日本人英語学習者の誤り分析

ーエラータグセットの構築を中心としてー

和泉絵美[†] 齋賀豊美[†‡] Ook Chung[‡] 井佐原均[†‡]

[†] 通信総合研究所
[‡] 通信・放送機構

1. はじめに

国際化・情報化が進む今日において，適切なコミュニケーションを図るために，言語障壁を克服することは重要な課題である．国際社会における公用語としての（様々な議論はあろうが）英語の習得は，言語障壁克服の近道のうちのひとつである．特に日本人は英語が苦手と言われる．そこで，日本人の英語能力を実際の言語データに基づいて客観的に評価し，「真の英語運用能力の習得のために，日本人は何を学ばねばならないのか」を明確しようと考えた．そこで得られた知見を，将来的にコンピュータによる学習支援環境の構築に役立てたいと考えている．本研究では日本人が特に苦手とする，英語を「話す」能力に注目し，日本人英語学習者発話コーパスを作成・分析している．

2. 本コーパスの概要

本コーパスの対象データは，一般に向けて実施されている SST (Standard Speaking Test) という1対1（受験者と試験官）のインタビューテストである．15分のインタビューの中で，受験者はイラスト描写・ロールプレイ・コマ割りされたイラストに対するストーリーテリングの3つのタスクをこなす．基礎データとしてその音声データを書き起こし，談話タグ（スピーカーターン・フィラー・言い直し・話者間のオーバーラップなど）の付与を行う．各発話データは SST 独自の基準に基づいて，9段階 (SST レベル 1〜9) に判定される．

既存の学習者コーパスは，書き言葉を対象としたものがほとんどで，本コーパスのような話し言葉中心のものは貴重と言える．また，学習者データが年齢や学習年数によって分けられているものが多いのに対し，9 段階の能力別に分けられた本データは，学習者の発達段階を観察する対象としてより信頼性が高い．

3. 学習者の誤り分析

本研究では，学習者の英語能力向上過程のモデル化に際し，学習者言語の分析のメインとして学習者の誤り分析を実施することにした．学習者コーパスの特徴で，他の一般コーパスのそれと最も異なるのは，学習者による誤りが含まれているという点である．学習者にとって適切な教育法を見出すための英語発話モデル構築において，学習者の誤り分析が有効であることは，第二言語習得研究の分野でも古くから言われてきたことである．本コーパスは９つの学習段階別に分かれているため，各レベルに特有の誤り，また各レベル間の傾向の比較によって，学習者の発達段階をより正確に記述するのに適していると言える．

学習者の誤りは，文法・語彙・音韻・語用・談話など多くのレベル・種類に亘る．さまざまな観点からなるべく多くの種類の誤りを分析するために，本研究では以下のような誤り分析手法を採用した．

まず，体系的な分類が比較的容易な文法的・語彙的誤りに対しては，独自のエラータグセットを構築し，人手でタグ付与を行う．エラータグについては 4. にて詳しく述べる．

エラータグ付与以外にも，誤り分析の一環として２つの補助的コーパスを作成し，学習者データとの比較を行う．まず一つは，学習者の英語を日

本語に訳し直した「日本語訳コーパス(back-translation corpus)」である．一般に“back-translation”とは，ある言語に翻訳された文を元の言語に訳し直し，最初の翻訳の精度を測る手法である．一方，今回は「学習者の誤った英語」を「誤った日本語」に訳し直すのではなく，たとえ誤りを含んだ英語でも，話者の意図をできる限り推測し，正しい日本語に訳すことを原則とする．この日本語訳データとエラータグ付きの学習者英語データを合わせて観察することによって，「日本人の英語は母語（日本語）の構造にどの程度引っ張られているか」「直訳的な表現をしがちな事柄は何か」といった，外国語習得に際して母語の干渉度合いを測ることが狙いである．学習者の誤りの原因は，母語の干渉によるものとそれ以外のものに大別されると言われており，母語の干渉による誤りを明確にすることで，今後の指導法の改善に役立つと考えられる．

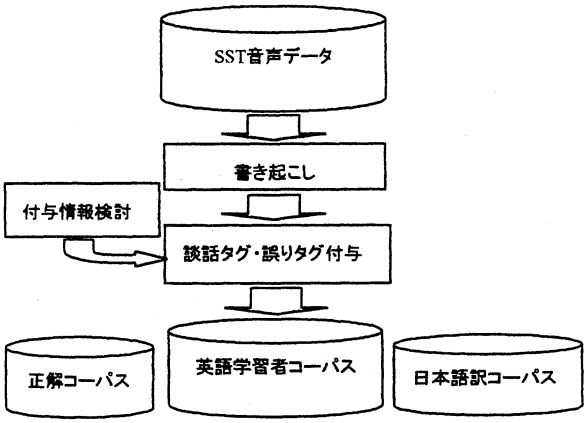
エラータグ付与は文法的・語彙的誤りといった形式的・局所的な誤りを対象としているため，会話の進め方や連語（コロケーション）の使用傾向といった，発話の「ネイティブらしさ＝（英語としての）自然さ」を測定することができない．そこで，こういったコミュニケーション全体に関わる誤りや不自然箇所の抽出をするために，2つめの補助的コーパスである「正解コーパス（＝英語母語話者の発話データ）」を作成する．正解コーパスの作成手順は以下の通りである．

① 「日本語訳コーパス」を元に，学習者の発話内容を簡単なシナリオで表す．
↓
② ①を英訳する．
↓
③ 英語母語話者にSSTインタビューを受けてもらう．ただし，英語母語話者は自由に発話するのではなく，上記シナリオに則った発言をする．
↓
④ ③で得られたデータを学習者データと同じ仕様で書き起こし，談話タグを付与する．

自由発話ではなく，シナリオに沿ってコントロールされた発話を収集する理由は，なるべく2コーパスの内容を近しくし，比較し易くするためである．同じトピックについての話の進め方の違い，使用する表現の違いなどを分析するのが主な狙いである．

コーパスの全体像は以下のようになる．エラータグ・日本語訳コーパス・正解コーパスによって，学習者の誤りをさまざまな観点から分析し，なるべく多くの種類の誤りを体系化することを目指す．

図1. コーパスの全体像



4. エラータグセットの構築

次に，本研究で独自にデザインしたエラータグセットについて説明する．先にも述べたように，誤り分析は古くから行われてきたが，そのほとんどはごく小規模のデータを対象にしたものであり，今回のような大規模データ中の誤りを計算機処理している研究は少ない．エラータグ付与についても，ごく一部の文法項目などについてのみ行われているが，さまざまな種類の誤りを網羅するようなタグセットは世界的にもまだ少なく，標準となるものも存在しない．本研究では，網羅性のあるタグセットを構築すべく，以下のような手順を踏んだ．

4.1 タグセットの構成

まず，文法・語彙に関わる誤りを以下のように

分類した．現在のカテゴリ数は37である．（表1）これからの分析段階で，カテゴリ間の統合・分割・新カテゴリの導入などを必要に応じて行い，より頑健な体系化を行う．

表1. 誤りの分類基準

レベル1	レベル2	レベル3
形態素	名詞	活用の誤り
	動詞	活用の誤り
	形容詞	活用の誤り
		数の誤り
	副詞	活用の誤り
	その他	和製英語
		その他
文法	冠詞	なし
	名詞	単複の誤り
		格の誤り
	動詞	主語-動詞の人称不一致
		形の選択誤り
		定形・不定形の誤り
		時制の誤り
		態の誤り
		否定形の誤り
		疑問形の誤り
		助動詞の誤り
	形容詞	数量を表す形容詞の誤り
		比較級・最上級の活用および用法に関する誤り
	副詞	位置の誤り
		比較級・最上級の活用および用法に関する誤り
	代名詞	なし
語彙構文	名詞	可算・不加算の使い分けの誤り
		補部の誤り
		従属前置詞の誤り
	動詞	補部の誤り
		従属前置詞の誤り
	形容詞	補部の誤り
		従属前置詞の誤り
	前置詞	補部の誤り
語彙	なし	なし
その他	余剰な語の使用	なし
	項目の脱落	
	語順の誤り	
	種類が特定できない誤り	
	誤りとは言えないが不自然	

※ 語彙レベルの誤りにはレベル2・3なし．

※ その他の誤りにはレベル3なし．

次に，上記の誤りの種類ごとに標識を与える．一つの誤りに対するタグの選択は，表1に従ってレベル1（大カテゴリ）→ レベル2（品詞など）→ レベル3（システム・ルールなど）の順で項目を組み合わせ，決定する．（一部レベル1，2までのものもある．）

各タグはXMLに準拠し，開始タグ <レベル1_レベル2_レベル3 crr="訂正候補"> …（エラー部分）終了タグ </レベル1_レベル2_レベル3> という構成になっている．例えば，文法レベル（レベル1）で，動詞（レベル2）で，時制（レベル3）の誤りには，以下のようなタグが付与される．

図2. エラータグの一例[2]

<g_v_tns crr="X"> … </g_v_tns>

レベル 1 2 3

レベル 1 ＝ 大カテゴリ

(e.g. g = grammatical errors)

レベル 2 ＝ 品詞など

(e.g. v = verb)

レベル 3 ＝ システム・ルールなど

(e.g. tense = tns)

crr ＝ 訂正候補

このような基準で，データ中の誤りに対して人手でエラータグを付与していく．一部自動抽出が可能な誤りもある（例えば，複数形の名詞の前に不定冠詞 “a” が付いている，など）が，ほとんどすべての誤りは人手による検証が必要となる．タグ付けは主に英語をよく知る日本人によって作業されるが，どうしても母語話者の判断が必要な場合は，母語話者との共同作業を行っている．[1]

このタグセットに則って，試験的に少量のデータを対象にエラータグ付与を行ったところ，SSTレベル1〜9で，エラーの数は比例することが分かった．SSTレベルの判定は，文法などの形式的な誤りを細かく見るよりも，むしろ「コミュニケーションとして適切かどうか」を重視して行われているが，それでもやはり，文法・語彙といった形

式的な部分を基盤として身につけることも，円滑なコミュニケーションを可能にする大切な1コンポーネントと言えるのかもしれない．

4.2 問題点

上記のようなきちんと体系化された誤り分類基準に則って実際のデータにエラータグ付与しようとすると，しばしば文法書的分類と実例のギャップによって作業がうまく進まないことがある．ひとつの誤りに対して複数のカテゴリへの振り分けの可能性があることは多く，作業者間での揺れが生じる．今後タグ付け・分析が進む中で生じる問題を検討し，随時エラータグセットの強化を行う予定である．また，こういった体系的な規範に属さない誤りとはどんなものか，の観察も行いたい．

5. まとめ

今後の計画として，本稿で述べたような誤り分析を中心に，データに基づいた日本人英語学習者の習得過程モデルの構築を進めたい．その成果を元に第二言語習得・英語教育分野の研究者とのディスカッションを通じて良い学習法の検討を行い，最終的には自然言語処理技術との協調によって，学習支援システムの構築を目指したい．

参考文献

[1] Dagneaux, S. 1998: *"Computer-aided error analysis"*, System 26 pp163-174

[2] Isahara, H., Saiga, T., Izumi, E.: 2001 *"The TAO Speech Corpus of Japanese Learner English, Error Tagging Manual ver.1.0"*