Standard Speaking Test (SST) Corpus の公開と研究・教育利用

Exploiting the Standard Speaking Test Corpus for Research and Teaching

投野由紀夫 （明海大学外国語学部）
和泉絵美 （通信総合研究所）
金子恵美子 （(株)アルク）

1. はじめに

本ワークショップは，まもなく公開予定の Standard Speaking Test (SST) Corpus の構築に関する最終成果報告として，SST Corpus の概要，付帯するコーパス作成ツール群の機能および SST Corpus を使用した第二言語習得および言語教育への応用例を紹介することを目的とする。ワークショップとして計画するのは，実際に話し言葉コーパスを構築する際のさまざまな書き起こしガイドラインの策定，それに付随する書き起こしツールなどの一連の操作をデモなどを交えながら紹介する実践的な内容を含むからである。

2. Standard Speaking Corpus とは

2．1．プロジェクトの目的

本プロジェクトは通信・放送機構（TAO）が平成１０年度より開始した「先端技術移転加速型研究開発」の一環として行われ，特に「適合型コミュニケーション技術の研究開発」と称した平成１２年度からの３年プロジェクトにより SST コーパスの構築が行われた。これはちょっとした言葉遣いの誤りや言い間違いを，そのまま受け取ったり，意味を理解できないと反応するのではなく，話し手が何を言いたかったのかを推論して，適切なコミュニケーションを実現させるようなシステムが，ユーザーフレンドリなコミュニケーションに必要であることから，このような「適合型コミュニケーション技術」を実現するために，通信総合研究所が開発してきた，比較的少量の言語データから文法や辞書を効率よく獲得できるシステムを使い，日本人の英語を対象に以下の研究開発を実施したものである。

（1） 英語発話データベース（SST Corpus）の作成と公開
（2） データベース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

2．2．SST Corpus の特徴

SST Corpus は以下のような特徴を持つ：

- 被験者は日本人英語学習者
- 話し言葉データ
- １－９の英語発話能力レベル別（９が最上級）
- 大規模（１５分の発話データを１２００名分）

対象データは(株)アルクの実行するSSTと呼ばれる英語会話能力テストの１５分程度のインタビューがソースになっており，これが米国のACTFL OPI と同等の基準で日本の現状に合うような課題で会話能力を測定できるように構成されている。

３．コーパス構築の概要

SST Corpus の構築に際して，会話データの特性を活かせるような書き起こしガイドラインの策定，および基本情報を付与するためのタグセットの仕様を決めた。また書き起こしエディタを開発し，それにより会話データの書き起こし作業が一定レベルの水準と客観性を保ちながら効率的に行えるような工夫をした。

さらに一部のコーパスにはエラータグを付与することとし，そのエラータグの仕様も詳細に検討した。エラータグ付与以外にも誤り分析の一環として，以下のサブコーパスを作成した：

- 日本語訳コーパス（back-translation corpus）
- 正解コーパス （corrected corpus）

これらを適宜組み合わせて使用することにより，学習者の誤りのさまざまな観点の分析を行えるようなデータ作成を目指した。

４．本コーパスの応用例

SST Corpus の公開に先立ち，１００件のデータがモニター・コーパスとして関係する研究者に利用されて具体的な研究・教育への応用例が検討された。今回のワークショップではその具体例をいくつか紹介したい。

（A）冠詞習得傾向の分析
　　冠詞習得エラーの分析による中間言語プロセスの記述
（B）英語学習者の発話データからの誤り抽出とその傾向
　　レベルとエラーの相関関係の分析による習得段階を特定する発達指標を探る
（C）習熟度判定
　　機械学習の手法を用いた習熟度判定の可能性を検討

これらの例を示しながら，具体的なSST Corpus の利用方法に関して参加者と議論を深め，かつ実際にデータを使っていただけるような十分な情報提供を行いたいと考えている。

基本参考文献：

先端技術移転加速型研究開発プロジェクト「適合型コミュニケーション技術の研究開発最終報告書」（平成１５年３月）通信・放送機構