日英機械翻訳の精度改善と原文の読みやすさ向上のための日本語書き換えルールの作成と評価
—地方自治体ウェブサイト文書を対象に—

宮田 玲 †　立見 みどり ‡　Anthony Hartley‡
影浦 峡 †　井佐原 均 ‡
† 東京大学大学院教育学研究科　‡ 豊橋技術科学大学

1　はじめに

地方自治体のウェブサイトは，外国人居住者に対しても，地域の生活・行政・法律等に関する情報を提供する必要がある．全ての情報をコストのかかる人手翻訳で提供するのは難しいため，機械翻訳を利用して多言語情報を発信する自治体が増えている[1]．近年，機械翻訳の質は向上しつつあるが，特に言語構造の大きく異なる日英間の機械翻訳の精度は十分とは言えない．そこで，制限言語や前編集など，翻訳文書作成の上流工程における原言語テキストの統制が求められる [1].

本研究では，主に作文技術に関する書籍を参考にしながら，47 種類の日本語書き換えルールを抽出し，機械翻訳の精度改善への効果を予備的に確認した上で，最終的にルールを 22 種類に絞り込んだ．さらに，各ルールについて，愛知県豊橋市のウェブサイト[2]から抽出した日本語文書を対象として，日本語文の読みやすさの変化を定量的に評価した．

2　関連研究

日本語制限言語の初期の研究としては，長尾らが 1980 年代に提案した文の曖昧さを取り除くための制限文法が挙げられるが [2]，実用文への応用にまでは至っていない．その後，計算機科学の分野では，主に前編集・書き換えとして原文自動書き換えの研究 [3] が進められてきた．一方で，文書作成実務の場面ではテクニカル・ライティングなどの執筆ガイドラインの開発が進められてきた [4]．近年では，産業日本語など，産業文書（特に特許文書）の翻訳における日本語の規格化が進められている．産業日本語は,「人に理解しやすく，機械翻訳を始めとする言語処理技術を活用するコンピュータにも処理しやすい日本語」 [5] を目指しており，計算機科学（制限言語）と文書作成実務（テクニカル・ライティング）の総合的な知見が求められる．しかし，日本語の統制に関する要件は十分に解明されておらず，評価方法も定まっていない．例えば，制限言語や書き換えの効果を，定量的な測定を踏まえて評価している研究は， [3] [6] など一部に限られている．

本研究では，人手による書き換えを想定して，機械・人間両方の処理しやすさ・理解しやすさを志向した日本語書き換えルールを提案し，日本語文の品質の観点から，書き換えルールの効果を評価する．機械翻訳文の品質の観点からの評価も進めているが，現段階では十分にまとまっていないため本研究では報告しない．

3　書き換えルール

3.1　ルールの収集と作成

まず愛知県豊橋市のウェブサイトから，32530 の日本語文を抽出し，単語列や体言止めのものを除き，13727 文の豊橋市データセットを構築した[3]．続いて，人手による書き換えを想定して，以下の 2 つの方法から暫定的に 47 種類のルールを収集・作成した．

既存の執筆ルールの収集 [7]　スタイルガイド，テクニカルライティングの教本，作文技術に関する書籍（計 17 冊） [4, 8–23] の中から，機械翻訳精度に関わると想定されるルールを取り出した．

機械翻訳出力の分析に基づいたルール作成 [24]　豊橋市データセットを，Google 翻訳[4]と The 翻訳プロフェッショナル V15[5]（以下「The 翻訳」）にかけた結果を分析し，翻訳精度を下げると考えられる要素を規制する形でルールを作成した．

[1] 例えば，筆者らが調べた限り，都道府県庁所在地の 47 ウェブサイトの内，24 サイトで機械翻訳の導入が確認できた（2013 年 1 月 12 日現在）.
[2] http://www.city.toyohashi.aichi.jp/
[3] 文書のドメインは，広報文書，Q & A，組織情報（災害，多文化，子育て），報道発表資料，最近の話題である.
[4] http://translate.google.co.jp/
[5] 東芝ソリューション株式会社, http://pf.toshiba-sol.co.jp/prod/hon_yaku/index_j.htm

No	書き換えルール	書き換え前・後の日本語文サンプル
a	一文はできる限り 70 文字以内におさめてください。それ以上になる場合でも、100 文字以内にはおさめてください。	[前] 豊橋の民営クラブの多くは、昭和 50 年代より保護者同士の助け合いの中で先駆的に設置され、平成 5 年度に最初の公営クラブを開設する以前から運営が続けられています。 [後] 豊橋の民営クラブの多くは、昭和 50 年代より保護者同士の助け合いの中で先駆的に設置されました。平成 5 年度に最初の公営クラブを開設する以前から運営が続けられています。
b	箇条書きで書くときは、列挙項目の前後の文を完結させてください。	[前] その他では、 ・総合動植物公園での動物飼育の仕事 ・保健所で迷子になった犬、ネコの保護や狂犬病予防に取り組む仕事 があります。 [後] その他では、以下の仕事があります。 ・総合動植物公園での動物飼育の仕事 ・保健所で迷子になった犬、ネコの保護や狂犬病予防に取り組む仕事
c	文の中に、括弧書きで長い説明を入れないでください。	[前] ゴミ袋（40ℓ 程度）を二重にして、中に半分程度の水（風呂の残り水を使うと便利）を入れ、玄関などにすき間なく並べる。 [後] ゴミ袋（40ℓ 程度）を二重にして、中に半分程度の水を入れ、玄関などにすき間なく並べる。水は、風呂の残り水を使うと便利です。
d	主語と述語の関係を明確にしてください。	[前] 音楽科の 1、2 年生により演奏され、プログラムから演出、アナウンスまですべて生徒が作ったものです。 [後] 音楽科の 1、2 年生により演奏されました。プログラムから演出、アナウンスまですべて生徒が作ったものです。
e	修飾語と被修飾後の関係を明確にしてください。	[前] 豊橋・田原地域の地産地消の拠点として、安全安心で新鮮な地場農産物を販売する東三河で最大級の農産物直売施設です。 [後] 豊橋・田原地域の地産地消の拠点として、安全安心で新鮮な地場農産物を販売する、東三河で最大級の農産物直売施設です。
f	「が」を使って文をつなげるのは、「しかし」の意味を持つ場合だけにしてください。	[前] 朝市ですが、豊橋のどこでいつ開催されているか詳しく教えて下さい。 [後] 朝市は、豊橋のどこでいつ開催されているか詳しく教えて下さい。
g	「ので」の意味で「ため」を使わないでください。「ので」を使ってください。	[前] 地震で地盤が揺れると土砂が水とともに液体のように流れ動くため、地盤の液状化現象といわれます。 [後] 地震で地盤が揺れると土砂が水とともに液体のように流れ動くので、地盤の液状化現象といわれます。
h	「from」を意味するときは「〜から」を使ってください。「より」は比較のときだけ使用します。	[前] 平成２１年度より整備を進めてまいりました豊橋市南消防署西分署庁舎が、このたび竣工いたしました。 [後] 平成２１年度から整備を進めてまいりました豊橋市南消防署西分署庁舎が、このたび竣工いたしました。
i	1 つの文の中で複数の否定形を使わないでください。	[前] なお、期限内に手続きをしないと、受給資格があっても手当を受けることができなくなる場合がありますので、ご注意ください。 [後] なお、受給資格がある場合でも、手当を受けるには、期限内に手続きをする必要がありますので、ご注意ください。
j	可能や尊敬の意味で「〜れる」「〜られる」を使わないでください。	[前] 立っていられず、ブロック塀が壊れる。 [後] 立っていることができず、ブロック塀が壊れる。
k	複数の意味に解釈できる言葉ではなく、なるべく明確な意味を持つ言葉を使ってください。	[前] 木の枝は、60 センチ以下に束ねて出しましょう。 [後] 木の枝は、60 センチ以下に束ねて捨てましょう。
l	口語表現の「〜になります」表現を避けてください。	[前] 自然史博物館の展示物の中で最大の展示物になります。 [後] 自然史博物館の展示物の中で最大の展示物です。
m	「〜という」表現はなるべく省いてください。	[前] １ヶ月単位で入院費が安くなるという制度があると聞きました。 [後] １ヶ月単位で入院費が安くなる制度があると聞きました。
n	「ような」、「こと」、「もの」はなるべく省いてください。	[前] 近くの職場同士で協力し合うものとする。 [後] 近くの職場同士で協力し合う。
o	1 つの文の中で、同じ語句や重複した意味を持つ語句を使わないでください。	[前] 応募の際には各試験ごとの募集要綱で確認してください。 [後] 応募の際には各試験の募集要綱で確認してください。
p	「思われる」「考えられる」は必要なとき以外は省いてください。	[前] 民間企業も含め、今後駅前の開発に様々な事業者が関与する可能性も考えられます。 [後] 民間企業も含め、今後駅前の開発に様々な事業者が関与する可能性もあります。
q	サ変動詞にはなるべく「行う」を付けないでください。	[前] また、願書は、消防本部予防課、中消防署及び南消防署にて配布を行っています。 [後] また、願書は、消防本部予防課、中消防署及び南消防署にて配布しています。
r	「〜したり、」を使うときは列挙項目すべてに「したり」を付けてください。	[前] 屋根瓦やトタンがめくれたり壊れてないか [後] 屋根瓦やトタンがめくれたり壊れたりしていないか
s	項目を並べるときは、品詞や表現をそろえてください。	[前] 生徒・児童等はあらかじめ学校等で定められた方法によって帰宅、又は保護者に引き渡す。 [後] 生徒・児童等はあらかじめ学校等で定められた方法によって帰宅させるか、又は保護者に引き渡す。
t	なるべく標準的な和英辞典に載っている語を使ってください。	[前] ブラジル・パラナ州経済視察団一行が来豊し、市長表敬を行います [後] ブラジル・パラナ州経済視察団一行が豊橋市に来訪し、市長表敬を行います
u	サ変動詞をつなげた複合語を避けてください。	[前] 愛知県内の同システムに掲載参加している全自治体分を検索機能により見られます。 [後] 愛知県内の同システムに掲載・参加している全自治体分を検索機能により見られます。
v	誤字、脱字がないように注意してください。また、同音異義語や助詞の抜けにも注意してください。	[前] 毎年８月に中央図書館にて「平和を求めて」と題して、パネル・写真展行っています。 [後] 毎年８月に中央図書館にて「平和を求めて」と題して、パネル・写真展を行っています。

表 1: 書き換えルールとサンプル文

3.2 機械翻訳精度の検証とルールの集約

続いて，47 種類のルールの機械翻訳精度への効果を，以下の手順で検証した．

1. 47 種類の各ルールに違反するパターンを含む日本語例文とその書き換え例を，書籍の中から，2 セットずつ取り出す（合計 94 ペア，188 文）[6]
2. 全ての例文を上記 2 種類の機械翻訳にかけ，376 の翻訳英文を生成する
3. その結果を見ながら，筆者ら 2 名（1 名は日本語を母語とし，1 名は英語を母語とする）が独立に，日本語の書き換えにより翻訳精度が「向上した」「低下した」「変化なし」の判定を下す
4. 両者の評価をもとに，各翻訳文を定性的に診断して，翻訳精度の改善に効果が見込める書き換えルールを残す

以上の作業により，表 1 の a，b，d，l，m，t に対応するルールにおいて，全ての例文で翻訳精度の改善が確認できた．一方で，助詞の使い分けに関するいくつかのルールは，あまり高い効果が見込めなかったため，この段階で排除した．さらに，類似ルールを統合し，ルール間の重複を減らした．また豊橋市データセットを参照して，ルールに違反するパターンがほとんど見つからない場合[7]は，当該ルールを排除し，最終的にルールを 22 種類（a–v）にまで絞り込んだ（表 1）．

以下では，22 種類の書き換えルールが，日本語の読みやすさの向上にどれだけ貢献するかを，定量的な評価実験により検証する．

4 「読みやすさ」評価実験

4.1 評価用データ

豊橋市データセットを「1 文の長さが 70 文字以上」「1 文の長さが 70 文字未満」で二分し，それぞれセット L（2671 文），セット S（11056 文）とする．ルール a はセット L から，ルール b はセット L と S 両方から，それ以外のルール c–v についてはセット S から，それぞれのルールに違反するパターンを含む日本語文を 4–10 文抽出した（合計 120 文）．

続いて，抽出した 120 の日本語文（BJ とする）を筆者ら 2 名がルールにしたがってそれぞれ独立に書き換えた．その後，両者の書き換え結果をすり合わせ，120 の書き換え日本語文（AJ とする）を生成した．

[6]書籍の例文が足りない場合は，豊橋市データセットから，ルールに違反する文を抽出し，筆者らが書き換え例を作成した．

[7]例えば，「〜のように〜でない」の表現を規制するルールを作成したが，そもそもデータ中に該当する表現が見つからなかった．

4.2 評価手法

書き換え前後の日本語文ペアをランダムに提示した上で，「読みやすさ」の観点から，それぞれの文について「読みやすい・どちらかといえば読みやすい・どちらかといえば読みにくい・読みにくい」の 4 段階で，定量的な人手評価を行った．日本語を母語とする大学生 10 名に対し，オンラインアンケート上で評価・回答を依頼した．各評価者は，書き換え前後の日本語文，BJ，AJ の 240 文（120 ペア）を全て評価した．

4.3 評価結果

10 名の評価者により各文に付与された 「読みやすい・どちらかといえば読みやすい・どちらかといえば読みにくい・読みにくい」の 4 段階評価を 4，3，2，1 点に置き換え，さらに $AJ - BJ$ を算出し，書き換え前後での読みやすさの変化を判定した．合計 1200 の $AJ - BJ$ の値を，「0 より大：読みやすさ向上」，「0：読みやすさ変化なし」，「0 未満：読みやすさ低下」と分類し，それらをルールごとにまとめて集計した結果が図 1 である．

図 1 より，ルール b，f，t は，読みやすさの向上が 8 割近くに達しており，効果が高いことが示された．例えば，ルール b（表 1 参照）にしたがって書き換えることで，述語が箇条書きよりも前に提示され，文意が早い段階で把握できるようになったので，読みやすさが向上したと考えられる．

また一方で，内部で評価の割れたルールもあった．例えば，ルール q にしたがい「配布を行う」を「配布する」に書き換えたところ，読みやすさが向上したが，「保管・管理を行う」を「保管および管理する」に書き換えると読みやすさが低下した．後者のように，「行う」が複数の要素にかかる場合は，サ変動詞化することで，意味のかたまりが分割され（「保管」と「管理する」），内容が把握しづらくなるためであろう．

同一のルールであっても，具体的な文に適用する際には複数の書き換えの可能性がある．画一的ではなく，状況に応じた書き換えができるように，より精緻にルールを定義することが求められる．

5 おわりに

本研究では，既存の日本語文章執筆に関する知見を収集・整理し，さらに地方自治体の文書を分析することで，47 種類の日本語書き換えルールを考案した．また，機械翻訳精度への効果を検証しながら，22 種類のルールに集約した上で，評価実験により，日本語文の読みやすさへの効果を測定・評価した．日本語原文の

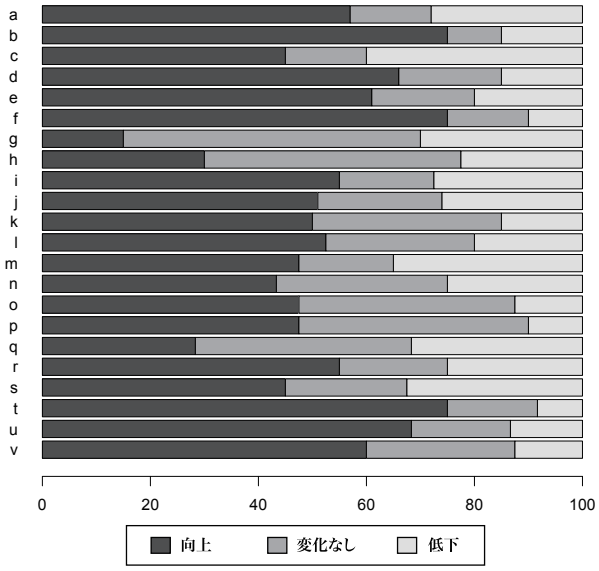

図 1: 日本語文読みやすさ評価結果

書き換えにより，多くのサンプルにおいて，読みやすさの向上が認められ，ルールの効力が明らかになった．

今後の課題として，まずは翻訳英文に対する定量的な人手評価実験を実施する予定である．今回は，筆者らの予備的な分析により翻訳文を評価したが，後編集しやすさの評価や第三者による人手評価実験を通して，より多面的かつ客観的にルールを評価したい．その上で，効果の高いルールを抽出し，また内部で評価が分かれたルールを詳細に検証する．さらに，対象ドキュメントの範囲を広げて評価実験を行い，ドメインに依存するルールと幅広いドメインに通用するルールを仕分ける作業が必要となる．なお今回は，筆者らが自ら原文を書き換えたが，第三者の執筆者による書き換え実験を行い，将来的には，書き換えを自動的に支援するツールの開発を進めたい．

謝辞 本研究の一部は，総務省の戦略的情報通信研究開発推進制度 (SCOPE)・地域 ICT 振興型研究開発「地域産業の国際競争力強化のための多言語情報発信支援の研究開発」並びに，国立情報学研究所共同研究「制限日本語と機械翻訳を用いたビジネス・技術文書多言語化の効率改善に関する研究」の枠組みで行われた．研究用の機械翻訳「The 翻訳プロフェッショナル V15」は，東芝ソリューション株式会社からご提供いただいた．

参考文献

[1] Arendse Bernth and Claudia Gdaniec. MTranslatability. *Machine Translation*, Vol. 16, No. 3, pp. 175–218, 2001.

[2] 長尾真, 田中伸佳, 辻井潤一. 制限文法にもとづく文章作成援助システム. 情報処理学会研究報告 (NL), Vol. 1984, No. 27, pp. 1–8, 1984.

[3] 白井諭, 池原悟, 河岡司, 中村行宏. 日英機械翻訳における原文自動書き替え型翻訳方式とその効果. 情報処理学会論文誌, Vol. 36, No. 1, pp. 12–21, 1995.

[4] 一般財団法人テクニカルコミュニケーター協会. 日本語スタイルガイド第 2 版. テクニカルコミュニケーター協会出版事業部, 2011.

[5] 松田成正. 平成 24 年度 特許版 ・産業日本語：新たにスタートした「特許ライティング支援システム」活動について. *Japio YEARBOOK 2012*, pp. 310–311, 2012.

[6] Anthony Hartley, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata. Readability and translatability judgments for 'controlled japanese'. In *Proceedings of the 16th EAMT Conference*, pp. 237–244, Trento, Italy, 2012.

[7] 吉田将. 科学技術文書を記述するための日本語の規格化—係り受け関係の制限について. 九州大学工学集報, Vol. 56, No. 3, pp. 205–211, 1983.

[8] 礒崎陽輔. 分かりやすい公用文の書き方 改訂版. ぎょうせい, 2010.

[9] 後藤禎典. [後藤式] 文章の技術 わかりやすい文が書ける明快ルール 100. PHP 研究所, 2005.

[10] 山本ゆうじ. IT 時代の実務日本語スタイルブック. ベレ出版, 2012.

[11] 高橋昭男. 技術系の文章作法. 共立出版, 1995.

[12] 長尾真, 牧野武則. コンピュータで翻訳する. 共立出版, 1995.

[13] 永山嘉昭, 雨宮拓, 黒田聡, 矢野りん. 説得できる文章・表現 200 の鉄則 第 4 版. 日経 BP 社, 2009.

[14] 篠田義明. ビジネス文 完全マスター術. 角川書店, 2003.

[15] 篠田義明. コミュニケーション技術. 中央公論社, 1986.

[16] 阿部紘久. 文章力の基本. 日本実業出版社, 2009.

[17] 三島浩. 技術者・学生のためのテクニカルライティング 第 2 版. 共立出版, 2001.

[18] 高橋麻奈. 入門テクニカルライティング. 朝倉書店, 2005.

[19] 小山透. 科学技術系のライティング技法. 慶應義塾大学出版会, 2011.

[20] テクニカルコミュニケーション研究会（編）. わかりやすいマニュアルを作る 文章・用字用語ハンドブック. 日経 BP 出版センター, 1995.

[21] 岩淵悦太郎. 悪文 第 3 版. 日本評論社, 1979.

[22] 本多勝一. 新装版 日本語の作文技術. 講談社, 2005.

[23] 弘前大学 人文学部社会言語学研究室. 「やさしい日本語」の作成ルール. `http://human.cc.hirosaki-u.ac.jp/kokugo/EJ3mokuji.htm` (2013/1/10 アクセス).

[24] 小倉英里, 工藤真代, 柳英夫. シンプリファイド・テクニカル・ジャパニーズ英訳を視野に入れて日本語を作る. 情報処理学会研究報告 (DD), Vol. 2010, No. 5, pp. 1–8, 2010.