学習者コーパスにおける誤り情報付与
－規範的アプローチから記述的アプローチへ－

和泉 絵美†‡　　内元 清貴†　　井佐原 均†‡

†独立行政法人情報通信研究機構 自然言語グループ　　‡神戸大学大学院 自然科学研究科

{emi, uchimoto, isahara}@nict.go.jp

1. はじめに

近年，新たな学習者コーパスがいくつも構築され，学習者言語のモデル化や学習システムの開発など，その利用可能性について活発な議論が行われるようになった．しかし，それらの利用法を実践するためにどのような情報付与を行うべきなのか，具体的な提案はなされていない．我々は，日本語を母語とする英語学習者の発話コーパス・The NICT JLE (Japanese Learner English) Corpus [1] を構築し，そのデータに含まれる形態素・文法・語彙誤りを対象に，規範的に分類したエラータグに基づく誤り情報付与を行った．これは，学習者の基本的な言語機能に関する運用能力を測るのに有益である．しかし，果たしてその学習者の発話が聞き手に通じるか通じないか，つまり聞き手に正しくメッセージが伝わるかどうか，という学習者のコミュニケーション能力を直接的に測ることができない．

本研究では，この不足を補うために，現行のエラータグセットをどのように拡張すべきか検討する．具体的には，英語母語話者による添削データにおいて，どのような誤りが指摘されているか把握し，それらの誤りに対して現行のエラータグセットでどこまで情報付与が可能なのか，また新たに補うべき要素は何であるか考え，エラータグの拡張に繋げる．

2. 学習者コーパスにおける誤り情報付与

まず，The NICT JLE Corpus とそれ以外の学習者コーパス研究における誤り情報付与の現状をまとめる．

2.1. The NICT JLE Corpus のエラータグ

The NICT JLE Corpus のエラータグは，47 種類の形態素・文法・語彙誤りを分類したものである．この三種類の誤りを対象とした理由は二つある．一つめは，本コーパスの作成を開始した当時，コーパスの応用目的として検討していた，学習者言語の機能的な言語項目での熟達過程の分析および構造的な誤りを含んだインプットの機械処理を実行するために必要であったこと．二つめは，文法ルールというすでにある規範的分類を用いることができ，タグセット体系への組み込みに適性が高いと考えたからである．タグのフォーマットは，図 1 の通りである．

開始タグ　　誤り箇所　終了タグ

<n_num crr= “X”>　・・・　</n_num>

↑　↑　↑

品詞 文法ルール 訂正候補

図 1 The NICT JLE Corpus のエラータグ

タグはすべて XML 形式で，開始タグと終了タグで誤り箇所を囲む．タグの見出しには，品詞（図 1 では n = 名詞）と文法ルールの名称（図 1 では num = number，単数形・複数形の選択誤り）を，そして訂正候補（図 1 で crr= “X” で示されている部分）も提示される．内容を理解または推測でき，明らかな誤りであると判断できる箇所には，いずれかの種類のタグを付与し，理解も推測も不可能な箇所には，「理解不可能」を表す特別なタグを付与した．

2.2. 誤り情報付与を導入している学習者コーパス

The NICT JLE Corpus 以外にも，ICLE (International Corpus of Learner English) [2]，JEFFL (Japanese EFL Learner) Corpus [3]，CCL (Cambridge Learner Corpus) [4] などが，誤り情報付与を行っている代表的な学習者コーパスである．これらすべてにおいて，The NICT JLE Corpus 同様，あらかじめ分類した誤りをタグ化し，データに付与するというエラータグ付与形式を採用している．誤りの分類基準は，文法ルール上の規範的な分類を取り入れているケース（例：動詞の時制誤り，名詞の可算・不可算性の誤りなど）と，表層構造上の逸脱形式とその品詞を組み合わせたケース（例：形容詞の脱落，動詞の置換，接続詞の余剰など）のどちらかである．

2.3. 現在のエラータグの利点と問題点

このことから，学習者コーパスにおける誤り情報付与は，規範的な文法ルールや品詞体系を元に，予想される逸脱形式を分類することにより，エラータグを設計する方式が主流であると言える．この方式によるエラータグ付与の利点と問題点を，The NICT JLE Corpus や，2.2 で挙げた他の学習者コーパスを用いて行われた応用研究を通して検討する．

2.3.1. 利点

Tono [3]は，日本語母語話者の書き言葉英語のコーパスである JEFFL Corpus を利用して，日本語母語話者が英語の主要文法形態素をどのような順序で習得するのか，解明を試みた．その後，我々も The NICT JLE Corpus のエラータグ情報を用いて同様の調査を行い，Tono の結果と相関が高い結果を得たが，一部で，言語産出の手段の違い（話し言葉か書き言葉か）によって引き起こされると推測される差異を確認した[5]．また，Granger[6]は，ICLE のフランス語母語話者データにおけるエラータグ情報のうち，動詞の時制エラーについて調査し，単純現在形と単純過去形が最も誤りやすいという結果を示している．また，我々はエラータグ付きデータを学習データとして，機械学習に基づく自動誤り検出実験を行い，冠詞の脱落型エラーの検出において，50%の再現率と 76%の適合率を得た[7]．このように，

規範的に分類されたエラータグ情報は，文法を中心とした基本的な言語機能に関する運用能力の実態を把握する手助けとなる．このような情報は，文法指導法の改善を試みる教師や，学習者の文法運用能力を詳細に記述しようとする研究者にとって有益であり，文法・語彙的正確性を測ることに重点が置かれる試験を目指して英語を学ぶ学習者などに還元できると考える．

2.3.2. 問題点

しかし，英語を実際のコミュニケーションの場におけるメッセージ伝達のために習得したいと考えている学習者にとってはどうだろうか．完璧に正確な英語を習得することがなかなか困難である以上，学習者は，自分の英語が文法的・語彙的に正しいかどうかよりも，誤りがあろうとなかろうと，果たして通じるのか通じないのか，という点を知りたいのではないだろうか．例えば，The NICT JLE Corpus のエラータグ付けは，英語が堪能な日本語母語話者によって行われた．なるべく多くの文法・語彙レベルの誤りを細かく特定していくためには，発話意図を少しでも多く理解する必要があるため，日本語母語話者の話す英語を，英語母語話者よりもより理解できるであろう日本語母語話者が作業を行った方がよいという判断からである．そして，2.1 で述べたように，日本語母語話者でさえ理解できなかった箇所にのみ「理解不可能」のタグが振られている．この手法の欠点は，本来日本語母語話者以外の聞き手には通じないかもしれない発話も，理解されたものとして「理解不可能」以外のタグが振られることになり，果たしてその発話が，実際のコミュニケーションの場面で通じるのか通じないのか判断できないという点である．

このことから，まずは英語母語話者による，内容を理解できる箇所とできない箇所の仕分けをする必要があるのではないかと考えた．またこれは，作業者が英語母語話者なのかそうでないか，という問題だけではない．例えば，現行のエラータグの分類上は同じカテゴリに属する誤りでも，その誤りが発話のどの部分で起こるかによって理解に及ぼす影響は変わるかもしれない．このことから，現在の文法上の分類に加えて，もしその誤りが単なる文法的なものにとどまらず，談話的なレベルにまで及ぶものであればその情報を追加し，最終的にその誤りの致命性（メッセージ伝達に及ぼす影響）を表すための，誤りに対する重み付けをする必要があるのではないかと考える．

そしてもう一点，学習者にとって気になることは，自分の英語がどう聞こえているかという点ではないだろうか．つまり，誤りでもなく，内容も理解してもらえているようだが，果たして「こなれた」英語になっているだろうか，という点である．誤りもせず，自分の言いたいことを正しく伝えられるようになったら，次はより母語話者らしい (Native-like) 英語の習得を目指したいと思う上級レベルの学習者も多いだろう．現在のエラータグ付けでは，明らかな誤りのほかに，誤りではないがより良い言い回しがある箇所にも一律エラータグを付与しているため，誤りであるのか誤りではないが不自然な箇所なのか, 区別ができない．この点も改善が必要である．

3. エラータグ拡張に向けた調査

2.3.2 で述べた現行のエラータグの問題点を踏まえて，エラータグ拡張の具体的な方針を決めるために，以下のような調査を行うことにした．

1) 「通じる発話」と「通じない発話」の仕分けを行い，それぞれの誤りの性質の違いを見出す．
2) 誤りとは違う，「非英語母語話者らしさが感じられる不自然な表現」を選ぶ，更に，不自然さの理由には，どのような種類があるか調査する．

3.1. 英語母語話者による添削

まず，The NICT JLE Corpus のうち，エラータグが付与されていないデータを英語母語話者に自由に添削してもらった．これは，試験官一名，受験者（学習者）一名による 15 分間のインタビューテストの模様を書き起こしたもので，発話の内容は，自己紹介・イラスト描写・ロールプレイ・ストーリーテリングである．対象としたデータは，中級者から上級者の 15 インタビュー，総語数 17,068 語（うちフィラー1,799 語，言い淀み 1,186 語），総文数 1,657 文である．添削の単位は特に指定せず，単語・句・文単位で，そして文のつながりに関する修正の場合は複数文に亘って添削された．何らかの修正が施された箇所に対しては，表 1 に示す三種類のコメントのうちのひとつを付与してもらった．また，それ以外にも特記する事柄があれば，追記してもらった．（例：「この言い回しは，相手に対して失礼に聞こえる」など．）

コメント 1	誤りだが，発話内容の理解に支障はない
コメント 2	発話内容を全く理解できない
コメント 3	誤りではなく，発話内容を理解できるが，英語母語話者には不自然に聞こえる

表 1 三種類の添削コメント

添削を行った英語母語話者は，日本在住経験 14 年の 60 才台の英国人である．英語教師経験はないが，英国の大学の日本語関連学科での約 20 年の教授暦を持つ. 日本語に関する専門知識があり，日本在住経験も長いことから，日本語母語話者が話す英語に慣れている人物であるが，あくまでも，作業対象の学習者発話が一般的な英語母語話者に通じるか通じないか，なるべく客観的に作業してもらった．

その結果，添削されたのは合計 959 箇所で，表 1 のコメントごとの内訳は表 2 の通りである．

コメント 1	724
コメント 2	57
コメント 3	178
合計	959

表 2 添削箇所の数の内訳

959 箇所のうちのほとんどが理解には支障をきたさないというコメント 1 が振られている．誤りがある発話でも，一応の内容理解が可能なケースが意外と多いことが分かる．一方，全く理解不可能であるとするコメント 2 が振られた箇所は 57 箇所であった．2.1 で述べた形式で，最初に行った日本語母語話者によるタグ付与作業では，理解不可能とされたのは 167 インタビュー中 101 箇所であったのに対し，今回の添削ではわずか 15 インタビュー中で 65 箇所もが理解不可能とされたことを見ると，やはり日本語母語話者の話す英語は日本語母語話者によってより理解されやすいこ

とが分かる．誤りではないが不自然であるとするコメント3はコメント2よりも多い178箇所であった．

3.2. 添削箇所の特徴

添削された箇所の特徴を考察するために，訂正語や特記事項を元に，誤りや不自然さの種類分け（形態素・文法・語彙・談話，それぞれどのレベルであるか）を行った．各コメントの内訳は表3の通りである．

	コメント1	コメント2	コメント3	合計
形態素	6	0	0	6
文法	429	0	52	481
語彙	286	43	78	407
談話	3	14	48	65
合計	724	57	178	959

表3 コメントごとの誤り･不自然さの種類

全体数では，形態素・文法・語彙・談話レベルのうち最も多かったのは，文法に関わるものだった．しかし，そのほとんどはコメント1に属しており，コメント2には一つも属していない．文法誤りは実際の内容理解に致命的な影響を持たないことが分かる．全体数として二番目に多かったのは，語彙に関わるものだった．これも半数以上はコメント1に属しているが，一部コメント2にも属しており，誤りの内容によっては理解に大きく影響することが分かる．文の繋がりや省略，照応など談話に関するものは，全体数としては少ない．しかし，全体数65箇所のうち14箇所がコメント2に属しており，全体数に占めるコメント2の割合が，他のどの種類の誤りよりも高く，内容理解に与える影響の大きさが伺える．コメント2に属する談話エラーとコメント3に属するそれの違いは，コメント2の談話エラーのほとんどが定形表現に基づかない，文脈によって柔軟に語を構成することを学習者が要求されるような内容のものであったのに対し，コメント3で談話に関わるもののほとんどが連語表現に関するものである，という点であった．

3.2.1. コメント1が振られた箇所の特徴

コメント1は半数以上が文法エラーであった．そして，そのほとんどが主語と動詞の数・人称の不一致や，冠詞エラーなどの局所的なものであった．語彙エラーは286箇所あったが，その内容は，ごく近い語義を持つ語彙間での誤用や，*be* 動詞など明示的な意味を持たない語彙の脱落といった，あまり深刻でないものがほとんどであった．

3.2.2. コメント2が振られた箇所の特徴

コメント2が振られた箇所のうち，特に目立った特徴を六点挙げる．

① 話題が飛躍し過ぎる発話（談話レベルの誤り）
ex) *I've been to the restaurant is first. I took lunch. The curry the restaurant serves is very much, so I was surprised and I'm now a little sleepy.*
推測される意味[1]：*初めて行ったレストランでランチにカレーを食べた．そのカレーはとても量が多く，びっくりした．(お腹がいっぱいで) 今少し眠い.*
→「お腹がいっぱいで」の部分が抜け落ちているため，なぜ眠いのか不可解に聞こえると推測される．

② *and, so, but, because* などの前後の繋がりが不明（談話レベルの誤り）

③ 代名詞・指示詞の照応が不明（談話レベルの誤り）

④ 語彙選択の誤り（語彙レベルの誤り）
コメント2では *be* 動詞など明示的な意味を持たない語彙の誤りがほとんどであったのに対し，コメント2では，発話の意味に大きく関わる語彙に関するものが多かった．

⑤ 主語･述語･目的語などの文の意味を担う語の脱落（文構造の誤りとも言えるが，語彙レベルの誤りと取る方が妥当と考える.）

⑥ 和製英語・直訳表現（語彙レベルの誤り）

3.2.3. コメント3が振られた箇所の特徴

コメント3が振られた箇所のうち，特に目立った特徴を六点挙げる．

① 冗長な表現（談話的不自然さ）
ex) T: *Can I call you Hanako?*
L: *Yes, please call me Hanako.*
better[2] → *Yes, please do.*

② (場面によっては) 社会言語学的に好ましくない表現（談話・社会）（談話（語用）的不自然さ）
ex) *What?*
better → *I beg your pardon?*

③ 言葉足らずによって細かいニュアンスが欠落したり，ぶっきらぼうに聞こえる表現（談話（語用）的不自然さ）
ex) T: Have you been busy lately?
L: No.
better → *No, not really.*

④ ③とは反対に，大げさな表現 (overstatement)（談話（語用）的不自然さ）
ex) （特に病み上がりなどということもなく，通常の挨拶の中で）T: *How are you?*
L: *I'm very fine.*
better → *I'm fine.*

⑤ 他の語彙の方が内容をより的確に表現できる場合（語を組み合わせるよりも，他の一語もしくは慣用句などの表現で表した方がよい場合，またその反対.）（語彙的不自然さ）
ex) *To go to high school in the mainland, I went out of the island.*
better → *... I left the island.*

⑥ また，表3からも分かるように，文法に関わるものがいくつかあった．それらは，冠詞，動詞の時制・相，語順など，文脈によって正用法が複数存在しうる項目であった．

3.3. 現行のエラータグでのタグ付け

次に，現行のエラータグを用いて，添削箇所へのタグ付けを試み，どこまで対応可能か，また新たに補うべき要素は何であるか検討した．結果としては，ほとんどすべての

[1] これは，日本語母語話者である第一筆者による推測である．文単体ではなく，前後関係も含めて推測した結果である．

[2] "better →" に続く表現は，添削者がより好ましいとして挙げたものである．

添削箇所に対して現行のエラータグを「ただ単純に付与する」ことは可能であった．その理由は二つある．一つは，現行のエラータグは誤り部分を訂正候補に置き換えることによって目標文へ復帰させる形式 (reconstruction) を取り，すべての品詞の置き換え・追加・削除ができる仕組みになっているからである．二つめは，談話の構成は文法的・語彙的選択と繋がっているため，談話を記述する用語は文法用語と必ずしも衝突しない[8]ためである．

しかし，このようなタグ付けでは，談話エラーを文法エラーのような表層構造上の逸脱として引き下げているに過ぎず，その誤りの本質を表せてはいない．1970 年代以降の第二言語習得研究における誤り研究の中心人物である Corder[9]が指摘しているように，学習者言語は，たとえそれが表面上逸脱のない発話だったとしても，母語話者と同じ言語体系を学習者が習得した証拠にはならない．つまり，その発話が生じた状況に意味的に関連していなければならないということである．表面上正確であるかどうかよりも，発話の場である状況のみが，その発話が誤りであるかどうかを決める鍵を持つということである．

三種類のコメントが振られた誤りごとに，エラータグでどこまで対応できているかというと，コメント 1 の誤りは，形態素･文法･語彙レベルの誤りがほとんどであったので，現行のエラータグの分類でその誤りの本質を表すことができた．コメント 2 の誤りは，談話レベルのものが多く，現行の分類ではその本質をほとんど表せていない．コメント 3 が振られた「不自然な表現」も，談話レベルの逸脱が原因であるものが多いが，3.2 で述べたように，連語表現のような定式化が比較的容易なタイプが多く，現行のエラータグでは，コロケーションの誤りを示すタグで対応できた．今後は，「誤り」と「誤りではないが不自然」を区別する必要があることと，なぜ不自然なのかという情報も追加すべきと思われる．

4. エラータグ拡張の方針

調査の結果，エラータグ拡張の四つの方針を定めた．

1) 談話レベルの誤りを細分類化し，エラータグセットに導入する．
2) 誤りではないが不自然な表現を指摘するタグを追加する．不自然さの原因も分類し，属性として追加する．
3) 語彙レベルの誤りのうち，内容理解に支障をきたすものとそうでないもの，それぞれの傾向を更に広く調査し，語彙誤りの致命度に関する情報の追加が可能かどうか検討する．
4) 誤りの言語学的レベル（形態素・文法・語彙・談話のいずれのレベルであるか．）を明示化する．

1)，2)，3) を行うには，談話エラーと語彙エラーの下位分類を行う必要がある．すでにいくつかの研究[10][11]において，具体的に非母語話者の談話エラーの分類が行われているが，広く認識されているような定まった分類はまだ存在しない．母語話者には見られない非母語話者特有のエラーの存在や，対象とする言語データの内容（発話手段・話題・状況など）が異なると，出現するエラーの種類にも可変性が生じることが予想される．形態素・文法・語彙エラーは，予想される逸脱形式に則って，あらかじめ規範的に分類することができたが，談話エラーの分類は，今回の調査で使用したような英語母語話者による添削データを元に，エラーの実態を記述的に調査することから始めるべきであると考える．

また，誤りの言語学的レベルによって内容理解に及ぼす影響が大きく違うことが分かったが，4) のように誤りのレベル情報を追加すれば，現在のエラータグデータでは不可能だった，発話の通じやすさ，つまり学習者のコミュニケーション能力に関する分析を行う際に大変有効な情報となるだろう．しかし，エラーによっては，文法・語彙・談話レベルのうちどのレベルに属するのか見極めるのが難しいものもいくつかあると予想され，実際のタグ付与作業はこれまで以上に慎重に進めなければならない．

5. まとめ

本稿では，The NICT JLE Corpus のエラータグを，学習者のコミュニケーション能力を分析できるようなものへと拡張することを目的に，英語母語話者による添削データを元に，誤りの種類と内容理解への致命性の関係に関する調査を行った．その結果，発話の通じやすさや発話の自然さに多大な影響を及ぼす談話エラーを新たな対象とするべきであることが分かり，エラータグの拡張を具体的にどのように進めるべきであるのか見出すことができた．今後は，更に添削データを収集し，談話エラーに関してデータに基づいた記述的な細分類化を行い，エラータグ体系への組み込みを実施する予定である．

参 考 文 献

[1] 和泉絵美，内元清貴，井佐原均. (2004). 『日本人 1200 人の英語スピーキングコーパス』 東京：アルク.

[2] Granger, S., Dagneaux, E., & Meunier, F. (Eds). (2002). *International Corpus of Learner English*. Brussels: UCL Press.

[3] Tono, Y. (2002). The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison Approach, Unpublished Ph.D. Thesis, Lancaster University, UK.

[4] http://uk.cambridge.org/elt/corpus/clc.htm

[5] Izumi, E., & Isahara, H. (2004). Investigation into language learners' acquisition order based on the error analysis of the learner corpus. In Proceedings of Pacific-Asia Conference on Language, Information and Computation (PACLIC) 18 Satellite Workshop on E-Learning, Japan. (in printing)

[6] Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In Hasselgard, H., & Oksefjell, S. (Eds). *Out of Corpora*. (pp. 191-202). Amsterdam: Rodopi.

[7] Izumi, E., Uchimoto, K., & Isahara, H. (2004). The overview of the SST speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. In *Proceedings of Language Resource and Evaluation Conference (LREC) 2004*, Portugal, 1435-1438.

[8] McCarthy, M. 著. 安藤貞雄他 訳. (1995). 『語学教師のための談話分析』 (pp. 44-128). 東京：大修館書店.

[9] Corder, P. 著. 武田良一他 訳. (1981).『中間言語入門』 東京：三修社.

[10] James, C. (1998). *Errors in Language Learning and Use: exploring error analysis*. (pp. 129-172). London: Longman.

[11] 宮田学（編）. (2002). 『ここまで通じる日本人英語　新しいライティングのすすめ』 (pp. 30-50). 東京：大修館書店.