日本人英語学習者コーパス作成とその利用可能性

齋賀 豊美[†‡] 和泉 絵美[‡] Ook Chung[†] 井佐原 均[†‡]
通信・放送機構[†]／通信総合研究所[‡]

1. はじめに

我々は、日本人が最も苦手とする話す英語の学習支援システム開発を目標に、その研究の基礎となるデータとして、日本人英語学習者の発話コーパスの作成を現在進めている。[1][2]

我々は、通信総合研究所の自然言語処理研究グループにおいてコーパス作成を進めているのだが、本プロジェクトの性質上、英語教育分野の研究者との意見交換を積極的に行っている。その際、英語教育の研究者たちが、自然言語処理技術に対してさまざまな期待を寄せているのを強く感じる。教師や教育研究者たちが築き上げてきた方法論や教材の開発に、新たな視点や手法を加えることが望まれているのである。

本稿では、本プロジェクトの概要と途中経過を報告し、他の有名な学習者コーパス・プロジェクトを紹介した後、我々の作成しているコーパスの幅広い活用を促進することを目的として、日本人の英語学習者による発話コーパスを公開した場合に、英語教育分野において期待されること、あるいは、純粋に自然言語研究の分野において、どのような研究に利用可能と考えられるかを述べる。

2. コーパス作成

2.1 概要

本コーパスは、日本人の英語学習者による発話を対象とする以外に、9 段階で評価された英語能力レベル別のデータを扱う大規模なコーパスであることを大きな特徴とする。対象とするデータは、㈱アルクの実行する SST (Standard Speaking Test)と呼ばれるもので、各被験者について、15 分程度のインタビュー形式のテストにおいて、自己紹介などの会話や、イラスト描写、ロールプレイ、ストーリーテリングの3つのタスクが行われる。

作成するコーパスは、メインとなるコーパスと、その一部データについて比較研究を行うための2つのサブコーパスからなる。

2.2 メインコーパス

音声からの書き起こしにおいては、発音の間違いは、文脈から単語が推測できれば無視しても良いこととした。同時に、すべての書き起こしテキストには繰り返しや言い直し、発話の重複、あいづちやフィラーなどの基本的な情報が付加されている。さらに、文法的、語彙的誤りについて、できる限り網羅性を持ったエラータグセットを作成し、人手によるエラータグ付与を行っている。なお、テキストに付加する情報を表すタグは、XML をベースとした表記方法を取っている。

総データ量は、今年度末時点で800件(約200時間、120万語以上)、プロジェクト最終年度である 2002 年度末までに 1200 件のデータの収集を予定している。

2.3 2つのサブコーパス

(1) 正解コーパス

ネイティブと学習者の使用語彙の差異などを分析するため、英語ネイティブの被験者に、擬似テストを受けてもらう。メインコーパスとの比較を可能とするため、自己紹介やタスクなどで触れるべき内容には指針が与えられる。

(2) 日本語訳コーパス

学習者がどれだけ日本語の構成、表現に引っ張られているか、といった母語の影響と外国語習得の関係を探るため、学習者が言いたかったであろう事柄を発話から推測し、日本語で表現する。

3. 英語学習者コーパス現状とその利用

3.1 学習者コーパスとは

学習者コーパスという用語はこの 10 年ほどで定着してきたものである。学習者コーパスとは、「ある言語を外国語として学ぶ学習者が話したり書いたりしたもののコーパス」を言う。

次節以降で、有名な学習者コーパス・プロジェクトを数例紹介する。

3.2 世界の学習者コーパス・プロジェクト

(1) International Corpus of Learner English (ICLE)

1990 年から始まった International Corpus of English (ICE)という世界18地域の英語変種コーパス構築プロジェクトの傘下で、英語の変種(variety)の1つとして英語学習者コーパスを作ろうというものである。現在世界の学習者コーパスの中でも、サンプリング方法の科学的なこと、規模と整備状況から見ても最も本格的なものである。Louvan 大学(ベルギー)の Ph.D. Sylviane Granger 氏が中心になり、多くの関連する研究発表を行っている。ICAME よりデータが公開される予定である。

(2) Longman Learners' Corpus (LLC)

Longman が所有する商用の学習者コーパス。約1000万語の規模は世界最大級。80年代初期からデータを収集している。データの採取方法はかなり不統一で、種類も雑多である。これらを世界70カ国以上から集めたデータベースをエクセター大学名誉研究員 Michael Rundell 氏が corpus manager として全体の統括をした。商用で本格的に出版物に学習者コーパスを利用しているのはロングマン社が最初である。

3.3 日本の学習者コーパス・プロジェクト

(1) Corpus of Japanese Learners of English

JACET'96 で呼びかけがあり、ハイパーメディア研究会のメンバーが中心で発足した日本人英語学習者コーパス作成のプロジェクト。中・高・大と連携したデータの広範囲な採取、WWW 上での共有化、エラータグ開発、音声コーパスへの取り組みなど全国規模で呼びかけようとした。1997年度からは東海大学教授の朝尾幸次郎氏が代表となって文部省科学研究費補助金研究のプロジェクトが3年計画で進行した。プロジェクト・メンバー中心で収集が行われ、100万語規模に近いデータは集まった。

(2) JEFLL (Japanese EFL Learners) Corpus

東京学芸大学で10年間にわたって行われた文部省科学研究費補助金研究による英作文プロジェクトで蓄積されたデータを母体とし、その後、明海大学助教授投野由紀夫氏をプロジェクトの中心として、中・高のボランティアの先生方を加え、さらにデータ増殖を大幅に行って現在の形態に至っている。中学2年から高校3年まで同一のトピックで自由英作文をさせたデータを電子化している。現在の規模は学習者コーパス部分のみで約50万語(作文45万語、会話5万語)。特に、インプットの影響を調べるために中学・高校英語教科書コーパス、また母語の特徴分析のために同一トピックを日本語で書かせた作文コーパスおよび研究用の一般日本語コーパスを比較コーパスとして同時に整備しているのが特徴的である。

4. 本コーパスの利用可能性

本章では、書き言葉の学習者コーパスの効用や研究の例とともに、話し言葉を対象とする本コーパス利用の可能性を述べる。

4.1 英語教育研究における可能性

(1) 英語学習プロセスの解明

1970 年代には、第 2 言語習得研究の分野で冠詞や複数の-s などの語形変化による文法的機能、すなわち、形態素の習得の順序を調査するために多くの研究が行われた。[3]その後もかなり具体的な習得段階モデルが提案されてきている。日本人学習者の習得順序モデルの構築によって、英語母語話者との習得プロセスの違いや、同じ日本人でも書き言葉と話し言葉についての違いが明確になる可能性がある。

(2) 母語干渉、日本語による表現

学習者とネイティブ・スピーカーの英語使用との差違を調査することで、ネイティブ・スピーカーとは異なる英語学習者特有のニーズがあることが浮き彫りになってくる。

教育現場ではどんなに工夫して生徒にふさわしい語彙や表現を提供しても、どうしても生徒の伝えたい特定の語彙や表現が抜け落ちていることがある。逆に、ある特定の表現や構文に関しては学年やタスクが変わっても学習者はいつでも苦労するというようなものがあるかもしれない。そういった面で、英語での表現活動の際に日本語にならざるを得ない部分にもっと注目してみるといい。本コーパスにも、伝えたいのにどうしても

表現できないために日本語を使用している実例が見られる。

(3) 英語教員の養成・研修

学習者コーパスによる実証データは、例えば英語教員養成プログラムに利用されて英語教師志望者の訓練のために用いられよう。学習者の習得プロセスや典型的な誤りの例を事前に見ることは教師に従来と異なる言語教育観を与えるかもしれない。

また新しい指導順序や指導法の実験などを行う際や現状の学習指導要領やシラバスの改訂の際にも、組織的に採取された大量の学習者データは重要な基礎資料となる。そして言語教育システム全体の改善のために有効に用いられる可能性が高い。

本コーパスには、どのような問い掛けが、生徒の自発的発話を促すかということを、受験者に多く話すことを促す試験官の技術から汲み出せる可能性も秘めている。

(4) 教授効果の予測

学習者コーパスによって、教師自身が英語学習の道筋についての詳しい情報を得ることができるようになると、ある構文や単語をどういう風に生徒が使うのか、典型的な誤りは何か、いつ頃まで気になる間違いは続くのか、日本語にひきずられる表現は何かといったさまざまな疑問に対して学習者データが何百、何千と言う具体例で答えてくれる。それによって英語教師の誤りへの性急かつ否定的な対応、神経過敏だった練習の方法、もっと重点を置くべきだった特定の語彙や表現への注意喚起など、教師が気づかなかった部分が変化していくかもしれない。

(5) 辞書、教科書開発

現在英米の英英学習辞典はほとんどが一億語以上の規模の現代英語コーパスをベースにしているが、次世代の学習辞典の特徴として彼らが真剣に検討しているのは、精選された学習者データから学習上のつまずきや母語の影響で間違いやすい語彙などの資料を系統的に得ることである。学習者コーパスの整備で、辞書も学習参考書も学習者のエラーや不自然な語彙の使用法について注意を喚起することが出来る。

ネイティブの使用語彙リストと共に学習者の表現語彙リストが頻度情報を持って整備されれば、もっと学習者のニーズに合った教科書作りが可能になろう。

それだけではない。学習者コーパスで導入から定着までの時間差や、項目間の難易度に関する客観的なデータが得られれば、まったく新しい教科書の提案が可能なのである。そもそも、日本人が触れたい話題や伝えたいと試みる内容は、ネイティブのそれとは異なるはずである。

(6) 学習者による直接利用

また、学習者コーパスは直接学習者に触れさせることで新しい教材としての可能性を持っている。例えば学習者自身が自分の表現したいことを他の同レベルの仲間がどのように英語で表現しているかを多数の用例で見ることができる。そしてそれをネイティブのコーパスで自然な英語表現と比較してみることで規則性や特徴を発見する、といった新しいタイプの学習・表現活動の可能性を持っている。

また日本語と対比させた日英対訳学習者コーパスを作ることで、日常彼らが表現したい内容を日本語と英語の表現を比較対照しながら表現辞典のように利用出来たりする。

4.2 言語処理研究における可能性

(1) 誤りを含むテキストの解析技術

従来の自然言語解析技術は、新聞コーパスなどを処理の対象として正しい入力に対して最大の精度を上げるように開発されてきた。本コーパスには、学習者による誤りが含まれており、極端な場合には、文法的あるいは意味的に成立しない、いわゆる非文まで含まれている。「誤りに強い」解析手法を開発するための基礎データとして本コーパスは利用可能である。この技術は、エラー診断・自動添削・レベルチェック システムの開発へ応用が期待される。[4]

(2) 音声認識システム

学習者コーパス・データが音声データの形式で収集でき、発音の誤りに関して系統的なパタンが獲得できれば、それを音声認識システムの基礎資料に使うことが出来る。現在の音声認識

システムの問題点は、発音の正解領域が狭いことと、学習者が母語を織り交ぜて話すことがあることを無視していることである。日本人のカタカナ英語といわれる発音から流暢で標準的な発音まで対応する技術が開発されれば、コンピュータを相手に英会話の訓練をすることや、海外に生活する場合にも日本人英語の発音にも対応する自動音声応答システムや英語のコマンドを理解して動くロボットなど、より優しいシステムの開発が可能になるかもしれない。

(3) 機械翻訳・作文支援システム

学習者コーパスは人間が外国語を使う際にどのような点で困難を感じ、実際にどのような誤りを犯すかの総合的な資料を与えることができる。これをもとにすれば、正しい言語を対象とする機械翻訳システムを目指す一方で、人間が目標言語の文法や語彙に困難を覚える点を重点的に補佐してやるシステムを開発することが考えられる。それは作文支援システムにも応用が可能である。問題は実際の利用者が作文を行う際にどういう点に関してサポートが必要か、ということに関するエンド・ユーザーの分析が不足していることである。学習者の作文コーパスに加えて、より顕著に学習者の特徴を表すと考えられる発話コーパスにより、このような分野に対して、効果的な基礎資料を提供できる可能性がある。

(4) 人工知能

自然言語処理の分野では、自然言語の文法を自動獲得する研究や学習モデルを作成することが行われている。それらの技術を応用することで、学習者コーパスから言語モデルの獲得を行い、それを段階を経てどのようにモデルの特徴が推移していくかを調べることで、今までとは異なる新しい知見が得られる可能性がある。

英語学習者の学習プロセスをある程度一般的に人工知能モデルで表現できれば、人工知能の研究そのものとしても画期的であるし、その学習プロセスにどのような指導や訓練をほどこすことによって、よりプロセスを促進できるか、といった言語教育面での応用にも発展していく可能性が含まれている。

5. おわりに

音声認識、話し言葉対話システムなどの技術が進歩することにより、遠い将来には、コンピュータに向かってコンピュータと対話することによって、スピーキング能力を磨くことが可能になるかもしれない。

我々は、今年度までに収集したデータの解析をはじめると共に、4章で述べた各分野において本コーパスの利用をより容易に可能にするための整備をして行きたいと考えている。

謝辞

本研究において、データの収集と利用、書き起こし基準の決定に際しては、平野琢也(㈱アルク)、金子恵美子(㈱アルク)、金子朝子(昭和女子大学)、投野由紀夫(明海大学)、
成田真澄(㈱リコー)の各氏の協力が不可欠でした。ここに感謝いたします。

参考文献

[1] 齋賀豊美、井佐原均:日本人学習者の英語発話コーパスの作成—概要と開発環境—、言語処理学会第7回年次大会発表論文集、pp.541-544、2001
[2] 和泉絵美、井佐原均:英語学習者発話コーパスにおける誤り分析—エラータグとその応用—、言語処理学会第7回年次大会発表論文集、pp.545-548、2001
[3] ロッド・エリス著、金子朝子訳:第二言語習得序説、研究社出版、1996
[4] 井佐原均、投野由紀夫、平野琢也:日本人学習者のレベル別英語発話コーパスの作成、言語処理学会第6回年次大会発表論文集、pp.32-34、2000