様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 6 月 11 日現在

研究種目：基盤研究(B)
研究期間：2007～2009
課題番号：19320077
研究課題名（和文） 「非母語話者による日本語話し言葉コーパス」構築の更なる充実と分析・研究
研究課題名（英文） Research Analysis and Further Advancement of “The Corpus of Spontaneous Japanese Spoken by Non-native Speakers” Project
研究代表者
土岐 哲（TOKI SATOSHI）
京都外国語大学・外国語学部・教授
研究者番号：10138662

研究成果の概要（和文）：本課題は、『非母語話者による日本語話し言葉コーパス(CSJ-NNS)』に収録するデータをさらに充実させ、構築されたコーパスを広く一般に公開する目的で行われた。音声データの収録や各種アノテーションの付与のみならず、個人情報を消去するなどして公開に向けて作業を行い、コーパスを頒布した。また、本コーパスと『日本語話し言葉コーパス(CSJ)』とを比較することにより、従来の日本語非母語話者音声研究に新たな視座を提示した。

研究成果の概要（英文）：The aim of this research project is to build upon and consolidate the data that was previously compiled in “The Corpus of Spontaneous Japanese Spoken by Non-native Speakers (CSJ-NNS)” Project, for the purpose of public distribution. In order to protect the privacy of individuals, their personal information and identity has not been disclosed to the public. In addition, from the comparison with the CSJ, these findings present new research possibilities regarding non-native Japanese speech analysis.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	4,500,000	1,350,000	5,850,000
2008 年度	6,100,000	1,830,000	7,930,000
2009 年度	3,000,000	900,000	3,900,000
年度			
年度			
総　計	13,600,000	4,080,000	17,680,000

研究分野：日本語教育学、音声学
科研費の分科・細目：言語学・日本語教育
キーワード：(1)日本語教育、(2)音声学、(3)コーパス、(4)日本語非母語話者、(5)話し言葉、(6)モノローグ、(7)日本語話し言葉コーパス

１．研究開始当初の背景

非母語話者が産出した資料を基にして作られたコーパスは「学習者コーパス(Learner Corpus)」と呼ばれ、中間言語研究や言語教育への応用が期待されている（大曽ほか 2003)。

母語話者のコーパスと同様、学習者コーパスも英語が先駆的役割を果たしているが、日本語学習者コーパスも整備されつつある。研究開始の時点で入手可能であった主な日本語学習者コーパスは、以下の通りである。

①「日本語学習者による日本語作文と、その母語訳との対訳データベース ver.2」
②「日本語学習者コーパス(日本語学習者の作文コーパス)」
③「KY コーパス」
④「インタビュー形式による日本語会話データベース」

上記①・②は書き言葉の学習者コーパス、③・④は話し言葉の学習者コーパスである。ただ、研究開始時においても話し言葉学習者コーパスが存在していたとはいえ、それらは書き起こしテキストを中心として整備されたものでであった。したがって、日本語学習者の話し言葉における特徴的な文法事項や談話構造の分析などは可能であったが、日本語学習者の話し言葉音声を音声学的に分析することは不可能な状態であった。そこで、我々は音声学的分析にも耐えうる日本語学習者コーパスを構築することを計画し、平成17 年度から 18 年度にかけて『非母語話者による日本語話し言葉コーパス』(以下、英略称 CSJ-NNS＝Corpus of Spontaneous Japanese Spoken by Non-native Speakers) の構築作業を行った(「『非母語話者による日本語話し言葉コーパス』の構築と分析・研究」基盤研究 (A)　課題番号：17202011　研究代表者：土岐哲)。

本課題は、平成 17 年度から平成 18 年度にかけて構築した CSJ-NNS をより充実させるため、CSJ-NNS 構築時に得られた国内外の人的ネットワークを活用して、さらに広範なデータを収集すると同時に、データを拡充した CSJ-NNS の分析を行うものであった。

２．研究の目的

(1) 本研究の目的

本研究の主たる目的は、以下の四つである。

①　既に構築した日本語非母語話者（以下「非母語話者」）の産出した自発発話音声のコーパス CSJ-NNS を、更に拡充させる。

②　①のコーパスを使用し、非母語話者が産出した音声について、日本語母語話者（以下「母語話者」）自身の産出した音声と詳細に比較対照し、より広い視野に立って具体的に検証する。

③　①・②で得られた成果を、日本語音声教育の現場に、より現実的な目標および方法を考察し、提示する。

④　③の成果の一部は、教育現場のみならず、広く日本語社会一般にも発信し、非母語話者の音声についてのより柔軟かつ公平な理解を求める。

(2) 本研究課題が目指したもの

「学習者コーパスは、母語話者コーパスとの比較をすることで、更に価値を増す」(大曽ほか 2003) との指摘もあるように、学習者コーパスは母語話者コーパスと比較対照可能であることが望まれる。正確な比較対照のためには、母語話者・非母語話者とも同じ条件でのコーパスである必要がある。そこで我々が構築、データの拡充を目指した学習者コーパスは、基本的に『日本語話し言葉コーパス(以下、英略称 CSJ＝Corpus of Spontaneous Japanese)』に準拠することにした。そのことにより、非母語話者の話し言葉コーパスである CSJ-NNS と母語話者の話し言葉コーパスである CSJ とが、話者以外同条件での比較対照が可能となる。

(3) 日本語学習者音声の実態把握のための基礎的資料の構築

非母語話者の産出する日本語音声は、それぞれの母語背景などが影響し、母語話者が産出するそれとは異なる特徴を見せることが従来から指摘されている。それは、「分節音の不自然さ」・「発話のリズムの不自然さ」・「アクセント実現の不自然さ」・「イントネーションの不自然さ」など、多岐にわたる。しかし、それらの「不自然さ」の中には、非母語話者の自発発話音声のみならず、現に母語話者の自発発話にも観察されるものもある。それにも関わらず、非母語話者の日本語音声のみが「不自然である」かのように問題視されることが多い。その実態を明らかにするためには、まず、「問題」の現象がどのような場合に観察されるか、母語話者・非母語話者の双方の場合について、大規模データをもとに詳細な分析・比較対照を実施する必要がある。

もし、母語話者の音声コーパスである CSJ に準拠した非母語話者の音声コーパスが存在すれば、非母語話者の音声について、大規模データを基に種々の現象を実証的に分析することが可能になることは勿論、ほぼ同一条件で、母語話者の音声と非母語話者の音声を比較対照することが可能となる。その結果、従来の、少人数の非母語話者のデータを基に、特定事象のみを対象とした研究では、単に「学習者の偏り」、「中間言語的現象」、「誤用」とされてきたものが、本研究によって見直しを迫られる知見も出てくる可能性が高い。そのような知見は一部の関連先行研究でも指摘されているものの、本研究によって得られる新知見は、従来の日本語音声教育のあり方、ひいては非母語話者の日本語音声に対する日本語社会のより冷静な理解などへの貢献も大いに期待できる。

以上のように、CSJ-NNS 自体の分析は勿論、同一言語項目について、CSJ-NNS での分析結果と CSJ での分析結果とを比較することにより、中間言語(学習者の産出する言語)の研究の発展、音声教育・話し言葉教育への新知見の還元が期待できる。

３．研究の方法

(1)音声収録

データは国内外（日本各地・ヨーロッパ各地）の学会等において非母語話者が日本語で発表する場合と、国内外の非母語話者による日本語スピーチ大会等において収録した。

収録に際しては、講演を母語話者が聞き、多角的な印象評定を行った。話者にはヘッドセットマイクを着用してもらった。これは、講演中の話者が姿勢や視線を動かした場合でも唇とマイクの距離を一定に保つためである。マイクからの音声はDAT（Digital Audio Tape）に録音した。

(2) DAT(Digital Audio Tape)へのダビング

本研究では、音声データの収集・処理が非常に大切である。何らかの事件、事故によりそのデータが失われることのないよう十分に注意した。

(3)音声のファイル化、及び非言語音の記述

データ処理の段階で、DATの音声をダウンサンプリングして計算機に格納した。

(4)ポーズに基づく基本単位への分割

書き起こし作業の基本となる単位（基本単位）の開始・終了時刻を同定し、音声との対応をとる。基本単位として、文などの単位が利用されることもあるが、CSJ-NNSが対象とする自発音声ではその認定が困難であるため、200ミリ秒以上のポーズにはさまれた時間区分を基本単位とした。認定作業では、計算機上に音声波形とサウンドスペクトログラムを表示し、音を聴取しながら基本単位の始端、終端位置をラベリング用ウィンドーに記録していく。講演者の発話（言語音）に加えて、非言語音（笑いや咳など）や、雑音についても同様の基準で基本単位を認定し、その開始・終了位置を同定した。

(5)収録音声の書き起こしテキスト作成

基本単位認定作業で得られた時間情報をもとに音声を文字化した。この作業では計算機上で繰り返し音声を聞き、漢字かな混じりで表記する「基本形」、発音を忠実に表記した「発音形」からなる「書き起こしテキスト」を作成した（図1参照）。

```
0005  00017.055  ー  00017.897  L:
気息の                &キソクノ
度合い                &ドアイ
0006  00018.312  ー  00019.155  L:
いわば                &イワバ
有気か                &(W ユキ;ユーキ)カ
無気か                &ムキカ
0007  00019.501  ー  00021.128  L:
そして                &ソシテ
調音の                &チョーオンノ
強さ                  &ツヨサ
```

図1　書き起こしテキストの例
（左が「基本形」右が「発音形」）

(6)形態素解析

書き起こしテキストに基づき、品詞情報などをKWIC（key word in context）形式で提示する「形態論情報」を作成した(図2参照)。なお、計算機による自動形態素解析に先立ち、人手により書き起こしテキストを短い単位に分割した。また、計算機で自動的に付与された情報には、人為的あるいは解析誤りに起因する種々のエラーが含まれるため、手作業により修正した。

図2　形態論情報の例

(7)分節音ラベル付与

書き起こしたテキストをもとに、音声データに対して分節音ラベルを付与した（図3参照）。

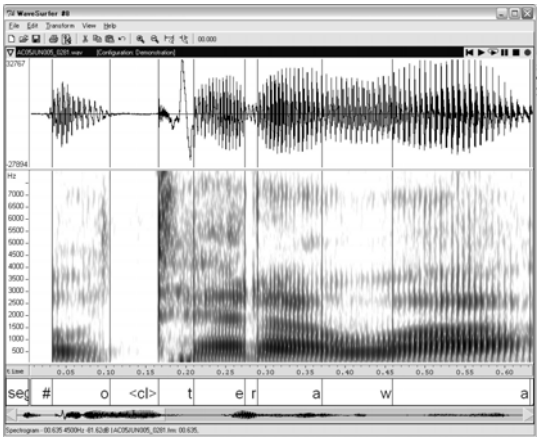

図3　分節音ラベル付与例
（上から音声波形，広帯域スペクトログラム，時間，分節音ラベル）

４．研究成果

(1)構築されたコーパスの特徴、及び格納されているデータ

①非母語話者の音声

CSJ-NNS は格納される音声が非母語話者の音声であることを最大の特徴としている。1.でも述べた通り、非母語話者のコーパスはいくつか存在するものの、音声学的分析が可能で、かつ母語話者のデータ(CSJ)と比較可能なコーパスは他に存在しない。

②自発発話

CSJ-NNS は非母語話者(中国語話者・韓国語話者・英語話者が中心)の自発発話を収録している。具体的には、学会やスピーチ大会などにおいて日本語で発表している音声である。中には原稿読み上げ形式の発表の音声も含まれているが、母語話者のそれにおいても自発音声の特徴と考えられるフィラーや言いよどみが頻出すると言われており(籠宮ほか 2000)、非母語話者においてはさらにその傾向が強いと思われる。したがって、このような音声も一部ではあるが CSJ-NNS の収録の対象としている。

③モノローグ

CSJ-NNS はモノローグを中心としたコーパスである。これは CSJ と同じ発話形式にし、比較を可能にするためである。

④付加情報

CSJ-NNS には高精度な転記テキスト、形態論情報及び音声学的ラベルを付与している。また収録されるデータが非母語話者であることをふまえ、学習歴、母語話者による印象評定等も付加情報として加えた。これにより、言語研究に十分耐え得るコーパスとなっている。

(2) CSJ と CSJ-NNS の比較を通して

CSJ の分析結果と、CSJ-NNS の分析結果とを比較し、母語話者・非母語話者それぞれの自発音声に見られる特徴・傾向の異同に関する考察を試みたところ、場合によっては非母語話者よりも母語話者の方が「発音のなまけ」という「規範からの逸脱」を起こしている現象もあることがわかった。またその一方で、「規範からの逸脱」の傾向が母語話者と非母語話者とが同様である場合も存在することが判明している（土岐ほか 2009)。

もっとも、母語話者の音声に観察される「規範からの逸脱」というのは、無秩序に生起しているわけではない。非母語話者の音声に観察される「規範からの逸脱」を考える際には、必ず同時に、母語話者の音声に観察される「規範からの逸脱」現象はどのようなものであるのかを記述する必要があることにも留意が必要である。教育上の目標とすべき姿は、当然のことながら、母語話者の生活感からくる勘だけではなく、確かなデータの裏づけによるものでなければならないからである。特に音声に限って考えただけでも、母語話者自身でさえはっきりとは意識できていないこと、それがまた、非母語話者にとっては、いち早く気がかりな問題となるような場合も少なくはない可能性もあろう。

上述の結果は、非母語話者の音声は常に「規範から逸脱した、訂正されるべきものである」といったある種の「思い込み」を再考するきっかけになるものと考えられるが、この点が非母語話者の日本語音声について、我々が最も重要視している基盤的考え方である。

謝辞

CSJ-NNS に音声を提供してくださった方々、及びご協力くださった各種学会・機関に，心より感謝申し上げます。

参考文献

大曾美恵子（2006)「日本語コーパスと日本語教育」『日本語教育』130 号 pp.3-10

大曽美恵子・滝沢直宏（2003)「コーパスによる日本語教育の研究―コロケーション及びその誤用を中心に―」『日本語学』Vol.22 No.5 pp.234-244

籠宮隆之・菊池英明・小磯花絵・前川喜久雄（2000)「大規模話し言葉コーパスにおける発話スタイルの諸相―書き起こしテキストの分析から―」『日本音響学会 2000 年秋季研究発表会講演論文集』 pp.107-108

小磯花絵 (2007)「『日本語話し言葉コーパス』を用いた助詞「の」の撥音化にかかわる言語内的要因の分析」『話し言葉コーパスに基づく言語変異現象の定量的分析』平成 16-18 年度科学研究費補助金（基盤研究(B)）研究成果報告書

土岐哲・江崎哲也・岡田祥平(2009)「『非母語話者による日本語話し言葉コーパス』の可能性」『日本語教育』第 142 号 pp.14-24

前川喜久雄（2004)「『日本語話し言葉コーパス』の概要」『日本語科学』15 pp.111-133

前川喜久雄（2006)「概説」『国立国語研究所報告 124 日本語話し言葉コーパスの構築法』pp.1-21 国立国語研究所

前川喜久雄（2008)「『日本語話し言葉コーパス』の設計と実装」『日本語学』27(5) pp.54-62 明治書院

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）
①土岐哲、江崎哲也、岡田祥平、「『非母語話者による日本語話し言葉コーパス』の可能性」、『日本語教育』、第 142 号、査読なし、2009、pp. 14-24
②土岐哲、「調査者に応じた被調査者のモード切替と音声の質的変化」、『論集』、Ⅳ号、アクセント史資料研究会、査読なし、2008、pp. 1-9
③岡田祥平、「自発的な〈標準語〉音声における〈促音〉化の概観―『日本語話し言葉コーパス』に基づく分析―」、『音声言語』、Ⅵ、近畿音声言語研究会、査読あり、2008、pp. 105-124
④岡田祥平、「独話と対話の連続性について考える」、『龍谷大学国際センター研究年報』、第 17 号、査読なし、2008、pp. 3-20

〔学会発表〕（計 15 件）
①土岐哲、「日本語の音声、ネイティブ・スピーカーは、つねに「正しい」か？―音韻論的意識と音声実現の乖離、学習者の聞こえ等を巡って―」、『日本語教育学会中国地区研究集会』招待講演、鳥取大学（鳥取県鳥取市）、2009 年 12 月 19 日
②山本麻実、趙國、山下洋一、「アクセント結合規則を利用した CRF に基づくアクセント型自動ラベリング」、『日本音響学会 2009 年秋季研究発表会』、日本大学（福島県郡山市）、2009 年 9 月 16 日
③岡田祥平、「現代日本語の音声言語におけるモーラの出現頻度―『日本語話し言葉コーパス』を使用した調査結果―」、2008 年(平成 20 年)度第 22 回日本音声学会全国大会、明海大学（千葉県浦安市）、2008 年 9 月 15 日
④江崎哲也、岡田祥平、ユンヨンファ、岩男考哲、「『非母語話者による日本語話し言葉コーパス』の概要」、『日本語教育学会 2008(平成 20)年度春季大会』、首都大学東京（東京都八王子市）、2008 年 5 月 25 日

〔図書〕（計 6 件）
工藤 浩、小林 賢次、真田 信治、鈴木 泰、田中 穂積、土岐 哲、仁田 義雄、畠 弘巳、林 史典、村木 新次郎、山梨 正明、『改訂版日本語要説』、ひつじ書房、2009、348(119-157)

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
①『非母語話者による日本語話し言葉コーパス』ver. 2、2010
②『非母語話者による日本語話し言葉コーパス』ver. 1、2009
③江崎哲也、「日本語の音声と母語別発音練習」、『第二回日本語ボランティア養成講座』、山梨日本語ボランティアの会、山梨県甲府市、2008 年 1 月 12 日

６．研究組織
(1)研究代表者
土岐 哲（TOKI SATOSHI）
京都外国語大学・外国語学部・教授
研究者番号：10138662

(2)研究分担者
鹿島 央（KASHIMA TANOMU）
名古屋大学・留学生センター・教授
研究者番号：60204377
中西 久実子（NAKANISHI KUMIKO）
京都外国語大学・外国語学部・准教授
研究者番号：30296769
山下 洋一（YAMASHITA YOICHI）
立命館大学・情報理工学部・教授
研究者番号：80174689
江崎 哲也（ESAKI TETSUYA）
山梨大学・留学生センター・講師
研究者番号：40420343
岡田 祥平（OKADA SHOHEI）
大阪大学・文学研究科・助教
研究者番号：20452401

(3)連携研究者
なし