外国語母語話者が作成する日本語技術文書を対象とした訂正履歴の分析

鄭育昌　長瀬友樹
株式会社富士通研究所　スピーチ＆ランゲージテクノロジー研究部
{cheng.yuchang, nagase.tomoki}@jp.fujitsu.com

1　はじめに

インターネットが地球規模で使えるようになり、国境を越えてSNSやWebサービスを利用する人が増えている。企業のみならず社会全体のグローバル化が急速に進んでおり、日本人が外国語の文章を書いたり、外国人が日本語の文書を書いたりする機会が増加している。ところが、外国語母語話者が正確で自然な日本語で作文することは簡単ではない。たとえば、日本語を長年学習して日本語検定1級を取得した人でも日本人では決して間違わない誤りを犯すことがしばしばある。昨今、SNSやブログで不自然な日本語に遭遇することがあるのは、これらが外国人によって執筆された文章であることが考えられる。

本稿では、中国人の執筆した日本語文書の訂正履歴を分析して、日本語文書の誤用パターンの分類を行い、各分類項目別の発生頻度の分布を明らかにする。そのうえで、事例に基づいて外国人にとって苦手な日本語表現について考察を行う。日本語の誤用に関する先行研究として、日本語の学習者によるエッセイ、作文など文芸作品での誤用に注目した研究[南保ら, 2007]があるが、本研究では技術文書を対象として分析をおこない、技術文書に特有の誤用についても言及する。

2　技術文書の訂正履歴の概要

本研究で分析対象とした日本語文書の訂正履歴は、中国語を母語とする人が書いた技術文書を日本人が校正したときの作業記録である。下記のようなデータを最小単位として含んでいる。

1)　A システムマニュアル.docx　pp.5　10 行目
今回テストする時に 6009 に指定した。
→今回テストする時に 6009 を指定した。

上記の一行目は当履歴に対応する文書ファイル名および位置の情報である。二行目は元文書における記述、三行目（矢印の行）は人手による校正した文である。本例では、校正箇所は「に」と「を」の助詞変更１箇所のみである。１文中に複数の校正箇所を持つ場合もある。

表 1 に訂正履歴の概要を示す。中国語を母国語とする人が作成した技術文書を対象として日本人が誤りを指摘した履歴であり、９千件超えの履歴を含んでいる。表 1 に、本研究で使用した訂正履歴データの概要を示す。

表 1: 訂正履歴データの概要

元文書数	395
執筆者人数	10 人
校正箇所数	9644

3　誤り種類の定義と分類作業

3.1　技術文書と技術文書の校正

技術文書の訂正履歴を分析するため、まず本研究の対象である技術文書の特徴を明確化する必要がある。技術文書とは、技術者が自己の技術的見解を示すための文書であり、一般的には論文、仕様書、マニュアル、報告書、操作手順書などを指す。技術文書は一般の新聞記事やビジネス文書と異なり、「読者にとって必要な技術情報をわかりやすくしかも効率的に伝達すること」が目的である[浅岡, 2006]。そのため、技術文書の校正には、日本語の言葉使いが文法的な正確さのみならず、技術文書に相応しい表現の校正も求められている。

一般的な技術文書の校正のチェックリストを以下にあげる[浅岡, 2006]（一部抜粋）：

用字と表記の校正
-誤字、脱字、余字を確認
-当て字、俗字を使わない
-漢字、ひらがな、送り仮名を確認
-単位、量と数字の表記
用語の校正
-難解な専門語を使わない
-用語を統一
文構造の校正
-文構造に問題ないか
-句読点が適切か
-修飾語の順序が適切か
-助詞の使い
-出だしと係り結び
-能動と受動を区別
-不適切な中止法を使わない
文章表現の校正
-文の長さを注意
-くどい説明を使わない
-足りない情報を追加

上記のチェックリストは日本人の執筆した日本語の技術文書を対象に「日本語の誤り」と「技術文書として不適切な表現」の内容を含んでいる。外国語母語話者が作成した技術文書校正する場合、外国人

特有の日本語の誤用についても校正のチェックリストとする必要がある。次節では上記のチェックリストを参考にして、外国人の日本語誤用の観点から訂正履歴の分類を試みる。

3.2 誤り分類の定義

3.1 節の議論により、訂正履歴の誤り分類は「日本語の正しさ」と「技術文書の適切さ」を同時に考慮すべきである。外国語母語話者による日本語誤りの分析については、[大木ら, 2011]の研究があるが、小規模の技術文書コーパス（6 文書）を対象に分析したものであり、「技術文書としての適切さ」が考慮されていなかった。

我々は訂正履歴の分類定義に「技術文書の適切さ」と「日本語の正しさ」を表現するため、3.1 節の議論と[大木ら, 2011]の調査結果を参考し、校正箇所の誤りを分類するタグを定義した（表 2）。まず、校正箇所の誤りに関して、文書の意味理解に及ぼす影響によって４つのカテゴリを定義する。さらに各カテゴリを誤りの種類によって細分化する。表 2 に誤り分類の種類と定義をまとめる。次節で述べる誤り分類作業はこの定義にしたがって誤り箇所に対するタグ付けを行う作業である。

カテゴリは誤りが文書の理解に与える影響を基準に４つにわけ、カテゴリ１，２は「文書の内容を理解するに大きな支障がある」の誤り分類であり、カテゴリ３，４は「技術文書に相応しくない表現」の誤り分類である。各カテゴリに対して、以下に詳細を述べる。

カテゴリ 1：表記、言葉の理解

カテゴリ１は表記と言葉の理解と処理に影響が出る誤りである。このカテゴリの誤りは文書内容の理解に大きく影響し、構文解析などの処理にも支障が出る。例えば、以下の事例はカテゴリ 1 の分類に当たる：

A) プロ***バ***ティ → プロ***パ***ティ （誤発音）
B) インタフェ***イ***ス → インタフェ***ー***ス (長音)
C) 指摘***と***おり → 指摘***ど***おり （濁音）
D) ***キ***ャンプ → ***ジ***ャンプ （誤入力）
E) 利用できない***こ***増加 → 利用できない増加（余字）
F) 資料を***中文***に翻訳した →資料を***中国語***に翻訳した（日中混同）
G) ***極性***設定 → ***属性***設定（意味誤り）
H) 提供***禁る***前に → 提供***する***前に（読み漢字変換）

カテゴリ 2：文法の理解

カテゴリ２は助詞、動詞の扱いについての誤りである。[大木ら, 2011]の調査結果では、助詞の誤りが最も多かった。我々は大木らの調査結果を参考に助詞と文法に関する分類をカテゴリにまとめる。

このカテゴリの履歴は、訂正前の文が日本語として非文ではないが、文書の内容が誤解される可能性が高いもの、もしくは助詞の誤用により言語処理に支障が出るものである。例えば、以下の事例はカテゴリ 2 の分類に当たる：

I) 位置飛び***を***発生しない →位置飛び***が***発生しない（助詞変更）
J) 電源 Off する → 電源***を*** Off ***に***する （助詞追加）
K) 今回***は***制御部は校正不要 → 今回制御部は校正不要（助詞削除）
L) 最大接続数に***達する***場合 →最大接続数に***達した***場合 （動詞時制）
M) 伝票が***ヒットされる*** →伝票が***ヒットする*** （能動受動）
N) サポートが***充実***といえない → サポートが***充実している***といえない（名詞句と動詞句）

カテゴリ 3：表現の選択

カテゴリ３は単語、フレーズカテゴリの校正であり、3.1 章の用字、用語、表記校正に対応するものである。このカテゴリの校正は意味的に誤解する恐れが少ないが、技術文書として不適切な言葉使いや表記の訂正である。以下の事例はカテゴリ 3 の分類に当たる：

O) 品質を***保証***した → 品質を***確保***した （語彙意味）
P) 品質***制御***→ 品質***管理***（専門語）
Q) 上記 prmt を***読みだし*** → 上記 prmt を***読み出し***（漢字、ひらがな）
R) 次の***二つ***の種類 → 次の***２つ***の種類 （数字表記）
S) 空車状態を***切替え*** → 空車状態を***切り替え***（送り仮名）
T) コードの位置が合理的***じゃない*** → コードの位置が合理的***でない***（口語）
U) シールド***しました***ので → シールド***した***ので（敬体常体）

表 2：誤り分類の定義と頻度

カテゴリ		誤りの分類	頻度
1	表記誤り	誤字、脱字、余字	316
		英字スペル	49
		日中混同	142
		濁音、長音、誤発音	239
	言葉使い	意味誤り	255
		読み漢字変換誤り	59
2	助詞の使用	助詞追加	720
		助詞削除	401
		助詞校正	2907
	動詞の使用	動詞時制とアスペクト	205
		能動受動	290
	品詞区別	名詞句と動詞句の混同	573
3	意味変化なし	漢字、ひらがな、送り仮名	674
		口語	187
		数字と単位	123
		敬体常体	76
	意味変化あり	専門語	267
		語彙意味	896
4	意味変化なし	冗長短縮	350
		文構造	809
	意味変化あり	情報追加	106

カテゴリ 4: 文構造の変更

カテゴリ 4 は文全体の意味や構造を校正するものである。文書全体の理解、および背景知識が必要である。このカテゴリは 3.1 章で議論した「技術文書のわかりやすさ」に対応して、更に技術文書に相応しい文書作成に追求する訂正である。以下の事例はカテゴリ 4 の分類に当たる：

V) 開発中心は、64 ビット対応である。→開発中心は 64 ビット対応である。（文構造）

W) 参考に*なれる意味がある* → 参考に*なる*（冗長短縮）

X) 意味は同様 → 意味は*左記*同様（情報追加）

3.3 誤りの分類作業

誤り分類作業とは誤り箇所に対するタグ付け作業である。訂正履歴の分析と分類は機械による言語処理の結果に基づいて行なっている。まず、訂正履歴文に対して形態素解析を行って、校正前後の形態素列の差分を校正箇所とする。分類作業は、各校正箇所に対して、3.2 節で述べる誤り分類の定義にしたがって、人手でタグを付与する。例えば 2 章の事例に対する誤り分類作業は図 1 で示したツールを使用する。タグ付け作業者は校正履歴の一覧から作業項目を選ぶと、校正前後の形態素解析の結果が表示される。本事例は以下のような形態素列がある：

校正前：今回/テスト/する/時/に/6009/*に*/指定/した。

校正後：今回/テスト/する/時/に/6009/*を*/指定/した。

上記の「/」は形態素の区切りを示す。図 1 左下の形態素列に色付けの部分はタグ付け対象の形態素差分である。本事例は校正前の助詞「に」と校正後の助詞「を」が差分になり、この形態素差分ペアーは一つの校正箇所である。図 1 の右下部分はタグ付け作業領域、タグ付け作業者は 3.2 節の定義にしたがい、校正箇所のタグをリストから選ぶことで誤り種類のタグを付与する。

4 誤り傾向と分析

本節では分類済みの訂正履歴から誤りの分布と特徴から外国語母語話者が技術文書を作成する際犯しやすい誤りの傾向をまとめる。

まず、カテゴリ別の分布をみると、カテゴリ 2 の誤りが全体の半分以上の約 53%（5096/9644）を占めている。助詞の誤用に対する訂正（4028 個、約 42%）が最も多いことがわかった。助詞は一般的に意味が抽象的かつ多義であるため、外国語母語話者にとって使い方の把握が困難であると言われる[今枝ら, 2003]。この助詞の校正数から、技術文書を作成する場合、外国人にとって助詞扱いが大変困難であることがわかった。

助詞別の誤り頻度の分布を表 3 に示す。最も誤りの多い助詞使いは主語判定の「は、が」である。「は、が」と「が、を」が訂正履歴全体の 15%を占め、主語の判定と目的語の判定は外国語母語話者にとって特に間違いやすいことがわかった。技術文書の複文が多い場合、外国人が主語を区別することが困難であることが考える。

一方、助詞使いの誤りは助詞のみの誤用に限らず、同一文中の他の部分に連動することがある。例えば、以下の事例 Y)では、助詞「が、を」の使い方と動詞の能動/受動を同時に見て構成しなければならない。このように、助詞の扱いは助詞に対する知識のみならず、複数の文法要素の考慮が必要であるため、外国語母語話者には難しいと伺える。

Y) 時間*が*要*され*、 → 時間*を*要*し*（「が、を」、「能動受動」）

執筆者が書いた助詞の校正の他、助詞が欠落している現象が多いことから、執筆者が母語（中国語）の影響で助詞を入れ忘れる傾向が伺える。例えば、

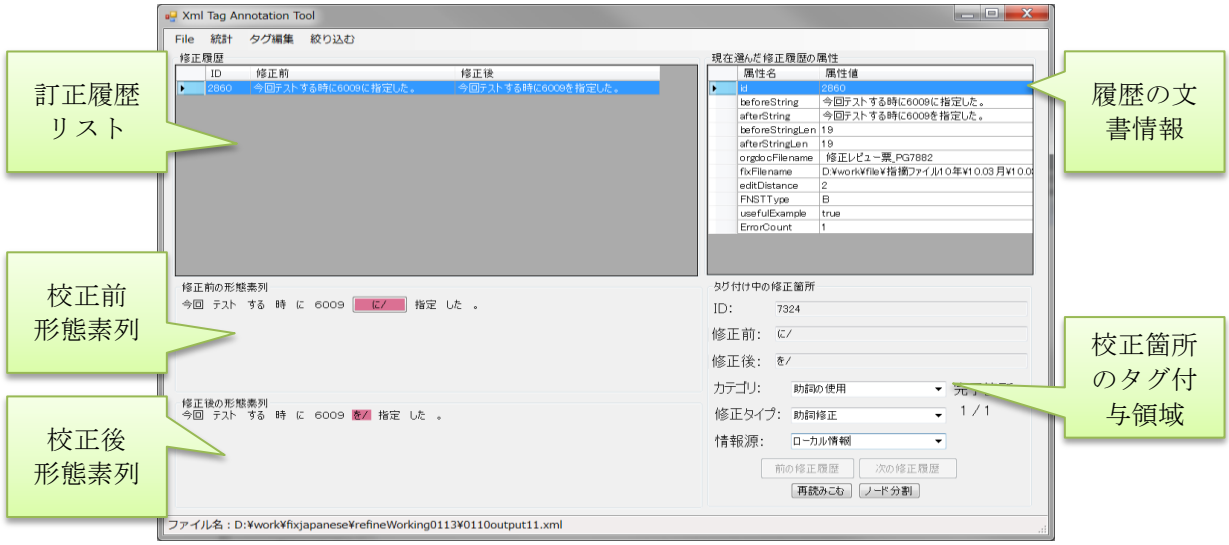

図 1：誤り種類の分類タグ付け作業のスクリーンショット

3.2 節の事例 J)では執筆者が中国語の習慣で連続する語彙に助詞を入れ忘れると考える。

カテゴリ 2 の動詞の扱いに関する誤りは、「動詞句名詞句」の訂正が多い。これは、中国語には動名詞の使いに語形変化がないため、日本語に必要な語形変化を意識せず、名詞のままに使用することによると考えられる（3.2 節の事例 N)はその一例である）。

動詞時制とアスペクトについて、中国語母語話者にとって日本語の時制の扱いが難しいと言われる（3.2 節の事例 L)はその一例である）。訂正履歴の中に「現在形→過去形」の校正が多いことから、中国人が過去形を現在形に間違う傾向があるだろう。

次に、カテゴリ 3 の校正は約 23%（2223/9644）を占める。「意味変化なし」項目の「漢字、ひらがな、送り仮名」、「数字と単位」、「口語」と「敬体常体」などの誤り種類は意味の変化がなし（約全体の 10%を占める）、表記の揺れと考える。[浅岡, 2006]にある「漢字、ひらがな、送り仮名」表記使いの規則を外国語母語話者に教えることにより回避可能だと考える。

一方、「意味変化あり」の項目（「専門語」、「語彙意味」）は、文脈依存の類似用語の選択だと考える。このような事例が多いことは執筆者が該当する類似言葉に疎くて、正確に選択できなかったと考える。

カテゴリ 4 は全体の 13%(1256/9644)を占める。「冗長短縮」は必要がなく、文書がわかりづらくなる記述を削除するものである。「冗長短縮」の件数が「情報追加」の件数より多いことから、3.2 節の事例 W)のように執筆者ができる限り情報を補う傾向が伺える。これにより文書が冗長になる。

カテゴリ 1 は全体の約 11%（1060/9644）であり、誤字、脱字および英字のスペルミスは合計 365 個あり、3.2 節の事例 E)は代表事例である。脱字とスペルミスの事例は特別な特徴が認められないが、誤字の一部事例に特徴がある。例えば、3.2 節の事例 D)では、キーボード配置の「j」と「k」が隣り合わせるため、「じ(ji)」を「き(ki)」に誤入力したと考える。外国語母語話者が普段日本語で入力しないと考えられ、比較的にタイプミスが起きやすいだろう。

一方、ひらがなとカタカナの濁音、長音、誤発音の誤りは 239 個がある。3.2 節の事例 A)～C)が代表である。濁音と長音の使い方は外国語母語話者にとって把握しにくいことが伺える。誤発音の事例について、その原因は外国語母語話者が外来語の英語発音を直接にカタカナで表記することと考える。

「意味誤り」と「日中混同」の事例では、執筆者が語彙の意味を理解していないことが原因であるが、「日中混同」では中国人の執筆者が一部の語彙を中国語の語彙を使う傾向が伺える。例えば、3.2 節の事例 F)がその一例である。

表 3:助詞の誤り種類分布

助詞誤り種類	数
は、が	1019
が、を	493
に、を	169
て、に	158
を、の	113
ほかの助詞校正	955
助詞追加	720
助詞削除	401
合計	4028

5 まとめ

本稿では、中国人の執筆した日本語技術文書の訂正履歴をもとに、外国語母語話者による日本語誤用パターンの分析を行った。その結果、助詞の誤りが全体の42%で最も多く、特に「は」と「が」の区別で誤りが多いことがわかった。次に多いのが「を」と「が」の区別であり、正しい選択には能動態/受動態の使い分けとも関連して、外国人にとって難しい問題と思われる。中国語の言語特徴が動詞の時制判断、語彙の選択、および入力誤りに影響して誤りを生じることがわかった。さらに技術文書として、文の長さの校正と表記正規化の校正が外国人にとって難しいこともわかった。

今回は中国人の執筆した文書の誤りを分析したものであり、今回の結果が中国人以外の外国語母語話者が今回の結果と同じ傾向を示すかどうかはわからない。今後中国語以外を母語とする人に対して同様の調査を行い、母語による日本語誤りの傾向を分析したい。

参考文献

[1] 大木環美, 大山浩美, 北内啓, 末永高志, 松本裕治. 非日本語母国話者の作成するシステム開発文書を対象とした助詞の誤用判定. 第17回言語処理学会年次大会, pp. 1047-1050, 2011

[2]浅岡伴夫. 技術文書の作り方・書き方―SE・製造技術者・理工系学生のための.シーエーピー出版, 2006

[3] 南保亮太, 乙武北斗, 荒木健治. 文節内の特徴を用いた日本語助詞誤りの自動検出・校正(語学学習支援・自動校正). 情報処理学会研究報告2007-NL-181, 2007

[4] 今枝恒治, 河合敦夫, 石川祐司, 永田亮, 桝井文人. 日本語学習者の作文における格助詞の誤り検出と訂正. 情報処理学会報告2003-CE-68, 2003