言語資源活用ワークショップ 2016

Abstract 集

2017 年 3 月 6 日 (月)　『語彙資源活用シンポジウム』
2017 年 3 月 7・8 日 (火・水)　『言語資源活用ワークショップ 2016』

大学共同利用機関法人 人間文化研究機構
国立国語研究所 コーパス開発センター 編

アンケートのお願い

シンポジウム・ワークショップ終了後以下のアンケートにご協力ください。

- 2017 年 3 月 6 日 (月)　『語彙資源活用シンポジウム』
  `https://goo.gl/aavudL`

- 2017 年 3 月 7・8 日 (火・水)　『言語資源活用ワークショップ 2016』
  `https://goo.gl/3aVhMd`

アンケートの URL は当日会場にて Physical Web 対応のスマートフォンアプリでも受信できます。

Programme:語彙資源活用シンポジウム

2017 年 3 月 6 日 (月)

【セッション 1】 (2F 講堂)

10:10-10:15　**趣旨説明**
……浅原正幸 (国立国語研究所)

10:15-10:45　**『UniDic』の拡張計画**
……岡照晃 (国立国語研究所)

10:45-11:15　**単語分かち書き用辞書『mecab-ipadic-NEologd』を公開して得た知見について**
……佐藤敏紀 (LINE)

11:15-11:45　**拡張型 NLP『JMAT』における実利用に向けた形態素解析のリソースチューニング**
……北浦雅子・紀伊馬章 (ジャストシステム)

11:45-13:00　休憩

【セッション 2】 (2F 講堂)

13:00-13:30　**『JUMAN++』の大規模語彙獲得へ向けた取り組み**
……森田一 (京都大学)

13:30-14:00　**『分類語彙表』の特徴と問題点**
……山崎誠 (国立国語研究所)

14:00-14:15　休憩

【セッション 3】 (2F 講堂)

14:15-14:45　**『日本語歴史コーパス』に出現した新規語の『UniDic』への登録について**
……鴻野知暁 (国立国語研究所)

14:45-15:15　**『日本国語大辞典』の編集方法—これまでとこれから**
……佐藤宏 (小学館)

15:15-15:45　**中型国語辞典『大辞林』編集と見出し語の収集・選定について—未知語・新語を中心に**
……山本康一 (三省堂辞書出版部)

15:45-16:00　休憩

【パネルセッション】 (2F 講堂)

16:00-17:00　パネルセッション・総合討論

Programme:言語資源活用ワークショップ 2016

2017 年 3 月 7 日 (火)

10:00-10:15　■挨　拶 (2F 講堂)　前川喜久雄

10:15-11:05　■口頭発表　A グループ (2F 講堂)

[O-A-1]

国語教科書と高校生作文の複文構造比較—従属節の構造と節形式の量的比較—

..... 松本理美 (立命館大:学生)

[O-A-2]

友人への「断り」に対する評価に関する質的考察 —日本語母語話者と中国人日本語話者の評価を通して—

..... 滕越 (東京大:学生)

11:05-11:55　■招待講演 (2F 講堂)

[I-1]

講演・講義の音声認識と字幕作成へのコーパスの活用

..... 秋田祐哉 (京都大学)

12:00-13:00　休憩

13:30-15:00　■『国語研日本語ウェブコーパス』検索系『梵天』デモ (2F セミナー室 238 室)

13:00-14:15　■ポスター発表　A グループ (2F フロア・多目的室)

[**P-A-1**]

もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら

. . . . . 今田水穂 (文部科学省)

[**P-A-2**]

コーパス構築における発話アライメントの現状

. . . . . 石本祐一 (国語研)

[**P-A-3**]

発話文への発話者情報付与の基本設計 —『現代日本語書き言葉均衡コーパス』収録の小説を対象に—

. . . . . 宮嵜由美・柏野和佳子・山崎誠 (国語研)

[**P-A-4**]

夢梅本『倭玉篇』全文テキストデータベースの構築

. . . . . 高橋大希・劉冠偉 (北海道大:学生)・池田証壽 (北海道大)

[**P-A-5**]

『日本語諸方言コーパス』の構築について

. . . . . 木部暢子・佐藤久美子・中西太郎 (国語研)
中澤光平 (与那国町与那国語辞典編集業務嘱託員)

[**P-A-6**]

相談における談話構造　—修辞機能と脱文脈化の観点からの分析—

. . . . . 田中弥生 (国語研・東京大:学生)

[**P-A-7**]

『UniDic』と『分類語彙表』の見出し対応表データの構築

. . . . . 近藤明日子 (国語研)・田中牧郎 (明治大)

[**P-A-8**]

『名大会話コーパス』の比較に基づく教室談話における「中途終了型発話」の特徴

. . . . . 矢田真菜 (東京学芸大:学生)

14:15-14:20　休憩 (ポスター切替)

14:20-15:35　■ポスター発表　B グループ (2F フロア・多目的室)

[P-B-1]
『多言語母語の日本語学習者横断コーパス』の母語話者データにおけるタスクと産出語彙の関連
..... 小西円 (国語研)

[P-B-2]
『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行
..... 加藤祥・浅原正幸・山崎誠 (国語研)

[P-B-3]
「大規模日常会話コーパス」プロジェクト—コーパスに基づく話し言葉の多角的研究—
..... 小磯花絵 (国語研)

[P-B-4]
日本語語構成情報データベースの構築
..... 淺尾仁彦 (情報通信研究機構)

[P-B-5]
発話文自動生成のための日本語表現文型辞書の作成
夏目和子 (名古屋大)・刀山将大 (名古屋大:学生)・佐藤理史 (名古屋大)

[P-B-6]
スマホで古辞書 —『篆隷万象名義』の IDS 検索を例に—
.....劉冠偉・李媛 (北海道大:学生)・池田証壽 (北海道大)

[P-B-7]
機械翻訳用超大規模辞書データ資源
..... 春遍雀來 (日中韓辭典研究所)

[P-B-8]
モンゴル語アクセント研究のためのデータベース
.....玉栄 (内モンゴル大・国語研)・西川賢哉・前川喜久雄 (国語研)

[P-B-9]
多重の読みを持つテキストのコーパス化
..... 小木曽智信 (国語研)

15:35-15:45　休憩

15:45-17:25　**■口頭発表　B グループ (2F 講堂)**

[O-B-1]

次元形容詞にみる母語話者らしい日本語形容詞の使用

. . . . . 西内沙恵 (国語研・立教大)

[O-B-2]

日本語コーパスの包括的検索環境の実現に向けて

前川喜久雄・浅原正幸・小木曽智信・小磯花絵・木部暢子・迫田久美子 (国語研)

[O-B-3]

機能語用例文データベース『はごろも』の今後の展開

. . . . . 堀恵子 (東洋大・筑波大)・内丸裕佳子 (岡山大)・加藤恵梨 (朝日大)
小西円・山崎誠 (国語研)・江田すみれ (日本女子大)
建石始 (神戸女学院大)・中俣尚己 (京都教育大)・李在鎬 (早稲田大)

[O-B-4]

日本語学習者コーパスの教育応用における留意点—『多言語母語の日本語学習者横断コーパス』に見る母語話者 L1 産出データの安定性検証を中心に—

. . . . . 石川慎一郎 (神戸大)

18:00-19:30　**■懇親会**

2017 年 3 月 8 日 (水)

10:10-11:00 **■口頭発表　C グループ (2F 講堂)**

[O-C-1]

漢語の仮名表記―実態と背景―

. . . . . 間淵洋子 (明治大:学生)

[O-C-2]

『日本語歴史コーパス』短単位アノテーション作業効率化に向けた形態素解析用辞書『UniDic』の段階的特殊化の検討―近松コーパスを例として―

. . . . . 岡照晃 (国語研)

11:00-11:50 **■招待講演 (2F 講堂)**

[I-2]

言語資源の設計・再設計と言語資源を活用した実習授業の設計

. . . . . 松吉俊 (電通大)

11:50-13:00 **休憩**

13:00-15:30 **■『国語研日本語ウェブコーパス』検索系『梵天』デモ (2F セミナー室 238 室)**

13:00-14:15　**■ポスター発表　C グループ (2F フロア・多目的室)**

[P-C-1]

全文検索システム『ひまわり』における言語分析支援機能の拡張

. . . . . 山口昌也 (国語研)

[P-C-2]

児童生徒の「手」作文に於ける経年変化の計量的分析

阿部藤子 (東京家政大)・今田水穂 (文部科学省)・宗我部義則 (お茶の水女子大付属中)
冨士原紀絵 (お茶の水女子大)・松崎史周 (日本女子体育大)・宮城信 (富山大)

[P-C-3]

『日本語日常会話コーパス』収録の進捗状況

. . . . . 田中弥生・柏野和佳子・角田ゆかり (国語研)
伝康晴 (千葉大)・小磯花絵 (国語研)

[P-C-4]

『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消

. . . . . 鈴木類 (茨城大:学生)・古宮嘉那子 (茨城大)・浅原正幸 (国語研)
佐々木稔・新納浩幸 (茨城大)

[P-C-5]

形態素解析ソフトウェア 『Web 茶まめ』の改良と Web API の試作

. . . . . 川口寛治・薦田龍輝 (東京電機大:学生)・堤智昭 (東京電機大)

[P-C-6]

『現代日本語書き言葉均衡コーパス』を用いた「～ ていく」「～ てくる」構文の意味分析

. . . . . 加藤麟太郎 (東京大:学生)・藤井聖子 (東京大)

[P-C-7]

明治初期教科書『物理階梯』のコーパス作成による語彙の考察

. . . . . 田中牧郎 (明治大)・島田むつみ・髙橋雄太 (明治大:学生)

[P-C-8]

話し言葉コーパスの転記タグ :『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

. . . . . 西川賢哉 (国語研)

[P-C-9]

『日本語日常会話コーパス』の転記基準と作業工程

. . . . . 川端良子 (国語研・千葉大:学生)・臼田泰如・西川賢哉 (国語研)
徳永弘子 (国語研・東京電機大)・小磯花絵 (国語研)

14:15-14:20　**休憩 (ポスター切替)**

14:20-15:35 **■ポスター発表　D グループ (2F フロア・多目的室)**

[P-D-1]
『現代日本語書き言葉均衡コーパス』と『分類語彙表』を利用した漢字 3 文字略熟語の抽出
……山崎誠 (国語研)

[P-D-2]
名詞項構造付与データの構築
……竹内孔一 (岡山大)

[P-D-3]
『名大会話コーパス』中納言版・ひまわり版公開データの作成
……柏野和佳子・西川賢哉・小磯花絵 (国語研)

[P-D-4]
『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション—基準策定，仕様書作成の必要性について—
……松本理美 (立命館大:学生)・浅原正幸 (国語研)・有田節子 (立命館大)

[P-D-5]
『日本語話し言葉コーパス』における発声様式の自動分類
森大毅 (宇都宮大)・藤本雅子 (国語研)・浅井拓也 (北陸先端大:学生)・前川喜久雄 (国語研)

[P-D-6]
近代文語文の通時的変化の分析　—語種率・品詞率に着目して—
……近藤明日子 (国語研)

[P-D-7]
結合の強度を測る指標としての Log-r の有用性：日・英語のバイグラムデータに基づく MI、LLR などとの比較
……藤村逸子 (名古屋大)・青木繁伸 (群馬大)

[P-D-8]
語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成
……長谷川守寿 (首都大)・西尾広美 (国語研)

[P-D-9]
固有表現抽出におけるアノテーション手法の比較
……鈴木雅也 (茨城大:学生)・古宮嘉那子 (茨城大)
岩倉友哉 (富士通研)・佐々木稔・新納浩幸 (茨城大)

15:35-15:45　休憩

15:45-16:35　**■口頭発表　D グループ (2F 講堂)**

[**O-D-1**]

『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析

.....宮内拓也 (国語研・東京外大:学生)・浅原正幸 (国語研)
中川奈津子 (千葉大・学振)・加藤祥 (国語研)

[**O-D-2**]

読み時間と情報構造について（ちょっとながめ）

.....浅原正幸 (国語研)

16:35-17:00　**■クロージング (2F 講堂)**

目次

Abstract

[O-A-1]
国語教科書と高校生作文の複文構造比較―従属節の構造と節形式の量的比較―

松本理美 (立命館大:学生)

コーパスという言語資源を活用した文体研究は，語彙，品詞，文法などに関するものなど，数多く見られるが，複文構造や従属節の研究への活用が十分であるとは言えない。これは，現時点で，解析器による従属節への情報付与技術の発展に対し，データ分析技術の普及が追いついていないことも起因していると考える。また，複文に着目した文体研究において，高校生作文や学校教科書を対象としたものは，管見の限りない。そこで，本研究では，文章中の従属節に着目し，各種学校の国語教科書と高校生作文における文体特徴を比較することを試み，文章カテゴリーごとに従属節の出現割合を求め，副詞節については，意味別に接続形式を出現頻度でランキングした。従属節の分析からは，国語教科書と高校生作文において，名詞修飾節と副詞節の出現割合に大きな差が見られ，副詞節の意味別接続形式ランキングからも文体特徴を捉えることができた。

【口頭発表】3/7(火) 10:15-10:40

〔利用する言語資源〕 高校生作文・国語教科書

[O-A-2]
友人への「断り」に対する評価に関する質的考察 ―日本語母語話者と中国人日本語話者の評価を通して―

滕越 (東京大:学生)

異文化間の「断り」に関しては，中間言語語用論などの分野で，「言語や社会的規範の違いにより衝突が起きやすい」と論じられることが多い。本研究では，個人差に焦点を当て，評価の視点から研究を進めた。『BTSJ コーパス』から 5 つの「友人の依頼への断り」の音声データを選択し，日本語母語話者 3 名と中国人日本語話者 3 名に，断られる側の視点に立って，5 つの音声の好ましさをプロトコル分析とインタビューを通して評価してもらった。その結果，録音ごとに評価が比較的一致しているものとばらけているものがあり，特に評価のばらつきが大きかった 2 つの録音は，評価者の「友人への断り」における基本的態度が，「合理性・効率性重視」か，「心情・気遣い重視」かで評価が分かれていた。また，今回のデータからは，評価のばらつきと評価者の母語との関連性は見いだせなかった。

【口頭発表】3/7(火) 10:40-11:05

〔利用する言語資源〕 『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』

[I-1]
講演・講義の音声認識と字幕作成へのコーパスの活用

秋田祐哉 (京都大)

講演・講義を記録して映像や音声を公開する取り組みが，大学をはじめとして広く行われている．これらに字幕を付与することは，専門的な内容の理解や，障碍者などの視聴支援に有用である．また，事後的にではなく，講演・講義のその場でリアルタイムに字幕を提供することは，情報保障の重要な手段である．我々は音声認識を用いてこのような字幕を効率的に作成するための研究を進めており，実際の学会講演や大学講義に対する字幕の付与も実施している．講演・講義のような，いわゆる「話し言葉」には冗長な表現が多く含まれることから，音声認識結果をそのまま字幕とすることは必ずしも適当ではなく，読みやすく整形することが望まれる．本講演では我々の字幕作成の取り組みを紹介するとともに，音声認識や整形処理を構成するために『日本語話し言葉コーパス』(CSJ) や『現代日本語書き言葉均衡コーパス』(BCCWJ) などをどのように活用しているかについて述べる．

【招待講演】3/7(火) 11:05-11:55

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』・『日本語話し言葉コーパス (CSJ)』

[P-A-1]
もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら

今田水穂 (文部科学省)

『児童・生徒作文コーパス』と『現代日本語書き言葉均衡コーパス』(BCCWJ) を用いて，児童が BCCWJ と同等の水準で漢字を使用した場合に，各漢字の頻度がどの程度になるかを推定し，その結果をワードクラウドを用いて可視化した。また，その結果を用いて，学年ごとの推定頻度の比較 BCCWJ における漢字頻度との比較，教科書コーパスについて同様に漢字頻度を推定したものとの比較を行い，推定頻度と学年の相関，児童作文に固有の高頻度漢字，小学校配当外の高頻度漢字，小学校配当の低頻度漢字などを調べた。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕『UniDic』・『現代日本語書き言葉均衡コーパス (BCCWJ)』・『児童・生徒作文コーパス』

[P-A-2]
コーパス構築における発話アライメントの現状

石本祐一 (国語研)

音声コーパスの構築にあたり，音声信号に対し発話・音韻・韻律などの各種ラベルを付与する必要がある。これらのラベルは音声分野の知識を有した作業者による目視や聴音を基に付与されることがほとんどであり，大規模コーパス構築において大きな負担となっている。特に近年研究対象となることが多い自発発話では，言い誤りや言い澱み，曖昧な発声などの現象が頻繁に生じるため，自動ラベリングを困難にしている。本稿では，転記テキストのラベリングに焦点を絞り，既存の音声認識によるシステムを応用した自動アライメントの現状について報告する。自発発話が収録されている「日本語話し言葉コーパス (CSJ)」および「日本語日常会話コーパス (CEJC)」を用いてシステムの性能評価を行い，自動アライメントの今後の課題について述べる。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕 『日本語話し言葉コーパス (CSJ)』・『日本語日常会話コーパス (CEJC)』

[P-A-3]
発話文への発話者情報付与の基本設計 ―『現代日本語書き言葉均衡コーパス』収録の小説を対象に―

宮嵜由美・柏野和佳子・山崎誠 (国語研)

現在, 国立国語研究所音声言語研究領域では,『日本語日常会話コーパス』(以下, CEJC) の開発が行われている。多様な話し言葉の会話行動の収録を目指す上記プロジェクトの理念と同様，本プロジェクトの目指す, 書き言葉における会話場面の「発話」への話者情報付与も重要な“日本語の会話”の一端を担うものである。すでに公開されている『現代日本語書き言葉均衡コーパス』(以下, BCCWJ) の約 6 割を占める書籍のサンプルには, 会話場面における大量の発話文が存在する。発話文は地の文とは言語的に異なる特徴を持つことが多いため, 分析に当たっては別に扱うことが妥当であるが, 現在の検索環境では難しい。　そこで, 本稿では, BCCWJ 収録の小説を対象に, 小説特有ともいえる発話部分特定の問題点 (かぎ括弧で括られない例や非現実場面での発話など) を提示する。機械抽出のみでは同定の難しい発話箇所と発話者情報付与について, その基本設計の「発話認定箇所」基準を中心に提案する。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-A-4]
夢梅本『倭玉篇』全文テキストデータベースの構築

高橋大希・劉冠偉 (北海道大:学生)・池田証壽

本発表は，夢梅本『倭玉篇』の全文テキストデータベースの構築と，その利用について述べるものである。『倭玉篇』は中世に生まれ近世まで広く用いられた漢和字書である。多種の写刊本が現存し，それらについて研究が行われてきたが，その多くは部首配列や特定の部を対象にした部分的なものであった。そこで『倭玉篇』の，特に和訓に関する全体的な研究を目的として，慶長 10 年（1605) 刊行の夢梅本『倭玉篇』の全文テキストデータベースを構築した。この字書は『大広益会玉篇』を中心とした中国辞書を編纂基盤としており，すでに構築されているデータを用いて効率的に入力作業を進めることができる。構築したデータベースは約 22,000 字を収録する掲出字テーブルと，約 24,000 の和訓を収録する和訓テーブルからなり，中世末期の字訓対応の資料としても価値がある。また，このデータは『平安時代漢字字書総合データベース』の和訓データの整備にも使用される予定である。

【ポスター発表】3/7(火)13:00-14:15

〔利用する言語資源〕『平安時代漢字字書総合データベース（HDIC)』

[P-A-5]
『日本語諸方言コーパス』の構築について

木部暢子・佐藤久美子・中西太郎 (国語研)
中澤光平 (与那国町与那国語辞典編集業務嘱託員)

近年，大量の言語データの整備と言語コーパスの構築が世界各国で進み，それに基づく言語研究が盛んになっている。現代日本語に関しては『現代日本語書き言葉均衡コーパス (BCCWJ)』,『日本語話し言葉コーパス (CSJ)』,『国語研日本語ウェブコーパス (NWJC)』，古典語に関しては『国語研歴史コーパス (CHJ)』，学習者の日本語に関しては『中国語・韓国語母語の日本語学習者縦断発話コーパス (C-JAS)』,『多言語母語の日本語学習者横断コーパス (I-JAS)』，等々のコーパスが整備され，研究資源として大きな役割を果たしている。しかし，方言に関しては，大規模な日本語方言コーパスがまだ存在しない。このような状況を踏まえ，国立国語研究所共同研究プロジェクト「危機言語・方言」（2016~2021 年度）では，研究所に所蔵されている諸方言の音声データを活用して，方言を横断的に検索する『日本語諸方言コーパス』の構築を開始した。現在，「各地方言収集緊急調査」（1977 年から 1985 年にかけて文化庁が行った事業による方言音声データ）のうち，方言テキストの検証作業が終わっている 48 地点（各地点平均 30 分）のデータを使用し，パラレルコーパスの方式で共通語から諸方言を検索するためのデータ整備を進めている。発表では，『日本語諸方言コーパス』の基本設計やデータ整備上の問題点，コーパスの試験的な活用について報告する。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕『日本語諸方言コーパス』

[P-A-6]
相談における談話構造　―修辞機能と脱文脈化の観点からの分析―

田中弥生 (国語研・東京大:学生)

本発表は，選択体系機能言語理論における談話分析手法の一つである修辞ユニット分析（Rhetorical Unit Analysis）によって，相談談話の構造を分析するものである。「修辞機能」と「脱文脈化程度」という，従来の相談談話分析にはない観点からその構造を確認する。『談話資料 日常生活のことば』（現代日本語研究会編）に収録され，「場面１」が「相談」である発話文を分析対象とする。先行研究では，ラジオ番組の医療相談や心理相談，また，インターネット上の相談コーナーともいえる Q&A サイト Yahoo!知恵袋などの談話構造の分析が行われてきたが，日常的な相談場面の分析はまだあまり行われていない。日常の生活における相談場面における談話構造を明らかにすることを検討する。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕『談話資料 日常生活のことば』

[P-A-7]
『UniDic』と『分類語彙表』の見出し対応表データの構築

近藤明日子 (国語研)・田中牧郎 (明治大)

国立国語研究所の大規模コーパスの構築に利用されている形態素解析辞書の元データである電子化辞書『UniDic』の見出し（語彙素）と，日本語の代表的なシソーラスの一つである国立国語研究所（編）『分類語彙表増補改訂版』の見出しとを対応づけた表形式データ（2017 年公開予定）の構築について報告する。この対応表により，『UniDic』に基づき形態素解析したコーパスへの語義情報の付与および語義情報を利用したコーパスの活用が可能となる。対応表は (1) 見出しの読み，(2) 見出しの表記，(3) 類の対応に基づき，人手による作業を経て構築した。その結果，『分類語彙表』の見出し約 64,000 と『UniDic』の語彙素約 50,000 の多対多の対応付けが実現している。また，構築の過程で明らかになった，見出しの単位設計の相違により対応付けの困難な『分類語彙表』見出しの存在や対応表を用いた大規模コーパスへの語義情報付与に向けての課題についても報告・検討する。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕『UniDic』・『分類語彙表』

[P-A-8]
『名大会話コーパス』の比較に基づく教室談話における「中途終了型発話」の特徴

矢田真菜 (東京学芸大:学生)

教室談話における「中途終了型発話」の特徴を明らかにすることを目的とし，『名大会話コーパス』による日常会話と比較した。「中途終了型発話」とは，最後まで言い切らない発話末形式のことである。指標としては，話者の関係と発話数の関係，発話末形式の生起割合，選定語の談話機能，補償行動を挙げた。「ポライトネス理論」に基づき，どのようにフェイスへの配慮が行われているかに着目した。結論として，以下のことがいえた。(1) 日常会話の二者間会話では発話権が均等に分布したのに対し，教室談話では教師の発話権が多く分布したことから，教師の発話権の多さが教室における教師の権力性を表していると考えられる。(2) 日常会話よりも教室談話のほうが「中途終了型発話」の生起割合が多く，「中途終了型発話」がフェイスへの配慮から生起することをふまえると，教室談話では日常会話よりもフェイスへの配慮が尊重されていると考えられる。

【ポスター発表】3/7(火) 13:00-14:15

〔利用する言語資源〕『名大会話コーパス』

[P-B-1]
『多言語母語の日本語学習者横断コーパス』の母語話者データにおけるタスクと産出語彙の関連

小西円 (国語研)

学習者コーパスを用いた研究は，学習者データと母語話者データを比較することによって行われることが多い。そのため，母語話者データの特徴を把握しておく必要がある。本研究では,『多言語母語の日本語学習者横断コーパス』(I-JAS) の母語話者データのうち，ストーリーテリング（以下，ST）2 種とロールプレイ（以下，RP）2 種を対象に，タスクの異なりが産出語彙にどのような影響を与えるか，その要因は何かについて考察した。考察にはコレスポンデンス分析の結果を用いた。その結果，タスク形態が独話か対話かによって，多くの品詞が異なる分布を示した。また，名詞や動詞は，タスク形態だけでなく，話題によっても分布が異なっていた。ST1 と ST2 は異なる話題を扱ったものとみなすことができ，名詞や動詞に分布の差があるが，RP1 と RP2 は扱う言語機能は異なるものの，話題という点からはほぼ同一のものとみなされ，名詞や動詞に分布の差があまり見られないことがわかった。一方で，感動詞や助詞はタスク形態だけでなく，機能によって分布に差が出る傾向が見られた。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕『多言語母語の日本語学習者横断コーパス (I-JAS)』

[P-B-2]
『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーションの試行

加藤祥・浅原正幸・山崎誠 (国語研)

国立国語研究所では,『現代日本語書き言葉均衡コーパス』(以下，BCCWJ）に分類語彙表番号を付与する作業を開始した。アノテーション作業対象として，コアデータに含まれる新聞のサンプル 54 個（部分集合 A）から順次作業に着手している。具体的には,『分類語彙表増補改訂版』(2004）の分類語彙表番号を，人手で『UniDic』の語彙素に対応させたデータ（近藤・田中 2017）により，BCCWJ の解析単位に対応可能性のある分類語彙表番号を枚挙し，短単位と長単位のそれぞれについてアノテーション作業を行う。作業者は, 枚挙された分類語彙表番号の選択肢から, 該当する意味分類が選択可能であれば選択し, 選択できない場合や, 語彙素に対応する分類語彙表番号がない場合には, 新たに適切な番号を付与する。本発表では, 短単位と長単位それぞれの番号付与作業基準と作業状況を報告する。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕 『UniDic』・『分類語彙表』・『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-B-3]
『日常会話コーパス』プロジェクト—コーパスに基づく話し言葉の多角的研究—

小磯花絵 (国語研)

国立国語研究所では，2016 年 4 月から「大規模日常会話コーパスに基づく話し言葉の多角的研究」プロジェクトを開始した。このプロジェクトでは，さまざまなタイプの日常会話 200 時間をバランス良く収録した大規模な日常会話コーパスを構築し，それに基づく分析を通して，日常会話を含む話し言葉の特性を，レジスター・相互行為・経年変化の観点から多角的に解明することを目指す。本発表では，プロジェクト全体で推進する研究，およびそのために整備・公開する複数の言語資源の全体像について触れた上で，本プロジェクトの中核を占める『日本語日常会話コーパス』を取り上げ，コーパスの設計について報告する。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕 『日本語日常会話コーパス (CEJC)』

[P-B-4]
日本語語構成情報データベースの構築

淺尾仁彦 (情報通信研究機構)

本研究では，形態素解析辞書『UniDic』への語構成情報の付与について紹介する。語構成情報とは，例えば名詞「招き猫」は，動詞「招く」と名詞「猫」の複合語であるといった情報を指す。日本語について語構成の情報が付与された公開データベースは，複合動詞など特定のカテゴリに限定されたものを別とすれば，管見のかぎり存在しない。このデータベースでは，『UniDic』に対して語構成情報をできるかぎり網羅的に付与し，品詞・語種・アクセントなど『UniDic』に元々含まれている情報と組み合わせることにより，「名詞＋動詞の複合名詞」，「アクセントが無核の動詞の名詞化で，アクセントが有核のもの」といった複雑な条件での検索を行うことができ，語彙論・音韻論・形態論などの多様な分野で言語資源として活用可能である。合わせて，開発中の検索インタフェースの紹介を行う。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕『UniDic』

[P-B-5]
発話文自動生成のための日本語表現文型辞書の作成

夏目和子 (名古屋大)・刀山将大 (名古屋大:学生)・佐藤理史 (名古屋大)

発話文の自動生成の実現基盤となる日本語表現文型辞書を作成した。この辞書は，依頼や勧誘といった発話の目的（発話意図）に対して，それを伝達する際に使用する複数の言語形式（表現文型）を整理したもので，現在，50 の発話意図に対して，のべ 675 件の表現文型が収録されている。たとえば，発話意図【依頼-実行】には，表現文型「V-てくださらない？」，「V-てくれんか？」，「お願い，V-て」などの 31 種類の表現文型が収録されている。この辞書の特徴は，それぞれの表現文型に，話し方の特徴を表す情報が付与されている点にある。たとえば，「V-てくださらない？」には，「女性的-2，大人っぽい-1，婉曲的-2，丁寧-1」という情報が付与されている。これらの情報を利用することにより，話者の特徴に応じた表現文型の選択が可能となる。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕『日本語表現文型辞書』

[P-B-6]
スマホで古辞書 ―『篆隷万象名義』の IDS 検索を例に―

劉冠偉・李媛 (北海道大:学生)・池田証壽 (北海道大)

近年，スマートフォンやタブレットのようなモバイル端末が普及し，日常生活を変えつつあり，日本語教育・日本語研究にも使えるようになると予想される。しかしながら，構築・公開が盛んである古典籍・古文書のデータベースは PC 向けが多く，PC 以外の端末で利用する際は表示サイズのずれや機能障害がよく発生する。そこで，モバイル端末でデータベースを利用しているユーザを想定した利便性が高い言語資源データベースの Web インタフェースを開発したい。漢字字形の構造情報を用いて古辞書のテキスト・画像を検索することによって文字の同定に使える Web アプリがまだないので，篆隷万象名義の掲出字について IDS 検索と画像表示を可能にするツールを試作した。本アプリによって，漢字のパーツで篆隷万象名義に掲載している文字の画像をスマートフォンなどの携帯端末で検索でき，写本の解読・翻刻する際に役立つと期待している。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕『平安時代漢字字書総合データベース（HDIC)』

[P-B-7]
機械翻訳用超大規模辞書データ資源

春遍雀來 (日中韓辭典研究所)

情報交流の国際化に伴い多言語情報の充実は今や喫緊の課題である。特に固有名詞や POI (points of interest) は膨大な数量に加え頻繁な名称変更にも対応する必要があるため，正確で充実した多言語辞書データ資源が必須だ。そこで，機械翻訳の作業効率と精度を格段に向上させる，超大規模辞書データ資源 (Very Large Scale Lexica: VLSL) の構築例として，固有名詞・専門用語等を含む日中韓英辞書データベースや多言語固有名詞辞書データベースを紹介する。VLSL は情報検索・形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野に応用が可能で更なる展開が期待される。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕日本語，中国語（簡体字・繁体字)，韓国語，英語，アラビア語，インドネシア語，ベトナム語，タイ語，ヒンディー語，ロシア語，ドイツ語，ポルトガル語，スペイン語，フランス語，イタリア語

[P-B-8]
モンゴル語アクセント研究のためのデータベース

玉栄 (内モンゴル大・国語研)・西川賢哉・前川喜久雄 (国語研)

モンゴル語のアクセントは，音韻論的には弁別的でないといわれているが，音声学的な特徴については研究者によって意見が分かれている。従来は，第一音節に固定ストレスアクセントを認める研究が主流であったが，1980 年代以降，実験音声学の影響によって，アクセントは第一音節に固定されておらず，その変異には音節構造が関係しているとの主張が広がってきた。本発表では，モンゴル語アクセントの音声学的特徴を把握するために，筆者らが設計と実装を進めている音声データベースについて報告する。このデータベースは，音節構造，母音の長短，隣接子音等に配慮した単語リストを複数の話者が発音したサンプルに，種々の音響特徴量を付与したものとなっており，モンゴル語のアクセントが種々の韻律的特徴（長さ，強さ，高さ）および分節的特徴とどのような関係にあるかを解明するために利用できる。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕 モンゴル語

[P-B-9]
多重の読みを持つテキストのコーパス化

小木曽智信 (国語研)

日本語のテキストには，本文漢字の通常の読みを示すのではない特殊な読みをもつ振り仮名（たとえば「強敵」と書いて「とも」とふりがなを振る類）や，掛詞（「ながめ」を「眺め」「長雨」の両用に読む類から，語形の一部から別の語を連想させる類まで），各種の洒落など，意図的に多重の読みを持たされたテキストが少なくない。また，書き手においては唯一の読みが定められるものであっても，読み手にとっては複数の妥当な読みがありうる場合がある（漢字の複数の読み方から，仮名文字列の解釈の仕方まで）。従来のコーパスではこのような多重の読みは切り捨てられ，選択されたただ一つの読みを配置することが多かった。本発表では，このような多重の読みを持つテキストについて，主として『日本語歴史コーパス』の事例を整理して示すとともに，そのあるべきコーパスアノテーションの方法について論じる。

【ポスター発表】3/7(火) 14:20-15:35

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』・『日本語歴史コーパス (CHJ)』

[O-B-1]
次元形容詞にみる母語話者らしい日本語形容詞の使用

西内沙恵 (国語研・立教大)

日本語非母語話者（以下，NNS）は，形容詞用法の習得過程において「* 大きいの魚」といった名詞修飾にノ格を挿入する誤用や，形容動詞との活用の混同，時制の間違い，また連用用法での活用間違いなどを経ることが知られている。加えて，「？古い先生」など意味の面での誤用も少なくない。しかし，これらの文法規則こそが，日本語形容詞の使用における特性の全てだろうか。非文ではないが，日本語らしくないと感じられることがあるのはなぜか。本研究では，『現代日本語書き言葉均衡コーパス』で得られた実例をもとに，現代日本語の次元形容詞「高い」がとる構造を分析した。その構造とは，被修飾名詞に求められる格の要素とそれに伴い表出する意味である。さらに，『多言語母語の日本語学習者横断コーパス』で得られた NNS と日本語母語話者（以下，NS）の言語使用を比較したところ，NNS に文脈依存的な発話が目立つ一方で，NS には助詞句を明示する発話が多かった。「高い」がとる構造と NS・NNS 間の使用における差異から，形容詞使用の特性を明らかにする。

【口頭発表】3/7(火) 15:45-16:10

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』・『多言語母語の日本語学習者横断コーパス (I-JAS)』

[O-B-2]
日本語コーパスの包括的検索環境の実現に向けて

前川喜久雄・浅原正幸・小木曽智信・小磯花絵・木部暢子・迫田久美子 (国語研)

国立国語研究所コーパス開発センターでは，従来個別に開発・提供されてきた各種日本語コーパスの検索環境を統合し，複数のコーパスを横断的に検索可能な包括的検索環境を整備する計画を進めている。既に公開済みのコーパス群だけでなく，第 3 期中期計画期間に種々の研究プロジェクトで開発ないし拡張を予定しているコーパス群の一部も検索対象に含める。本発表では，検索対象となる予定のコーパスを紹介した後に包括的検索環境の実現に向けてどのような問題があるかを検討し，解決の方向性を探る。

【口頭発表】3/7(火) 16:10-16:35

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』・『日本語歴史コーパス (CHJ)』・『多言語母語の日本語学習者横断コーパス (I-JAS)』・『日本語話し言葉コーパス (CSJ)』

[O-B-3]
機能語用例文データベース『はごろも』の今後の展開

堀恵子 (東洋大・筑波大)・内丸裕佳子 (岡山大)・加藤恵梨 (朝日大)
小西円・山崎誠 (国語研)・江田すみれ (日本女子大)
建石始 (神戸女学院大)・中俣尚己 (京都教育大)・李在鎬 (早稲田大)

機能語用例文データベース『はごろも』は，web 上で機能語の一部を検索すると，意味，項目の難易度，典型例 (作例)，話し言葉と書き言葉の用例などが見られるツールで，2015 年秋に公開した。利用者からは，文法項目が日本語教育の観点から調べられる便利なツールと評価される一方，用例だけでは理解が難しいこともある，意味用法の説明が難しい，項目の意味・機能から文法項目が選べるといいなどの声を聞く。そこで，今後の改定の方針として，(1) 見出し項目を精査，(2) 文法項目に文法機能の情報をつける，(3) 意味用法の記述を精査し，階層のある分類とする，(4) 文法項目の前接の形式を明示する，(5) 学習者作文コーパスなどから学習者の正用，誤用の文を学習者のレベルと共に示す，の 5 点挙げ，3 チームで作業を進めている。2017 年度末までは改訂版を公開する予定である。

【口頭発表】3/7(火) 16:35-17:00

〔利用する言語資源〕『日英新聞記事対応付けデータ (JENAAD)』・『Kyoto University and NTT Blog コーパス (KNBC)』・『現代日本語書き言葉均衡コーパス (BCCWJ)』・『CASTEL/J CD-ROM v1.5』・『日本語会話データベース』・『宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)』『名大会話コーパス』・『BTS による多言語話し言葉コーパス—日本語会話 1 』

[O-B-4]
日本語学習者コーパスの教育応用における留意点—『多言語母語の日本語学習者横断コーパス』に見る母語話者 L1 産出データの安定性検証を中心に—

石川慎一郎 (神戸大)

『多言語母語の日本語学習者横断コーパス』(I-JAS) を初めとする大型の日本語学習者コーパスの整備が進んだことで，母語話者と学習者の言語運用を比較し，学習者の逸脱性を客観的に明らかにした上で L2 教育の質的改善を図る可能性が拓かれつつある。しかし，こうした研究を実践する際には，母語話者データおよび学習者データの性質を十分に理解し，得られた結果を慎重に解釈する必要がある。本研究では，日本語学習者コーパスの教育応用を考える際に留意すべき問題点を概観した後，とくに母語話者による L1 産出データの安定性の問題を取り上げ，I-JAS を使った検証を行う。検証の結果，母語話者の L1 産出であっても，その正確性や言語特性については想像以上の多様性が存在することが示された。

【口頭発表】3/7(火) 17:00-17:25

〔利用する言語資源〕『多言語母語の日本語学習者横断コーパス (I-JAS)』

[O-C-1]
漢語の仮名表記―実態と背景―

間淵洋子 (明治大:学生・学振)

本発表は，本来漢字で表記されるはずの漢語が平仮名や片仮名で表記される事象を取り上げ，『現代日本語書き言葉均衡コーパス』（以下，「BCCWJ」と表記）を用いて，その実態と背景を明らかにすることを目的とする。BCCWJ の網羅的な漢語の表記実態調査に基づいて求めた，個々の語の仮名表記率から，語の表記において仮名表記が主たる表記になっている語，仮名表記がある程度一般的である語を特定した。その上で，仮名表記の定着度合いに，(1) 語自体の出現状況（語彙レベルが高い漢語ほど仮名表記率は低い），(2) 常用漢字を基準とする字体カテゴリ（常用漢字表外音訓を含む語は仮名表記率が高いが，表内字でも仮名表記率の高い語がある），(3) 音声変位形の有無（「格好/カッコウ」にたいする「カッコ」のような音声変位形を持つ語は仮名表記率が高い），(4) 意味分野（動植物や近接する食物や家事の分野では仮名表記率が高い），(5) 品詞（副詞用法を持つ語は仮名表記率が高い），(6) レジスター（Web 媒体のテキストでは仮名表記率が低い）などの関連性が見られることを明らかにした。また，仮名表記選択に最も強い影響を与えると思われる字体特徴にかかわらず，意味分野や品詞において特定の語彙群が同様の傾向を見せるのは，表記の選択や嗜好に，類似性に基づく合理化作用が働くことに起因する可能性を主張した。

【口頭発表】3/8(水) 10:10-10:35

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』

[O-C-2]
『日本語歴史コーパス』短単位アノテーション作業効率化に向けた形態素解析用辞書『UniDic』の段階的特殊化の検討―近松コーパスを例として―

岡照晃 (国語研)

国立国語研究所では，日本語の通時コーパス『日本語歴史コーパス』（以下 CHJ）の構築を進めている。CHJ の特徴として，国語研の規定する言語単位 : 短単位での形態論情報がアノテーションされていることがある。CHJ 構築では，形態論情報の人手アノテーションの効率化のため，各時代専用に形態素解析器『MeCab』用の解析用辞書『UniDic』のコストを学習し，自動解析後，人手修正するという作業方針を採っている。現状では基本，各時代ごとの辞書を 1 個用意するという粒度であるが，地の文と会話文での文体・文法の差から，文語版，口語版をそれぞれ用意している時代もある。本発表では，人手の形態論情報修正をさらに効率化するため，各資料により特化した解析用辞書構築について述べる。具体的には類似ドメインの汎用解析用辞書から始め，段階的に学習用コーパスを限定していく追加学習を繰り返すことで，対象となる資料に特化した解析用辞書を構築する。今回は，近世上期上方資料である近松門左衛門の世話物浄瑠璃を対象とし，『洒落本解析用 UniDic』から『近松解析用 UniDic』の構築について述べる。

【口頭発表】3/8(水) 10:35-11:00

〔利用する言語資源〕『UniDic』・『日本語歴史コーパス (CHJ)』

[I-2]
言語資源の設計・再設計と言語資源を活用した実習授業の設計

松吉俊 (電通大)

この講演では,「言語資源と設計」に関して，主に次の 2 つのトピックについて話す。1 つめのトピックは,『現代日本語書き言葉均衡コーパス』の一部に対して，事象のモダリティの情報を付与したコーパスである。このコーパスは，2016 年 3 月より中納言の関連データ配布サイトから入手可能である。コーパスの内容，作成目的，活用事例等について紹介する。このコーパスのモダリティラベル体系は，最初の設計から 7 回も改訂され，今に至っている。各々の再設計が必要になった理由について，言語学的観点と工学側からの要請を軸に話したい。2 つめのトピックは，計算言語学や自然言語処理に興味を持ってもらえるような実習授業の設計である。私は，2014 年度から 3 年間，情報系の大学 3 年生を対象としたコンピューター演習授業を担当している。言語資源を活用し，学生自身がテキストアノテーションを体験する授業の内容について紹介したい。
【招待講演】3/8(水) 11:00-11:50

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』・青空文庫・Wikipedia

[P-C-1]
全文検索システム『ひまわり』における言語分析支援機能の拡張

山口昌也 (国語研)

本稿では，筆者が開発している全文検索システム『ひまわり』の言語分析支援機能の拡張について述べる。元来,『ひまわり』は言語資料の検索と閲覧を目的に設計されたコンコーダンサであり，検索結果を分析するための機能を十分に備えていなかった。しかし，検索対象の資料の規模が大きくなると，大量の検索結果を単に表示するのではなく，集約して分析する必要性が生じる。また，検索結果の統計的な分析には，資料に含まれる文字数といった，基本的な情報を計測できなければならない。そこで，(1) 検索結果の集約機能，(2) 統計的分析のための基礎データの収集機能を『ひまわり』に実装した。拡張された機能を用いることにより，例えば『名大会話コーパス』の各会話中の発話数，文字数，単語数，特定の単語の出現数といった情報を収集できるようになる。
【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕 全文検索システム『ひまわり』・『名大会話コーパス』・国会会議録・青空文庫

[P-C-2]
児童生徒の「手」作文に於ける経年変化の計量的分析

阿部藤子 (東京家政大)・今田水穂 (文部科学省)・宗我部義則 (お茶の水女子大付属中)
冨士原紀絵 (お茶の水女子大)・松崎史周 (日本女子体育大)・宮城信 (富山大)

本発表は児童生徒らの文章作成能力の経年変化を計量的分析によって明らかにすることを目的とする。その基礎資料として作文を電子化した「「手」作文コーパス」を構築した。本コーパスの資料は 1992 年及び 2016 年に児童生徒らが書いた「手」を題とする作文である (両資料は，同一の国公立大附属小中学校で同条件で作成されたものである)。両資料の調査時期にはおよそ四半世紀 (24 年) の隔たりがあり，本発表の目的はその間の児童生徒らの文章作成能力の変化の有無を明らかにすることにある。予備調査を行った結果，1 サンプル当たりの文章量 (総字数)，語数，文節数等で両資料間に明確な差異を見いだすことはできず，文章の量的観点からは大きな経年変化は見られないことが分かった。一方で，現場の教師らから「以前に比べて子ども達が作文が書けなくなった」という指摘を聞くこともあり，使用語彙の種類や品詞の偏り，文末形式等の文体的特徴の違いを数量的差異として抽出し，2 つの資料の異動を観察する。その結果に基づき先の教師らの指摘の妥当性を検討する。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕 「手」作文コーパス（1992 年・2016 年）

[P-C-3]
『日本語日常会話コーパス』構築における会話収録方法と進捗状況

田中弥生 (国語研・東京大:学生)・柏野和佳子・角田ゆかり (国語研)
伝康晴 (千葉大)・小磯花絵 (国語研)

2016 年度から構築が始まった「大規模日常会話コーパス」プロジェクトによる『日本語日常会話コーパス』の収録手続きの概要と進捗状況について報告する。本プロジェクトでは，日常場面の中で自然に生じた会話を対象とする。そのため，性別・年代などの点からバランスを考慮して調査協力者を選別し，収録機材等を 2~3 カ月程度貸し出して調査協力者自身に日常会話を収録してもらう方法を採用している。本発表では，こうして定めた収録方法の概要を述べるとともに，これまでに終了した 13 名の調査協力者による約 200 時間の収録について進捗状況や生じた問題などを報告する。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕 『日本語日常会話コーパス (CEJC)』

[P-C-4]
『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消

鈴木類 (茨城大:学生)・古宮嘉那子 (茨城大)・浅原正幸 (国語研)
佐々木稔・新納浩幸 (茨城大)

all-words の語義曖昧性解消とは，文章中の全多義語の語義を一意に決定するタスクである．単語の語義はその周辺の文脈によって決まることから，周辺の単語同士が類似している場合その中心にある語義曖昧性解消の対象単語同士の語義も類似していると考えられる．そこで本研究では，単語の分散表現を用いて対象単語の周辺単語群と対象単語の各語義候補における類義語の周辺単語群の間の距離を測り，その距離を用いて対象単語の語義を予測した．そして，単義語の語義と多義語の予測で得た語義を基にして『分類語彙表』の概念（語義）の分散表現を作成し，“ 単語の分散表現＋概念の分散表現”を用いて周辺単語群間の距離を測りなおして再び語義を予測し，さらにこれを繰り返し行った．『現代日本語書き言葉均衡コーパス』に『分類語彙表』のコードが付与されたコーパスを用いて実験を行ったところ，単語の分散表現のみを用いた予測では 54.2%，単語と概念の分散表現を用いた予測では最大で 59.0% の正解率となった．

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕『分類語彙表』・『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-C-5]
形態素解析ソフトウェア 『Web 茶まめ』の改良と Web API の試作

川口寛治・薦田龍輝 (東京電機大:学生)・堤智昭 (東京電機大)

国立国語研究所では，近代文語文や近代以前の古典・古文資料の分析が可能な，『UniDic』を用いた形態素解析支援アプリケーションである『Web 茶まめ』を公開している。『Web 茶まめ』は，日本語研究者の研究支援や，教育機関における学習利用を目的として，手軽に形態素解析を行える環境の提供を目的としている。そのため『Web 茶まめ』は，Web ブラウザを用いて簡単な GUI 操作で，特別なソフトウェアを導入することなく形態素解析が可能である。本稿では，以下の二点を報告する。一点目は『Web 茶まめ』の機能拡張についての報告である。『Web 茶まめ』を公開した以降に寄せられた意見や指摘を元に，『Web 茶まめ』の機能拡張や改善を行った。二点目は，WebAPI の試作についての報告である。Web 上で形態素解析を行う新たな利用形態として，『Web 茶まめ』の WebAPI を試作した。外部の Web サイトやアプリケーションは，試作した WebAPI を用いて『Web 茶まめ』サーバと HTTP 通信を行うことで，『Web 茶まめ』の機能を利用可能となる。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕『UniDic』

[P-C-6]
『現代日本語書き言葉均衡コーパス』を用いた「～ていく」「～てくる」構文の意味分析

加藤麟太郎 (東京大:学生)・藤井聖子 (東京大)

『現代日本語書き言葉均衡コーパス』(以下，BCCWJ) を用いて「～ていく」「～てくる」構文の物理的移動の用法の意味分析を行った。焦点をあてた研究問題は，「～ていく」「～てくる」の前に共起する動詞の意味特性と「～ていく」「～てくる」構文の意味特性との関係は，どの程度規則的でありどの程度予測可能かという問題である。BCCWJ からの「～ていく」「～てくる」構文の無作為抽出のうち，物理的移動を表す用法をそれぞれ 497 用例，580 用例抽出し，森田 (1998) に基づく仮説を立て，意味コーディングをした。コーディングに基づく定量的予備分析において，まず多くの用例がある程度予測可能な規則的傾向を示すことが明らかになったが，本稿では，左記傾向の定量的分析の報告に加えて，コーパスにみられる例外的用例の方に着目し， 例外的用例の定性的分析を示した上で，新たな意味分類を提案する。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-C-7]
明治初期教科書『物理階梯』のコーパス作成による語彙の考察

田中牧郎 (明治大)・島田むつみ・髙橋雄太 (明治大:学生)

明治初期の小学校用の物理学の教科書『物理階梯』のコーパスを作成し，同時期の啓蒙雑誌『明六雑誌』のコーパスのそれと比較することを通して，語彙の考察を行った。まず，『物理階梯』の語彙は，『明六雑誌』の語彙に比べて，異なり語数において和語の比率が高いことがわかった。また，『明六雑誌』と比較した場合の『物理階梯』の特徴語を抽出して，その性質を考察すると，物理学のテーマに関連してよく用いられる「テーマ語」，物理学を体系的に論じるために必要とされる「専門語」，物理学に限らず学術的な内容を叙述するのに適した「学術語」の３種に分類できた。その３種を，『増補改訂分類語彙表』の部門別に集計すると，テーマ語は「生産物および用具」に，専門語は「自然物および自然現象」に，学術語は「抽象的関係」に，それぞれ特に多いことなどがわかった。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕 『UniDic』・『日本語歴史コーパス (CHJ)』・『物理階梯』のコーパス（自作）

[P-C-8]
話し言葉コーパスの転記タグ:『多言語母語の日本語学習者横断コーパス』と『日本語話し言葉コーパス』の比較

西川賢哉 (国語研)

近年の話し言葉コーパスにおいては，発話を書き起こした転記テキストに，タグ（転記タグ）が付与されることが多い。本発表では，『多言語母語の日本語学習者横断コーパス』(I-JAS) および『日本語話し言葉コーパス』(CSJ) を対象に，そこで用いられているタグの種類・形式・目的・実際の用例を整理したうえで，両者の比較を行なう。比較の結果，(i) 一方にしか存在しないタグもあるが，両コーパスでほぼ同様の機能を果たすタグも少なからず存在する（例えば，フィラー，語断片，発音誤りを表すタグ），(ii) ただし，同様の機能を果たすタグとはいえ，タグの適用範囲（転記テキストのどこからどこまでにタグを付与するか）や，タグの適用対象（タグをそもそも付与するか否か）など，細かい点では違いもある，ということが判明した。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕『多言語母語の日本語学習者横断コーパス (I-JAS)』・『日本語話し言葉コーパス (CSJ)』

[P-C-9]
『日本語日常会話コーパス』の転記基準と作業工程

川端良子 (国語研・千葉大:学生)・臼田泰如・西川賢哉 (国語研)
徳永弘子 (国語研・東京電機大)・小磯花絵 (国語研)

本稿は，平成 28 年度から構築を進めている『日本語日常会話コーパス』の転記基準と転記作業工程を紹介する。本コーパスには，日常場面で自然に生じるさまざまなタイプの会話 200 時間がバランス良く収録される予定である。日常会話には，極めてくだけた表現も頻出する。こうしたデータを多人数で書き起こしをするためには，文字化をするための基準を明確に定める必要がある。また，大量の会話を限られた期間で書き起こすために，効率的に作業をするための工夫が必要になる。本発表では，これまでに収録された会話を転記しながら策定した転記基準と効率的に作業を行うために用いている方法を紹介する。

【ポスター発表】3/8(水) 13:00-14:15

〔利用する言語資源〕『日本語日常会話コーパス (CEJC)』

[P-D-1]
『現代日本語書き言葉均衡コーパス』と『分類語彙表』を利用した漢字 3 文字略熟語の抽出

山崎誠 (国語研)

「政財界」「国内外」などの漢字 3 字で構成される「略熟語」と呼ばれる形式は，先行研究が少なく実態が明らかでない。国語辞書にも掲載されることが少ない。本発表では，現代日本語にはどのような略熟語が存在するかを『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）と『分類語彙表』を使って自動的に抽出することを試みた。具体的には，BCCWJ から，前後が非漢字という条件で漢字 3 文字連続を抜き出し，それらを構成する漢語の頻度および分類語彙表における意味番号を付与したデータを作成した。そこから，出現頻度が一定以上で，構成要素となる漢語の分類番号が一致するものとして 874 語を抽出した。内訳は「政財界」タイプ 656 語，「国内外」タイプ 297 語，重複が 79 語であった。目視で確認したところ，抽出された 3 字漢語には，略熟語でないものも多く，精度を高めるにはさらに別の条件が必要であることが分かった.

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕 『分類語彙表』・『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-D-2]
名詞項構造付与データの構築

竹内孔一 (岡山大)

含意認識タスクなど言語処理での文間の表現を取り扱う際，名詞の意味的な関係を捉える必要がある。言語学の分析から名詞の中には名詞の意味を補完する外部情報が必要なものが分かっており，生成語彙における特質構造 (クオリア構造) として記述することが提案されている。また言語資源では NomBank に代表されるように名詞の項構造を事例とともに構築されている。本研究では，先行研究で提案された特質構造を利用した名詞の項構造データを基に言語処理の観点からより形式化した構築法を提案する。具体的には名詞の項構造の例文を構築するとともに，項を同定し，述語との関係を項構造を通して結び付ける記述枠組である。述語のデータとして述語項構造シソーラスを利用し，NTCIR の RITE-2 で出現した名詞を対象に項構造の例文および対応する述語と項の関係を記述したデータを構築した。本稿では，記述枠組，および具体的に構築した名詞項構造データの事例を説明すると共に，付与での問題点や現状について記述する。

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕 『述語項構造シソーラス』

[P-D-3]
『名大会話コーパス』中納言版・ひまわり版公開データの作成

柏野和佳子・西川賢哉・小磯花絵 (国語研)

『名大会話コーパス』は，科学研究費基盤研究 (B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（研究代表者 : 大曾美恵子，平成 13 年度 ～15 年度）の一環として作成された，120 会話，合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである。国立国語研究所に移管後，文字化テキストを公開し，続けて『中納言』版，『ひまわり』版を作成し，公開している。本発表では，『名大会話コーパス』の概要と特徴を述べる。また，『中納言』版，『ひまわり』版公開データの作成に際して行った，形態素解析結果の人手修正の内容について報告する。

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕 『名大会話コーパス』

[P-D-4]
『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション―基準策定，仕様書作成の必要性について―

松本理美 (立命館大:学生)・浅原正幸 (国語研)・有田節子 (立命館大)

本発表では，「現代日本語書き言葉均衡コーパス」に対する節の意味分類情報アノテーションについて報告する。多様な形式を持ち，文脈の中でその意味が解釈される日本語文中の従属節の意味分類については，人手による分類が不可欠である。そこで，我々は「鳥バンク」基準互換 (池原 2009) の節の意味分類情報アノテーションを進めている。しかし，現行の作業においては，節の認定，タグ付け箇所，作業者の言語感覚に頼るところが大きい意味分類判断など，作業上の揺れも多く，改善が求められる。作業効率と信頼性の向上に繋がる基準策定と仕様書作成が必要であり，そのためには現行作業での問題点を整理することが必須であると考える。本発表では，人手による節境界アノテーション・節の意味分類タグ付け作業についての基準策定と仕様書作成が今後のコーパス開発に資することを主張し，現在の作業における問題点に焦点を当てた考察を行う。

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』

[P-D-5]
『日本語話し言葉コーパス』における発声様式の自動分類

森大毅 (宇都宮大)・藤本雅子 (国語研)・浅井拓也 (北陸先端大:学生)・前川喜久雄 (国語研)

喉頭音源由来の声質の違いは，話者のパラ言語メッセージならびに心的・認知的状態を伝えるシグナルであり，自発音声コーパスに求められる重要な情報であるが，そのアノテーションは音声学の専門家でなければ難しくコストが大きい。本研究は，機械学習による声質の自動アノテーションの可能性を探ることを目的とする。本研究では，非流暢性にも関連する従来よく用いられてきた発見的な音響特徴量に加え，近年音声からの感情認識で広く用いられるようになった大規模な特徴量セットの効果を検証した結果を報告する。

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕『日本語話し言葉コーパス (CSJ)』

[P-D-6]
近代文語文の通時的変化の分析　―語種率・品詞率に着目して―

近藤明日子 (国語研)

明治期において，論説文・報道文等の実用文は文語体であることが主流であり，明治初期には漢文訓読文をはじめとする複数種類の文語体が行われたものが，しだいに融合し，明治 30 年代に普通文と呼ばれる標準的な文語体が確立・定着する変遷が知られている。しかし，その変遷の詳細については未だ明らかでない点も多い。そこで本発表では，その実態の一端を明らかにすることを目的として，国立国語研究所（2016）『日本語歴史コーパス明治・大正編 I 雑誌』（短単位データ 1.0）の中の文語体で書かれた非文芸記事のデータを利用し，基本的な文体指標である語種率・品詞率の通時的変化の分析を行う。そして，分析の結果判明した漢語率・名詞率の増加等の傾向の背景について考察する。

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕『日本語歴史コーパス (CHJ)』

[P-D-7]
結合の強度を測る指標としての Log-r の有用性：日・英語のバイグラムデータに基づく MI，LLR などとの比較

藤村逸子 (名古屋大)・青木繁伸 (群馬大)

2 語からなるコロケーションは一般に共起頻度と 2 語の結合力によって特徴づけられる。本研究は, 結合力の指標として Fujimura & Aoki (2016) において提案した Log-r を, 同じ目的の指標として言及されることの多い MI, LLR, t-score, Dice, Jaccard と比較し, 簡素な指標である Log-r の有用性を主張する。データは『現代日本語書き言葉均衡コーパス』と英語の大規模新聞コーパスから網羅的に採取した多量のバイグラムを用いる。横軸にバイグラムの共起頻度をとり，縦軸に各指標値をとった散布図を作成して各指標の特徴を視覚的に描き，散布図間の比較によって指標間の差異を明示する。
【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕『UniDic』・『現代日本語書き言葉均衡コーパス (BCCWJ)』・英語新聞・フランス語新聞

[P-D-8]
語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成

長谷川守寿 (首都大)・西尾広美 (国語研)

現在，多くの幼稚園で日本語を母語としない保護者（以下 NNS 保護者）が見られるが，中には日本語学習の機会がなく日本語が十分に理解できないケースもある。そのような場合，幼稚園の配布文書が理解されず，情報伝達がうまくいかずに保育活動に支障をきたすという問題も発生している。そこで我々は，将来的に教師と NNS 保護者を結ぶ「保護者に伝わるやさしい日本語」のテキスト化をめざし，まず語彙・文型調査の前段階として『幼稚園の配布文書コーパス』を作成している。コーパスの作成では，より精度の高い語彙・文型調査が行えるよう，OCR ソフトの認識誤りを人手だけで修正するのではなく，形態素解析システム（unidic-mecab-2.1.2）も活用して誤りを発見して修正し，さらに正確に語に区切れない場合は表記の変更・記号の追加を行っている。発表では，形態素解析システムをテキスト入力に用いる特定目的のためのコーパス作成法について報告する。
【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕『幼稚園の配布文書コーパス』・『UniDic』

[P-D-9]
固有表現抽出におけるアノテーション手法の比較

鈴木雅也 (茨城大:学生)・古宮嘉那子 (茨城大)
岩倉友哉 (富士通研)・佐々木稔・新納浩幸 (茨城大)

本稿では，非専門家による固有表現抽出のタスクとしてのアノテーションを題材に，ふたつの手法について比較を行った. ひとつは既存の固有表現抽出器によるアノテーション結果に対し，人手で修正を行う手法であり，もうひとつは人手で一からアノテーションを行う手法である. 実験には現代日本語書き言葉均衡コーパス (BCCWJ) を利用し，手法ごとに 1 テキストに対し 2 人の非専門家を割り当てて，アノテーションを行った. 評価には，アノテーションにかかる時間，一致率，Gold Standard との比較による正解率，それぞれの手法で作成されたコーパスを訓練事例とした場合の正解率を利用し，ジャンルごと，及び，全ジャンルのマイクロ平均とマクロ平均を算出した. 本実験の結果から，全ジャンルのマイクロ平均とマクロ平均で比較した場合には既存のアノテーション結果を用いた手法の方が良い結果となるが，既存の固有表現抽出器の訓練事例から離れたジャンルで同様に比較した場合には人手でアノテーションを行う手法の方が良い結果となることが明らかになった.

【ポスター発表】3/8(水) 14:20-15:35

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』

[O-D-1]
『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析

宮内拓也 (国語研・東京外大:学生)・浅原正幸 (国語研)
中川奈津子 (千葉大・学振)・加藤祥 (国語研)

冠詞がない言語を母語とする者にとって，冠詞がある言語を習得する際の冠詞選択は難しいものである。冠詞選択は，一般に定性や特定性などの情報構造が大きな影響を与えると考えられる。言語処理の分野では英語母語話者が産出した大量のテキストから，英語学習者の冠詞の誤りを検出する手法が提案されているが，日本語母語話者が産出する他言語の冠詞選択を検討する場合，日本語における名詞句の情報構造を考慮する必要がある。さらに機械翻訳において日本語文を冠詞のある言語に翻訳する際にも，日本語の情報構造が問題となってくる。本稿では，機械翻訳での冠詞選択の問題に関する基礎研究として，『現代日本語書き言葉均衡コーパス』(BCCWJ) のテキスト内の名詞句に対して情報構造に関わる文法情報のアノテーションを行った結果を報告する。

【口頭発表】3/8(水) 15:45-16:10

〔利用する言語資源〕『現代日本語書き言葉均衡コーパス (BCCWJ)』

[O-D-2]
読み時間と情報構造について（ちょっとながめ）

浅原正幸 (国語研)

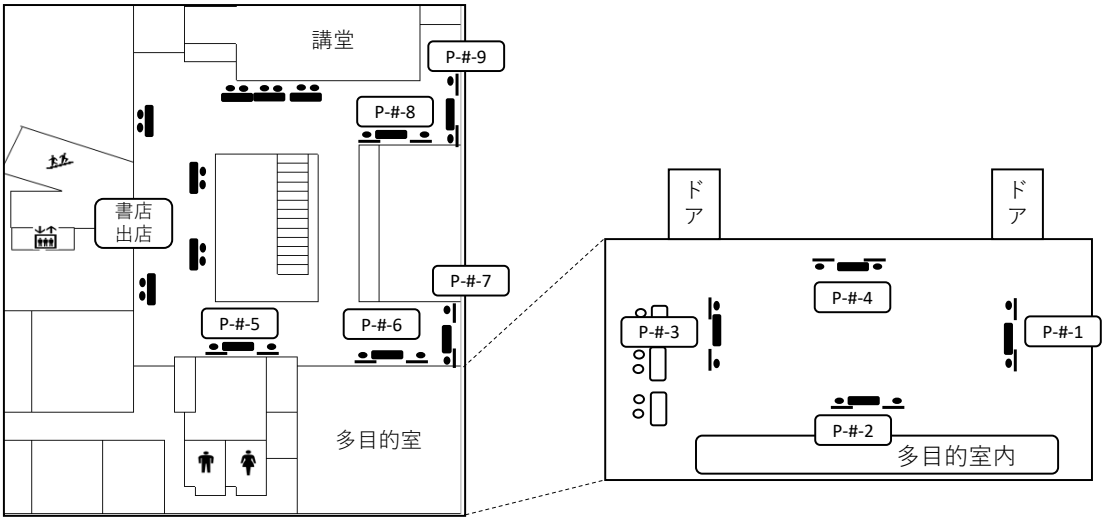
本研究では『現代日本語書き言葉均衡コーパス』に対して付与された，文の読み時間データ『BCCWJ-EyeTrack』と，名詞句の定性などの情報構造アノテーションデータの対照分析を行った。日本語母語話者 24 人分のデータを線形混合モデルにより分析した結果，特定性 (specificity)・有情性 (sentience)・共有性 (commonness) が文の読み時間に影響を与え，それぞれ異なったパターンの読み時間の遅延を引き起こすことがわかった。特に共有性においては新情報 (hearer-new)・想定可能 (bridging) が識別可能なレベルで異なった。このことは，ある名詞句が言語受容者にとって新情報なのか想定可能なのかを読み時間データから推定することができる可能性を示唆しており，文書要約のユーザ適応などの応用に利用することが期待できる。

【口頭発表】3/8(水) 16:10-16:35

〔利用する言語資源〕 『現代日本語書き言葉均衡コーパス (BCCWJ)』

Information

ポスター設営図・出店



- P-#-1, P-#-2, P-#-3, P-#-4 は電源が利用できます。
- 無線 LAN は eduroam をご利用ください。
- ポスター掲示用ついたてサイズは横 150cm× 縦 210cm です。
- ポスター掲示用資材は国語研で準備いたします。
- 書店各社の出店がございます。

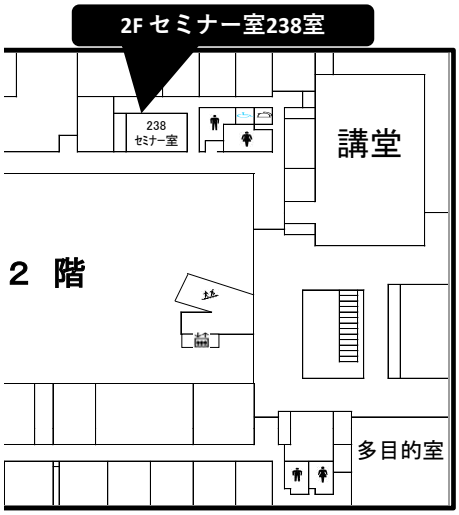
『国語研日本語ウェブコーパス』検索系『梵天』デモ

国立国語研究所コーパス開発センターは「超大規模コーパス」プロジェクト (2011-2015 年) の成果物として,『国語研日本語ウェブコーパス』とその検索系『梵天』を公開いたしました。

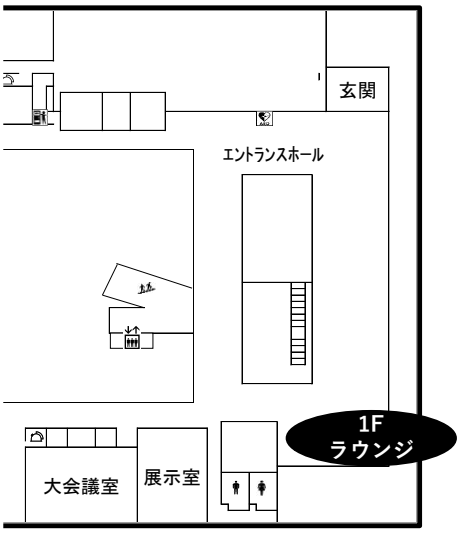
- 『国語研日本語ウェブコーパス』ウェブページ
  `http://pj.ninjal.ac.jp/corpus_center/nwjc/`
- 検索系『梵天』
  `http://bonten.ninjal.ac.jp/`

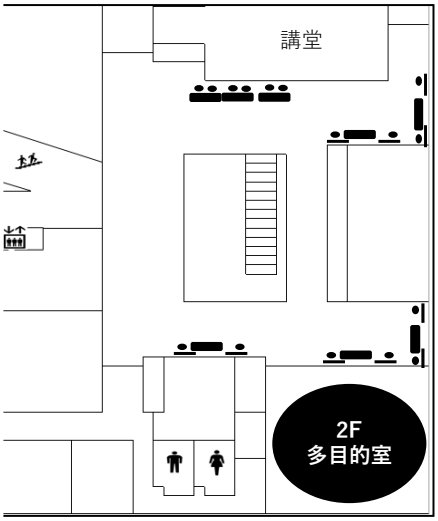
以下の時間帯に 2F セミナー室 238 室で『国語研日本語ウェブコーパス』の検索系『梵天』のデモを行います。

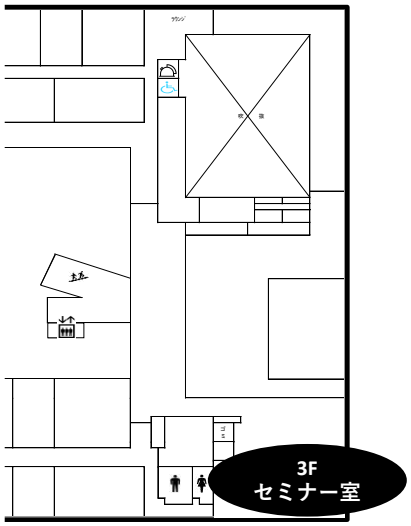
- 3 月 7 日 (火)　13:30-15:00
- 3 月 8 日 (木)　13:30-15:00



ランチスペース



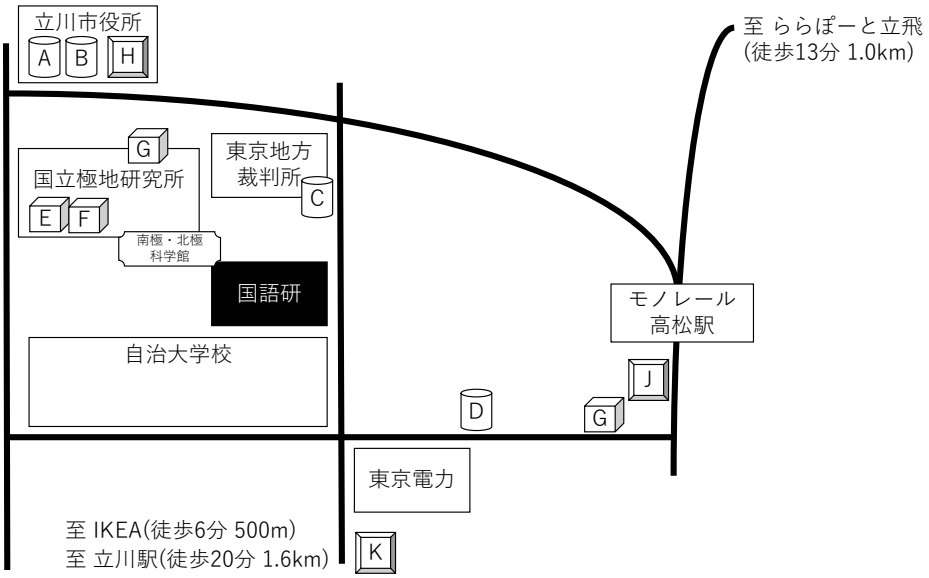




- **2F** 講堂内では飲食しないでください。
- **ゴミの分別にご協力をお願いいたします。**ゴミ箱に弁当ガラを捨てる際にはワリバシを別のごみ箱に分別して捨ててください。
- ランチスペースは 1F ラウンジ・2F 多目的室・3F セミナー室の 3 カ所です。
- 2F 多目的室はポスター発表会場を兼ねており，椅子が片付けられておりますが，適宜椅子・机を移動して食事していただいて結構です。移動した椅子は元通りに戻しておいてください。
-

ランチマップ

国語研近隣は，食事できる場所が限られております。



- お食事処
  - A 立川市役所食堂 (市役所内 3F) 11:00-15:00
    ラーメン・スパゲッティ 300 円程度，日替定食 500 円程度
  - B Café ハーモニー (市役所内 1F) 10:00-16:00
    トースト 300 円程度，ピラフ 600 円程度
  - C 東京地方裁判所食堂 (裁判所内 B1) 8:00-20:00
  - D 中華料理 瑞京 11:00-14:30, 17:00-23:00
- お弁当
  - E チコマート お弁当屋 (極地研内 1F) 11:00-13:00
    各種お弁当 300 円-500 円
  - F ハイジ カレー弁当 (極地研内 1F) 11:00-13:00
    各種 500 円
  - G 瑞京 中華弁当 (ワゴン 2 か所) 11:00-13:00
    日替わり 2 種 500 円
- コンビニエンスストア
  - H ポプラ (市役所内 1F)
  - J ミニストップ (モノレール高松駅南)
  - K セブンイレブン (東京電力南)

単独

彼の		迚も	弄る	れる			然う 然う	狸
まで	私				狐			或る
だ	否	猫	色んな		憂鬱			頑張る
		より	其れ		らしい	今晩は	黒ずむ	為
為る				仮令			だけ	
新しい	只管		ほど					せる
		同じ		遡る	やら		固より	其処
		無い	川獺		うぇい	所謂		
	を	ない			漸く			御早う

名詞　動詞　形容詞　副詞　連体詞　感動詞　代名詞　助動詞　助詞

空白に「品詞」を埋めましょう。
埋めるべき品詞は UniDic 品詞体系に基づく「名詞」（「名詞-普通名詞-副詞可能」「名詞-普通名詞-形状詞可能」を含む）・「動詞」・「形容詞」（「名詞-普通名詞-形状詞可能」を含む）・「副詞」（「名詞-普通名詞-副詞可能」を含む）・「連体詞」・「感動詞」・「代名詞」・「助動詞」・「助詞」の九つです。
語彙主義的な観点で「可能性に基づく品詞」の単語は，用法主義的な観点で品詞を一つ想定してください。

こちらの表は単語に対する品詞付与スペースとしてご利用ください

名詞　動詞　形容詞　副詞　連体詞　感動詞　代名詞　助動詞　助詞

言語資源活用ワークショップ 2016 Abstract 集

発行日：平成 29 年 2 月 28 日
発行者：国立国語研究所コーパス開発センター
連絡先：〒190-8561 東京都立川市緑町 10-2　　lrw@ninjal.ac.jp