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－

齋賀　豊美　　井佐原　均
通信・放送機構／通信総合研究所
{saiga | isahara}@crl.go.jp

1　はじめに

高度情報化社会が進展し、国際的コミュニケーションの手段である英語の必要性が強く認識されるようになっている。われわれは、読む、書く、聴く、話すといった英語能力のうち、日本人が最も不得手とする「話す」に注目し、言語習得のモデル化や発話支援・学習支援の基礎となるデータとして、日本人学習者による英語発話コーパスの作成を進めている。

本研究は、通信・放送機構の進める「適合型コミュニケーション技術の研究開発プロジェクト」の一環として行われているものである。このプロジェクトでは、日本人の英語を対象に（1）英語発話データベースの作成と公開、（2）データベースの作成に伴う自然言語処理技術の開発、（3）データベースおよび処理技術の有効性の実証、を目的とした研究開発を実施している。

本稿では、まず、作成するコーパスの特徴とデータとなる発話の収録について説明する。次に、現在行われている音声データの書き起こしと、コーパス化する上で付与する基本的な情報、そして、書き起こし作業において使用するために独自開発したエディタについて述べる。

ここで作成される大量の英語発話データは、自然言語処理技術、実際の英語教育場面での知見などを組み合わせることにより、言語獲得過程のモデル化や、発音、談話、語彙、文法などの各分野での英語教育の研究で広く利用されることが期待される。

2　コーパスの特徴と音声データの収録

2.1　コーパスの特徴

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、英語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。

従来、英語学習者コーパスとしては書き言葉を対象とするものがほとんどであった。話し言葉の英語学習者コーパスは極めて少ない。存在するコーパスは小規模のものしかなく、学習段階別のものではない。これらの理由から、本プロジェクトで行うコーパス作成は、世界的に見ても初めての試みである。[1]

2.2　対象データ

ここで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。

試験官は研修、認定試験を受けた日本人である。試験官が外国人でないことは、受験者が過度の緊張を強いられずに済むという利点となっていることが予想できる。

2.3 SST

SSTは5つのステージから構成される。

ステージ1：ウォームアップ
ステージ2：イラスト描写
ステージ3：ロールプレイ
ステージ4：イラスト描写
ステージ5：ワインドダウン

ステージ1では簡単な挨拶や自己紹介などを行い、ステージ5では簡単なおしゃべりを行う。ステージ2－4のタスクでは、受験者が主体的にできるだけ多く話すような構成になっている。

2.4 収録

SSTは、受験者1人、試験官1人の2人によって、対面形式で行われる。発話は、モノラルカセットテープ2本に収録される。2本あるのは、録音失敗に備えてである。ただし、本プロジェクトのために、途中からはDATを2本のうちの1本として取り入れている。

受験者の性別、年齢、海外生活経験期間、職業、TOEICやTOEFLの得点などが、同時にデータとして残される。

3 書き起こしテキストの作成

英語発話コーパスの開発は、(1)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、(2)学習者コーパスに付加するべき情報の検討、(3)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。この章では、音声データからの書き起こしと、基本的な情報の付加について述べる。

3.1 書き起こしガイドライン

書き起こしに際して、いくつかの取り決めを行った。まず、英語はなるべく聞き取ったとおりに筆写する。ただし、例えば、[r]と[l]の違いなど発音などの細かい特徴は無視してよい。単語は、文脈から認識可能であれば、発音が多少悪くても、正しい単語のスペリングを用いてよい。略字、略号など、もし、文字を1字ずつ言った場合は、文字と文字の間にスペースをあける。もし、続けて一語として言った場合は、続けて書く。日付や番号は必ず、発話どおりに文字で書く。文区切りは、前後の発話から筆記者が判断した部分で適宜入れてもらう。ピリオド、カンマ、疑問符など通常の記号を用いてよい。場面上重要な情報があれば、非言語的な出来事でも必要に応じて表す。

3.2 情報の付加

音声データからの書き起こし作業に際して、同時に、以下に述べるような基本的な情報をタグとして付加することを行った。

- 繰り返しや言い直し
- 繰り返しや言い直しであるが、何を言っているかあいまいな部分
- 聞き取れない部分
- 日本語が用いられた場合
- ポーズ
- 発話の重複
- あいづちやフィラーなど
- 咳などの非言語音
- 笑いながら話している部分

これらを表すタグは、XMLをベースとした表記方法を取っている。

3.3 タグの仕様

われわれのタグセットは、ある1点においてXMLに準拠していない。それは、タグの交差

を許している点である。同一話者発話内に限ってではあるが、例えば、日本語での発話と、受験者と試験官の発話重複とを表現するためには
<JP>nanteiunokana <OL> renshudakara </JP> training</OL>.
のように交差することになる。また、この先、付加する予定であるエラータグについても、交差して付加する自由度を持たせるべきだと考えたからである。ここで、タグ<JP></JP>は日本語、<OL></OL>は重複を表す。

4　書き起こしエディタ

各種タグを実際に人手で付与することは、非常に時間がかかる上、記入間違いも多く起こるという意味で非常に効率が悪いので、タグ付けを容易に行うことができるような環境が必要となる。その上、われわれのタグセットはXMLに準拠していない点があるため一般的なタグエディタではフォーマットチェックが不可能である。そこで、われわれはタグ付けのためのエディタを独自開発した。

このエディタを開発する上で考慮した点は以下の事柄である。

- 発話そのものの入力など、人手でしかできない事以外は極力簡単な操作で編集が可能になるようにする。
- パソコン操作に不慣れな人にもわかりやすい使い勝手を持ち、熟練した人にも使いやすいようユーザによるカスタマイズが多くの操作で可能であることなど、不特定多数の作業者がそれぞれ自分にあった操作ができること。
- エラータグなど、タグセットの追加や削除などの変更などに対応できること。

1番目と2番目の考慮点に関しては、タグの付与に関して、マウスでボタンをクリックすることによる方法、ショートカットキーを使う方法、一度に複数の単語を一気にタグ付けする方法（図1）、操作手順上次に来ることが予想される特定のタグは自動挿入を可能にするなどいろいろな選択肢を用意して対応した。

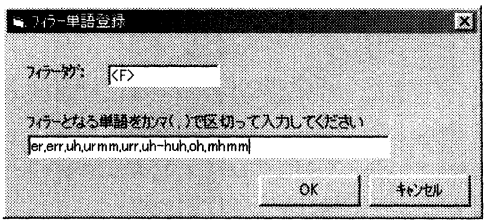

図1：フィラーとなる複数の単語を登録するダイアログ

受験者に関する情報などを記述するヘッダ部分のタグや、SSTのステージなど構造を表すタグは、必須であり挿入する位置が決まっている。これらのタグは、テンプレートで用意され、タグ挿入の手間を大幅に省いている。このテンプレートは簡単なファイル編集によって変更が容易に行える。

視認性を高めるため、タグに好みの色を割り当てることができるようにもした。また、操作に関して気付きにくいテクニックは、できるだけウインドウ内にメッセージを書き込むよう努めた。

書き起こし作業における記入間違いには、単純なタグの間違い、タグの挿入位置に関する規定違反、英語のスペルミスなどが考えられる。登録されていないタグの検出や、個々のタグがその用途に応じた規定通りに使用されているかのチェックなどはフォーマットチェック機能で行える。また、テキストのみ（あるいは任意のタグとテキスト）を抽出して表示する機能を持つため、スペルチェック機能を持つ他のエディタを併用すれば、スペルを随時チェックすることも可能となっている。（図2）

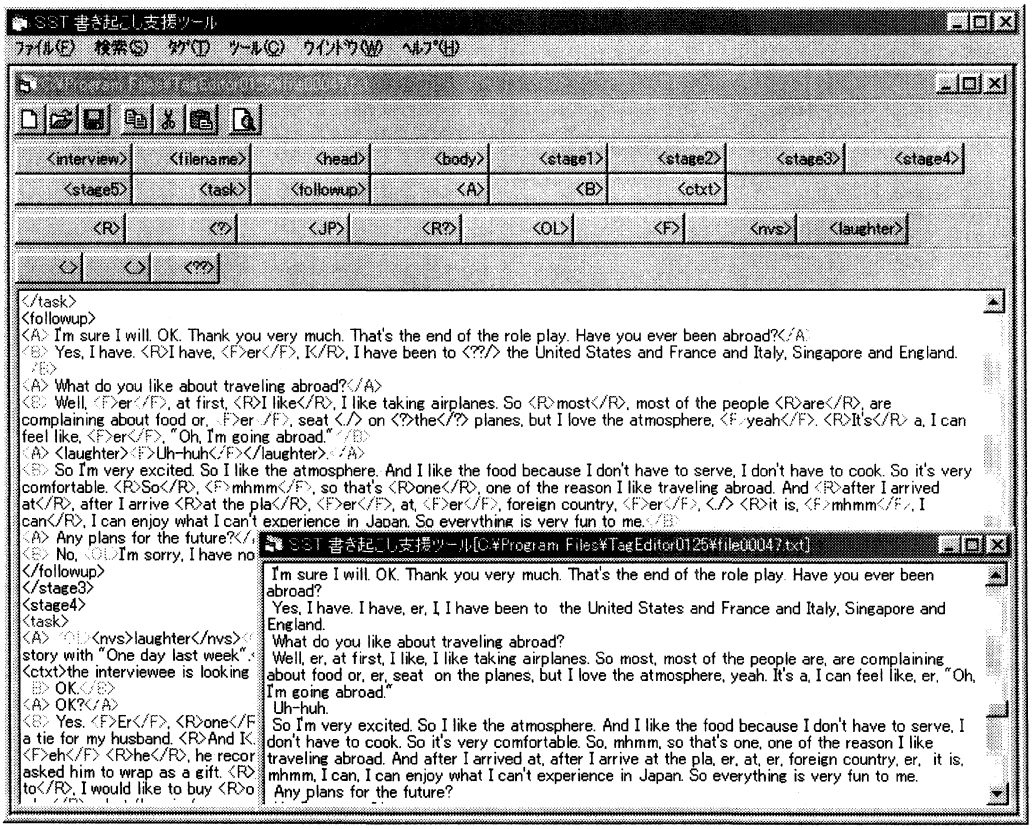

図2：編集画面およびテキストのみを表示するウインドウ

5　おわりに

本プロジェクトでは、現在、書き起こし作業を進めると共に、発話に含まれる誤り情報を付加するためのエラータグの体系化を進めている。このコーパスは、将来、広い分野での活用を目指して、出来るだけ広く知見を集めながら作成を進めていく予定である。

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）、投野由紀夫（ランカスター大）、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]　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