Controlled Document Authoring in a Machine Translation Age

Rei Miyata

First published in 2021

ISBN: 978-0-367-50019-1 (hbk)
ISBN: 978-1-003-04852-7 (ebk)

1 Introduction

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

1 Introduction

1.1 Background

In this digital age, we have witnessed an increasing proliferation of information that is digitally created and disseminated online. In conjunction with this, rapid advances in translation technologies, such as machine translation (MT), have promoted the multilingualisation of digital text. Not only companies but also governments have increasingly adopted commercially or freely available MT systems to create translations in multiple languages to reach a wide audience. End users themselves can take advantage of online MT services to obtain information communicated in languages that they cannot understand.

In the context of Japanese municipalities, which is the main focus of this book, a variety of information is published online regarding not only regional events and tourism, but also certain procedures that must be complied with when living in the municipalities (e.g. registering residency with the local city hall; sorting and recycling garbage; taking action in the case of emergencies). In general, such texts are produced in the official language(s) of the country—in our case, Japanese. There are, however, many foreign residents who do not have the Japanese language skills necessary to understand official documents written in Japanese. Although some of the larger municipalities provide human translations of their websites into various other languages spoken by local communities, the target languages are limited, usually to English alone. Moreover, the scope of the translated versions is often much more restricted than the original Japanese documents since, as Carroll (2010, p.386) points out, 'to expect local governments with limited resources to translate their entire websites into one or more foreign languages would be unrealistic'. In most small municipalities, resources are so scarce that they cannot even provide English translations. Under such circumstances, municipalities typically rely on MT tools, or else the residents themselves rely on MT, such as Google Translate, to grasp the meaning of texts.

Several issues arise. The original Japanese documents are typically created by non-professional writers, since local councils often do not have the financial resources to hire trained authors. Despite recent advances, MT tools are still known to be imperfect. The documents are often embedded in HTML (since they are web pages), which may further complicate automatic translation by

fragmenting sentences. As a result, the translated texts are often misinterpreted or not understandable by their intended audience.

1.2 Problems

We identified that residents might encounter several difficulties when they read governmental or municipal documents, at three different levels. We outline each of them below.

1.2.1 Document-level issues

In Japanese municipal documents, we often find cases in which individual sentences make sense, but directions provided by the document as a whole are confusing. Figure 1.1 provides an example of this.[1] The figure explains the process for registering personal seals in Shinjuku City, Tokyo, which is one of the largest municipalities in Japan and provides human-translated information in multiple languages.

When we read this document, we may be at a loss, wondering whether we are eligible for seal registration or not, since the eligibility conditions for registering a personal seal are stated only at the bottom of the document. Furthermore, the conditions are not expressed in a clear manner, for example: 'Those Who Are Not Eligible for a Residence Records'. Similarly, the requirements for re-registering a seal are vague: 'If You Move Out of Shinjuku City'; 'If You Leave Japan'. Moreover, the sub-section 'Personal Seals Registration Certificate', which is a distinct and separate task from seal registration, offers no explanation as to why we would need to obtain such a certificate and whether it is required or optional, although this is alluded to at the end of 'Personal Seal (*Inkan*)'.

1.2.2 Sentence-level issues

Suppose you have not yet gained a sufficient operational command of Japanese to understand instructions when they are provided in Japanese only. While some information on municipal websites is provided in languages such as English, Chinese, Korean and Portuguese, these translations are usually created in part by MT. Many municipalities provide information vital for completing necessary administrative procedures or tasks in daily life in Japanese only. In such circumstances, having recourse to free, online and thus readily available MT may seem an attractive option. However, while the quality of MT output is reasonable for many language pairs in many practical situations, MT systems does not always produce satisfactory results. Occasionally, you will encounter MT outputs similar to the following:

(a) Garbage let's take away.
(b) From July 2013, you will not be able to use only the specified garbage bag.

Seal Registration

■ **Personal Seal (*Inkan*)**
In Japan, personal seals are used as a symbol of agreement or approval, like a signature, to verify official documents, such as contracts.
You can order a personal seal for your name at a stamp engraving outlet and register the imprint at the City Office. When necessary, you can request a personal seal registration certificate that certifies that the personal seal is registered.
■ **Personal Seal That Cannot Be Registered**
- Stamps with letters that do not combine to form part of your full name, last name, or first name as registered in your residence record
- […]
- Stamps that are inappropriate for registration (for example, stamps without an outer rim, cracked stamps, ready-made stamps, ring stamps, etc.)

■ **Personal Seal Registration Procedures**
Please bring the personal seal you wish to register along with your valid residence card [...] and complete the application procedures in person. […] Some registration restrictions apply, such as those on age (must be 15 years of age or older).
[...] bring the following items to the service counter where you filed your application:
- The response sheet
- [...]

When registration has been completed, you will be issued a personal seal registration card. [...]
■ **Personal Seals Registration Certificate**
To apply for a certificate, please complete application procedures [...] and show your personal seal registration card. You will be issued a personal seal registration certificate, which certifies that your personal seal has been registered. [...]
■ **When Notification Is Necessary (for Personal Seal Registration)**
(a) If you lose your personal seal […] → Notification of Discontinuation
[...]
■ **If You Move Out of Shinjuku City (Personal Seal Registration)**
If you have completed personal seal registration but are moving out of Shinjuku City, [...].
■ **If You Leave Japan (Personal Seal Registration)**
If a person with personal seal registration leaves Japan, the personal seal registration becomes invalid and is deleted
Even if you move back to the same address, you must complete personal seal registration again.
■ **Those Who Are Not Eligible for a Residence Record**
Anyone who is not eligible for a residence record [...] cannot register a seal.

Figure 1.1 Personal seal registration procedure (excerpted from the website of Shinjuku City)

The Japanese input for (a) is ‘ごみは持ち帰ろう/*Gomi wa mochikaero*’, a sensible translation of which is ‘Please take your garbage home’. The MT error stems from differences between the Japanese and English languages when expressing public requests. The Japanese input for (b) is ‘2013年7月からは、指定ごみ袋しか使えません/*2013-nen 7-gatsu kara wa, shitei-gomibukuro sika*

tsukae-masen', which is correctly translated as 'From July 2013 you can use only the specified garbage bags'. In this case, the MT system has mis-translated the Japanese construction 'しか...ない/*shika ... nai*', which is somewhat akin to a double negation.

While, as a native or non-native speaker of English, you may be able to guess what (a) means, you may equally be misled into interpreting this notice as a call for community volunteers to clean up garbage in, for instance, a nearby park. In the case of (b), the meaning of the MT output is the exact opposite of the Japanese original. You are therefore at risk of completely misinterpreting the message or, even if you suspect its true intended meaning, being left in a state of uncertainty. Such misunderstandings and doubts can pose significant problems for you when living as a non Japanese-speaking resident in Japan.

Given that MT systems are being and will continue to be widely used in this domain, along with the cost of having all municipal information translated solely by human translators, improving the quality of MT output to a reliable level is an urgent task.

From the perspective of MT technologies, MT systems dealing with Japanese as source language (SL) or target language (TL) are state-of-the-art level. To date, Japanese natural language processing (NLP) researchers have invested considerable energy in MT research and at times have been world leaders in developing this technology, such as the paradigm proposed for example-based machine translation (EBMT) (Nagao, 1984; Sato and Nagao, 1990). Moreover, there are many commercially and freely available Japanese MT systems developed in Japan which are based on different architectures and technologies: rule-based machine translation (RBMT), which uses (manually constructed) rules to transform SL into TL; EBMT, which uses analogical reasoning based on translated examples; statistical machine translation (SMT), which relies on statistical learning from large aligned bilingual text corpora; and neural machine translation (NMT), which also uses large bilingual corpora to build a neural network model that consumes source text (ST) and generates target text (TT) in an 'end-to-end' manner. Although we have recently witnessed the great improvement of MT since the advent of NMT, current MT systems still face difficulties in dealing with certain types of linguistic patterns, such as long complex sentences. Alternative—or complementary—approaches to improving the performance of MT include imposing restrictions on the form and/or length of SL texts, using controlled language (CL). If we can diagnose what MT can and cannot do and embed MT within the overall framework of information flow, we anticipate being able to use MT for producing reliable outputs.

1.2.3 Terminology issues

Finally, terminology issues cannot be ignored in terms of accurate and consistent understanding of both ST and TT (Wright and Budin, 2001; Warburton, 2015b). We sometimes observe that several different terms refer to the same concept within a website or across websites of municipalities. For example, in a municipal website, '印鑑証明書/*inkan-shomei-sho*' and '印鑑登録証明

書/*inkan-toroku-shomei-sho*' are used in source Japanese side and their English translations are respectively 'personal seal proof certificate' and 'seal registration certificate'. These variations may be confusing for those who have not sufficient domain knowledge about Japanese municipal procedures.

Moreover, from the point of view of MT technologies, terminology influences the output quality. One example MT output awkwardly translated is 'Burned trash'. The Japanese input is '燃やすごみ/*moyasu gomi*', a proper translation of which is 'Combustibles'. The particular MT system fails to capture the term '燃やすごみ', and erroneously processed the verb '燃やす/*moyasu*' (burn) as a past participle. The problem of technical terms can be addressed by maintaining terminology for MT dictionaries.

1.3 Solution scenario

We thus have different, but related, problems. We might face both ill-organised document structure and poor MT output, which may well aggravate the situation as readers have no reliable context to aid them in 'guessing' the actions being described. This observation led us to realise the necessity of pursuing a unified solution to the overall issue of multilingualisation of municipal procedural information, namely, introducing *controlled authoring*, where the control is applied consistently and seamlessly at both the sentence- and document-level.

Controlled authoring is 'the process of applying a set of predefined style, grammar, punctuation rules and approved terminology to content (documentation or software) during its development' (Ó Broin, 2009, p.12). According to the ISO standard, controlled authoring is defined as an 'authoring that uses limited vocabulary and textual complexity to produce clear documents' (ISO, 2012). Though these two definitions focus on linguistic and terminological control of authoring processes, we notice that they presuppose the existence of 'documentation' or 'documents'. Hence, we can reasonably extend the idea of controlled authoring to cover document-level control, and thus we propose the notion of *controlled document authoring*, that is, authoring that uses formalised document structure, limited grammar, lexicon and style, and approved terminology to produce well-structured documents that are clear and consistent.

As we will elaborate in later chapters, we find that an integrated approach of embedding controlled sentences within a well-designed document structure has clear benefits in further improving MT output. Take, for instance, '文書を印刷する/*Bunsho o insatsu-suru*', which may naturally appear as a task title or as a step in a procedure. A given MT system may translate this as 'To print the document', which is appropriate wording for a title but not for a step in a process, where 'Print the document' (imperative) is needed. If we know the functional element in which a Japanese expression occurs, we can exploit this knowledge to pre-process expressions where necessary by transforming them so that the MT system is coerced into producing a contextually appropriate English translation. Since the pre-processing would be an internal operation which does not change the Japanese text seen by readers, the readability of the ST would not be degraded

by TL-oriented writing rules designed to improve MT output quality, which can happen (e.g. Hartley et al., 2012).

While the idea of contextual MT, i.e. the unification of document elements, CL and MT, has been already proposed (Bernth, 2006; Hartley, 2010), its feasibility and applicability have not yet been fully investigated. We address this research challenge with a special focus on the task of translating municipal procedural documents, which is the most significant contribution of this book.

1.4 Research questions

The main research question to be answered in this book is as follows:

RQ Can controlled document authoring help non-professional writers to create well-structured source texts that are machine-translatable and human-readable?

This research question can be divided into two aspects: framework and application. From the point of view of framework, we specify our questions as follows:

RQ-F1 Can municipal documents be well formalised?
RQ-F2 To what extent can a Japanese CL improve the quality of source text (ST) and target text (TT)? (**sentence-level CL**)
RQ-F3 Can the combination of controlled language and document structure further improve the TT quality without degrading ST quality? (**document-level CL**)
RQ-F4 Can municipal terms be comprehensively captured and well controlled?

To answer **RQ-F1**, we employ an existing document standard used for technical documentation. If we can properly formalise the municipal documents based on this standard, we can conclude that they are well formalised. To answer **RQ-F2**, we formulate CL rules specifically intended for the municipal domain and evaluate their effectiveness in terms of ST readability and MT output quality. To answer **RQ-F3**, we contextualise CL rules into the municipal document structure and diagnose the MT outputs. To answer **RQ-F4**, we construct Japanese–English bilingual controlled terminologies and evaluate them in terms of coverage and quality of control.

Based on the results obtained through answering the questions **RQ-F1** to **RQ-F4**, we propose and implement an authoring support system, **MuTUAL**, which is designed to help non-professional municipal writers create controlled documents that are both machine-translatable and human-readable. The core module of the system is the controlled authoring assistant, which automatically checks conformity to the CL and controlled terminology when users are drafting and rewriting ST. Hence, in terms of application, this book asks following questions:

RQ-A1 How accurately does the system detect CL rule violations in text?
RQ-A2 Is our system usable for non-professional writers?
RQ-A3 Does the use of the system help improve the quality of ST and TT?

To answer **RQ-A1**, we implement CL violation detection rules and benchmark their detection performance using a test dataset. To answer **RQ-A2** and **RQ-A3**, we conducted a usability evaluation to see whether our proposed system can improve the user's writing performance and output text quality.

1.5 Scope

The ultimate goal of this research project is to provide an integrated authoring environment that makes use of off-the-shelf MT systems to enable writers to create and publish documents and their multilingual equivalents in the Japanese municipal domain. Tackling all possible varieties of texts available on municipal websites and also multiple target languages is not a realistic goal at this stage. As such, this book focuses on the task of Japanese-to-English translation of municipal documents regarding daily life as a starting point, with a specific focus on procedural documents.

According to Oda (2010, p.22) and OpenUM Project (2011, p.9), Japanese municipal websites typically feature the following kinds of information:

1. Legal information (including constitution, law, government ordinance and ministry ordinance)
2. Official information (including municipal bylaw, regulation and notice)
3. Public-related information
    - 3-1. Information for residents (**municipal-life information**)[2]
    - 3-2. Information for business operators
    - 3-3. Information for tourists
    - 3-4. Policy information of administrations

Municipal-life information usually pertains to content which is directly related to citizens' daily life, and there is a growing need for multilingualisation of this content. In particular, procedural documents are of most importance as they enable residents to avail of municipal services (such as child allowance) and carry out necessary municipal procedures (such as tax payment). From a methodological standpoint, we assume procedural documents are well-suited for document formalisation, and we can make use of existing document structures developed in the field of technical writing and business documentation. Therefore, as a pilot study, we investigate municipal procedural documents, which offers a point of reference for future work.

The reasons why we choose English as the target language are as follows:

- English is still an overwhelmingly popular choice when translating Japanese municipal texts, followed by Chinese, Korean and Portuguese (Carroll, 2010).

- MT systems often use English as a pivot language. For example, Google Translate appears to produce Japanese-to-Vietnamese translation by first translating Japanese into English and then translating English into Vietnamese.[3] Thus, improving English-language MT output quality leads to secondary improvements in MT output quality for many other languages.

It is worth noting in advance that the framework and system environment that we propose in this book are applicable to other text domains/types and language pairs.

1.6 Chapter organisation

This book consists of four parts, divided into a total of ten chapters (see Figure 1.2).

Part I, Chapters 1 and 2, explains the background to our research. Chapter 2 summarises existing research on document formalisation, MT and its practical implementation, CLs, terminology management and authoring environments.

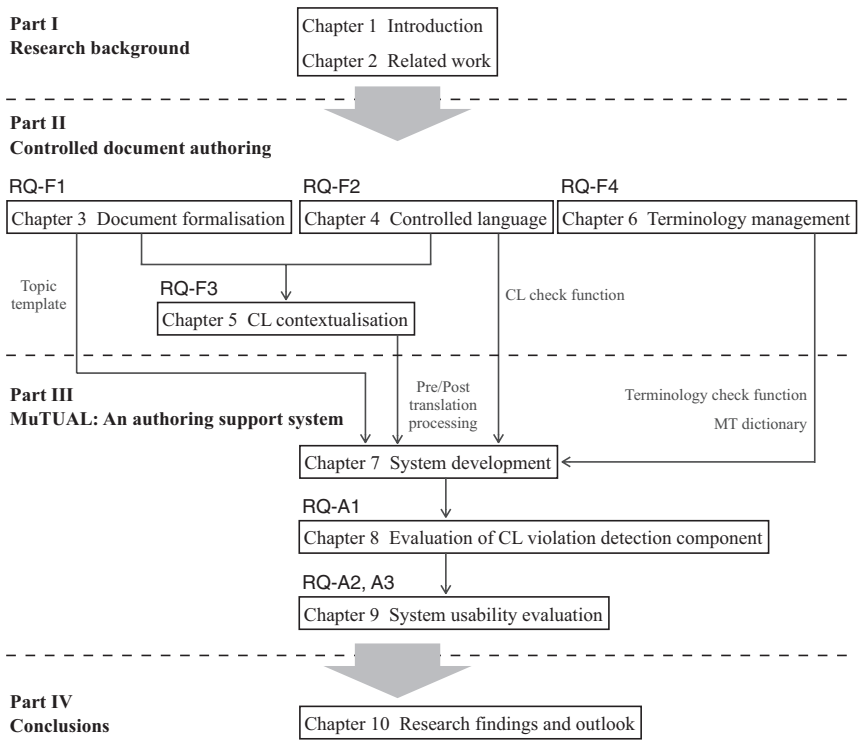

Figure 1.2 Chapter organisation with research questions

Part II, Chapters 3–6, presents our research on controlled authoring at different textual levels. In Chapter 3, as a document-level study, we present (with examples in English) an analysis of procedural texts from Japanese municipalities, and show how a standard document structure can be specialised to cover these texts. In Chapter 4, as a sentence-level study, we design and evaluate Japanese controlled-language rules for improved machine-translatability and source readability, focusing on texts featuring municipal-life information. Chapter 5 details how to combine document structure with controlled language, which is the most innovative idea proposed by this study, and suggests mechanisms to further improve text quality. In Chapter 6, as a terminology-level study, we manually construct and evaluate Japanese–English controlled terminologies for the municipal domain.

Part III, Chapters 7–9, proposes an authoring environment, MuTUAL, that exploits the framework established in the previous chapters to support the writing of municipal procedures such that they are easily translatable by automated tools. Chapter 7 demonstrates the concept, module organisation and intended use scenario of MuTUAL, and describes the implementation of each module. We first evaluate the precision and recall performance of a subcomponent of the system in Chapter 8. Based on the result, we conduct a user study to evaluate the usability of the core authoring module in Chapter 9.

Part IV, Chapter 10, summarises research findings and concludes the book. We discuss the results of the previous chapters, pointing out the major contributions and limitations of the study, and sketch out our future plans towards a practical implementation of the system in real-world scenarios.

Notes

1 Shinjuku City, Seal Registration, www.foreign.city.shinjuku.lg.jp/en/todoke/todoke_6/
2 In this book, we use ‘municipal-life information’ instead of ‘information for residents’.
3 This is not announced by Google, but we can reasonably infer that Google Translate uses English as a pivot language from the fact that a Japanese-to-Vietnamese MT output is almost the same as the Japanese-to-English-to-Vietnamese MT output.