日本人学習者のレベル別英語発話コーパスの作成

井佐原均（通信総研） 投野由紀夫（ランカスター大学） 平野琢也（アルク）

1. はじめに

これからの高度情報化社会の中で、日本人にとっては言語障壁の問題は、避けて通ることが出来ない。読む、書く、聴く、話すといった英語能力の中で、日本人が最も不得手とする「話す」に注目して、その能力を実際の言語データに基づいて客観的に評価し、発話支援・学習支援という場面で何をするべきかを明確にすることが、日本人の情報発信能力の向上へのブレークスルーになると考える。このような研究の基礎となるデータとして、われわれは日本人の英語学習者の発話データのコーパス化を開始することにした。

ここで作成される大量の実言語データと、言語データに基づく自然言語処理技術、システム化技術、実際の英語教育場面での知見を組み合わせることにより、英語学習を支援する環境の開発が可能となろう。また、このデータは、言語獲得過程のモデル化や、英語学、英語教育の分野の研究で広く利用されることが期待される。

従来より、日本人の英語使用者を対象に、英語生成支援環境と英語精読支援環境の開発が行われていた [1, 2]。このような研究開発とその成果の利用における言語データの重要性に注目して、昨年８月に開催された第１２回国際応用言語学会世界大会(AILA'99)においてコーパスと英語教育に関するシンポジウム (Narita, M., McEnery, T., Barlow, M., Tono, Y. and Isahara H.: The Roles of Corpora in Language Teaching and Language Engineering.) が開催され、多くの参加者を得た。その後、シンポジウムの講演者を中心に、英語教育に有効となる言語データの収集について検討を重ね、大量の日本人英語学習者の発話データの収集と、その研究用の公開に向けてのプロジェクトを開始することにした。現在、試行的に、小規模ながら発話データの書き起こしによる「話者レベル付き英語発話コーパス」の作成を開始している。

2. 英語学習者コーパスの現状

従来、英語学習者コーパスとしては、ベルギーの Catholic University of Louvain を中心にした大学３年生の英語学習者のデータを集めた ICLE (International Corpus of Learner English) がある。これは、学習者の１５の母語ごとに、２０万語を収集しているが、対象は英作文データであり、また、学習段階（英語能力）別のコーパスではない。話し言葉の英語学習者コーパスは世界的にも極めて少なく、同大学が学内で立ち上げている話し言葉コーパス収集のプロジェクトがあるが、これも学習段階別ではない。

学習者コーパスとしては、他に、Longman Learner's Corpus（商用：作文、英語能力別)、Cambridge Learner Corpus（未公開：作文、英語能力別)、HKUST（香港科学技術大学）corpus（未公開：作文、大学生)、などが知られているが、どれも作文が対象で会話コーパスは大型のものが存在しない。

日本では、大学英語教育学会ハイパーメディア研究会の有志による科研プロジェクトが３年計画で終了して、約１００万語の作文コーパスが収集されてきた。ただし、大学生のデータが多いことと、対象が作文データであることなど、今回の我々の試みとは異なる。JEFLL (Japanese EFL Learner Corpus)プロジェクトでも、作文・発話データをコーパス化しているが、話し言葉の部分はまだ５万語程度で非常に小さい。

話し言葉としては、CHILDES の幼児の言語習得コーパスが有名だが、これは第１言語習得が中心で、第２言語のデータはほとんどが、バイリンガルの子供のデータにかぎられている。

以上の状況から、英語能力によって分類された大規模な発話コーパスは世界的に見ても初めての試みである。

3. 発話者のレベルによって分類された英語発話コーパスの開発

ここで対象とするデータは 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が行われ、その結果が判定者によって９段階にランク付けされる。このテストは米国で ACTFL (全米外国語教育協会) の下で大規模に行われている OPI (Oral Proficiency Interview) を元に、日本の英語教育事情を考慮して開発されたものであり、結果の判定は出来るだけ客観的であるように構成されている。

英語発話コーパスの開発は、(あ) 音声データからの書き起こし、及び情報付加による発話データベースの作成、(い) 学習者コーパスに付加するべき情報の検討、(う) 学習者コーパスの作成支援としての、誤りを含むテキストに対する解析技術の研究、からなる。

3.1. 音声データからの書き起こしと情報付加

音声データとしては、(株) アルクの実行する SST テストの被験者データを用いる。このテストの受験者は年１５００人程度と予想され、一人が約１５分のインタビューを受ける。これにより、年に４００時間程度のデータを入手できる。５年間で２０００時間のコーパス化が目標である。学習者のレベルが９段階に分かれることから、２０００時間という量は十分ではないが、一応の規模に達すると言えよう。

3.2. 発話コーパスに付加すべき情報の検討

従来の情報 (タグ) 付き英語コーパスは正しい英文に対して情報を示すタグ (品詞、構文的役割、意味、など) を付与したものであった。今回対象とする学習者コーパスは学習者の発話の誤りを含むデータであり、そこで用いるタグは複雑なものとなる。たとえば、発話においては「言いよどみ」が存在するために、それを表すタグをつけることが行われるが、そのようなタグを単純に作るだけでは、I want some…something という発話の後半部分を something の言いよどみと考えるか、some Xs からの言い直しと考えるか、という振れを吸収できない。人間の判断を入れないようなタグの体系を作ることが必要となる。

本研究では、初年度において、単純な書き起こしデータを作成する傍ら、学習者コーパスに必要となる、音韻、単語、構文、意味等のタグの体系を作成する。このタグ体系に基づいて、発話コーパスを作成する。

3.3. 誤りを含むテキストの解析技術の研究

従来の自然言語解析技術は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。しかしながら、今回対象とする学習者コーパスの場合は、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。一方大量のデータ作成をすべて人手で行うことは効率が悪い。このため、誤りのある部分も含めて全体的な解析が出来る枠組みと、誤りを含む部分を詳細に解析する枠組みとを組み合わせた、「誤りに強い」解析手法を開発する。

4. 発話コーパスの応用

今回作成を開始した発話コーパスを用いれば、(あ) 日本人の英語発話モデルの構築と、発話者のレベル分けに基づく、発話モデルの変化のモデル化、(い) これらのモデルに基づく英語教育法の提案、(う) 英語学習支援環境の開発などが可能となろう。

4.1. 英語発話モデルの構築と、発話モデルの変化のモデル化

従来、英語教育研究の一環としての学習者のモデル化は、小規模のデータに基づく、教師あるいは研究者による主観的な分類がほとんどであった。今回の発話コーパスを用いれば、大規模な言語データを元に、各レベルの学習者の持つ典型的な言語運用モデルを構築し、比較することが可能になる。モデル化は基本的に以下のような過程で進められよう。

(1) 各レベルの発話者の発話モデルの作成

それぞれの英語能力における発話データからそこに共通する誤りを検出しモデル化する。

(2) モデルの変化のモデル化

英語能力が上がるにつれて、発話モデルがどのように変化していくかを検討し、その変化のモデルを作成する。これは、学習者が第2言語としての英語を獲得する過程のモデルとなる。

4.2. 言語モデルに基づく英語教育法の提案

第2言語の習得過程がモデル化できれば、次の段階はそのモデルを出来るだけ早く達成するような学習手法の開発が課題となる。習得過程モデルから効果的な学習法を予測し、それをモデルに適用するという繰り返しで教育法を提案すると共に、実際の学習者を対象に心理実験を行うことにより、教育・学習法の開発を行うことが可能となろう。

4.3. 英語学習支援環境の開発

これまで、英語生成支援環境と英語精読支援環境の開発を行ってきた。対象を発話とした場合には、これらのような実際の言語利用の過程を支援するものではなく、発話能力を向上させる過程を支援する学習システムが目標となろう。

5. おわりに

このコーパス作成は、大規模な実データを元に英語の発話能力の獲得過程を客観的に解析し、英語教育の向上に寄与しようとする試みであり、テーマの大きさから、短期間で達成できるものは限定されたものとなろう。しかしながら、本研究で行われる話し言葉に注目した言語データの集積と研究開発により、将来的に、言語の獲得・運用といった人間の言語能力のモデル化の研究が進み、効率のよい言語学習法も開発されるであろう。また、言語学への寄与も考えられる。今回の対象は英語であったが、他の言語についてもデータの集積を行い、ここで開発されるモデル化の手法を適用することにより、言語一般のもつ特徴を導き出すこともできよう。

我々が今後も日本語を使い続けつつ、積極的に情報発信を行うには、英語能力の充実が必須である。特に人間同士のコミュニケーションにおいては、「書く」「読む」「聴く」「話す」のうちで、「話す」能力が充実しなくては、フレンドリーなコミュニケーションは不可能である。今回作成を開始する発話コーパスと、それを用いた研究成果は、このような「話す」能力の向上に直接的に寄与するものである。

英語能力の問題は頻繁に語られる問題でありながら、客観的な取り扱いはほとんどなされてこなかった。どこかで一度、きちんとしたデータ集積の上で、何が出来るかを明確にする必要があると思われる。我々が小規模にはじめた発話コーパスの作成は、そのような思いから開始したものである。幸い、データ収集については目処がたった。これをさらに大規模に行うことは我々のみでなく、この分野の研究全体の活性化と発展につながると確信する。

もちろん、英語教育はこれまでにも多くの研究や実践が行われてきており、我々だけで十分な成果が出せるものではない。研究を行うに際しては、できるだけ広く知見を集めたいと考えている。

謝辞

今回のコーパス作成計画の立案および実施に当たっては、(株) リコーソフトウェア研究所の成田真澄研究員と、昭和女子大学の金子朝子教授には大変お世話になっている。ここに記して感謝する。

参考文献

[1] Narita, Masumi: A Tool for Assisting Japanese Researchers in Writing English Abstracts. Proceedings of EUROCALL'99. pp. 69-70. (1999)

[2] 井佐原均: 談話構造と読解過程に注目した英文精読支援システムの開発. 言語処理学会第4回年次大会発表論文集. pp. 673-675. (1998)