機械翻訳の活用を見据えた文書構造と言語表現の対応づけ —自治体手続き型文書を対象とした予備的報告—

宮田 玲† Cécile Paris‡ Anthony Hartley# 影浦 峡†
†東京大学大学院 ‡オーストラリア連邦科学産業研究機構
#東京外国語大学
rei@p.u-tokyo.ac.jp

1 はじめに

機械翻訳（MT）の現実的な運用を，原言語テキストの執筆から目標テキストの作成までを翻訳工程に含めて考えると，主な介入点としては原文執筆・前編集・MT エンジン・後編集を挙げることができる [1]。この中でも翻訳の上流工程，つまり原文書作成のコントロールの観点から包括的な翻訳効率の改善を目指す手法は，翻訳学における「ローカリゼーション」のパラダイムの中の「国際化」の文脈で取り組まれてきた [2]。特に多言語での文書展開を想定している場合，なるべく原文の段階で統制することで，後編集を含めたトータルのコストを大幅に下げることが期待できる。加えて，原文の段階での情報構成やテキスト品質の改善が併せて求められており，原文執筆に関する方法論は，主にテクニカルライティング（Technical Writing: TW）の分野で取り組まれてきた [3]。

以上のような背景から筆者らはこれまで自治体のウェブサイト文書を対象に，制限言語（Controlled Language: CL）ルールの構築を進めてきた [4, 5]。TW の知見を参照しながら，MT 文（英文）の品質のみならず，原文（日本語文）の品質の向上を目指した CL のルールセットを策定・評価した結果，原文品質の大幅な向上が確認できた一方で，翻訳品質は微増にとどまった。言語構造の大きく異なる日英 MT を実用レベルにまで性能を引き上げるには，TW で定義されるような比較的ゆるやかな文章規則では不十分であったことが示唆された。また用いる MT の種類により CL の効果が異なるという結果から，MT 一般に適用可能な CL の策定を目指すだけでなく，個別の MT にチューニングした CL を定義していくことが必要であると明らかになった。

これまで提案されてきた CL の操作対象はあくまで語彙・文法・スタイルといったセンテンスレベルの言語表現であり，文書構造の議論は十分になされていない [6]。これに対して筆者らはまず，神門の機能構造分析 [7] や技術文書用の標準規格である DITA (Darwin Information Typing Architecture) [8] の枠組みを参照しながら，自治体の手続き型文書[1] の構造の記述・定式化を試みてきた [9]。しかし，このような作業と先述の TW や CL といった言語表現の統制とは独立に議論しており，両者の対応づけについてはこれまで報告していない。文書構造がいかに言語表現の幅を規定し，さらに MT の品質を改善しうるかについて，事例を踏まえた説明が求められる。

本稿では，自治体手続き型文書を対象に，具体的な事例を挙げながら，DITA の枠組みを用いて文書構造の側から言語表現をコントロールし，さらに MT を活用していく方略を示す。

2 DITA の活用

2.1 DITA の概要と基本構成要素

DITA とは，技術文書の作成・出版のための XML ベースの規格で，モジュール化した情報をまとめ上げて文書を構成するという特徴を持つ [10,11]。DITA には，基本概念として「トピック（topic）」と「マップ（map）」が定義されている。トピックとは，DITA における情報の基本単位であり，それ自体で意味をなす独立したユニットである。トピック（汎用トピック）は，次のような基本構造を持つ。

- タイトル（title）：トピックのテーマを記載する
- 要約文（short description）：トピックの目的やテーマの簡単な説明を含み，プレビューや検索にも使われる
- プロログ（prolog）：トピックに関するメタデータを記載する
- トピック本体（body）：トピックの実際の内容を記載する
- 関連リンク（related links）：補助的な情報への関連リンクを記載する

[1] 例えば，印鑑登録の仕方や転出届の出し方など，自治体の各種手続きを遂行する際に住民が参照するための文書を指す。

またマップとは，出力媒体や使用目的に合わせて，複数のトピックの順序や階層を定義しながら，最終的な制作物に編成する仕組みである。ユーザーマニュアルと技術者向けマニュアルで使用するトピックを一部変えるなど自己完結的な情報のトピックを柔軟に組み合わせながらドキュメントを生成することで，トピックの再利用が促進される。

以下本稿ではマップの議論には踏み込まず，いかに自治体手続き型文書をトピックとしてまとめていくかを考えていく。

2.2 タスク・トピック

DITA ではあらかじめ 3 つのトピックの型（Type）が定義されている。

1. Concept 型（コンセプト・トピック）
2. Task 型（タスク・トピック）
3. Reference 型（レファレンス・トピック）

コンセプト・トピックは,「これは何のことか」(what) という問いに答えるための情報型で，概念の説明に使われる。タスク・トピックは,「どうやって」(how to) という問いに答えるための情報型で，手続き・手順の記述に使われる。レファレンス・トピックは，手続き実行の時などに参考になる情報の記述に使われる。

これらの型は，先述の汎用トピックの「トピック本体」の部分を，特殊化して定義される。DITA は，これらのトピックに書かれるべき文書の機能的な要素（以下，機能要素）を定義しており，この枠組みにしたがってトピックを執筆することで，必要な情報を漏れなく体系的に含めることができる。本研究で扱う自治体手続き型文書は主にタスク・トピックにより構成されており，その「トピック本体」部分では，以下の構造が定義されている [2]。

- 事前条件（prereq）：タスクを実行するのに必要な，事前条件を記述する
- 背景情報（context）：タスクの背景情報や予備知識を記述する
- 手順（steps）：タスクを完了するために，ユーザーが行う一連の手順を記述する
- 期待結果（result）：タスクを完了したときの，期待される結果を記述する
- 実行例（example）：タスクの実行例，または，タスク実行の説明を補助するための例を記述する
- タスク完了後の操作（postreq）：タスクを完了した後に，次に行うべきこと記述する

[2] [12] の説明を一部変えて引用。

2.3 タスク・トピックの具体化

しかし，このタスク・トピックはあくまでタスクに関する一般的な型を定義したものであり，実際に自治体の各種手続きを執筆する際の指針として利用するには不十分である。筆者らはこれまで，自治体国際化協会，新宿，浜松市の生活情報から選定した手続き型文書の機能要素を抽出・整理した上で，DITA のタスク・トピックとの対応づけを行ってきた。今回それをさらに整理しなおし，DITA タスク本文の各要素を，自治体手続きに合わせて具体化した（表 1）。

表 1: DITA タスク本文の具体化

DITA(初期設定)	詳細機能要素
事前条件（prereq）	個人条件 イベント条件 アイテム条件
背景情報（context）	説明（概要，目的，効力，罰則，関連概念）
手順（steps）	必要なものを持参する 申請場所へ行く 様式を提出する （手数料を払う）
期待結果（result）	得られる結果（所要期間，交付物，連絡）
実行例（example）	[不要]
タスク完了後の操作（postreq）	関連手続きへの誘導

表 1 について説明する。自治体手続きの基本構造は,「ある初期状態にある個人（住民）が一定の条件を満たした時に，特定の行政手続きを遂行することで，別の状態に変化すること」であるといえる。これを DITA のタスク・トピックに照らして，文書として具体的に表現することが，ここでの目標である。

まず,「ある初期状態にある個人（住民）が一定の条件を満たした時に」という部分は，DITA の「事前条件」に該当し，自治体手続きにおいては，大きく「個人条件」「イベント条件」「アイテム条件」の 3 種類に分けることができる。「個人条件」とは,「15 歳以上の人」「外国籍の人」「新宿区に居住している人」といった個人の社会的属性に関する条件を指す。「イベント条件」とは,「日本に来た時」「結婚した時」「新宿区に転入した時」といった出来事に付帯した条件を指す。「アイテム条件」とは，例えば「印影の大きさが一辺 8mm の正方形に収まるもの，または一辺 25mm の正方形に収まらないもの」といった登録できない印鑑の条件など，物体に関する制約条件を指す。

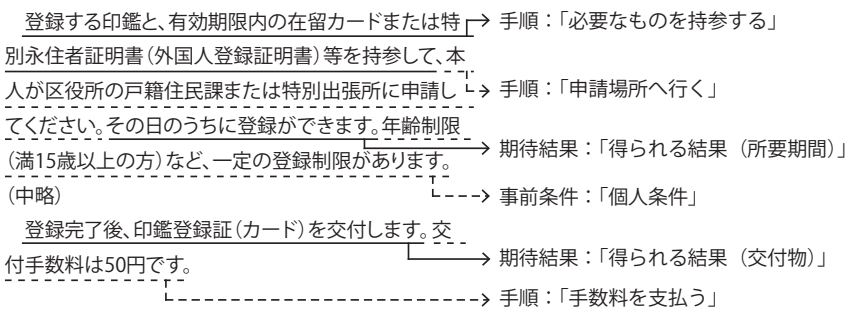

図 1: 「印鑑登録」文書の DITA による分析（一部省略）

次の「背景情報」は，手続きの遂行においては必ずしも必要とは限らないが，読み手の理解を補助し，円滑な手続きを促進する上では重要な役割を持つ。例えば,「手続きは何のために行うのか（目的）」「手続きをしないとどのようなペナルティが課されるのか（罰則）」といった情報がここでは提示される。

続いて，事前条件を適切に満たした上で，自治体手続きを確実に完遂するための「手順」を詳細化する。これまで調査した範囲では，表 1 に示したの 4 つの要素が中心的であることが明らかになった。

DITA では，これらの手順を確実に遂行できた場合の「期待結果」も明示される。ある手続きが完了した場合，いつ，何が起こるのか，という「得られる結果」に関する情報は，読み手があらかじめ手続きの終了条件を認識する上でも有用である。

「実行例」については，自治体手続きにおいては，筆者らが調べた範囲で，実行例が示されることはほとんどなかったので，基本的には不要であると考えられる。

最後に,「タスク完了後の操作」については，自治体手続きではしばしば関連した手続きが付記されることがある。例えば，印鑑登録について説明する文書では，併せて「印鑑登録証明書」の交付手続きについて書かれることが多い。これは，関連情報として有用である一方で，当該手続き（ここでは印鑑登録）の遂行上，混乱を招くことも予想される。そのため,「タスク完了後の操作」として関連手続きをまとめて別置し，あくまで別の手続きであることを強調した上で，適切に読み手を誘導することが有効だろう。

以上のように詳細化した DITA の機能要素を実際の自治体手続き型文書[3] に適用した例が図 1 である。ここから例えば,「事前条件」が先頭ではなく「手順」や「期待結果」の間に書かれていることや,「手順」や「期待結果」がそれぞれ文書中に点在しており一箇所にまとまっていないことが読み取れる。

[3]新宿区, 生活情報「印鑑登録」http://www.city.shinjuku.lg.jp/foreign/japanese/guide/todoke/todoke_7.html

3 言語表現の対応づけ

以上の作業により自治体手続き型の文書構造が暫定的に確定し,「何をどのような順序・構成で書けばよいか」の指針が定まった。ここで MT はあくまでテキスト表層上の言語表現のみを扱うことを踏まえると，引き続き「具体的にどのように書けばよいか」という言語表現の形を定義していくことが必要である。手順としては，表 1 で定式化した DITA の文書構造に応じて，(i) これまで書かれた言語表現パターンを把握することと，(ii) 書かれるべき言語表現パターンを定義すること，の 2 段階に分けられる。

(i) では，例えば，DITA の「事前条件」要素内の「イベント条件」では「〜という時」「〜した場合」といった条件節の表現パターンが抽出できる。また「手順」要素では文末が「してください」「しましょう」「します」といった複数のパターンが見られる。文書構造に応じてどのような言語表現がこれまでとられてきたかを記述的に整理することがまず求められる。

(ii) では，(i) で整理した複数のありうる表現パターンの中でも，ある一定のパターンのみを許容することで，原文で使われる言語表現の幅を抑えることができる。例えば,「手順」要素では平叙文を用いて文末は「〜する」の形にする，といったルールを定義できる。

しかしこの作業は，原文の一貫性・理解しやすさに寄与しようとも，必ずしも MT 性能の向上を担保するわけではない。例えば,「手順」要素で平叙文を使った原文を MT [4] にかけた結果は以下の通りである：

> [原文] 身分証明書（運転免許証，外国人登録証など）を持参する。
>
> [MT] The ID (driver's license and foreigner registration card) is brought.

ここで MT は受身形で訳しているが，この場合は英文では単純な命令形を用いることが望ましいだろう。

[4]株式会社高電社の「J-SERVER プロフェッショナル翻訳ゲートウェイ V3」を利用した。

そこで，原文の文末を命令形の「〜しなさい」に書き換えたところ，以下の通り命令形で訳された：

> [書き換え文] 身分証明書（運転免許証，外国人登録証など）を持参しなさい。
>
> [MT] Bring the ID (driver's license and foreigner registration card).

原文では平叙文を維持したまま，翻訳文では命令形，といった訳し分けを実現するためには，このような前処理工程が必要となるが，この操作は概ね自動化できる。原文執筆が完了した段階では，必ずしも MT に最適化されていなくとも，自動的な前処理を経て，MT 性能を引き出すことが可能である。ここで重要なのは，原文の言語表現パターンが自動的な一括操作が可能な程度に統制されていることであり，これを改めて CL ルールとして策定していくことが必要である。

4 おわりに

本稿では，MT 性能を引き出すための CL ルールの拡張方針を，主に DITA による文書構造の定式化，言語表現パターンの定義，MT の前処理工程の一連の流れに沿って説明してきた。現段階では予備的な文書構造の定式化と言語表現の対応づけの例を示したのみで，今後以下の課題に取り組む予定である。

文書構造に関しては，表 1 で示した DITA 構造が実際の自治体手続き型文書をどの程度カバーするのかという適用可能性について検証する必要がある。既存文書を図 1 のような形で分析・診断しながら，適宜 DITA の各要素を改良していくことが求められる。

言語表現パターンの幅を調査するためには，レジスター分析を行う [13]。DITA の要素ごとに言語表現パターンを抽出・類型化した上で，特定のパターンのみを許容する CL ルールを策定していく。

なおこれらの課題に取り組むにあたっては，既存文書の文書構造・言語表現を必ずしも踏襲する必要はなく，一定の経験により裏付けされた TW などの文章技術を参照しながら，規範的に望ましい文書デザインを進めることが肝心である。

謝辞

本研究は KDDI 財団の調査研究助成「自治体文書の多言語化支援システムの開発」の枠組みで行われた。共同研究の遂行にあたっては，JSPS 外国人研究者招へい事業（短期）「多言語展開を考慮した文書オーサリング支援環境の構築と MT の活用」の助成を得た。研究用の MT「J-SERVER プロフェッショナル翻訳ゲートウェイ V3」は，株式会社高電社からご提供いただいた。

参考文献

[1] Hutchins J. Current Commercial Machine Translation Systems and Computer-based Translation Tools: System Types and their Uses. *International Journal of Translation*, Vol.17, No.1-2, pp.5-38, 2005.

[2] Pym A. 翻訳理論の探求. 武田珂代子 訳, みすず書房, 2010.

[3] 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.

[4] 宮田玲ほか. 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. 言語処理学会第 19 回年次大会, pp.710-713, 2013.

[5] Tatsumi M. et al. Towards Acceptable Quality Machine Translation without Post-Editing for Municipal Websites: An Evaluation of Japanese Controlled Language Rules. *MT Summit XIV: QTLaunchPad Workshop on Human-Centric Machine Translation and Evaluation*, 2013.

[6] 井佐原均ほか. 企業の多言語情報発信を支援する取り組み：国際化をにらんだ産業文書の効率的作成へ向けて. 言語処理学会第 18 回年次大会, pp.369-372, 2012.

[7] 神門典子. 構成要素カテゴリを用いた原著論文の内部構造分析. 情報処理学会研究報告, Vol.1992, No.32, pp.39-46, 1992.

[8] OASIS. Darwin Information Typing Architecture (DITA) Version 1.2. `http://docs.oasis-open.org/dita/v1.2/spec/DITA1.2-spec.html` (accessed 2015-1-8)

[9] 宮田玲, Hartley A, 影浦峡. 自治体ウェブサイト文書の多言語展開を支援するシステム環境. 言語処理学会第 20 回年次大会, pp.812-815, 2014.

[10] Carey L., Schlotfeldt M., Bellamy J. *DITA Best Practices: A Roadmap for Writing, Editing, and Architecting in DITA*, IBM Press, 2012.

[11] Hackos J. T. DITA 概説書. DITA コンソーシアムジャパン 訳, エスアイビーアクセス, 2010.

[12] DITA コンソーシアムジャパン. タスク・トピックの構造. `http://dita-jp.org/webhelp-feedback/about_task_topic/task_topic_structure.html` (accessed 2015-1-8)

[13] Biber D., Conrad S. *Register, Genre, and Style*, Cambridge University Press, 2009.