Developing and Validating a Revision Scheme to Account for Both Source and Target Texts: The Case of Japanese to Chinese Translation of Hybrid Texts

Beibei He
Rikkyo University

Documents that feature technical content with a marketing function have become a trending genre in the translation industry, towards which clients tend to have high quality expectations. In order to provide a systematic and consistently applicable revision scheme for such hybrid texts, this paper develops and validates a revision scheme customized for the translation of hybrid corporate documents from Japanese to Simplified Chinese. The paper also describes the identification of problematic source features that inhibit translation quality. Thus, the customized scheme accounts for not only the target texts but also the source texts. Using iterative cycles of annotation and discussion of disagreement, moderate inter-assessor agreement on the categorization of revisions is achieved. Identifying translation errors using the customized scheme is a precursor to identifying source text features that correlate to errors in the translations.

Keywords : quality assessment, revision scheme, error detection, problematic source features, translation practice

1. Introduction

In Japan's translation industry, Simplified Chinese (ZH-CN) has become the major target language for translation from Japanese (JA). Based on the author's experience as a reviser and QA checker working in the industry, the quality of the draft ZH-CN translation of JA tends to entail time-consuming and expensive revision that

involves consultation with the client. The same observation can be seen in TAUS (2017a: 10) which points out "[m]any people in the human translation (HT) space say translators are not good enough to tackle today's challenges, and more and more errors are detected than in the past." The quality issue becomes even more outstanding when it comes to hybrid texts which have both technical content and a marketing function, a trending genre in the translation industry. Texts that have more than one function are referred to as 'hybrid' texts in both translation studies (Munday 2016: 114-117) and in the translation industry. In its April 2018 journal, Amelia[1], a well-known translator network in Japan, highlighted this trending genre by featuring interviews on translation as content marketing with language service providers (LSPs) in Japan, such as SDL Japan and Lionbridge Japan. According to Amelia, the need for translating content for promoting product sales has been increasing over the past few years, creating a new and promising domain with, however, insufficient competent translators. The volume is relatively limited compared to manuals, but clients tend to have high expectations towards quality because these documents have a direct influence on sales performance. According to the interviews, typical hybrid source types are: articles and videos for presenting products, use cases, advertisements and E-learning content.

This paper presents the analysis of 11 publicly available ZH-CN translations of JA CSR (Corporate Social Responsibility) reports which can be clearly seen as belongings to this same hybrid genre, as explained further in Section 2.1. It does so through quality assessment performed by three assessors. The result reveals an unacceptably high incidence of translation errors, demonstrating this is a systemic issue in JA to ZH-CN translation. What also emerges is that JA to ZH-CN translation quality issues in this hybrid genre originate to a significant degree in problematic source texts (STs). Motivated by the presence of errors in the target texts (TTs) and problematic features in the STs, this study develops a novel revision scheme that accounts for both STs and TTs. Mossop defines 'revising' (or revision) as "reading a translation in order to spot problematic passages, and making any needed corrections or improvements" (Mossop 2014: 1). This study adopts

1) https://www.amelia.ne.jp/

Mossop's definition of 'revising' that covers both 'corrections' and 'improvements and thus prefers the label 'revision scheme' to the more common 'error scheme'. It refers to problematic text spans (a continuous/discontinuous sequence of words/characters) in TTs as translation 'errors' and to problematic text spans in STs as 'features', both of which "involve checking linguistic correctness as well as the suitability of a text's style to its future readers and to the use they will make of it" (Mossop 2014: 1). Therefore, categories in the customized revision scheme cover not only TTs but also STs.

Section 2 describes the CSR reports which form the basic data for the study and sets out the research questions. Section 3 discusses available error schemes and the process of merging the two schemes finally selected, which are the TAUS Harmonized DQF-MQM Error Typology[2] and the Framework for Standardized Error Marking[3] of the American Translators Association (ATA). Section 4 introduces the process of validating the customized revision scheme. Section 5 discusses the results observed from the validation process. Section 6 offers a conclusion to the paper.

2. Data selection and research questions

2.1. CSR reports as data

As mentioned above, hybrid texts have become a 'hot' genre in the industry. However, expectations of high quality and a shortage of competent translators have made it too hot to handle. In addition, existing error annotation schemes widely used in the industry are not suitable for this genre because most were designed for assessing solely technical content. This becomes a major obstacle when it comes to quality assessment, since some of the translation issues that arise cannot be accurately captured using standard benchmarks. This paper proposes an error

2) https://www.taus.net/evaluate/qt21-project#harmonized-error-typology
3) https://www.atanet.org/certification/aboutexams_error.php

annotation scheme that can account for hybrid texts. One example of such documents that I have personal experience of is CSR reports.

CSR stands for Corporate Social Responsibility. A CSR report, according to Global Reporting Initiative[4] (GRI), is a report "published by a company or organization about the economic, environmental and social impacts caused by its everyday activities", which "also presents the organization's values and governance model, and demonstrates the link between its strategy and its commitment to a sustainable global economy". Japan is one of the leading countries in CSR reporting as a result of encouragement from the Japanese government and the Japanese Business Federation. Most corporations in Japan produce CSR reports in accordance with international guidelines such as ISO and GRI standards. GRI's Sustainability Reporting Guidelines prescribe topics for the report, among which topics such as economic performance, marketing and labelling and general disclosures (disclosure of activities, brands, products and services, etc.) have a strong advertising function. Different companies produce different content, but their CSR reports share the same headings and scope, the same type of readers, the same functions of presenting technical innovations and of marketing. Thus, they can be seen to belong to the same genre. Although the aim of CSR reports is not an increase in sales of a certain product, the optimistic outlook towards the future development of the company justifies likening these documents to marketing texts. Moreover, large corporations tend to entrust the translation of CSR report into other languages to localization companies because it is a complicated task which requires specialized (1) resources (e.g., multi-lingual translator resources), (2) expertise (e.g., knowledge and skills regarding desktop publishing and web content management) and (3) tools (e.g., translation memory tools and DTP tools).

The following JA text and its published English (EN) translation exemplify the hybridity of CSR reports. The company names are anonymized as 'XXX' and 'YYY'.

4) https://www.globalreporting.org/information/sustainability-reporting/Pages/default.aspx

JA: XXXは、写真分野で培ったコラーゲン技術を基盤に開発した、細胞培養に必要な細胞外マトリックス（リコンビナントペプチド：RCP）「cellnest」を研究用試薬として発売しているほか、2014年当時、国内で唯一、自家培養表皮や培養軟骨等の再生医療製品を上市していたYYYを連結子会社化。

EN: XXX is marketing “cellnest” (recombinant peptide or RCP), an extracellular matrix essential for cell culture developed with Fujifilm’s collagen technology accumulated in photography business. In addition, in 2014 we acquired YYY, the only company in Japan marketing autologous cultured epidermis and autologous cultured cartilages at that time.

Based on the three functions of language (informative, expressive, appellative) identified by Bühler (1934/1965), Reiss (1977/1989) categorizes text types (or genres) into four types: informative, expressive, operative and audio-media texts. When a text features the characteristics of more than one text type, it is regarded as a ‘hybrid’ type (Munday 2016: 116). As shown in the example above, a CSR report not only transmits information regarding the company which produced the CSR report (thus informative), but also fulfills the operative function by attempting to earn trust from the readers, promote the companies’ products or services and attract investment.

I take CSR reports as data first because of their hybridity, and second because their publication annually is a statutory requirement for Japanese corporations. Moreover, since they are publicly available with back numbers, research based on a substantial corpus becomes feasible. Finally, I have a professional familiarity with the genre. Using a 24,001-character parallel corpus of CSR reports published by 11 companies, I validate – through satisfactory inter-reviser agreement – a novel combination of revision categories from DQF-MQM and ATA. This scheme is customized to accommodate the hybridity of CSR reports, in other words, to account for both technical and marketing elements.

2.2. Research questions

This research addresses the following questions:

RQ1: What types of translation errors can be found in JA to ZH-CN translations of CSR reports published by major Japanese companies, and how frequently do these errors occur?
RQ2: Can they be accommodated in an existing error annotation scheme?
RQ3: If not, can a new scheme be devised which can obtain a satisfactory level of agreement between assessors?

Section 3 explains why the DQF-MQM and ATA error typologies were selected as the foundation for establishing a novel revision typology and the process of mapping them into my customized scheme, which is referred to hereafter as the 'CSR' scheme.

3. Developing a revision scheme to accommodate the hybridity of CSR reports

Offering an overview of the existing error classification schemes used in the translation industry, Secară points out that "[t]he commitment to deliver error-free translations to clients, the enormous amount of materials to be translated, and the growing competition between translation providers resulted in an increasing interest in quality assurance." (Secară 2015: 39). According to her, the motivation for creating explicit and applicable correction scales and translation error typologies since the 70's was to reduce such factors as time, money, human effort and subjectivity and to introduce a more systematic type of analysis. Also, the acceptability of a translation is often based on a threshold of a certain number of errors.

Many schemes exist, but “there are no generally accepted objective criteria for evaluating the quality of both translations and interpreting performance. Even the latest national and international standards in this area [···] do not regulate the evaluation of translation quality in a particular context. [···] The result is assessment chaos.” (Institut für Angewandte Linguistik und Translatologie 1999, cited in Williams 2001: 327). The Multilingual e-Learning in Language Engineering (MeLLANGE) project analyzed, evaluated and restructured the error categories implemented in well-known models such as: ITI (Institute of Translation and Interpreting, UK); SAE J2450 (a quality metric developed by the Society of Automotive Engineers in collaboration with General Motors); BlackJack (developed by the British translation agency ITR to overcome and improve existing systems, such as J2450) and ATA. On this basis, an error typology was designed and tailored to assessing learners’ translations (Castagnoli et al. 2006). A simplified version of the MeLLANGE typology, the MNH-TT typology (Babych et al. 2012), was adapted by conflating various subcategories and reducing the number of issue types, again for teaching/learning purposes. Noting that the MNH-TT typology in itself did not necessarily guarantee consistent human assessments, Fujita et al. (2017) demonstrated the advantages of a decision tree as a ‘navigation’ tool for consistent human decision making.

Existing error classification schemes differ in their granularity and their organization of error types, because the scope and granularity of errors depend on the purpose of translations and the aim of human assessments (e.g., formative or summative) (Fujita et al. 2017). The current paper develops a revision scheme for assessing professional ZH-CN translations of hybrid texts in JA. Taking into account that the typology alone does not necessarily guarantee consistent human assessments (Lommel et al. 2015), this research measures the reliability of the CSR scheme for annotation using a measure of inter-annotator agreement, which is an indication of the consistency with which assessors agree in their annotations. It further develops a decision tree to assist consistent decision making among assessors.

It is worth noting that the existing industrial error typologies focused on technical documents are mostly based on the assumption that all STs are translatable, and

thus they generally overlook any deficiencies in the STs. Therefore, they provide no mechanism for describing what is wrong in the STs. In the face of this shortcoming, the CSR scheme introduces a list of error categories to account for STs (see Table 6).

Section 3.1 introduces the DQF-MQM and ATA schemes, while Section 3.2 describes the process of merging the two selected schemes into a novel revision scheme customized for ZH-CN translation of JA CSR reports.

3.1. DQF-MQM and ATA error typologies

Both the TAUS DQF-MQM and ATA error frameworks are widely used error typologies which are available online at no cost.

ATA certification is one of the most respected and recognized credentials, and the Framework for Standardized Error Marking is its certification exam grading system, which is used to assess the language skills of a would-be professional translator. These skills cover comprehension of the ST, translation techniques and writing in the target language. ATA evaluators refer to a list of error categories (see Table 1) to identify translation errors and assign error points for each error. A translation with a final score of 18 or higher is marked fail (Koby and Champe 2013: 166).

On the other hand, the Multidimensional Quality Metrics (MQM) is "a flexible system that organizes many types of translation errors (called 'issues' in MQM) into a hierarchy that supports multiple levels of granularity" (Lommel et al. 2015: 4). MQM was developed as part of the (EU-funded) QTLaunchPad project based on an examination of a large variety of existing translation quality metrics. According to Mariana et al. (2015: 139), it draws most heavily from the LISA (Localization Industry Standards Association)[5] QA model and was designed to be applicable to a professional production environment - the translation industry - as well as a testing

5) Existing from 1990 to February 2011, LISA was a trade body for hardware and software publishers and companies involved in the translation of computer software and other documentation into multiple natural languages. Among its activities, it proposed methodologies and standards that would enable its members to achieve high quality as well as interoperability for tools developed according to these standards.

environment. The Dynamic Quality Framework (DQF), created by TAUS simultaneously with the development of MQM, "was based on industry best practice, with a focus on the issues commonly checked by translation service providers" (Lommel et al. 2015: 5). Compared to the MQM hierarchy which was comprehensive and detailed, the DQF hierarchy is smaller and flatter. (Lommel et al. 2015: 4-5). In light of relating the MQM and DQF schemes to avoid confusion among the growing number of users, the developers of the two frameworks, who were also partners of the QTLaunchPad project, agreed to make substantive changes to both frameworks to bring them into harmony. The DQF-MQM scheme, the result of this harmonization, is displayed in a hierarchy with 5 types of severity level which is an indicator of the importance of an issue with an accompanying numerical representation (Lommel et al. 2015: 7): critical, major, minor, neutral and kudos. Kudos is used to praise excellent achievements.

The current paper proposes a revision annotation scheme customized for JA to ZH-CN translations of CSR reports, taking into account that CSR reports are markedly different from the technical documents targeted by the industrial error schemes in so far as they have, in addition to technical content, a marketing function. These two widely used schemes were chosen because the DQF-MQM error typology is useful for its emphasis on errors related to localization aspects while ATA has a focus on textual analysis, which is well suited to the marketing aspect of CSR reports. In prior research that associates these two schemes, Mariana et al. (2015) match the ATA categories to the error categories in the MQM framework using 29 versions of student translations of a short news article (144 words) in a single genre. However, the 'mapping' process is simply to find MQM labels for the ATA categories for the purpose of grading student translations. It is not their aim to construct a novel scheme, unlike CSR, which focuses on a different, hybrid genre using professional translators with a different profile, and with a much larger volume of data.

Section 3.2 describes the process of mapping the two selected typologies into a novel scheme while introducing new categories that exist in neither ATA nor DQF-MQM.

3.2. Mapping of DQF-MQM and ATA schemes

Given their standing and complementarity, the first step is to map these two selected typologies into a novel merged scheme, removing error types that do not apply to JA to ZH-CN translations of CSR reports and merging similar error types to a single category. Table 1 shows the mapping of the ATA and DQF-MQM schemes.

The [CSR scheme] column in Table 1 shows the initial list of categories adopted for establishing the revision scheme — the CSR scheme — customized for assessing ZH-CN translations of JA CSR reports. Since this CSR scheme does not exclude the possibility of source deficiencies which may inhibit translation quality, it prefixes each error category with 'TT' to indicate that it is a translation error. The [ATA category label] column and [ATA category label] column show error categories that exist in their schemes respectively.

The categories fall into six clusters. Discussion of specific cases follows Table 1.

- No. 1-8 are error categories that the CSR scheme adopted from both ATA and DQF-MQM schemes. There is an equation established between ATA and DQF-MQM for each error category.
- No. 9-15 are error categories that the CSR scheme adopted from the ATA scheme which are absent from DQF-MQM.
- No. 16-21 are error categories that the CSR scheme adopted from the DQF-MQM scheme which are absent from ATA.
- No. 22-25 are original categories added to the CSR scheme by the author.
- No. 26-29 are error categories that the CSR scheme did not adopt from the ATA scheme. These categories exist in the ATA scheme but are absent from DQF-MQM.
- No. 30-35 are error categories that the CSR scheme did not adopt from the DQF-MQM scheme. These categories exist in DQF-MQM but are absent from ATA.

Table 1. Mapping of the ATA and DQF–MQM schemes

No.	ATA category label	CSR scheme	DQF-MQM category label
1	Addition	TT-Accuracy-Addition	Accuracy-Addition
2	Omission	TT-Accuracy-Omission	Accuracy-Omission
3	Mistranslation Misunderstanding	TT-Accuracy-Misrepresentation	Accuracy-Mistranslation
No.	**ATA category label**	**CSR scheme**	**DQF-MQM category label**
4	Punctuation	TT-Fluency-Punctuation	Fluency-Punctuation
5	Spelling/Character Capitalization (a sub-category of Spelling/Character) Diacritical marks / Accents (sub-category of Spelling/Character)	TT-Fluency-Spelling/Character	Fluency-Spelling
6	Grammar Syntax (a sub-category of Grammar) Word form / Part of speech (a sub-category of Grammar)	TT-Fluency-Grammar	Fluency-Grammar
7	Usage	TT-Fluency-Awkward/Unidiomatic	Style-Awkward Style-Unidiomatic
8	Other Errors	TT-Other errors	Other
9	Faithfulness	TT-Accuracy-Faithfulness	
10	Faux ami	TT-Accuracy-False friend	
11	Ambiguity	TT-Accuracy-Ambiguity	
12	Unfinished	TT-Accuracy-Unfinished	
13	Literalness	TT-Accuracy-Literalness	
14	Text Type (includes the former categories of Register and Style)	TT-Fluency-Register	
15	Cohesion	TT-Cohesion	
16		TT-Accuracy-Over-translation	Accuracy-Over-translation
17		TT-Accuracy-Under-translation	Accuracy-Under-translation
18		TT-Accuracy-Untranslated	Accuracy-Untranslated
19		TT-Fluency-Inconsistency	Fluency-Inconsistency

No.	ATA category label	CSR scheme	DQF-MQM category label
20		TT-Design	Design-Local formatting Design-Markup Design-Length Design-Truncation/text expansion Design-Missing text
21		TT-Locale convention	Locale convention-Address format Locale convention-Date format Locale convention-Currency format Locale convention-Measurement format Locale convention-Shortcut key Locale convention-Telephone format
22		TT-Accuracy-Transliteration	
23		TT-Accuracy-Calque	
24		TT-Terminology	
25		TT-TM	
26	Illegibility		
27	Indecision		
No.	**ATA category label**	**CSR scheme**	**DQF-MQM category label**
28	Terminology		
29	Verb Tense		
30			Accuracy-Improper exact TM match
31			Fluency-Grammatical register
32			Fluency-Link/cross-reference
33			Terminology-Inconsistent with termbase Terminology-Inconsistent use of terminology
34			Style-Inconsistent style Style-Company style Style-Third-party style
35			Verity-Culture-specific reference

For error categories No. 1-8, an equation can be established between ATA and DQF-MQM. For instance, [TT-Accuracy-Addition] was adopted from [Addition] in the ATA scheme, which equates to [Accuracy-Addition] in the DQF-MQM scheme. For No. 3, [TT-Accuracy-Misrepresentation] in the CSR scheme replaced ‘mistranslation’ (ATA and DQF-MQM) and ‘misunderstanding’ (ATA) with ‘misrepresentation’ to emphasize the importance of making judgments based on

textual evidence, because ‘misunderstanding’ implies that assessors need to guess what was going on in the translator’s mind. Also, ‘mistranslation’ is too broad to be regarded as a subcategory. For simplification, No. 5 was merged with [TT-Fluency-Spelling/Character], No. 6 with [TT-Fluency-Grammar] and No. 7 with [TT-Fluency-Awkward/Unidiomatic].

For categories adopted from the ATA scheme, [Faithfulness], [Faux ami] (renamed to [False friend] for better understandability), [Ambiguity], [Unfinished], [Literalness] and [Register] (taken from the ATA’s former category) are from personal experience error categories that are likely to be observed in professional JA to ZH-CN translations. For the same reason, [Accuracy-Over-translation], [Accuracy-Under-translation], [Accuracy-Untranslated], [Fluency-Inconsistency], [Design] (all sub-categories merged), [Locale convention] (all sub-categories merged) were adopted from DQF-MQM.

Regarding original categories added by the author:

- [TT-Accuracy-Transliteration] is an error type unique to JA to ZH-CN translation, where a JA word is converted to the simplified version of the same kanji in ZH-CN, which does not in fact exist in ZH-CN. For example, ‘介護’ (nursing-care) is translated as ‘介扩’ using the simplified version of ‘介護’, even though ‘介扩’ does not exist in ZH-CN.
- [TT-Accuracy-Calque] is another error type often observed in professional JA to ZH-CN translation, translating the whole sentence word by word inheriting the structure of the ST.
- [TT-Terminology] is a combination of terminology related categories in ATA and DQF-MQM. By ATA’s definition, a terminology error occurs when a term appropriate to a specific subject field is not used when the corresponding term is used in the ST. DQF-MQM, on the other hand, defines terminology error as (a) violation of specified glossary and (b) inconsistency of translating the same term. TT-Terminology in the CSR scheme refers to all three cases.
- [TT-TM] in the CSR scheme indicates errors such as where a translator has applied the wrong translation memory (TM) or translated a sentence without referring to a 100% match in the TM, which may violate the client’s

instruction depending on the agreement between clients and translation agencies on how to work on 100% matches. [TT-TM] also includes the TM-related category (No. 30) in DQF-MQM 'Improper exact TM match' which indicates errors such as when a translator fails to fix a 100% match (a match where there is no difference between the ST in the document and the ST in the translation memory) when the suggested translation does not apply to the current context.

From the ATA scheme, the CSR scheme did not adopt [Illegibility], [Indecision], [Terminology], [Text Type] (includes the former categories [Register] and [Style]) and [Verb Tense] for the following reasons.

- An [Illegibility] error applies only to handwritten texts, when graders cannot read what the candidate has written.
- An [Indecision] error occurs when the candidate gives more than one option for a given translation unit, which rarely occurs in professional settings.
- [Terminology] has been adopted in a broader sense.
- [Register], a sub-category of [Text Type], was adopted in the CSR scheme as [TT-Fluency-Register], while [Style], the other sub-category of [Text Type] (e.g., tone, method of exposition) was also subsumed under [TT-Fluency-Register].
- [Verb Tense] including modality is considered a grammar error in the CSR scheme.

From the DQF-MQM scheme, the CSR scheme did not adopt the following categories for the following reasons.

- [Accuracy-Improper exact TM match] has been adopted in a broader sense.
- [Fluency-Grammatical register] is subsumed under [TT-Fluency-Register].
- [Fluency-Link/cross-reference] is considered as a pre-processing or post-process issue rather than a translation error. If the translator has translated the tags wrongly, it should be an accuracy error.
- [Terminology-Inconsistent with termbase] and [Terminology-Inconsistent use of terminology] are subsumed under [TT-Terminology].

- [Style-Inconsistent style], [Style-Company style] and [Style-Third-party style] are subsumed under [TT-Fluency-Inconsistency].
- [Verity-Culture-specific reference] should be regarded as a deficiency in the ST rather than a translation error.

In summary, the CSR scheme is a combination of error categories from the ATA and DQF-MQM schemes with the addition of some original categories. Section 4 describes the process of validating the CSR scheme through satisfactory inter-assessor agreement.

4. Validating the combined revision scheme

To address the research questions articulated in Section 2.2, an experiment was conducted to validate the CSR scheme with an acceptable level of agreement. The validation process is a loop of iterative annotation, discussion, resolution and refinement.

In the experiment, I took 11 JA CSR report extracts with their official ZH-CN translations published by 11 companies[6] that won Year 2017 Environmental Communication Awards hosted by the Japan Ministry of the Environment. The average word count of each extract is 2,200 JA characters. Three assessors, A (the author), B and C – all of whom meet the requirements for qualified translators as well as revisers defined by ISO 17100: 2015[7] – were employed to use the CSR scheme to annotate translation errors or ST features. I measured the inter-assessor agreement level using the online calculator ReCal3 (Reliability Calculator for 3 or more coders)[8] that computes inter-assessor reliability coefficients for nominal data coded by three or more assessors, and stopped the iterative process once it had yielded a satisfactory level of agreement. ReCal3 calculates four of the most popular

6) The list of award-winning companies: http://www.env.go.jp/press/files/jp/104695.pdf (Japanese only)
7) https://www.iso.org/obp/ui/#iso:std:iso:17100:ed-1:v1:en
8) http://dfreelon.org/utils/recalfront/recal3/

reliability coefficients for multiple coders assessing nominal data: average pairwise percent agreement, Fleiss' Kappa, Cohen's Kappa, and Krippendorff's Alpha. This research uses Cohen's Kappa (Cohen 1960) to indicate pairwise agreement level as well as average pairwise agreement level, and Krippendorff's Alpha (Krippendorff 2011) to indicate the agreement levels between all three assessors. According to TAUS (2017b: 10), "[t]he inter-annotator agreement measurement is a good valuation to calculate how much participating raters agree on their answers. The Kappa coefficient is a widely used measurement, as it takes into account how much agreement would be expected by chance alone." The same guideline interprets Kappa scores as follows:

0.00-0.20: slight agreement
0.21-0.40: fair agreement
0.41-0.60: moderate agreement
0.61-0.80: substantial agreement
0.81-1.00: almost perfect agreement

Using the initial CSR scheme shown in Table 2, three assessors performed the following iterations in accordance with the OntoNotes method (Hovy et al. 2006).

Step 1. Assessor A identifies and marks text spans that require revisions in TTs.
Step 2. All three assessors label error type based on the initial CSR scheme.
Step 3. Assessor A collects disagreements among assessors including comments or questions marked by the assessors during the labelling process, discusses disagreements with the other two assessors until they are resolved, updates the CSR scheme by removing or adding revision categories, clarifying or modifying the definition of each category and developing a decision tree to support consistent decision making.
Step 4. Repeat Steps 1 to 3 to annotate errors using the latest scheme.
Step 5. Terminate the iteration once a satisfactory agreement, i.e. moderate agreement (0.41 to 0.60), is achieved.

Step 6. Assessor A revisits all the errors previously agreed by the three assessors and makes a final proposal based on the latest scheme and definition, before sharing the result with B and C, and resolving any residual disagreements through discussion.

Section 4.1 introduces the final CSR scheme. Section 4.2 describes the inter-assessor agreement achieved in the experiment.

4.1. Final CSR scheme

The CSR scheme was finalized through the iterative process described above. As shown in Table 2, four categories were deleted from the initial CSR scheme with t he remainder unchanged. The deleted categories are [TT-Accuracy-Over-translation], [TT-Accuracy-Under-translation], [TT-Accuracy-Faithfulness] and [TT-Accuracy-Literal ness], which are marked in bold.

Table 2. Comparison between the initial and final CSR schemes

Initial CSR scheme	Final CSR scheme
TT-Accuracy-Addition	TT-Accuracy-Addition
TT-Accuracy-Omission	TT-Accuracy-Omission
TT-Accuracy-Misrepresentation	TT-Accuracy-Misrepresentation
TT-Accuracy-Over-translation	
TT-Accuracy-Under-translation	
TT-Accuracy-Untranslated	TT-Accuracy-Untranslated
TT-Accuracy-Faithfulness	
TT-Accuracy-False friend	TT-Accuracy-False friend
TT-Accuracy-Ambiguity	TT-Accuracy-Ambiguity
TT-Accuracy-Unfinished	TT-Accuracy-Unfinished
TT-Accuracy-Literalness	
TT-Accuracy-Transliteration	TT-Accuracy-Transliteration
TT-Accuracy-Calque	TT-Accuracy-Calque
TT-Fluency-Punctuation	TT-Fluency-Punctuation

TT-Fluency-Spelling/Character	TT-Fluency-Spelling/Character
TT-Fluency-Grammar	TT-Fluency-Grammar
TT-Fluency-Register	TT-Fluency-Register
TT-Fluency-Inconsistency	TT-Fluency-Inconsistency
TT-Fluency-Awkward/Unidiomatic	TT-Fluency-Awkward/Unidiomatic
TT-Cohesion	TT-Cohesion
TT-Terminology	TT-Terminology
TT-TM	TT-TM
TT-Design	TT-Design
TT-Locale convention	TT-Locale convention
Initial CSR scheme	**Final CSR scheme**
TT-Other errors	TT-Other errors

[TT-Accuracy-Over-translation] and [TT-Accuracy-Under-translation] were adopted from DQF-MQM, according to which an over-translation error occurs when the TT is more specific than the ST, while an under-translation error occurs when the TT is less specific than the ST. The previous version of DQF-MQM permitted explicitation and implicitation when necessary, but this exception disappeared in its latest scheme as of April 20, 2018. The initial CSR scheme included both categories because they are observed error types. However, they have caused disagreements in many cases because whether it is an [Under-translation] error or an implicitation necessitated by a translation strategy depends in some cases on an individual assessor's subjective judgment. The same applies to over-translation and explicitation. [TT-Accuracy-Faithfulness] was another contentious category that resulted in disagreements. According to the ATA's definition, a faithfulness error occurs when the TT does not respect the meaning of the ST as far as possible: one should translate the meaning and intent of the ST, not rewrite it or improve upon it. Whether it is a [Faithfulness] error by ATA's definition or a necessary translation strategy is, again, subject in some cases to an individual's subjective judgment. [TT-Accuracy-Over-translation] and [TT-Accuracy-Under-translation] are often confused with [TT-Accuracy-Faithfulness] by the assessors, because when the TT is more or less specific than the ST, even if there is a legitimate reason, it

matches the ATA' definition of [Faithfulness]. These three categories were removed from the CSR scheme for the following reasons.

- In order to fulfil the marketing function of hybrid texts, one is not only required to translate the meaning and intent of the STs but also expected to produce a translation that is appealing to the target readers. Therefore, one may need to depart from the ST in simplifying or elaborating the mode of expression.
- As explained above, since the CSR scheme accommodates the possibility of source deficiencies which may inhibit translation quality, an attempt to compensate for a problem in the ST by providing a translation based on a personal re-construal of the ST is deemed acceptable or even recommendable. This may not be the case when grading a certification exam, but is often the case when translating a document for an agency.
- When it comes to hybrid texts, clients tend to improve upon the translations without referring to the ST, and translation agencies are expected to respect such edits even if the original translation had delivered the meaning and intent accurately.

A focus on the marketing function of translations and the accommodation of source deficiencies are the essential differences between the CSR scheme and the ATA/DQF-MQM schemes.

[TT-Accuracy-Literalness] was adopted from the ATA scheme. By ATA's definition, a literalness error occurs when a translation that follows ST word for word results in awkward, unidiomatic, or incorrect renditions. Although the syntax and morphology of modern JA are largely different from ZH-CN, there is a certain degree of resemblance in vocabulary and syntax between these two languages, since JA borrowed heavily from ZH-CN at various times in the past. Moreover, the grammar and vocabulary of both languages have been heavily influenced by EN and other European languages historically. It is beyond the scope of this research to discuss the relationship between JA and ZH-CN. But in brief, the similarities and differences between JA and ZH-CN make it difficult for assessors to decide whether there is a literalness error in the first place, i.e. whether the interference observed

in the translation is due to an inherent similarity between these two languages thus not a translation error or whether it is a mistranslation. Disagreements emerged when assessors tried to apply this category. Similarities to the ST in word order or syntax found in the translation may lead to different types of translation errors. If a translation deemed literal can deliver the meaning of the ST, but the expression is awkward or unidiomatic to the native speaker, it is a fluency error rather than an accuracy error. If a translation deemed literal has misrepresented the meaning of the ST, it should be counted as an accuracy error. Also, since multiple text spans that require revision can be marked within one sentence, interference should be further categorized to fit different types of text spans (word level or above-word level) and to reflect the complex relationship between JA and ZH-CN. It is for this reason that [TT-Accuracy-Transliteration] (word-level) and [TT-Accuracy-Calque] (above-word level) were introduced in the original CSR scheme.

As mentioned in the description of the procedure taken for validating the CSR scheme, in the iterative process a decision tree was developed as a navigation tool to resolve uncertainties about error categorization.

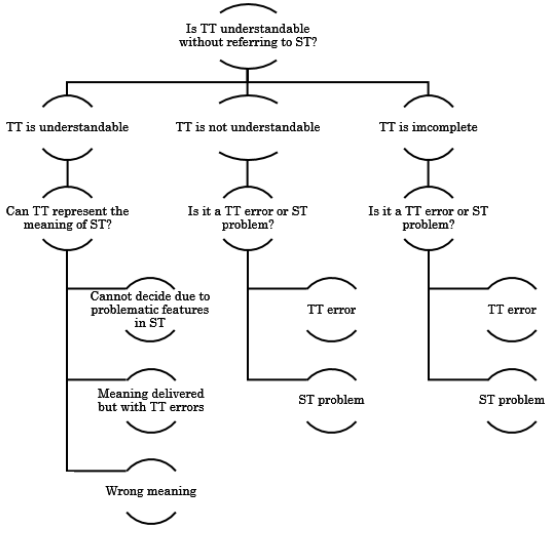

Figure 1. Decision tree

Figure 1 shows the top three levels of the decision tree. Further branches under the above levels define specific error types. This model puts great emphasis on readability and understandability of the TT because of the above-mentioned purpose of a hybrid text. Also, it takes into account deficiencies in ST which may inhibit translation quality and induce translation errors (see Table 6). This CSR scheme does not weigh the severity of each error type because severity is another dimension, which is not the focus of this research.

Section 4.2 discusses how the inter-assessor agreement level was measured and the results.

4.2. Inter-assessor agreement

As shown in Table 3, the assessors labelled error types in text spans previously identified by assessor A in extracts 1 to 3 with an average length of 2,200 JA characters. The procedure changed after extract 4 in that each extract was divided into two for the reason that labelling a whole extract took approximately four hours for each assessor while having discussions to solve disagreements took one-two hours between assessors A-C, A-B and then A-C. For assessors who are professional translators, taking such a long time out of daily work on a regular basis is difficult. Moreover, labelling a great number of errors at once can lead to fatigue. Hence, the workload was reduced to half for each extract.

Extracts 1 to 3 were annotated one by one without discussions in order to allow the assessors to get used to the practice, find problems and questions and establish patterns of annotation. After the first three extracts were done, we started discussing disagreements and refining the revision scheme accordingly. In addition, the decision tree was put to use from extract 4-1, from which point the inter-assessor agreement level started to show improvement. The value of average pairwise Cohen's kappa (Cohen 1960) is nearly the same as Krippendorff's alpha (Krippendorff 2011), which is calculated over all three assessors, when both are rounded to two decimal places.

In addition to extract numbers ('extract'), word count of each extract ('chars'), value of average pairwise Cohen's kappa ('av. kappa'), value of pairwise Cohen's

kappa ('pairwise kappa') between A-C, A-B and B-C, and value of Krippendorff's alpha ('alpha'), three columns are added. They are: (1) 'cases', which indicates the number of text spans that were identified, (2) 'decisions', which indicates the number of error categories that were labelled by three assessors, and (3) 'mean alpha', which shows the mean alpha value of previously annotated extracts. The bottom row shows the average of each column from extract 4-1 to 11-2 excluding outlier extracts 7-1 and 7-2, which are subject to abnormally high disagreement. It should be emphasized that assessors can mark 'no error' if they make a judgment that the text span marked by A needs no revision.

Table 3. Improvement of inter-assessor agreement level with training over time

extract	chars	cases	decisions	av. kappa	pairwise kappa			alpha	mean alpha
					A & C	A & B	B & C		
1	2070	97	291	0.20	0.29	0.20	0.13	0.20	
2	2150	121	363	0.30	0.32	0.31	0.26	0.30	0.25
3	2267	110	330	0.23	0.21	0.24	0.25	0.22	0.24
4-1	1054	46	138	0.39	0.41	0.55	0.22	0.39	0.28
4-2	1106	53	159	0.53	0.64	0.51	0.44	0.53	0.33
5-1	1034	52	156	0.56	0.61	0.59	0.50	0.56	0.37
5-2	1030	53	159	0.44	0.56	0.45	0.32	0.44	0.38
6-1	1155	53	159	0.45	0.54	0.46	0.36	0.45	0.39
6-2	1084	36	108	0.50	0.63	0.47	0.39	0.50	0.40
7-1	1087	41	123	0.31	0.49	0.21	0.22	0.30	0.39
7-2	1080	39	117	0.36	0.45	0.35	0.27	0.36	0.39
8-1	1138	52	156	0.43	0.51	0.49	0.30	0.43	0.39
8-2	1183	49	147	0.44	0.56	0.46	0.31	0.44	0.40
9-1	1028	48	144	0.63	0.70	0.71	0.49	0.63	0.41
9-2	855	49	147	0.49	0.65	0.48	0.35	0.49	0.42
10-1	1128	78	234	0.45	0.43	0.55	0.38	0.45	0.42
10-2	1120	71	213	0.55	0.61	0.62	0.41	0.55	0.43
11-1	1091	62	186	0.51	0.58	0.46	0.48	0.51	0.43
11-2	1105	55	165	0.51	0.54	0.61	0.39	0.51	0.44
Av. 4-1 (7-1/2) - 11-2	**1079.36**	**54.07**	**162.21**	**0.49**	**0.57**	**0.53**	**0.38**	**0.49**	

Figure 2 illustrates the improvement of the pairwise inter-assessor agreement level with training over time. There were setbacks in between because each extract is a unique text, and some extracts (as noted with 7-1 and 7-2) are subject to disagreement. Furthermore, the lines vary in harmony.

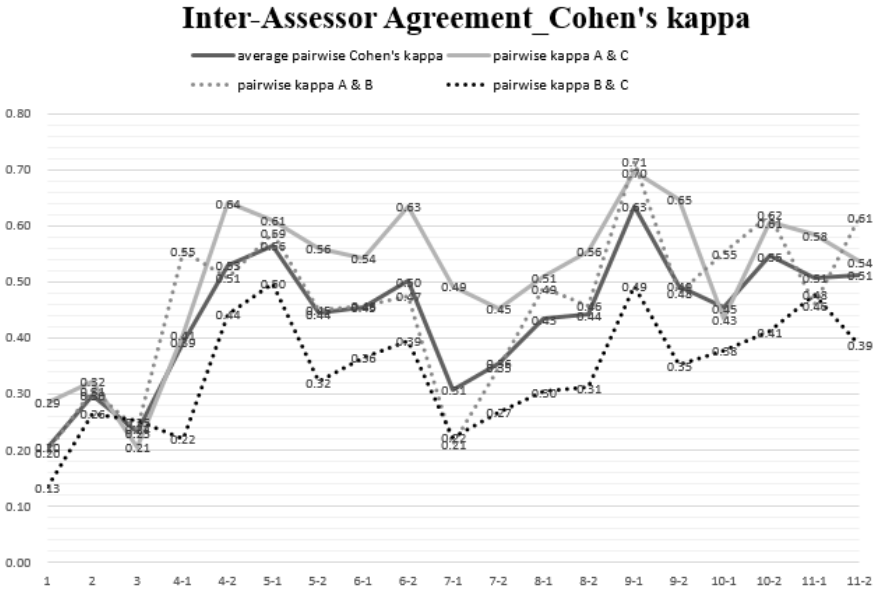

Figure 2. Inter-assessor agreement (Cohen's kappa)

As shown in Table 3 and Figure 2, 35 pairwise kappa values out of 57 (61%) in total are above 0.41, of which 20 (35%) values are above 0.5. The mean value of alpha also stabilized at 0.41. Therefore, it is safe for us to say that we have achieved an agreement level of 0.41 to 0.60 which is considered as moderate agreement. This compares favorably with Lommel et al. (2015), who used the MQM scheme to measure inter-assessor agreement in identifying and classifying erroneous text spans in MT outputs. They obtained only low kappa values ranging from 0.18 to 0.36, which they attributed it to the lack of decision-making tool. On this basis, we decided to terminate the iteration. It is worth mentioning that achieving high agreement level is difficult in general and even more so when it

comes to a hybrid text which has both technical and marketing functions. This is because the STs are more expressive and drafted using more complicated sentence structures, which increases the complexity and difficulty of identifying and labelling errors.

A similar iterative procedure was taken by Fujita et al. (2017) who achieved high Cohen's kappa values, up to 0.831. However, compared to my research, their development set is much smaller – only three ST/TT pairs with an average of 781 words for each ST, whereas the CSR scheme was developed on the basis of 24,001 JA characters. As in the development of the CSR scheme, Fujita et al. (2017) identified issues in advance, therefore assessors only had to classify errors. However, in contrast to Fujita et al. (2017), the assessors in the present study were also allowed to mark a given span as 'no error' and to mark additional text spans which they considered problematic. Thus, in the case of developing the CSR scheme, the tasks assigned to the assessors are more complex. For the same reason, it is understandable that Fujita et al. (2017) achieved better kappa values because they are agreement levels for classification of already identified issues. While two annotators consistently classified 289 issues out of 340, achieving a notably high agreement (kappa=0.794), this was just a single pair of assessors. Therefore, it is not clear if the agreement would be repeated across a higher number of annotators. Mariana et al. (2015) also point out that inconsistency of annotation even among professional translators is a common issue.

In this section, I described the process of validating the CSR scheme, compared the original and final CSR schemes and explained the reasoning behind the modifications. I further showed the details with regard to the improvement of inter-assessor agreement. The type and frequency of each category revealed in the results of the experiment are explained in Section 5.

5. Results

5.1. Error type observed

In the experiment 1,087 agreed errors were identified in the official ZH-CN translations of 11 JA CSR report extracts with a total word count of 24,001 characters. That is 45.3 errors per 1,000 characters, which falls far below the industry standard. The permitted error count per 1,000 words is usually pre-defined between clients and LSPs. In the author`s experience, in a professional setting 1% (error count per source word count) is deemed moderate. Referring to the Canadian Language Quality Measurement System (Sical), the translation quality assessment model developed by the Canadian government's Translation Bureau, Williams (2001: 330) points out "[i]n theory, then, a fully acceptable translation of 400 words could contain as many as 12 errors of transfer, provided no major error was detected."[9)]

Of 1,087 agreed errors 530 were labelled as accuracy errors, 517 as fluency errors, 39 as cohesion errors and 1 as terminology error, shown in Table 4.

Table 4. Type and frequency of agreed TT errors

Type	Frequency
TT-Accuracy-Misrepresentation	374
TT-Fluency-Grammar	195
TT-Fluency-Awkward/Unidiomatic	187
TT-Fluency-Register	85
TT-Accuracy-Calque	46
TT-Accuracy-False friend	43
TT-Cohesion	39
TT-Fluency-Punctuation	35
TT-Accuracy-Addition	18
TT-Accuracy-Omission	18
TT-Accuracy-Transliteration	17
TT-Fluency-Spelling/Character	12
TT-Accuracy-Ambiguity	8
TT-Accuracy-Untranslated	6
TT-Fluency-Inconsistency	3
TT-Terminology	1

9) A general ratio applied in the translation industry counts 400 EN words as 1000 JA characters.

The significant number of accuracy errors points to a low level of translation competency in terms of comprehension of the source language and of writing skills in the native language. [TT-Fluency-Inconsistency] is rare because the sample is small, therefore it is unlikely to contain repeated ST sentences. However, repetition of same ST sentences is common in published CSR reports, which average 30,000-60,000 characters. Moreover, translators are usually provided with a style guide together with the translation request. But no such reference materials were available for reference during the experiment to identify and label errors. Likewise, there is only one [TT-Terminology] error because no glossary file was available for reference during the experiment. Therefore, the question whether the translation of a given technical term or proper noun complies with the specified terminology does not apply.

Table 5 demonstrates that the errors were evenly distributed across all extracts rather than being concentrated in particular extracts.

Table 5. Error distribution in each extract

Extract	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11
Character	2070	2150	2267	2160	2064	2239	2167	2321	1883	2248	2196
Number of errors	83	103	112	99	105	86	80	101	97	149	112
Errors per 1,000 chars	40.1	47.9	49.4	45.8	50.9	38.4	36.9	43.5	51.5	66.3	51.0

The range 36.9-66.3 errors per 1,000 characters falls far below the industry quality standard. Especially given that the source documents have a marketing function, this issue is very likely to leave the target readers with a negative impression.

Section 5.2 introduces revision categories motivated by problematic ST features labelled in the iterative annotation process.

5.2. Problematic ST features observed

As previously stated, the CSR scheme takes into account deficiencies in the ST

because problematic source features can inhibit translation quality by inducing translation errors. We observed errors that could not be clearly categorized because the ST exhibits a serious deficiency. The ST categories shown in Table 6 were validated as part of the process described in Section 4.

Table 6. Problematic ST features

Category	Frequency
ST-Incomprehensibility	9
ST-Cohesion	8
ST-Illogicality	8
ST-Ambiguity	6
ST-Incompleteness	4
ST-Awkward/Unidiomatic	4
ST-Other errors	1

The example below is an instance marked as ST-Incomprehensibility. The JA ST with the company name anonymized as 'XXX' is followed by an EN gloss, the published ZH-CN translation, its EN gloss, and the published EN translation.

JA	今後、FCVの普及によって、ドライバーが**クリーンなエネルギーを自由な場所に移動させ**、「つかう」時代がやってくるとXXXは考えています。
Gloss of JA	XXX believes that, in future, in pace with popularization of FCV, the time that drivers move green energy to free place and 'use' will come.
ZH-CN	XXX所追求的是随着今后FCV的普及，驾驶员可以**自由的使用和移动清洁能源**的时代到来了。
Gloss of ZH-CN	What XXX is pursuing is that in pace with popularization of FCV, the time that drivers can use and move green energy has come.
EN	XXX believes that the time will come with the spread of FCVs in the future when individual drivers **freely carry around clean energy and "use" it at any given time and any given place**.

There is more than one issue in the ZH-CN translation, but the most problematic part is the translation of ‘クリーンなエネルギーを自由な場所に移動させ、「つかう」’. The EN gloss of the JA text span in question is ‘move green energy to free places’, and that of the corresponding ZH-CN translation is ‘use and move around green energy freely’. The assessors cannot decide if it is a translation error because ‘自由な場所に移動’ **(move to free places)** is incomprehensible. In the published EN translation, the translator has translated this span as “individual drivers freely carry around clean energy and ‘use’ it at any given time and any given place” which presents more information that could be obtained from the JA ST, as the gloss indicates. Spans in the STs that have such features are marked as [ST-Incomprehensibility].

On the other hand, the iterative annotation process has revealed that, although understandable and natural to native speakers, some ST features consistently induce translation errors. Examples are given below.

- Adjectives such as ‘適切な’ and ‘確実な’
- Verbs such as ‘支援’ and ‘取り組む’
- Nouns such as ‘地域’ and ‘活動’
- …ため、…ように
- Noun1やNoun2、Noun3など
- Centered dot
- Noun1やNoun2をVerb1・Verb2する
- ‘AについてのB’ is shortened as ‘AのB’

Section 5 discussed the results observed from the experiment conducted to validate the CSR scheme and introduced a list of problematic source features that inhibit translation quality. The final section summarizes the findings and achievements of this research and indicates possible future research.

6. Conclusions

Through an iterative error annotation process, we have developed an error typology that accounts for not only TT errors but also problematic features of STs. Its consistent applicability has been statistically validated, achieving a moderate inter-assessor agreement level. The CSR scheme accommodates texts with both technical and marketing functions, i.e. it is optimized for hybrid texts. It is also customized to JA to ZH-CN translation. Following the procedure adopted by Fujita et al. (2017), the text spans that require revision were identified before the assessors performed error categorization. Achieving a satisfactory level of agreement on identifying erroneous spans is beyond the scope of this research.

The number of accuracy and fluency errors in JA to ZH-CN translations of CSR reports identified in the experiment provides evidence consistent with a common industry observation that ZH-CN translators in general lack competency (TAUS 2017a). Section 5.1, provides answers to research question RQ1 concerning the type and frequency of the errors found in our corpus of hybrid texts.

The answer to question RQ2 – which asks if the errors can be accounted for by an existing scheme – is negative. Neither the ATA nor the DQF-MQM scheme alone or in combination can accommodate the errors observed in ZH-CN translations of JA CSR reports. However, the CSR scheme, created by combining elements from the ATA and DQF-MQM and adding categories tailored for JA-ZH-CN translation, has demonstrated its applicability through the validation process adopted in the experiment.

In answer to research question RQ3 – concerning the feasibility of achieving a satisfactory level of inter-assessor agreement – Section 4.2 demonstrates that the final CSR scheme achieved an agreement level in the range 0.41–0.60, which is deemed only moderate according to TAUS (2017b: 10). We argue that this is nonetheless a positive answer in the domain of translation evaluation, despite the fact that for most inter-assessor agreement tasks (in other domains) agreement of at least 0.85 is required for a measure to be considered reliable. In translation

evaluation, in contrast, “it is well known that human judgments of translation show a high degree of variance: in WMT [Workshop on Statistical Machine Translation] testing, the inter-annotator agreement [···] did not exceed 0.40 (kappa) and intra-annotator agreement (i.e., the agreement of raters with themselves when faced with the same assessment task multiple times) did not exceed 0.65 (Bojar et al. 2013)” (Lommel et al. 2015). The results achieved in the present research outperform the WMT results by a considerable margin.

In the process of error identification and labelling, it became clear that when certain features appear in the ST, the possibility of translators making errors in translation increases. Therefore, although the number of translation errors is significant, we should not attribute all errors to translators’ lack of competency without first analyzing the STs for features that correlate with recurrent translation errors. As pointed out by Nyberg et al. (2003: 245), “[b]oth humans and computers may experience difficulties in understanding and translating natural language, due to its inherent ambiguity and complexity.” When discussing writing and developing content for translation, Esselink (2000: 27) states that “[t]he most important thing in a written text is that the text must be understood by non-native readers. [···] the text must be written with translation in mind, so the translator can work quickly and accurately, without the need for clarifications, rewrites, or cultural modifications.”

Hence, a necessary next step is to focus on finding ways of identifying translation errors that relate consistently to problematic features of the STs. Formulating and validating rewrite rules to eliminate such features to improve translation quality is foreseen as the biggest challenge. Development of error typologies that apply to other Asian language pairs such as Korean-Chinese in the genre of hybrid texts may be timely. The methodology applied in this study may offer a possible solution to other studies focused on other language pairs.

References

Babych, B., Hartley, A., Kageura, K., Thomas, M., and Utiyama, M. (2012). MNH-TT: a collaborative platform for translator training. *Translating and the Computer, 34,*

Bojar, O. and Macháček, M. (2013). Results of the WMT13 metrics shared task. *The Eighth Workshop on Statistical Machine Translation, 45-51.*

Bühler, K. (1934/1965). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Stuttgart: Gustav Fischer.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N., and Volanschi, A. (2006). Designing a learner translator corpus for training purpose. Paper presented at the 7th International Conference on Teaching and Language Corpora.

Chesterman, A. (ed.) (1989). *Readings in translation theory*. Helsinki: Finn Lectura.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37-46.

Esselink, B. (2000). *A practical guide to localization*. Amsterdam: John Benjamins.

Fujita, A., Tanabe, K., Toyoshima, C., Yamamoto, M., Kageura, K., and Hartley, A. (2017). Consistent classification of translation revisions: a case study of English-Japanese Student Translations. *the 11th Linguistic Annotation Workshop (LAW*), 57-66.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. *The Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL) Short Papers,* 57-60.

Koby, G. and Champe, G. (2013). Welcome to the real world: professional-level translator certification. *The International Journal of Translation and Interpreting Research* (5)1, 156-173.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-reliability. Retrieved from https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers on 20 April 2018.

Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., and Popović, M. (2015). QT21 deliverable 3.1: harmonised metric. Retrieved from http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf on 26 April 2018.

Mariana, V., Cox, T., and Melby, A. (2015). The multidimensional quality metrics (MQM) framework: a new framework for translation quality assessment. *The Journal of Specialised Translation* 23, 137-161.

Mossop, B. (2014). *Revising and Editing for Translators* (3rd edn.). London/New York: Routledge.

Munday, J. (2016). *Introducing Translation Studies: Theories and Applications* (4th edn.). New York: Routledge.

Nyberg, E., Mitamura, T., and Huijsen, W.-O. (2003). Controlled language for authoring and translation. In Sommers, H. (ed.), *Computers and Translation: A Translator's Guide*. Philadelphia: John Benjamins.

Reiss, K. (1977/1989). Text types, translation types and translation assessment (A. Chesterman, Trans). In Chesterman, A. (ed.) (1989), 105-115.

Secară, A. (2015). Translation evaluation: a state of the art survey. In Proceedings of the

eCoLoRe/MeLLANGE Workshop, 9-44.
TAUS (2017a). The translation industry in 2012. A report from the TAUS Industry Summit, Amsterdam. Retrieved from https://www.taus.net/think-tank/reports/event-reports/the-translation-industry-in-2022 on 25 April 2018.
TAUS (2017b). Readability evaluation guidelines. Retrieved from http://info.taus.net/download-readability-evaluation-guidelines on 25 April 2018.
Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta* 46(2), 327-344.

This paper was received on 30 April 2018; revised on 11 May 2018; and accepted on 30 May 2018.

Author's email address
beibei.h.s@rikkyo.ac.jp

About the author
Beibei He is a PhD student pursuing her degree in Translation Studies at Rikkyo University. After working as a professional Japanese to Chinese translator/reviser and QA (quality assurance) checker for past ten years, she currently works as a language process consultant with a global LSP in Japan.