室で賛助会員等による書籍等の展示・販売を行います。展示室には茶菓もご用意いたしますので、ご休憩場所としてもご利用いただけます。
- 昼食について：第 1 日（10 月 6 日）は、学生食堂が利用できます。第 2 日（10 月 7 日）は学内の食堂は営業しておりませんので、近隣の飲食店をご利用ください。
- 非会員の参加について：会員ではない方も、「当日会員」としてご参加いただけますので、お誘い合わせの上ご参加下さい（参加費 2,000 円、2 日間共通）。懇親会（下記）へもぜひご参加下さい。大会当日に入会受付もいたします（年会費：一般 6,000 円、学生 3,000 円）。
- 大会第 1 日の学術プログラム終了後の懇親会は、インフォーマルな雰囲気の中、参加者同士さまざまな意見交換、情報収集ができる場です。大会ご出席の方々には、ぜひ奮ってご参加いただけましたら幸いです。なお、会場準備の都合上、参加ご希望の方には事前の予約をお願いしております。ご協力のほどよろしくお願い申し上げます。
    - 英語コーパス学会第 44 回大会懇親会
    - 日時：10 月 6 日（土）18:00–20:00
    - 場所：イタリア食堂 TOKABO 神楽坂店（東京都新宿区神楽坂 2-2 志満金ビル 4 階）
    - 会費：5,000 円

  ※懇親会参加ご希望の方は、参加申込 Web フォーム（https://goo.gl/forms/cCOJmwiadShD9iN83）から 9 月 28 日（金）までにお申し込み下さい。期限前でも 65 名の申し込みがあった場合には締め切りますので、お早めにお申し込みください。