日本語コーパスの包括的検索環境の実現に向けて

著者	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
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	170-179
発行年	2017
URL	http://doi.org/10.15084/00001472

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（国立国語研究所音声言語研究領域）†
浅原 正幸（国立国語研究所コーパス開発センター）
小木曽 智信（国立国語研究所言語変化研究領域）
小磯 花絵（国立国語研究所音声言語研究領域）
木部 暢子（国立国語研究所言語変異研究領域）
迫田 久美子(国立国語研究所日本語教育研究領域客員教授)

Toward the Realization of a Comprehensive Searching Environment for Japanese Corpora

Kikuo Maekawa , Masayuki Asahara, Toshinobu Ogiso, Hanae Koiso, Nobuko Kibe, and Kumiko Sakoda (NINJAL)

要旨　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。

１．はじめに

国立国語研究所が当時未開拓であった日本語言語資源の整備事業に着手したのは1990年代末であった。その後、一連の事業で開発した種々の日本語コーパスは、幸い、国内外において幅広い研究領域の研究者の支持を集めることとなり、現在では言語資源整備が国立国語研究所の中核的な事業のひとつとして社会的に認知されるに至っている。

しかしながら、これまでに公開してきた各種コーパスは、それぞれ独立に検索系が開発されており、複数のコーパスを横断的に検索することができない点に運用上の制約が認められる。現在、広く利用されているコーパス検索用ウェブアプリ『中納言』も検索対象のコーパスごとに異なるバージョンを提供している。検索ロジックはほぼ同一だが検索に利用できる情報の選択肢はコーパスごとに異なっている（2節参照）。

そこで、2016 年度から 2021 年度にわたる第３期中期計画期間におけるコーパス開発センターの目標設定に際して、この問題の解消を主要な活動目標として設定することにした。この目標を達成することで、時間的、地理的変異を含む日本語コーパスが出現し、研究所がこれまでに進めてきた日本語言語資源の整備事業を一端集大成することができると考えている。

以下、２節では包括的検索環境の対象とする予定の一連のコーパスの仕様を紹介する。その後、３節で仕様にどのような問題があるかを検討した後、４節で今後どのような課題を解決する必要があるかを検討し、現時点で考えられる対応策について論じる。

† kikuo@ninjal.ac.jp

２．対象となるコーパス群

２.１　公開済みのコーパス群

最初に既に構築が終了するか、ある程度まとまった規模に達していて、国立国語研究所コーパス開発センターから公開されているコーパス群を観点に紹介する。

２.１.１　『日本語話し言葉コーパス』(略称 CSJ)

現代の標準日本語話者の自発音声コーパスである（Maekawa et al. 2000, 小磯編 2015)。規模は短単位で 752 万語。時間にして 650 時間の音声を収録している。音声認識での利用（すなわち言語モデルと音響モデルの構築）を念頭において設計されているので、内容の95%は独話である。具体的には各種学会での口頭発表と日常的な話題についての一般的なスピーチ（模擬講演）が大部分を占める。残る 5%は、独話と比較するために対話音声と朗読音声に充てられている。

アノテーションとしては、各種のタグが付与された音声の転記テキスト（発音形と基本形の 2 種類、転記単位ごとの音声信号との時間アライメント情報を含む。4.2.1 参照)、短単位と長単位による二重形態論情報、節境界ラベル等を提供している。またコアと呼ばれるサブセット（50 万語、44 時間）に対しては、X-JToBI 方式による分節音・イントネーション情報、文節係り受け構造、談話境界情報なども提供されている。コアに含まれるサンプルの形態論情報は手作業で精度を向上させている。メタ情報として、講演種別の他に、話者の属性情報（性別、年代、出身地など）を提要している。

CSJ は 2004 年の一般公開以来、DVD（第 4 刷からは USB メモリ）で頒布されている。専用の検索系は公開していないが、2011 年には、コア部分のすべてのアノテーションが RDB (SQLite)で利用可能になり、DVD 版ユーザーには無償で提供されている。また 2016 年には、コーパス全体の短単位形態論情報が『中納言』（次節参照）で検索可能になった。現在は DVD 版のユーザーのみを対象とした試験公開であるが、近日中に一般公開(無償、要登録)も開始する予定である。

２.１.２『現代日本語書き言葉均衡コーパス』（略称 BCCWJ）

現代日本語の書き言葉を対象とした均衡コーパスで、規模は短単位で 1 億語である。書籍・雑誌・新聞・広報誌・ブログ・ネット掲示板・国会会議録・法律・詩歌など多様なレジスターから抽出されたサンプルから構成されており、サンプルはすべて著作権処理済みである（Maekawa et al. 2014, 山崎編 2014)。

アノテーションとして最も重要なのは、長短両単位による形態論情報である。コア（100 万語）に含まれるサンプルの形態論情報は精度が高い。他に、文字・表記に関するタグと文書構造に関するタグも提供されている。前者にはルビ文字列、原文の誤表記、外字などの情報が、後者には「記事＞クラスター＞段落＞文」のような文書の階層構造、図表、引用、注記などの情報が含まれるが、提供されるタグの範囲はレジスターによる異動がある。メタ情報として豊富な書誌情報を提供しているのも特徴である。筆者属性のほか、原本のタイトル、巻号、出版社、出版者、ISBN、サンプル抽出位置などが提供されている。

2010 年以来、DVD 版で全データを頒布しているが、他にウェブ上で 2 種類の検索系を無償公開している。『少納言』ではユーザー登録なしに全テキストの文字列検索が可能であり、正規表現も部分的に利用できる。検索結果は書誌情報の一部とともに表示される。1 検索に対するヒット数が 500 を超える場合は、全検索結果から無作為抽出された 500 サンプルだけが画面に表示される。

『中納言』は形態論情報を検索するためのウェブインターフェースで、短単位ないし長単位の N グラム（N は 11 まで）を検索できるコンコーダンサーである。形態論情報としては、表層の文字列（書字形）の他に、語彙素(lemma)、語彙素読み、品詞（3 階層)、活用形、活用型などを指定できる。『中納言』では検索結果を上限 20 万サンプルまでダウンロードできるので、著作権保護の観点から、利用者登録をお願いしている。登録・利用は原則無償

である。
BCCWJ の公開後に作成され、公開されたアノテーション情報もある（関連 URL 参照）。文節係り受けアノテーション情報は、1 億語全体を自動解析したデータが提供されている。他に、述語項構造、述語項構造シソーラス、日本語フレームネット、時間情報・時間的順序関係、文体指標、節境界、拡張固有表現、単語係り受け構造、「れる・られる」の用法などのアノテーションが、コーパスの一部に対して提供されている。

２．１．３ 『太陽コーパス』

明治後期から大正期(1895～1925 年)の有名な総合雑誌『太陽』（博文館）から 5 年分を抽出した全文コーパスである（国立国語研究所編 2005)。2005 年の公開時には、タグ付きテキストコーパスとして頒布され、形態論情報は付与されていなかった。しかし、その後、近代語の自動形態素解析技術が実用に達したので、2016 年には短単位解析結果がウェブ上で公開された。検索系は『現代日本語書き言葉均衡コーパス』の項で紹介した『中納言』である。規模は短単位で 1100 万語（文字数で 1450 万語）である。今後は、同じく近代語を対象とした雑誌コーパス群（『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』とともに、後述する『日本語歴史コーパス』「明治・大正編 I 雑誌」の一部を構成することになる。

２．１．４ 『日本語歴史コーパス』（略称 CHJ）

上代（奈良時代）から近代（明治・大正時代）までの日本語の歴史を通時的に研究するためのコーパスである（小木曽 2016)。2012 年より構築済みの部分から公開を開始し、現在では「平安時代編」（仮名文学 16 作品、約 86 万短単位）・鎌倉時代編 I 説話・随筆（5 作品、約 71 万短単位)、室町時代編 I 狂言（虎明本狂言集、約 24 万短単位)、明治・大正編 I 雑誌（上述の雑誌、約 1254 万短単位）が公開されている。BCCWJ と同様に短単位と長単位の二つの単位で形態論情報を付与しているが、現在のところ近世（江戸自体）以降のデータについては短単位のみである。残された貴重な資料を活用するため、「鎌倉時代編」の『今昔物語集（本朝部)』の一部と「明治・大正編」の雑誌の一部を除き、全体に人手による修正を施している。
検索インターフェースとして BCCWJ と共通の『中納言』によって公開を行っている。検索結果の各行から、外部のサービスにリンクがはられており、各作品の本文や原文の画像データなどが確認できるようになっている。たとえば、小学館の『新編日本古典文学全集』を底本とする作品はジャパンナレッジで公開されている当該ページにリンクがあり、本文・注釈・現代語訳を参照することができるほか、『今昔物語集』や近代雑誌では、原文の画像データが確認できる。

２．１．５『国語研日本語ウェブコーパス』（略称 NWJC）

BCCWJ の量的不足を補うためにウェブ上の日本語を母集団として構築された短単位 253 億語規模のウェブコーパスである。クローリング技術によって、約 1 億 URL の日本語ウェブページを繰り返し収集することで安定してアクセス可能なウェブページを決定した。公開データは、2014 年 10-12 月期に収集したデータである。
NWJC ではウェブ言語データの深刻な問題であるコピーサイトの問題を軽減するために、文単位の重複性排除を行っている。文単位の異なりを取ることによる文型パターンとしてのデータベース化を行っている。
現在提供されている形態論情報は UniDic 体系の短単位形態論情報のみである。また自動解析の結果をそのまま提供しており、CSJ や BCCWJ のように手作業で修正したサブセット（コア）は NWJC には設定されていない
NWJC の特徴として、データ全体に文節係り受け構造自動解析結果が提供されている。
NWJC は、ウェブ上の新しい検索系である『梵天』を用いて検索する。『梵天』には、文

字列検索、品詞列検索（形態論情報検索＝『中納言』の短単位検索と同等）にくわえて、係り受け検索の機能も実装されている。検索系『梵天』の実装においては 1 文中に同一語彙素が 2 回以上出現する場合、ヒットするのは最左用例だけという制約がある。ただし正確な頻度情報を必要とするユーザーには別途作成した語彙表を提供する。文字列検索機能は、前述の『少納言』と同様、ユーザー登録なしで一般公開されているが、システムに高い負荷がかかる品詞列検索ないし係り受け検索を行うユーザーには、事前に講習会を受講してもらっている。登録・利用は無償である。

２.１.６『多言語母語の日本語学習者横断コーパス』（略称 I-JAS）

I-JAS（International Corpus of Japanese as a Second Language）は日本語学習者のコーパスである。12 の異なった言語（英語、中国語、韓国語、インドネシア語、ロシア語、タイ語、フランス語、スペイン語、ドイツ語、ハンガリー語、ベトナム語、トルコ語）を母語とする日本語学習者の発話と作文のデータであり、海外の学習者と日本国内の学習者の両方が対象となっている。

これまでに公開されている日本語学習者コーパスの問題点をふまえて設計されており、多面的な角度から利用できるように、さまざまなタスクのデータと日本語母語話者のデータが含まれている。

2016 年 5 月、第一次データとして 12 言語を母語とする海外日本語学習者、国内の教室環境学習者と自然環境学習者各 15 名、日本語母語話者 15 名の計 225 名分を公開した。最終的には 2020 年春までに学習者 1000 名、日本語母語話者 50 名の計 1050 名分のデータ公開を目指している。

I-JAS の特徴としては、次の 5 点が挙げられる。(1) 既存のコーパスに比べ、規模が大きいこと。(2) データ内容が豊富であること。発話には 4 種類（ロールプレイ・ストーリーテリング・絵描写・対話）のタスク、作文は 3 種類（ストーリーライティング・メール・エッセイ）のタスクが設定されている。(3) 学習者全員が共通の日本語能力テストを受け、その評点が明示されていること。そのため、地域や機関が異なっていてもレベルの基準が統一され、比較が可能となる。(4) 多面的な利用が可能なデータ形式であること。発話データは書き起こしを行い、テキストだけでなく、検索システムの利用が可能である。また、発話の音声データも公開している。(5) 学習者の背景情報があること。全ての学習者の言語環境や学習歴などの情報が含まれている。

２.１.７『名大会話コーパス』

約 100 時間分の雑談を文字化したコーパスであり、姫路獨協大学（構築当時は名古屋大学）の大曾美恵子氏が構築されたコーパスである。一時期、国立国語研究所から『日本語自然会話書き起こしコーパス』の名称で公開されていたが、2016 年に短単位形態論情報を付与したデータを『中納言』で公開するにあたり、旧名称を復活させた。規模は短単位で 114 万語（句読点などを除く）である。

２.２　構築中のコーパス群

次に、現在構築作業が進行中のコーパス群を紹介する。前節に紹介したコーパスの拡張作業も含まれる。

２.２.１『日本語諸方言コーパス』(略称 CJD)

日本語諸方言の自然談話のコーパスである。北海道から沖縄まで全国の自然談話が横断的に検索でき、あわせて音声とテキストのダウンロードができる形で公開する。

資料としては、1977～1985 年に文化庁が行った「各地方言収集緊急調査」のデータを使用する。全体は、全都道府県 224 地点、1 地点につき 30 時間程度の談話録音テープよりなる資料で、内容は当時 60 歳以上の地元出身者数人による自然談話である。一部は 『全国方

言談話データベース 日本のふるさとことば集成』（国書刊行会）として音声、テキスト、標準語訳が公開されているが、多くは未公開の状態。本コーパスでは公開分、未公開分を合わせて、2021 年度までに最低 75 時間分のデータを公開する。

本コーパスの特徴は、諸方言の談話を標準語で検索し、それに対応する方言形とそれを含む一定の発話単位を横断的に検索する点にある。言うまでもなく、方言と標準語は 1 対 1 で対応しない。そのため、標準語での方言検索には対応のずれといった問題が生じる。しかし、方言形で検索システムを構築するには、各地方言の形態素辞書を作らなければならず、それには膨大な時間と労力が必要となる。それに対し、標準語での検索には、すでにある日本語形態素解析用辞書を利用することができ、しかも、方言間のゆるやかな横断検索が可能となる。また、諸方言コーパスがどのように利用されるかを考えてみると、標準語での検索システムは必須のように思われる。したがって、標準語による検索方法をとることとした。

本コーパスの構築に向けて、現在、次のような手順で作業を進めている。①方言音声の転記テキスト（方言テキスト）のチェック。②発話単位の認定。③方言テキストに対する時間アライメント情報の付与。④方言テキストに対応する標準語テキストのチェック。①の方言テキストと④の標準語テキストは、文化庁の事業の際にすでに作成されたものがあり、これをもとにして、チェック作業を進めているが、標準語テキストについては、全面的な見直しが必要である。前述のように、方言と標準語は 1 対 1 で対応するわけではないので、標準語テキストによっては、本来、検出されるべき方言形が検出されなかったり、検索結果が変わってきたりする可能性があるためである。標準語テキストの付け方については、作業の過程で詳細なマニュアルを作成しており、それもあわせて公開する予定である。②の発話単位の認定については、基本的に 0.2 秒の無音という基準で発話単位を認定しているが（ただし 3.2 および 4.2.1 も参照）、それに加え、話者同士の発話の重なりや相づち、フィラー、間投詞等に対し各種タグ付けをして公開する。

２.２.２『日本語日常会話コーパス』(略称 CEJC)

現代日本語の日常会話を対象とするコーパスで、規模は 200 時間（推定 200 万語相当）を目指す（小磯ほか 2017）。本コーパスは機関拠点型基幹研究プロジェクト「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」（平成 28～33 年度、リーダー：小磯花絵）において現在構築中のものであり、平成 33 年度末の公開を予定している。

本コーパスが対象とするのは、収録のために集められた状況での会話ではなく、日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話である。日常会話の幅広いレジスターをカバーするようサンプルを選ぶには、私たちが普段、どのような種類の会話をどの程度行っているかを把握する必要がある。そこで予備研究として約 250 人を対象とする会話行動調査を実施し、その結果を参考にしながら、多様な種類の会話をバランス良く納めたコーパスを構築する（小磯ほか 2016）。

会話データは、性別・年代の点から均衡性を考慮して選別された協力者 40～50 人に収録機材等を 2～3 ヶ月ほど貸し出し、協力者自身に日常会話 15～18 時間程度を収録してもらう方法を中心に集める。この中から、多様な種類・場面の会話となるよう、1 協力者あたり 4～5 時間を選別してコーパスに格納する。

アノテーションとしては、各種のタグが付与された音声の転記テキスト（転記単位ごとの音声信号との時間アライメント情報を含む）に加え、発話単位情報、短単位と長単位による二重形態論情報、文節係り受け情報などを付与する予定である。またコア部分（約 20 時間）に対しては、国際標準化規格 ISO24617-2 に基づき日常会話用に整備した談話行為情報や、CSJ に付与されている X-JToBI を簡略化した方式に基づくイントネーション情報を付与する予定である。コアに含まれるサンプルの形態論情報・係り受け情報は手作業で精度を向上させる。メタ情報として、話者の属性情報（性別、年代、出身地、相手との関係性など）や会話の属性情報（会話の場面、形式、人数など）を提要する。

２.２.３『日本語歴史コーパス』の拡張

『日本語歴史コーパス』は、この 3 月に「鎌倉時代編 II 日記・紀行」として、『とはずがたり』や『海道記』など 5 作品の追加公開を予定している。来年度以降は、すでに試行版を公開中の「江戸時代編」の洒落本・人情本を拡充して公開するほか、「奈良時代編 I 万葉集」「室町時代編 II キリシタン資料」の公開を予定している（いずれも 2017 年度予定)。『万葉集』やキリシタン資料では、万葉仮名やローマ字で書かれた原文と漢字仮名交じり本文とのアライメントをとり（４.２.２参照)、当該部分の原表記を確認できるようにする予定である。さらに、続日本紀宣命（奈良時代編)、近松の世話物浄瑠璃（江戸時代編)、国定読本などの教科書や近代文学作品（明治・大正編)、和歌集などのコーパス化を行い、上代から近代までの日本語を通時的に研究することのできるコーパスとする計画である。

２.２.４『多言語母語の日本語学習者横断コーパス』の拡張

2017 年春に第二次データとして、韓国語、中国語、英語、トルコ語の海外日本語学習者各 35 名、国内の環境別日本語学習者を各 25 名、日本語母語話者を 35 名、合計 225 名のデータを追加公開する予定である。これにより、日本語と言語的な類似点の多い韓国語とトルコ語、類似点の少ない中国語と英語のデータが各 50 名分となり、母語と学習者の日本語レベルの観点からの分析も容易になる。

また、韓国語、中国語、英語、フランス語については、各言語の母語話者同士での発話データの収集を計画している。I-JAS の日本語学習者に実施したタスクのうち、ロールプレイとストーリーテリング、メールに関して、母語のデータと比較することによって、母語と学習者言語（日本語）でのコミュニケーション上の問題についての研究が可能となる。

３．仕様の問題点

上に紹介したコーパス群を主にアノテーション仕様の観点から比較することで、問題の所在を明らかにすることを試みた。

３.１　形態論情報

現時点ですべての対象コーパスに付与されているという意味で、最も基本的なアノテーションは、短単位形態論情報である。日本語の膠着語的性格を考えると長単位での解析も行われていることが望ましいが、現在、両単位による二重解析が施されているのは CSJ と BCCWJ にとどまる。

同じ短単位と言っても、コーパスの開発時期によって、細部で仕様が異なることがある。CSJ と BCCWJ の間にも無視しえない差が生じていたが、現在、『中納言』でオンライン公開されている CSJ の短単位情報は、BCCWJ の規定に沿う形で統一が図られている。

また、そもそも形態素解析作業で何が解析されるかも対象コーパスによって異なる。CJD(『日本語諸方言コーパス』)の場合、形態素解析が施されるのは標準語テキストであり、方言テキストは解析対象ではない（ただし 4.3 も参照)。検索系はヒットした標準語テキストに対応する方言テキスト（と音声）を出力する。I-JAS の場合、形態素解析されるのは、いわゆる誤用を修正された日本語テキストであり、検索系は誤用を含む転記テキストを出力する（例えば、I-JAS で語彙素「経営」を検索すると、通常の「経営」以外に「けえ」「けいえ」「けいえん」「けいいん」などと転記されたサンプルが表示され、画面上にはそれらが「経営」を意図した発話であることがタグで表示される)。

３.２　発話単位

CSJ、I-JAS、CJD、名大会話コーパスなど、自発的な話し言葉をあつかうコーパスでは、発話単位をどう認定するかが問題になる。現在はコーパスごとにバラバラの状態にある。これをある程度まで企画して統一できるかどうかは緊急性の高い検討課題である。また独話と対話でも認定基準が異なってくる可能性があり、その点の検討も必要である。

CSJ のように物理的な基準（0.2 秒以上のポーズで区切るのが原則）に依拠すれば、作業基準は明確になるが、言語学的な意味づけは時に困難になる。

発話単位は、検索結果の音声再生の単位となることが予想される（4.2.1 参照）。その観点からは、極端に長いもしくは短い単位が頻出することは避けたいという要請もある。

３.３　タグ

話し言葉の転記テキストには、様々なタグが埋め込まれることが多い。対象コーパスのなかでは、CSJ と I-JAS で多数のタグセットが利用されており、CJD にも今後種々のタグが付与される予定である。

CSJ と I-JAS のタグは目的がかなり異なっており、前者では転記の正確性を高めることに主眼が置かれているのに対し、後者では、学習者のいわゆる「誤用」を含む日本語を修正して、形態素解析可能なテキストに整形することが主要な目的となっている（本ワークショップにおける西川の発表参照）。

書き言葉コーパスのテキストにもタグが付与されている。BCCWJ のタグには先に 2.1.2 節で触れた。CHJ のテキストには、本文校訂の情報の他、会話などの本文の種別、話者の情報などが付与されているが、両コーパス間でのタグの共通性は低い。CHJ ではタグの一部が『中納言』での検索対象の絞り込みに利用されているが、BCCWJ では現在そのような機能は提供されていない。今後、CHJ が万葉仮名やローマ字の文献を扱うようになれば、CHJ のタグはさらに増加するものと予想される。

３.４ その他のアノテーション

係り受け構造アノテーションは、CSJ と NWJC にだけ付与されている。両者の仕様には相違がある。韻律に関する X-JToBI アノテーションは CSJ のコアにだけ付与されている。I-JAS には上昇イントネーションを示すタグがある。CEJC にも句末イントネーションの機能に関するタグが付与される予定であり、またサブセット（コア）に対して X-JToBI 的なラベリングを施す予定がある（2.2.2 参照）。

４．議論

４.１ 包括的検索系と個別検索系

われわれは今後、上に述べたように、仕様に種々の異同をもつコーパス群を対象とした包括的検索環境を設計することになる。その際、もっとも基本的な決定のひとつは、新しい包括的検索系と従来から存在する個別検索系（『中納言』や『梵天』）の関係であり、ことに、包括的検索系が検索対象のデータをどのような方式で保持するかという問題である。

ひとつの方式は、既存対象コーパスのデータは適宜修正し、今後構築する対象コーパスのデータは当初から統一を図って、包括的検索系独自のデータを構築し、それを検索対象にするというものである。この場合、完成後に、従来の個別検索系（現在公開されている様々な『中納言』や『梵天』）をどうするかという問題が生じる。今後慎重な検討を要する問題であるが、ここでひとつの方針を述べるならば、包括的検索系において、各対象コーパスおよび対応する個別検索系の仕様の相違を十分に吸収できない場合は、少なくとも一定期間、包括的検索系と個別検索系とを併存させることにしたい。

もうひとつの方式として、包括的検索系を個別検索系に対するラッパー(wrapper)として設計する可能性も考えられる。包括的検索系は、個別検索系に対する検索リスエストを発行して、その結果を受けとり、それを適宜整形して出力するという形のシステムである。この場合、実際に検索を実施するのは個別検索系であるから、当然それらを維持しつづけることになり、結果として、検索系を再開発する困難を回避することができる。反対に、この方式の問題点としては、既存システムの出力を完全に整形して統一することが、おそらく非常に困難であろうことが挙げられる。

４.２　コーパス開発に対する技術的支援

対象コーパス群のうち、現在構築中のコーパスについては、構築作業を効率化する必要がある。コーパス開発センターでは、現在以下の二点について重点的に技術支援を行っている。

４.２.１ 音声-テキスト間アライメント

話し言葉コーパスのうち、音声ファイルを持ち、検索系を通して音声を再生しようとするコーパスでは、形態論情報等の検索対象となる転記テキストと、音声信号との時間アライメント（対応づけ）をとる必要がある。その可能性をもつのは、現状では、CSJ、I-JAS に加えて、CJD と CEJC である（名大会話コーパスは音声ファイルが公開されていない)。

既存コーパスの中でこの情報をもっとも組織的に付与できているのは CSJ であるが、CSJ の開発においては、音声・テキストアライメントは転記テキスト作成作業の一環として人手で実施した。当然、高いコストが発生した。

近年、音声認識技術の飛躍的な発展によって、この作業を全自動で実施する可能性が現実のものとなりはじめている。特に標準語音声が対象となるコーパスにおいては、自動アライメント技術を活用できる可能性が高い。一方、方言音声や学習者音声を、既存の技術でどこまで処理できるかは今後の検討を要する問題であり、基礎研究の課題である（本ワークショップにおける石本の発表参照)。

アライメントの単位は短単位などにも設定できるが、あまり短い単位を音声再生しても、知覚が困難になることも多いので、3.2 で論じた発話単位がひとつの現実的な候補であると考えられる。

４.２.２　二テキスト間アライメント

CJD では、標準語テキストを検索して、ヒットした短単位を含む標準語テキストに対応する方言テキストを出力する。そのためには、標準語テキストと方言テキストのアライメントが必要になる。方言コーパスは、その出発点となった文化庁のデータ（2.2.1 参照）が対訳形式で作成されているので、一応のアライメントはとれているのだが、今後発話単位の認定（３.２参照）などで現在のテキストを改めた場合には、アライメントの再実施が必要になる。

もうひとつ、テキストアライメント技術の応用が期待されるのが CHJ である。古典本文と現代語訳の対応がとれれば、CHJ の応用可能性が高まると期待される。また万葉仮名で書かれた文献や漢文文献の場合は、万葉仮名ないし漢文と読み下し文のアライメントが必須である。

４.３　共同研究

上記ふたつのアライメント処理技術は、音声認識や自然言語処理の研究のなかで発展してきた技術であり、現在もそこで最先端の技術が開発されつつある。その成果を効率的に取り込むために、2016 年 10 月から研究所外の研究者との共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」を実施している。

この共同研究では、アライメント処理技術以外に、意味処理技術、音声特徴自動抽出技術、教師無し言語解析技術などの共同研究を実施するか実施予定である。その成果は、各種コーパスに対する意味情報（分類語彙表番号）の付与（本ワークショップにおける加藤らの発表参照）や、音声のピッチ情報の抽出精度向上、声質(voice-quality)情報の自動付与（本ワークショップにおける森らの発表参照)、韻律情報アノテーションの部分的自動化、方言テキストの形態素解析などの形で、コーパス開発へのフィードバックを試みる予定である。

５．まとめ

本稿では、国立国語研究所コーパス開発センターで構築を進めている、複数の日本語コー

パスを包括的に検索可能なシステムについて、検索対象となる予定のコーパス群の仕様を手短に紹介するとともに、包括的検索システムの設計に関わる様々な問題を探索し、検討した。今後は、方向で指摘した個別的な問題群について技術開発の現状を報告すると同時に、全般的な開発状況についても、適宜報告していく予定である。

謝　辞

本稿は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」ですが、同時に、筆者らが過去 2 年間自主的に開催してきた SMOKKA 研究会の成果も反映されています。同研究会の参加者・発表者に深く感謝します。

文　献

Masayuki Asahara, Kazuya Kawahara, Yuya Takei, Hideto Masuoka, Yasuko Ohba, Yuki Torii, Toru Morii, Yuki Tanaka, Kikuo Maekawa, Sachi Kato and Hikari Konishi (2016). "'BonTen' Corpus Concordance System for "NINJAL Web Japanese Corpus"', Proceedings. of COLING-2016. Demo Session. (To Appear).

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato and Hikari Konishi (2014). "Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan", *Alexandria*, 25:1-2, pp.129-148.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). "Spontaneous Speech Corpus of Japanese", In Proceedings of LREC-2000 (Second International Conference on Language Resources and Evaluation), Vol. 2, pp.947-952.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Makiro Tanaka, and Yasuharu Den (2014). "Balanced Corpus of Contemporary Written Japanese", *Language Resources and Evaluation*, 48, pp.345-371.

小磯花絵 編（2015）『話し言葉コーパス:設計と構築』（講座日本語コーパス第 3 巻）朝倉書店.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・伝康晴(2016).「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」『国立国語研究所論集』Vol.10, pp.85-106.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉(2017).「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会予稿集』

小木曽智信（2016).「『日本語歴史コーパス』の現状と展望」國語と國文學 93 (5), pp.72-85.

国立国語研究所編(2005)『雑誌「太陽」による確立期現代語の研究―「太陽コーパス」研究論文集―』（国立国語研究所報告 122）博文館新社刊.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016).「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』6:3, pp.93-110.

山崎誠(編)(2014)『書き言葉コーパス:設計と構築』（講座日本語コーパス第 2 巻）朝倉書店.

関連 URL

『日本語話し言葉コーパス』 http://pj.ninjal.ac.jp/corpus_center/csj/
『現代日本語書き言葉均衡コーパス』 http://pj.ninjal.ac.jp/corpus_center/bccwj/
『現代日本語書き言葉均衡コーパス』アノテーションデータ http://pj.ninjal.ac.jp/corpus_center/anno
『太陽コーパス』 http://pj.ninjal.ac.jp/corpus_center/cmj/taiyou/
『日本語歴史コーパス』 http://pj.ninjal.ac.jp/corpus_center/chj/
『国語研日本語ウェブコーパス』 http://pj.ninjal.ac.jp/corpus_center/nwjc/
『多言語母語の日本語学習者横断コーパス』https://ninjal-sakoda.sakura.ne.jp/lsaj/?cat=3
『日本語日常会話コーパス』https://www.ninjal.ac.jp/research/project-3/institute/spoken-language/
『少納言』(コーパス検索アプリケーション) http://www.kotonoha.gr.jp/shonagon/
『中納言』(コーパス検索アプリケーション) https://chunagon.ninjal.ac.jp/
『梵天』(『国語研日本語ウェブコーパス』検索系) http://bonten.ninjal.ac.jp/