翻訳教育での利用を意識した翻訳エラー分類体系の再構築

豊島 知穂　　田辺 希久子　　藤田 篤　　影浦 峡
関西外国語大学　神戸女学院大学　情報通信研究機構　東京大学

1　はじめに

翻訳の産出過程では一般的に，下訳に対するレビューをふまえて修正が行われる．翻訳学習者にとっても，学習者が自らの翻訳におけるエラーを指摘され，それをふまえて修正を行う，という機会を繰り返し持つことは有益である．特に翻訳教育の現場においては，翻訳に関わる概念などの深い理解のみならず，教員から学習者への教授の円滑化，教員に対する信頼の担保のために，翻訳の質やエラーの評価を一貫して行うための道具立てが不可欠である．これまでに，翻訳の際に生じる種々のエラーを包括的にカバーするような分類体系についての検討が行われてきた (2 節を参照されたい)．しかし，個々のエラーを分類するための手続きについては検討が不足しており，一貫した分類が容易ではない．

我々は，特に共同翻訳に関する翻訳教育を想定して，インストラクタと翻訳学習者の間，および翻訳学習者間で翻訳に関する知識を共有し，学習を効率化するための方法論について検討している．例えば，学習者の翻訳における個々のエラーを一貫して分類するための基準として，Babych ら [1] による分類体系に基づいて，決定木およびエラーの典型例および境界例を列挙した事例集を作成した．本稿では，それらの作成手順と新たに考慮した点，Toyoshima ら [3] が収集した英日翻訳結果におけるエラーの分類実験を通じて明らかになった課題について述べる．

2　先行研究・関連研究

翻訳におけるエラーの判定基準は，翻訳の用途や目的によって異なる [2]．同様に分類の粒度や優先すべきエラーの種類なども異なりうる．学習者の翻訳の評価や学習者に対するフィードバックなどの用途を想定して作成されたエラーの分類体系としては，MeLLANGE プロジェクトにおける 3 階層 44 カテゴリからなるものがある [2]．Babych ら [1] が開発した翻訳学習者向け共同翻訳プラットフォーム MNH-TT では，この分類体系を単純化した 2 階層 16 カテゴリの体系を用いている[1]．Toyoshima ら [3] は，学習者による英語記事 20 件の日本語訳に対して MNH-TT の体系に基づいてエラーを付与し，この体系が英日翻訳におけるエラーの分類・分析に資することを確認した．ただし，いずれの研究においても，個々のエラーを分類する手続きについては検討されておらず，一貫した分類が容易ではない．

翻訳エラーの分類体系には他にも，欧州の Quality Translation LaunchPad で開発された Multidimensional Quality Metrics (MQM)[2]，Translation Automation User Society (TAUS) で開発された Dynamic Quality Framework (DQF)[3] などがある．これらはいずれも，プロの翻訳者が産出する翻訳や機械翻訳の出力[4]に対する多面的かつ詳細なエラー分析を目的として設計されているため，翻訳教育で採用するには複雑すぎると思われる．

表 1: エラーの大まかな優先度.

Lv 1	訳が未完成である
Lv 2	起点言語文書の要素に対して過不足や誤解がある
Lv 3	目標言語の文法的・統語的な問題がある
Lv 4	目標言語文書に質的な問題がある
Lv 5	納品・公表するプロダクトとしての問題がある

3　エラーの分類基準の作成

我々は，Toyoshima ら [3] と同様に MNH-TT におけるエラー分類体系 [1] に沿って，所与のエラーを一貫して分類するための決定木[5]および事例集を作成した．なお，分類対象のエラーは，(a) 目標言語文書上のエラー箇所に，(b) 具体的な修正案または大まかな修正方針が指定されている (つまり検出済) と仮定する．

まず，Toyoshima らが収集したエラー事例から，各エラーカテゴリの典型例を抽出した．あわせて，各カテゴリの判定時に参照される情報 (起点言語文書，目標言語文書，用語集，ブリーフ) を整理するとともに，実務翻訳におけるレビュープロセス，各カテゴリの深刻さをふまえて，翻訳教育における大まかな優先度を表 1 のように定めた．Lv 1 と Lv 2 は起点言語から目標言語への翻訳の正確さ，残りの 3 つのレベルは目標言語における流暢さ・適切さの観点でのエラーである．

次に，Toyoshima らが収集したエラー事例の一部 (3 文書分) に対して，(i) 上記の優先度を考慮しながら複数名がエラーカテゴリを付与し，(ii) 分類が一致しなかっ

[1] Babych ら [1] は “revision categories” と呼んでいるが，分類対象は修正処理内容ではなく，あくまでエラーそのものである．

[2] http://www.qt21.eu/launchpad/content/training/

[3] https://evaluate.taus.net/evaluate/dqf/dynamic-quality-framework

[4] 言語対やドメインによっては機械翻訳の出力が下訳として使うことができるレベルになってきたため．

[5] 教育者・学習者間のコミュニケーションにおいて個々のエラーの分類の判定理由を明確に説明する必要があるため．MQM でも，エラーの分類のために決定木が作成されている．

表 2: エラーの分類体系: ラベルの色は MNH-TT における中分類との対応を表す．緑: content，青: lexis，赤: grammar，黄: text．

Lv 1	**未完成**
X4a	未翻訳
X6	曖昧さ未解消
Lv 2	**誤訳**
X7	用語の訳出エラー
X1	原文内容の欠落
X2	原文にない要素の付加
X3	原文内容の歪曲
Lv 3	**目標言語の文法的または統語的な問題**
X8	コロケーションのエラー
X10	前置詞や助詞のエラー
X11	活用のエラーや数・性などの不一致
X12	綴りエラー・誤変換
X13	句読法に関するエラー
X9	その他の文法的・統語的エラー
Lv 4	**目標言語文書の質の問題**
X16	結束性違反
X4b	直訳調
X15	表現のぎこちなさ
Lv 5	**納品・公表に際しての問題**
X14	レジスタ違反

た事例について議論を行う，という手続きを繰り返し，判定基準 (分類の決定木における設問) を明確化しつつ，境界例を蓄積した．

このようにして得たエラーの分類体系を表 2 に，各エラーを分類するための決定木を図 1 に示す．MNH-TT の分類体系からの主な変更点は次の 4 点である．

【X4 原文表現の押しつけ】の細分化: MNH-TT における【X4 原文表現の押し付け】は，未翻訳のエラーと直訳すぎるエラーの両方を含むが，これらは深刻さが異なる．そこで，Lv 1 のエラー【X4a 未翻訳】と Lv 4 のエラー【X4b 直訳】に分けた．

【X5 目標言語表現の押し付け】の廃止: MNH-TT では行き過ぎた工夫によって起点言語文書の意味が通じなくなった場合に，【X5 目標言語表現の押し付け】とするが，翻訳者本人でなければ行き過ぎた工夫と原文内容の誤解を正確に見分けることはできない．そこで，起点言語文書の意味が正しく伝わらない場合はまとめて【X3 原文内容の歪曲】と呼ぶことにした．

【X7 用語の訳出エラー】の優先: 下記の例 (1) のように，用語の対訳は事前に用意された用語集に従う必要がある．翻訳実習においても，翻訳に着手する前に用語集を作成するのが一般的である．学習者による問題の回避・解消をうながすために，起点言語文書と目標言語文書の両方にかかわる Lv 2 のエラーの中でも最も優先的に指摘することにした．例えば，【X12 綴りエラー・誤変換】，【X14 レジスタ違反】というカテゴリが存在するが，用語の綴りエラーや初出の用語に略称をつけるというレジスタ依存のルールに対する違反は，【X7 用語の訳出エラー】とする．

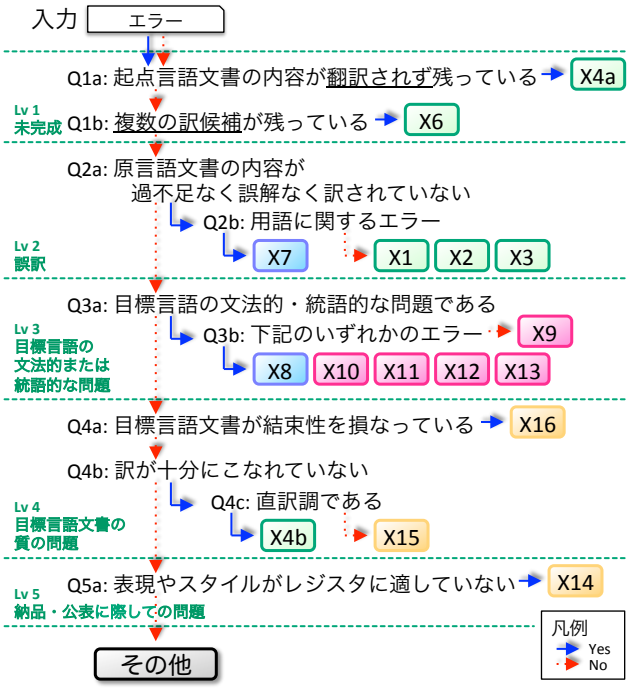

図 1: エラー分類のための決定木．

(1) ST: UCLA's Civil Rights Project has been tracking national trends.
TT: UCLA の市民権（⇒ 公民権）プロジェクトが合衆国全体の傾向を追っている。

【X15 不自然なスタイル】の呼称変更: MNH-TT に従い，目標言語文書における表現のぎこちなさ・冗長さを【X15 不自然なスタイル】と呼んでいた．しかし，エラー判定の際に，「スタイル」という名前に影響され，英数字の半角・全角の使い分けや，下記の例 (2) のような引用末尾の句点の打ち方のエラーなどの【X14 レジスタ違反】を誤ってここに含めてしまいがちだった．そこで呼称を【X15 表現のぎこちなさ】に変更した．

(2) ST: The bout, said Forsyth, "would add to the distraction, not only of the police, but of just people in general."
TT: フォーサイス氏は、「この試合は、警察だけでなく、まさに世間一般の気を逸らしてくれるでしょう。」（⇒ 」）と語った。

表 3: 作業者による分類結果の再現率，精度，F 値: 0.8 以上の値を緑色，0.6 以上 0.8 未満の値を黄色で示す.

エラー	著者らの分類	作業者 A					作業者 B				
		回答数	正答数	再現率	精度	F 値	回答数	正答数	再現率	精度	F 値
X4a	3	6	2	0.67	0.33	0.44	2	0	0.00	0.00	0.00
X6	0	0	0	0.00	0.00	0.00	0	0	0.00	0.00	0.00
X7	48	23	17	0.35	0.74	0.48	39	23	0.48	0.59	0.53
X1	62	91	60	0.97	0.66	0.78	91	56	0.90	0.62	0.73
X2	24	19	14	0.58	0.74	0.65	25	18	0.75	0.72	0.73
X3	272	260	208	0.76	0.80	0.78	203	175	0.64	0.86	0.74
X8	26	16	7	0.27	0.44	0.33	4	3	0.12	0.75	0.20
X10	28	33	24	0.86	0.73	0.79	23	17	0.61	0.74	0.67
X11	9	24	9	1.00	0.38	0.55	10	6	0.67	0.60	0.63
X12	15	18	14	0.93	0.78	0.85	15	13	0.87	0.87	0.87
X13	18	24	13	0.72	0.54	0.62	22	12	0.67	0.55	0.60
X9	9	16	6	0.67	0.38	0.48	16	6	0.67	0.38	0.48
X16	33	24	19	0.58	0.79	0.67	24	19	0.58	0.79	0.67
X4b	108	68	49	0.45	0.72	0.56	128	66	0.61	0.52	0.56
X15	17	19	3	0.18	0.16	0.17	18	6	0.35	0.33	0.34
X14	24	39	12	0.50	0.31	0.38	12	1	0.04	0.08	0.06
その他	0	16	0	0.00	0.00	0.00	64	0	0.00	0.00	0.00
Lv 1 のいずれか	3	6	2	0.67	0.33	0.44	2	0	0.00	0.00	0.00
Lv 2 のいずれか	406	393	337	0.83	0.86	0.84	358	305	0.75	0.85	0.80
Lv 3 のいずれか	105	131	85	0.81	0.65	0.72	90	62	0.59	0.69	0.64
Lv 4 のいずれか	158	111	80	0.51	0.72	0.59	170	98	0.62	0.58	0.60
Lv 5 のいずれか	24	39	12	0.50	0.31	0.38	12	1	0.04	0.08	0.06

4 エラー分類の一貫性

作成した決定木および事例集 (のべ 52 例) に基づいて，Toyoshima ら [3] が収集した全エラー事例を改めて分類した．分類作業に先立って新たなエラーを発見したため，分類対象は 763 事例となった．

まず，分類基準の作成に関わっていない 2 名[6]にエラーの分類作業を依頼した．その際，決定木および事例集に従うように指示したが，特定の用語集を参照するようには指示しなかった．2 名の分類結果の一致率は 68% (515/763)，Cohen の κ は 0.618 だった．分類先のカテゴリ数 (表 2 の 16 種類と「その他」) を勘案すると，これらの値は決して低くはない．ただし，Toyoshima らの結論に反して，作業者 A が 18 事例，作業者 B が 71 事例と，多数のエラーを「その他」と判定していた．また，作業者 2 名の分類結果が一致していても，それが我々の判断とは一致しない事例も見つかった．

そこで，作業者の分類の適否，分類を困難にする要因を分析するために，我々自身もエラーを分類した．著者のうち 2 名による分類結果の一致率は 86% (658/763)，Cohen の κ は 0.830 だった．この作業を通じて，さらにエラーを発見したり，一部の事例についてエラーの箇所あるいは修正案を変更したりしたため，分類対象は 781 事例となった．その後，議論をふまえて不一致の事例の分類を定め，全体の見直しも行った．

エラーの箇所および修正案に変更がなかった 696 事例について，著者らの分類結果を正解とみなして，作業者の分類の再現率，精度，F 値を計算した．結果を表 3 に示す．F 値が 0.80 以上となったのは【X12 綴りエラー・誤変換】のみだった．最も多く生じていた【X3 原文内容の歪曲】の分類精度は比較的高かったが再現率が低く，その次に多かった【X4b 直訳調】については両作業者とも F 値が 0.56 に留まった．

頻出した分類誤りを表 4 に示す．最も多かったのは【X3 原文内容の歪曲】と【X4b 直訳調】，すなわち原文内容を正しく理解できているかどうか (設問 Q2a) の判断の不一致によるものだった．例えば，我々は下記の 2 例を “risk”，“focus” に対する訳語の選択誤りとして X3 に分類したが，作業者は 2 名とも X4b に分類した．

(3) ST: We routinely ask of people to take on jobs that risk their families.
TT: 常日頃から、市民に対し家族を危険にさらす（⇒ 犠牲にしかねない）仕事につくよう求めてきました。

(4) ST: I asked him about the FBI’s focus on animal rights and environmental groups.
TT: FBI が動物保護団体や環境保護団体に焦点を置いている（⇒ を標的にしている）ことに関して彼に尋ねた。

[6]日本語を母語とし，英語起点言語文書および決定木・事例集の内容を十分に理解でき，用語集，ブリーフ，テキストタイプなどの翻訳に関わる概念に関する知識を有する者．

表 4: 頻出した分類誤り.

正解	誤答	Lv	作業者 A	作業者 B	合計
X3	X4b	異なる	12	44	56
X4b	X3	異なる	29	12	41
X3	X1	同じ	13	12	25
X8	X3	異なる	7	7	14
X7	X3	同じ	11	3	14
X3	X11	異なる	11	3	14
X3	X9	異なる	6	7	13
X7	X1	同じ	7	5	12

逆に，例えば例 (5) における 2 つのエラーを，我々は X4b に分類した．つまり，原文内容を誤解なく目標言語で表現することはできていると考え，設問 Q2a について “No”，文脈を考慮して修正が必要であると考え，設問 Q4b について “No”，ぎこちなさの要因は直訳調であると考え，設問 Q4c について “Yes” と判断した．一方，作業者 A はこれらを X3 に分類した．

(5) ST: I met a young college student and asked her .... She said proudly, “I emailed my professors and said I won’t be in class today; I’m going to get an education.”
TT: 私は若い大学生（⇒ 女学生）に会い、...を聞いた。彼女は堂々と「私は教授にもメールをして今日授業には出ないことを伝えました。私は教育を受けるつもりです（⇒ 今日はここで学びます）」と言った。

3 番目に多かった誤答は，次の例 (6)，(7) のような【X3 原文内容の歪曲】の例を【X1 原文内容の欠落】と分類したものだった．

(6) ST: The Parks moved to Detroit.
TT: パークス（⇒ パークス一家）はデトロイトへ引っ越した。

(7) ST: We have filed a class action for approximately a hundred sailors.
TT: およそ 100 人の兵士（⇒ 海軍・海兵兵士）のための集団訴訟を起こした。

エラー箇所と修正案の表層的な包摂関係のみに基づいて X1 に分類してしまった可能性がある．【X7 用語の訳出エラー】を【X1 原文内容の欠落】と誤答しているのべ 12 事例のうち 11 事例もこれに類する．いずれも同じ Lv 2 のエラーであるため，翻訳の質の大まかな評価が目的であれば精密な区別は必要ないかもしれない．一方，翻訳教育においては，学習者に対する説明の一貫性を担保するため，上記の例はやはり【X3 原文内容の歪曲】と呼ぶべきであろう．

その他には，【X7 用語の訳出エラー】や【X8 コロケーションのエラー】を，誤って【X3 原文内容の歪曲】に分類している事例が多かった．用語の認定基準やコロケーションの定義が明確でなかったことが原因であると思われる．起点言語文書から作成した用語集を参照するように指示する，コロケーションについては事例集を増補する，などの対策が考えられる．

エラーカテゴリによっては，判定の際に起点言語文書と目標言語文書全体を参照する必要があるが，局所的な情報のみに基づいて判定してしまったと思われる事例もあった．作業指示の改善が必要である．

5 おわりに

本稿では，学習者が作成した翻訳におけるエラーを一貫して分類するために作成した，決定木および事例集について述べた．また，781 件のエラーの実例の分類を通じて，判断の揺れが生じやすい箇所を明らかにし，その原因について考察した．

翻訳学習者は，様々なテキストタイプの文書を用いて翻訳技術を学ぶ．我々も，今回対象としたジャーナリズム記事 [3] のみならず，多様な文書の翻訳実習を通じて，今回作成した決定木および事例集をブラッシュアップする予定である．一方で，学習者に対する短時間でのフィードバックおよび教員の負荷の削減のために，エラーの自動検出にも取り組みたい．例えば，【X14 レジスタ違反】の一部については表層的なルールで精度よく検出できる．その他のエラーの検出については，機械翻訳に対する語レベルの Quality Estimation に関する試みが参考になる．

エラーの分類体系は，学習者のエラー傾向 [3] や学習過程における傾向の変化 [4] の分析にも有用である．別途の報告を予定している．

謝辞: 本研究の一部は科研費基盤研究 (A)（課題番号：25240051，代表：影浦峡）の支援を受けた．

参 考 文 献

[1] B. Babych, A. Hartley, K. Kageura, M. Thomas, and M. Utiyama. MNH-TT: a collaborative platform for translator training. In *Proceedings of Translating and the Computer 34*, 2012.

[2] A. Secară. Translation evaluation: A state of the art survey. In *Proceeding of the eCoLoRe/MeLLANGE Workshop*, pp. 39–44, 2005.

[3] C. Toyoshima, K. Tanabe, A. Hartley, and K. Kageura. Error categories in English to Japanese translations. In *Proceedings of the 21st Annual Meeting of the Association for Natural Language Processing*, pp. 1076–1079, 2015.

[4] 山本真佑花, 田辺希久子, 藤田篤. 翻訳学習者の学習過程におけるエラーの傾向の変化. 言語処理学会第 22 回年次大会発表論文集, 2016. (in this proceedings).