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

宮田 玲† 影浦 峡† Anthony Hartley‡
†東京大学大学院教育学研究科　‡東京外国語大学

1 はじめに

自治体は、日本語を母語としない外国人住民に対しても、生活や命に関わる情報を、正確かつ効率的に提供することが求められる。コスト・時間的な制約から、人手による高品質の翻訳文書を網羅的に用意することは難しく、機械翻訳（MT）を導入する自治体ウェブサイトが増えている。これまで、MT の品質向上を目指して、起点テキストの言語表現に一定の制約をかける制限言語（Controlled Language: CL）や前編集（Pre-editing）のアプローチから研究が進められてきている [1] [2] [3]。井佐原らは、MT を用いた産業文書の多言語展開を支援するために、用語レベル、文章レベル、文書レベルでのアプローチが重要であるとし、とりわけ翻訳に適した文書構造に関する議論が不足していることを指摘している [4]。現実の運用場面を見据えて MT をはじめとした言語処理技術を活用するためには、テキスト内の言語表現をフラットに扱うのではなく、文書構造上の位置と対応づけて処理・翻訳することが求められる。

加えて、このような高度な言語処理技術の活用を支援するためのシステム環境を併せて提示することが重要である。例えば、制限言語チェッカーなどのツールはいくつか提案・実用化されてはいるものの [5] [6]、文書構造にまで踏み込んだ多言語文書作成に関する統合的なシステム環境はあまり提案されていない [7]。

本研究では、とりわけ住民のニーズが高いと考えられる、自治体手続き型文書[1]を暫定的に対象として、文書構造から言語表現を捉えた執筆支援方略を検討し、文書作成と多言語展開を統合的に支援するシステム環境を提案する。

2 文書構造の定式化

2.1 自治体手続き型文書の機能構造

文書構造と一言に言っても、章・節・段落や見出しといったテキストの論理的構造や HTML で定義されるようなマークアップ構造が想定されるが、ここでは、例えば学術論文に広く見られる IMRAD (Introduction, Methods, Results And Discussion) 型式のような、特定の文書ジャンルにおいて広く観察される修辞的・機能的な構造（機能構造）にまず焦点を当てる。自治体手続き型文書に関してはこのような構造が明示的に提示されていないため、ジャンル研究におけるテキストの修辞的構成の分析 [8] [9] や神門の提唱する機能構造分析 [10] を参考に、一定の構造を取り出す作業からはじめた。自治体国際化協会[2]、新宿区[3]、浜松市[4]のウェブサイトから、合計 123 の自治体手続き型文書を収集し、1 点 1 点分析しながら、手作業で機能的な要素を洗い出した。さらにそれらを手続きの時系列に沿って、階層的に配置しながら、機能構造を整理した（図 1）。

2.2 DITA へのマッピング

以上のように定義した機能構造は、あくまで「自治体手続き型文書では、これまでどのような内容（機能要素）が書かれてきたか」について簡単な階層関係を導入しながら網羅的に書き下したものであり、実際の文書作成と多言語展開の場面に生かすためには、より具体的に文書構造として定式化する必要がある。しかし先述したように、自治体手続き型文書に関する合意のとれた構造は十分明らかになっておらず、依然として自治体の文書作成者の裁量に任されがちである。そこで、技術情報の執筆・出版の標準規格として実績のある DITA (Darwin Information Typing Architecture)[5]を導入し、その形式に自治体手続き型文書の機能構造をマッピングすることで、文書構造の定式化を試みた。とりわけ、自治体手続き型文書の構造と親和性が高いと考えられる、操作手順を記述するための DITA「タスク・トピック」に対して、図 1 の機能構造の各要素をマッピングした（表 1）。DITA が主に想定している製品マニュアルでの操作タスクと、自治体の手続きタスクが、「特定の対象の、特定の初期状態が、一連の行為ステップを経ることで、別の状態に変化する」という点で共通しているため、一部配置を換えながらも、無理のないマッピングを実行できた。

[1]「転出届」「在留期間の更新」など、自治体における各種の届出や申請に関する文書が含まれる。

[2]多言語生活情報, http://www.clair.or.jp/tagengo/

[3]生活情報, http://www.city.shinjuku.lg.jp/foreign/japanese/guide/index.html

[4]カナル・ハママツ, http://www.city.hamamatsu.shizuoka.jp/hamaj/index.html

[5]DITA XML.org, http://dita.xml.org/

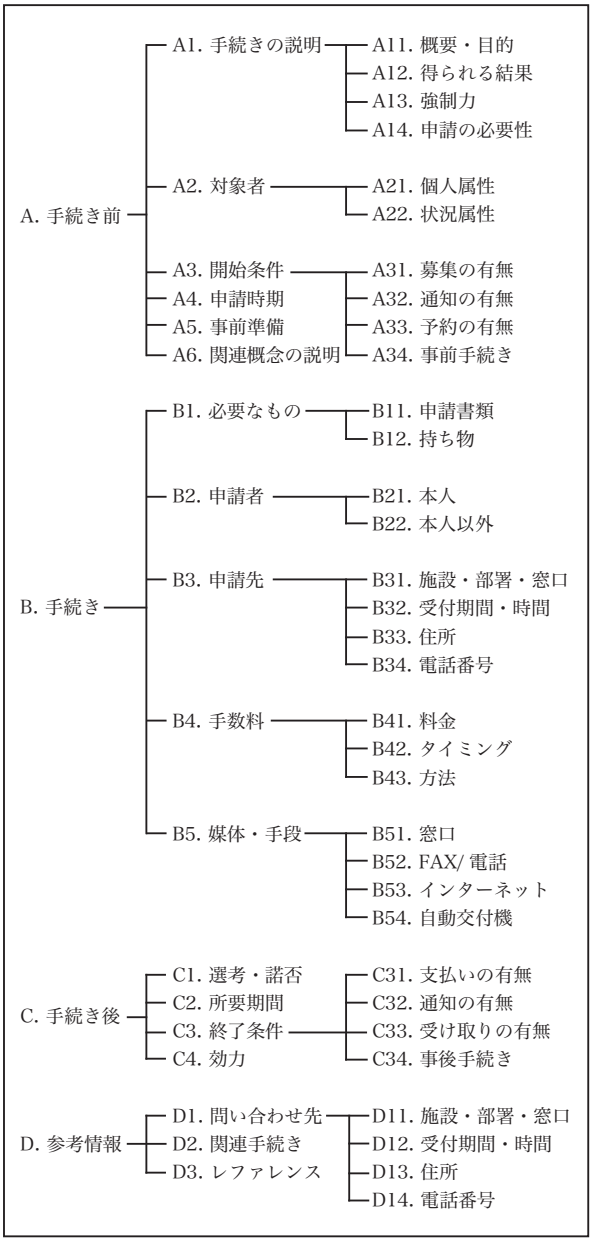

図 1: 自治体手続き型文書の機能構造

3　言語表現のコントロール

3.1　制限言語ルールの作成

筆者らはこれまで、テクニカルライティングの知見を集約しながら、起点テキスト（日本語）の読みやすさを保ちつつ、かつ MT の品質を改善するための制限言語ルールを構築してきた。ルールの構築と評価実験については別途報告しており [11] [12]、ここでは詳細を省略するが、文長の制限、箇条書きの使い方、修飾関係の明確化、二重否定の回避といった主に文法・構文・スタイルレベルの言語表現の制限項目を 22 種類定義している。また愛知県豊橋市のウェブサイトから抽出したセンテンスを対象とした定量的な評価実験により、各ルールの効果を検証した。テクニカルライティングの知見を応用したことで、日本語の読みやすさの大幅な向上が達成された一方で、MT の精度は部分的な向上に留まった。言語表現の操作のみによるアプローチの限界が示唆され、文書構造に応じたルールの精緻化とチューニングの必要性が明らかになった。例えば、DITA の「手順（steps）」要素では、日本語文の文末を動詞形に揃えるように制約をかけることで、原文のスタイルを統一できるだけでなく、MT では命令形を使って訳す、といった指定が可能となる。この他にも、機能構造の「A3. 開始条件」「A4. 申請時期」要素のように、「〜してから〜日以内に〜しなければなりません」といった形で構文パターンを明確に定義しやすい文書構造では、あらかじめ厳密な制限言語ルールを規定することが有効かつ可能である。

表 1: DITA マッピングの結果（一部省略）

DITA タスク本文	機能構造
事前条件（prereq）	「A2. 対象者」「A3. 開始条件」「A4. 申請時期」「A5. 事前準備」
背景情報（context）	「A1. 手続きの説明」「A6. 関連概念の説明」「D1. 問い合わせ先」
手順（steps）	
操作 1（step）	「B12. 持ち物」を用意する
操作 2（step）	「B2. 申請者」を指定する
操作 3（step）	「B3. 申請先」に行く
操作 4（step）	「B11. 申請書類」を書く
操作 5（step）	「B12. 持ち物」と「B11. 申請書類」を提出する（必要に応じて「B4. 手数料」を指示する）
期待結果（result）	「C1. 選考・諾否」「C2. 所要期間」「C31. 支払いの有無」「C32. 通知の有無」「C33. 受け取りの有無」
実行例（example）	該当なし
タスク完了後の操作（postreq）	「C34. 事後手続き」「C4. 効力」「D2. 関連手続き」

3.2　多言語用語集の整備

自治体ウェブサイトの文書には、地名や施設名の他に「外国人登録（alien registration）」や「公共職業安定所（Public Employment Security Office）」といった自治体固有の表現、「最低賃金法（Minimum Wage

Law)」といった法律用語が含まれている。起点テキストにおいても、目標テキストにおいても、一貫して的確にこれらの用語を使用する必要がある。執筆者による適切な用語の検索と選択、そしてMTによるそれらの用語の正確な訳し分けを担保するためには、統制された多言語用語集の整備が不可欠である。さらに、文書構造に応じて、用語リストを定義することも有効だろう。例えば、自治体手続きにおいて必要な「B12. 持ち物」は、全て列挙したとしても、一定の範囲で抑えることができるように、あらかじめ文書構造の側から部分言語的に用語リストを定義しておくことで、効率的な用語検索が可能となる。

このような正確な訳出や文書構造に応じた用語検索を実現するために、抽出した用語の出自となる文脈情報（どの文書のどの部分で使われているか）をなるべく保持することが必要である。本研究では、日英の対訳用語に対して、「主題[6]」「表示構造上の位置づけ（本文／見出し／図表）」「かな読み」「ローマ字読み」を付与した上で、文書構造に応じた用語リストも構築していく予定である。

4 システムの概要

これまで整理してきた文書構造と言語表現の要件を満たした形で多言語文書作成を支援するシステム環境について述べる（図2参照）。

構造化文書テンプレート 定式化した自治体手続き型文書の文書構造を穴埋め式のテンプレートとしてあらかじめ用意しておくことで、執筆者は必要な要素を漏れなく、記述することが可能である。また、各要素の入力ボックスに応じて、制限言語ルールや用語リストの詳細な定義が可能である。なおDITAに則り、表1に加えて、以下の要素をテンプレートに組み込んだ。

- タイトル（title）
- トピックの簡単な説明（short description）
- メタ情報（prolog）
  - 執筆者氏名（author）
  - 責任部署（publisher）

制限言語チェッカー 構造化文書テンプレートの各入力ボックスに文章を記述すると、制限言語ルールに違反した箇所をアラート表示するチェッカー機能を実装した（図2）。アラートと併せて、一部書き換え案を提示することで、執筆者はインタラクティブに書き換えを遂行できる。現段階では、既に構築した22種類の制限言語ルールのうち、次の12種類の実装を試みた。

- 一文はできる限り短くしてください。
- 文の中に、括弧書きで長い説明を入れないでください。
- 修飾語と被修飾語の関係を明確にしてください。
- 「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。
- 1つの文の中で複数の否定形を使わないでください。
- 口語表現の「〜になります」表現を避けてください。
- 「〜という」表現はなるべく省いてください。
- 「ような」、「こと」、「もの」はなるべく省いてください。
- 「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。
- サ変名詞にはなるべく「行う」を付けないでください。
- 「〜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。
- サ変名詞をつなげた複合語を避けてください。

また、文書構造に応じたスタイルルールとして、「タイトル（title）」要素において、文末を名詞形に統一するルールを試験的に実装した。

用語検索機能とMT辞書 まず用語検索機能として、執筆しながら自治体の固有表現や法律・医療などの専門用語をシームレスに検索・選択できるシステムを実装予定である。各用語に付与するメタデータをアクセスポイントとして検索対象に含めることで、なるべく漏れなく用語にアクセスできるようにすると同時に、テンプレートの入力ボックスに応じて、あらかじめ用語の検索範囲を絞り込むことで検索効率を高める。

また入力ボックスごとにMT辞書を登録することで、高精度の訳し分けを実現する。

5 おわりに

本研究で提案するシステム環境は、MTを使った多言語展開を見据えて、文書作成工程において執筆者を支援するものである。文書構造の側から、言語表現を捉えることで、MTをはじめとした言語処理技術の性能を最大限に引き出しながら、文書全体の品質を改善していく点が最大の特徴である。今後の課題としては、文書構造に応じた制限言語ルールと用語集のチューニングが大きく残されている。また本システムはあくまでプロトタイプの段階であり、対象文書の拡張や機能・インタフェースの改善と並行して、各種評価を実施することが必要である。

[6]「税金」「子育て」「教育」など文書のテーマに関する情報。

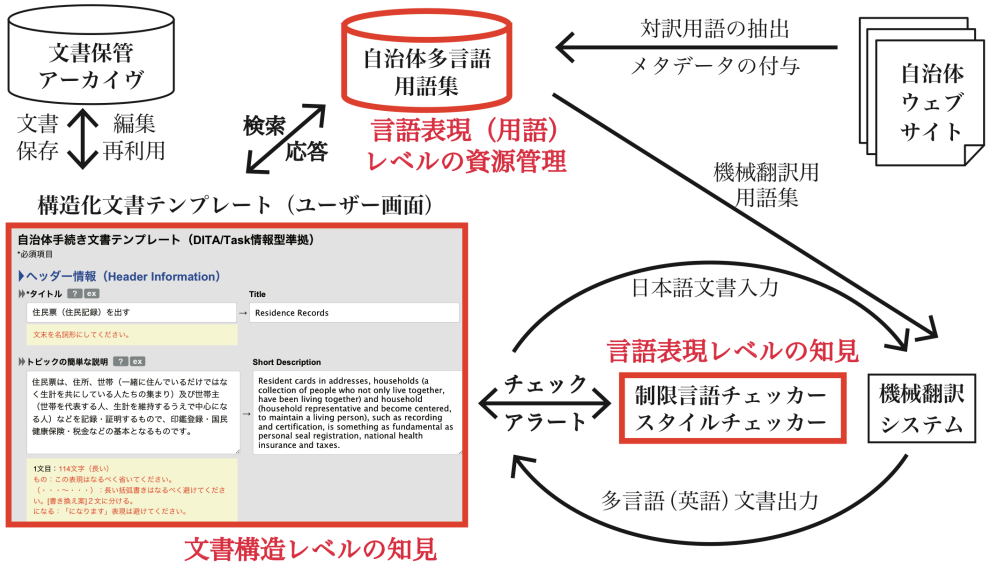

図 2: 多言語文書作成支援システムの全体構成

謝辞　制限言語ルールの作成は、立見みどり博士と豊橋技術科学大学の井佐原均教授との共同研究によるものである。また本研究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。

参考文献

[1] Roturier J. Assessing a Set of Controlled Language Rules : Can They Improve the Performance of Commercial Machine Translation Systems? *The 26th International Conference on Translating and the Computer*, pp.1-14, 2004.

[2] 吉見毅彦, 佐田いち子, 福持陽士. 頑健な英日機械翻訳システム実現のための原文自動前編集. 自然言語処理, Vol.7, No.4, pp.99-117, 2000.

[3] 渡邊豊英. 産業日本語プロジェクトの概要 特許・技術情報の利用性向上のために. 情報管理, Vol.53, No.9, pp.480-491, 2010.

[4] 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.

[5] 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, Vol.1984, No.27, pp.1-8, 1984.

[6] Nyberg E. et al. Controlled Langauge for Authoring and Translation. Somers H. ed. *Computers and Translation: A Translator's Guide*, John Benjamins, pp.245-281, 2003.

[7] Hartley A., Paris C. Multilingual Document Production From Support for Translating to Support for Authoring. *Machine Translation*, Vol.12, No.1-2, pp.109-129, 1997.

[8] Biber D., Conrad S., *Register, Genre, and Style*, New York: Cambridge University Press, 2009.

[9] Swales J. M., *Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press, 1990.

[10] 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.

[11] 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.

[12] 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.