Consistent Classification of Translation Revisions: A Case Study of English-Japanese Student Translations

Atsushi Fujita
NICT
atsushi.fujita@nict.go.jp

Kikuko Tanabe
Kobe College
kikukotanabe@gmail.com

Chiho Toyoshima
Kansai Gaidai University
c.toyoshima1113@gmail.com

Mayuka Yamamoto
Honyaku Center Inc.
yamamoto.mayuka@honyakuctr.co.jp

Kyo Kageura
University of Tokyo
kyo@p.u-tokyo.ac.jp

Anthony Hartley
Rikkyo University
A.Hartley@rikkyo.ac.jp

Abstract

Consistency is a crucial requirement in text annotation. It is especially important in educational applications, as lack of consistency directly affects learners' motivation and learning performance. This paper presents a quality assessment scheme for English-to-Japanese translations produced by learner translators at university. We constructed a revision typology and a decision tree manually through an application of the OntoNotes method, i.e., an iteration of assessing learners' translations and hypothesizing the conditions for consistent decision making, as well as reorganizing the typology. Intrinsic evaluation of the created scheme confirmed its potential contribution to the consistent classification of identified erroneous text spans, achieving visibly higher Cohen's κ values, up to 0.831, than previous work. This paper also describes an application of our scheme to an English-to-Japanese translation exercise course for undergraduate students at a university in Japan.

1 Introduction

Assessing and assuring translation quality is one of the main concerns for translation services, machine translation (MT) industries, and translation teaching institutions.[1] The assessment process for a given pair of source document (SD) and its translation, i.e., target document (TD), consists of two tasks. The first task is to identify erroneous text spans in the TD. In professional settings, when assessors consider a text span in a TD as erroneous, they generally suggest a particular revision proposal (Mossop, 2014). For instance, in example (1), a transliteration error is corrected.

(1) SD: Mark Potok is a senior fellow at the Southern Poverty Law Center.
TD: マーク・ポッドック (⇒ ポトク) 氏は南部貧困法律センターの上級研究員だ。
(Podok ⇒ Potok)

Henceforth, we refer to a marked text span reflecting the identification of a particular error or deficiency as an ***issue***. The second task is to classify each identified issue into an abstract ***issue type***, such as "omission" or "misspelling."

An inherent problem concerning translation quality assessment is that it inevitably involves human judgments, and thus is subjective.[2] The first task, i.e., identifying issues in TDs, relies heavily on assessors' translation and linguistic competence, as may the subsequent step of making a revision proposal for them, depending on the subtlety of the issue. It therefore seems impractical to create an annotation scheme that enables even inexperienced translators to perform this task at a comparable level to mature translators.

For regulating the second task, several typologies, such as those reviewed in Secară (2005), the Multilingual e-Learning in Language Engineering (MeLLANGE) error typology (Castagnoli et al., 2006), and Multidimensional Quality Metrics (MQM),[3] have been proposed. Existing issue typologies show diversity in their granularity and their organization of issue types, owing to the fact that the scope and granularity of issues depend

[1] These include both private companies and translation-related departments in colleges and universities.

[2] While automated metrics for MT quality evaluation are often presented as objective, many, including BLEU (Papineni et al., 2002), rely on comparison with a one or more human reference translations whose quality and subjectivity are merely assumed and not independently validated.

[3] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

on the purpose of translations and the aim of human assessments (e.g., formative or summative). However, the typology alone does not necessarily guarantee consistent human assessments (Lommel et al., 2015). For instance, while one may classify the issue in (1) as an "incorrect translation of term," it could also be regarded as a "misspelling."

In this paper, we focus on the quality assessment of learners' translations. Motivated by the increasing demand for translation, translation teaching institutions have been incorporating best practices of professionals into their curricula. When teaching the revision and review processes in such institutions, the assessor's revision proposal is normally not provided, in order to prevent learners believing that it is the only correct solution (Klaudy, 1996). Thus, issue type plays a crucial role in conveying the assessors' intention to learners, and its consistency is especially important, since lack of consistency directly affects learners' motivation and learning performance. Besides the consistency, the applicability of an assessment tool to a wide range of translations is also important. To the best of our knowledge, however, none of the existing typologies have been validated for translations between languages whose structures are radically different, such as English and Japanese. Neither have their applicability to translations produced by less advanced learners, such as undergraduate students, been fully examined.

Aiming at (i) a consistent human assessment, (ii) of English-to-Japanese translations, (iii) produced by learner translators, we manually constructed a scheme for classifying identified issues. We first collected English-to-Japanese translations from learners in order to assure and validate the applicability of our scheme (§3). We then manually created an issue typology and a decision tree through an application of the OntoNotes method (Hovy et al., 2006), i.e., an iteration of assessing learners' translations and updating the typology and decision tree (§4). We adopted an existing typology, that of MNH-TT (Babych et al., 2012), as the starting point, because its origin (Castagnoli et al., 2006) was tailored to assessing university student learners' translations and its applicability across several European languages had been demonstrated. We evaluated our scheme with inter-assessor agreement, employing four assessors and an undergraduate learner translator (§5). We also implemented our scheme in an English-to-Japanese translation exercise course for undergraduate students at a university in Japan, and observed tendencies among absolute novices (§6).

2 Previous Work

To the best of our knowledge, the error typology in the Multilingual e-Learning in Language Engineering (MeLLANGE) project (Castagnoli et al., 2006) was the first tool tailored to assessing learners' translations. It had been proved applicable to learners' translations across several European languages, including English, German, Spanish, French, and Italian. The MeLLANGE typology distinguished more than 30 types of issues, grouped into Transfer (TR) issues, whose diagnosis requires reference to both SD and TD, and Language (LA) issues, which relate to violations of target language norms. This distinction underlies the widespread distinction between adequacy and fluency, the principal editing and revision strategies advocated by Mossop (2014), and the differentiation between (bilingual) revision and (monolingual) reviewing specified in ISO/TC27 (2015). Designed for offering formative assessment by experienced instructors to university learner translators, it provided a fine-grained discrimination seen also in, for instance, the framework of the American Translators Association (ATA) with 23 categories.[4]

The MeLLANGE typology was simplified by Babych et al. (2012), who conflated various subcategories and reduced the number of issue types to 16 for their translation training environment, MNH-TT, which differs from MeLLANGE in two respects. First, it is designed for feedback from peer learners acting as revisers and/or reviewers, whose ability to make subtle distinctions is reduced. Second, it is embedded in a project-oriented translation scenario that simulates professional practice and where more coarse-grained, summative schemes prevail.[5,6] In our pilot test, however, we found that even the MNH-TT typology did not necessarily guarantee consistent human assessments. When we identified 40 issues

[4] http://www.atanet.org/certification/aboutexams_error.php

[5] SAE J2450, the standard for the automotive industry, has only seven categories. http://standards.sae.org/j2450_200508/

[6] The latest MQM (as of February 21, 2017) has eight top-level issue types (dimensions) and more than 100 leaf nodes. http://www.qt21.eu/mqm-definition/

Level 1: **Incompleteness**	Translation is not finished.
Level 2: **Semantic errors**	The contents of the SD are not properly transferred.
Level 3: **TD linguistic issues**	The contents of the SD are transferred, but there are some linguistic issues in the TD.
Level 4: **TD felicity issues**	The TD is meaning-preserving and has no linguistic issues, but have some flaws.
Level 5: **TD register issues**	The TD is a good translation, but not suitable for the assumed text type.

Table 1: Priority of coarse-grained issue types for translation training for novices.

in an English-to-Japanese translation by a learner and two of the authors separately classified them, only 17 of them (43%) resulted in agreement on the classification, achieving Cohen's κ (Cohen, 1960) of 0.36. This highlighted the necessity of a navigation tool, such as a decision tree, for consistent human decision making, especially given that the issue type serves as feedback to learners.

Multidimensional Quality Metrics (MQM) has been widely used in the MT community and translation industries. However, the consistency of classifying issues had not been guaranteed when only its issue typology was used. Lommel et al. (2014) measured the inter-assessor agreement of identifying erroneous text spans in MT outputs and classifying them, using the MQM typology comprising 20 types. Having obtained low Cohen's κ values, 0.18 to 0.36, and observed several types of ambiguities, they pointed out the lack of decision making tool. Motivated by this study, MQM later established a decision tree (Burchardt and Lommel, 2014). Nevertheless, MQM has not been validated as applicable to learners' translations, especially those between distant languages.

3 Collecting English-to-Japanese Translations by Learners

Our study began with collecting English-to-Japanese translations produced by learner translators. Assuming novice learner translators and the very basic competences to teach, we selected journalistic articles as the text type. As the SDs in English, 18 articles with similar conceptual and linguistic difficulties were sampled from the column page of a news program "Democracy Now!"[7] by a professional translator, who also had significant experience in teaching English-to-Japanese translation at universities. The average number of words in the SDs was 781. Then, 30 students (12 undergraduate and 18 graduate students) were employed to translate one of the SDs into Japanese.[8] All these participants were native Japanese speakers and had attended translation classes at a university in Japan. They were asked to produce a TD that served as a Japanese version of the original article.

The collected pairs of SD and TD were divided into the following three partitions.

Development: Three randomly selected document pairs were used to develop our scheme (§4).

Validation 1: Another 17 sampled document pairs were used to gauge the inter-assessor agreement of the issue classification task, given the identified issues (§5.1).

Validation 2: The remaining ten document pairs were used to examine the stability of identifying erroneous text spans in the TDs, as well as the inter-assessor agreement between a learner and an experienced assessor (§5.2).

4 Development of an Issue Classification Scheme

To alleviate potential inconsistencies, we structuralized the issue types in the MNH-TT typology (Babych et al., 2012), introducing a decision tree. We chose a decision tree as a navigation tool for human decision making, as in MQM (Burchardt and Lommel, 2014), because the resulting issues will be used not only by the instructors in order to evaluate the translation quality but also by learners in order to understand the diagnoses. We also considered that explicit explanation for decisions is critical in such scenarios.

We first determined the priorities of issue types through in-depth interviews with two professional translators, who also had ample experience in teaching English-to-Japanese translation at universities. These priorities were based on both the work-flow of the professionals and the nature of the issues they found in grading learners' translations. Table 1 shows the coarse-grained figure resulting from the two translators' agreement. Obvious incompleteness of translations are captured

[7] http://www.democracynow.org/blog/category/weekly_column/

[8] Some of the SDs were separately translated by more than one student.

Issue types in our typology		MeLLANGE	MQM
Level 1: **Incompleteness**			
X4a	Content-SD-intrusion-untranslated: The TD contains elements of the SD left untranslated in error	TR-SI-UT	Untranslated
X6	Content-indecision: The TD contains alternative choices left unresolved by the translator	TR-IN	n/a
Level 2: **Semantic errors**			
X7	Lexis-incorrect-term: Item is a non-term, incorrect, inconsistent with the glossary or inconsistent within the TD	LA-TL-{IN,NT,IG,IT}	Terminology
X1	Content-omission: Content present in the SD is wrongly omitted in the TD	TR-OM	Omission
X2	Content-addition: Content not present in the SD is wrongly added to the TD	TR-AD	Addition
X3	Content-distortion: Content present in the SD is misrepresented in the TD	TR-DI, TR-TI-∗, TR-TL-IN	Mistranslation
Level 3: **TD linguistic issues**			
X8	Lexis-inappropriate-collocation: Item is not a usual collocate of a neighbor it governs or is governed by	LA-TL-IC	n/a
X10	Grammar-preposition/particle: Incorrect preposition or (Japanese) particle	LA-PR	Grammar
X11	Grammar-inflection: Incorrect inflection or agreement for tense, aspect, number, case, or gender	LA-IA-∗	Grammar
X12	Grammar-spelling: Incorrect spelling	LA-HY-SP	Spelling
X13	Grammar-punctuation: Incorrect punctuation	LA-HY-PU	Punctuation
X9	Grammar-others: Other grammatical and syntactic issues in the TD	n/a	Grammar
Level 4: **TD felicity issues**			
X16	Text-incohesive: Inappropriate use or non-use of anaphoric expressions, or wrong ordering of given and new elements of information	n/a	n/a
X4b	Content-SD-intrution-too-literal: The TD contains elements of the SD that are translated too literally	TR-SI-TL	Overly literal
X15	Text-clumsy: Lexical choice or phrasing is clumsy, tautologous, or unnecessarily verbose	LA-ST-∗	Awkward
Level 5: **TD register issues**			
X14	Text-TD-inappropriate-register: Lexical choice, phrasing, or style is inappropriate for the intended text type of the TD	LA-RE-∗	Style (except "Awkward"), Local convention, Grammatical register

Table 2: Our issue typology, with prefixes (context, lexis, grammar, and text) indicating their coarse-grained classification in the MNH-TT typology (Babych et al., 2012). The two rightmost columns show the corresponding issue types in the MeLLANGE typology (Castagnoli et al., 2006) and those in MQM (Lommel et al., 2015), respectively, where "n/a" indicates issue types that are not covered explicitly.

at Level 1. While Level 2 covers issues related to misunderstandings of the SD, Levels 3 and 4 highlight issues in the language of the TD. Level 5 deals with violation of various requirements imposed by the text type of the translated document.

Regarding Table 1 as a strict constraint for the shape of the decision tree, and the 16 issue types in the MNH-TT typology as the initial issue types, we developed our issue classification scheme, using the OntoNotes method (Hovy et al., 2006). In other words, we performed the following iteration(s).

Step 1. Annotate issues in the TDs for development, using the latest scheme.

Step 2. Terminate the iteration if we meet a satisfactory agreement ratio (90%, as in Hovy et al. (2006)).

Step 3. Collect disagreed issues among assessors, including those newly found, and discuss the factors of consistent decision making.

Step 4. Update the scheme, including the definition of each issue type, the conditions for decision making, and their organization in the form of a decision tree. Record marginal examples in the example list.

Step 5. Go back to Step 1.

Three of the authors conducted the above process using the three document pairs (see §3), which resulted in the issue typology in Table 2 and the decision tree in Table 3. A total of 52 typical and marginal examples were also collected.

It is noteworthy that our issue typology preserves almost perfectly the top-level distinction of the MeLLANGE typology, i.e., the TR (trans-

ID	Question	Determined type or next question	
		True	False
Q1a	Is it an unjustified copy of the SD element?	X4a	Q1b
Q1b	Do multiple options remain in the TD?	X6	Q2a
Q2a	Is all content in the SD translated in proper quantities in a proper way?	Q3a	Q2b
Q2b	Is the error related to a term in the given glossary?	X7	X1/X2/X3
Q3a	Is it a grammatical issue?	Q3b	Q4a
Q3b	Is it predefined specific type?	X8/X10/X11 /X12/X13	X9
Q4a	Does it hurt cohesiveness of the TD?	X16	Q4b
Q4b	Does it hurt fluency?	Q4c	Q5a
Q4c	Is it too literal?	X4b	X15
Q5a	Is it unsuitable for the intended text type of the TD?	X14	Q6a
Q6a	Is it anyways problematic?	"Other issue"	"Not an issue"

Table 3: Our decision tree for classifying a given issue: we do not produce questions for distinguishing X1/X2/X3, and X8/X10/X11/X12/X13, considering that their definitions are clear enough.

fer) and LA (language) issues, described in §2. The priority of the former over the latter, implicitly assumed in the MeLLANGE typology, is also largely preserved; the sole exceptions are X7 (incorrect translation of term) and X4b (too literal). Table 2 also shows that our typology includes the following three issue types that are not covered by MQM.

- X6 (indecision) captures a student habit of offering more than one translation for a given text, which is not observed in professional translators.
- X8 (collocation) employs a more specific, linguistic terminology for diagnosing one subtype of X15 (clumsy).
- X16 (incohesive), which is also absent from the MeLLANGE typology but present in the ATA framework, appears not applicable in the common (commercial) situation where sentences are translated without reference to their context.

During the development process, we decided to identify and classify only the first occurrence of ***identical issues*** in a single TD. For instance, other incorrect translations of "ポッドック" for "Potok" in the same TD as example (1) will not be annotated repeatedly. This is because annotations are made and used by humans, i.e., assessors and learners, and persistent indications of identical issues may waste the time of assessors and discourage learners. This practice differs from ordinary linguistic annotation, especially that aiming to develop training data for machine learning methods, which requires exhaustive annotation of the phenomena of interest within given documents. Although there have been several studies on the use of partial/incomplete annotation, e.g., Tsuboi et al. (2008), our procedure is nevertheless different from these in the sense that we leave issues "unannotated" only when identical ones are already annotated.

5 Intrinsic Evaluation of the Scheme

It is hard to make a fair and unbiased comparison between different annotation schemes that target the same phenomena, employing the same assessors. We thus evaluated whether our issue classification scheme leads to sufficiently high level of inter-assessor agreement, regarding those poor results described in §2 as baselines, and analyzed the tendencies of disagreements and the distribution of issues.

5.1 Validation 1: Classification of Identified Issues

5.1.1 Inter-Assessor Agreement

We gauged the consistency of classifying identified issues by the inter-assessor agreement.

First, three of the authors who developed our scheme identified erroneous text spans in the 17 TDs (see §3) and made a revision proposal for each, through discussion. Then, four assessors were independently asked to classify each of the resulting 575 issues into one of the 16 issue types, "other issue," and "not an issue," following our decision tree in Table 3. Two of them were anonymous paid workers (A and B), while the others (C and D) were two of the above three authors. All four assessors were native Japanese speakers with a strong command of English and an understanding of our scheme and translation-related notions. While they were asked to adhere to our decision

ID	Background	Agreement ratio [%]				Cohen's κ			
		vs A	vs B	vs C	vs D	vs A	vs B	vs C	vs D
A	Bachelor of Engineering (now translation editor)	-	67.7	63.3	57.9	-	0.613	0.554	0.490
B	Master of Japanese Pedagogy (now translator)	67.7	-	67.1	61.4	0.613	-	0.592	0.523
C	Master of Translation Studies	63.3	67.1	-	86.6	0.554	0.592	-	0.831
D	Ph.D in Computational Linguistics	57.9	61.4	86.6	-	0.490	0.523	0.831	-

Table 4: Inter-assessor agreement on the 575 identified issues.

tree, no dictionary or glossary was provided.

Table 4 summarizes the agreement ratio and Cohen's κ (Cohen, 1960). The most consistent pair was C and D who agreed on 86.6% (498/575) of the issues and achieved almost perfect agreement, $\kappa = 0.831$, although it is indisputable that they had some advantages, having been engaged in developing the scheme and identifying the issues. Both of the two measures draw a clear distinction between the anonymous and identified assessors. As our analysis below illustrates, the anonymous workers made many careless mistakes, presumably because the human resource agency did not offer substantial incentive to pursue accurate and consistent annotations. Nevertheless, even the lowest κ value in our experiment, 0.490, was visibly higher than those achieved using the typologies with the same level of granularity but without a tool for consistent decision making (see §2).

Table 5 shows the most frequent disagreement patterns between each anonymous worker and the two authors (C and D) on the 498 issues about which the authors have agreed. The most typical disagreement was between X3 (distortion) and X4b (too literal). For instance, "has passed" in example (2) was mistakenly translated into "通過した ([bill] passed [legislature])," resulting in two exclusive subjects marked with nominative case marker "が," i.e., "各州政府 (state after state)" and "農業口封じ法 (Ag-Gag laws)."

(2) SD: State after state has passed so-called Ag-Gag laws.
TD: 各州政府がいわゆる農業口封じ法が通過した (⇒ を可決した)。
([bill] passed [legislature] ⇒ [legislature] passed [bill])

As the TD does not convey the original meaning in the SD, both C and D classified this issue into X3 (distortion). In contrast, both A and B regarded them as X4b (too literal), presumably considering that both of the original translation "通過した" and the revision proposal "可決した" were appropriate lexical translations for "has passed" when

A, B	C&D	A	B
X4b (Level 4)	X3 (Level 2)	37	8
X3 (Level 2)	X4b (Level 4)	11	24
X1 (Level 2)	X3 (Level 2)	13	10
X1 (Level 2)	X4b (Level 4)	6	5
X1 (Level 2)	X16 (Level 4)	6	4
X1 (Level 2)	X7 (Level 2)	5	4

Table 5: Frequent disagreements between anonymous workers (A and B) and two of the authors (C and D) among the 498 identified issues that C and D classified consistently.

separated from the context. The above results, and the fact that X3 (distortion) and X4b (too literal) also produced the most frequent disagreements between C and D (11 out of 77 disagreements), suggested that question Q2a in Table 3 should be defined more clearly. We plan to make this precise in our future work.

The other frequent disagreements concerned the issues classified as X1 (omission) by A and B, whereas C and D classified them as other types. For instance, both C and D classified the issue in (3) as X3 (distortion) since the original word "sailors" was incorrectly translated as "soldiers," and the issue in (4) as X7 (incorrect translation of term) since named entities compose a typical subclass of term.

(3) SD: We have filed a class action for approximately a hundred sailors.
TD: およそ 100 人の兵士 (⇒ 海兵兵士) のための集団訴訟を起こした。(soldiers ⇒ sailors)

(4) SD: President Ronald Reagan vetoed the bill, but, . . .
TD: レーガン大統領 (⇒ ロナルド・レーガン大統領) はその法案を拒否したが、. . .
(President Reagan ⇒ President Ronald Reagan)

These disagreements imply that the anonymous workers might not have strictly adhered to our decision tree, and classified them as X1 after merely

Issue type		n	undergrad. (6) avg.	s.d.	grad. (11) avg.	s.d.
Level 1	X4a	3	0.04	0.11	0.01	0.04
	X6	0	0.00	0.00	0.00	0.00
Level 2	X7	33	0.39	0.18	0.26	0.31
	X1	53	0.73	0.54	0.34	0.29
	X2	28	0.33	0.26	0.26	0.32
	X3	240	2.67	1.26	2.24	1.41
Level 3	X8	16	0.19	0.23	0.16	0.24
	X10	22	0.27	0.24	0.21	0.32
	X11	10	0.13	0.11	0.07	0.11
	X12	8	0.11	0.15	0.07	0.11
	X13	18	0.21	0.15	0.17	0.20
	X9	10	0.08	0.10	0.12	0.12
Level 4	X16	18	0.20	0.23	0.14	0.14
	X4b	92	0.87	0.69	1.05	0.93
	X15	14	0.09	0.07	0.21	0.25
Level 5	X14	28	0.34	0.14	0.25	0.32
Total		593	6.63	2.35	5.58	3.35

Table 6: Total frequency and relative frequency of each issue type (macro average and standard deviation over TDs).

comparing the marked text span with the revision proposal at the surface level.

5.1.2 Coverage of the Issue Typology

Through a discussion, the disagreements between C and D on the 77 issues were resolved and 18 newly identified issues were also classified. We then calculated relative frequency RF of each issue type, t, in each TD, d, as follows:

$$RF(t,d) = \frac{(\text{frequency of } t \text{ in } d)}{(\#\text{ of words in the SD of } d)/100}.$$

Table 6 summarizes the frequency of each issue type; the "n" column shows their total frequency across all TDs and the remaining columns compares macro average and standard deviation of the relative frequencies over TDs produced by each group of students. All the identified issues were classified into one of the 16 issue types in our typology, confirming that the MNH-TT typology had also covered various types of issues appearing in English-to-Japanese translations produced by learners. As reviewed in §2 and §4, both of our typology and the MNH-TT typology cover a broader range of issues than the MeLLANGE typology. Thus, we can even insist that our scheme is applicable to translations between several European languages that Castagnoli et al. (2006) have investigated. In our preliminary experiments on assessing English-to-Chinese and Japanese-to-Korean translations using our scheme, we have not observed any novel type of issues.

X3 (distortion) occurred significantly more frequently than the others. This is consistent with the previous investigation based on the MeLLANGE typology (Castagnoli et al., 2006), considering that X3 (distortion) in our typology corresponds to parts of the most frequent type, LA-TL-IN (Language, terminology and lexis, incorrect), and the second-ranked TR-DI (Transfer, distortion). The other frequent types were X4b (too literal) and X1 (omission), which are both listed in the two existing typologies in Table 2, and also frequently observed in the learners' translations between European languages (Castagnoli et al., 2006).

The annotation results revealed that the graduate students produced issues at Level 2 less frequently than the undergraduate students, while producing more Level 4 issues. Although the relative frequencies of issues vary greatly between individuals, we speculate that less experienced students are more likely to struggle at Level 2, i.e., properly understanding content in SDs.

5.2 Validation 2: Annotation by a Novice Learner Translator

We also evaluated our issue classification scheme in a more realistic setting: the comparison of an undergraduate learner translator with an experienced assessor.

The learner involved in this experiment, referred to as assessor E, was also a native Japanese speaker and had attended some translation classes at a university. The other assessor was D, who had participated in the first experiment. The two assessors separately identified erroneous text spans in the ten TDs (see §3) with a revision proposal, and classified them following our decision tree.

As a result, D and E respectively annotated 561 and 406 issues. Among these, 340 were for identical text spans, with not necessarily identical but similar revision proposals. They consistently classified 289 issues out of 340 (85.0%), achieving a substantial and notably high agreement, $\kappa = 0.794$. These are substantially higher than those achieved by the anonymous workers A and B (see Table 4), although they worked on different TDs. This fact indicates that the identified assessors in the first experiment (C and D) did not necessarily have an advantage. More importantly, this experiment verified the understandability of our scheme by actual learner translators. We expect that learner translators would be able

D	E	# of issues
X4b (Level 4)	X3 (Level 2)	6
X3 (Level 2)	X15 (Level 4)	6
X4b (Level 4)	X15 (Level 2)	5
X1 (Level 2)	X3 (Level 2)	3

Table 7: Frequent disagreements between a learner translator (E) and one of the authors (D) among the 340 issues they identified consistently.

to perform peer reviewing of their draft translations, once they have acquired a certain level of understanding of our scheme. Consequently, as Kiraly (2000) mentions, they would be able to effectively develop their translation skills through playing various roles in the translation work-flow, including that of assessor.

Typical disagreement patterns are shown in Table 7. Similarly to the first experiment, disagreement between X3 (distortion) and X4b (too literal) was frequently observed. E also classified as X15 (clumsy) 11 issues which D classified as X3 (distortion) or X4b (too literal). To answer question Q4c consistently, the literalness needs to be confirmed, for instance, by using dictionaries.

There were 221 and 66 issues identified only by D or E, respectively; 171 and 41 out of these were ignored by the other, including missed issues and accepted translations, reflecting the different levels of sensitivity of the assessors. The other 38 and 14 mismatches suggested the necessity of a guideline to consistently annotate ***single issues***. For instance, E identified one X3 (distortion) issue in (5), while D annotated two issues there: "情報が豊富な (with rich information)" as X2 (addition) and "お天気アプリ (weather application)" as X3 (distortion).

(5) SD: I put the question to Jeff Masters, co-founder at Weather Underground, an Internet weather information service.
TD: 情報量が豊富なお天気アプリ (⇒ 気象情報を提供するウェブサービス)、ウェザー・アンダーグラウンドの共同設立者であるジェフ・マスターズ氏に質問を投げかけた。(a weather application with rich information ⇒ a Web service which provides weather information)

6 Translation Exercise at a University

Having created and validated an annotation scheme, we should ultimately verify its usefulness in actual practice. We implemented our scheme in an English-to-Japanese translation exercise course for undergraduate students at a university in Japan.

6.1 Course Design

Two different types of English texts were used: travel guides from "Travellerspoint"[9] (henceforth, "Travel") and columns from "Democracy Now!" as in §3 (henceforth, "Column"). For each text type, the instructor of the course sampled three documents with similar conceptual and linguistic difficulties, and excerpted roughly the first 550 words of each document as SDs.

A total of 27 undergraduate students participated in the course held over 15 weeks, from April to July 2015. All of them were native Japanese speakers; eight had attended translation classes at a university, while the other 19 were absolute novices. Each student selected one of the sampled SDs for each text type. Before starting translation, they prepared a glossary and collected background information by themselves, and the instructor added any missing information. Each student first translated a "Travel" SD into Japanese over six weeks, referring to the corresponding glossary and background information, and then a "Column" SD in the same manner.

During the process of translating one SD, students' translations were assessed every two weeks (three times per SD); a teaching assistant identified erroneous text spans with a revision proposal, and classified them following our decision tree; and then the instructor double-checked them. While the identified erroneous text spans and the assigned issue types were fed back to the students, revision proposals were not shown (Klaudy, 1996). When the instructor fed back the assessment results to the students, she also explained our issue typology (Table 2) and decision tree (Table 3), also using the examples collected during the development process.

6.2 Observations

Through the course, 54 TDs were annotated with 1,707 issues, all of which fell into one of the 16 types in our issue typology. Table 8 summarizes the relative frequency of each issue type. X3 (distortion) occurred significantly more frequently than the others in translating both types of SDs, as in the results in Table 6 and previous work (Castagnoli et al., 2006). In other words, transfer-

[9] http://www.travellerspoint.com/

Issue type		Travel		Column	
		avg.	s.d.	avg.	s.d.
Level 1	X4a	0.03	0.07	0.03	0.07
	X6	0.00	0.00	0.00	0.00
Level 2	X7	1.14	0.79	0.53	0.39
	X1	0.47	0.42	0.49	0.31
	X2	0.09	0.13	0.16	0.19
	X3	2.20	0.95	2.91	1.03
Level 3	X8	0.11	0.11	0.13	0.18
	X10	0.19	0.24	0.18	0.27
	X11	0.07	0.09	0.12	0.15
	X12	0.15	0.16	0.11	0.11
	X13	0.09	0.16	0.17	0.21
	X9	0.22	0.34	0.03	0.08
Level 4	X16	0.07	0.11	0.03	0.07
	X4b	0.53	0.53	0.24	0.22
	X15	0.10	0.18	0.08	0.13
Level 5	X14	0.30	0.25	0.45	0.31
Total		5.76	2.17	5.65	1.84

Table 8: Relative frequency of each issue type (macro average and standard deviation over TDs).

ring content of the given SD is the principal issue for learner translators in general.

Table 8 also highlights that the relative frequencies of X7 (incorrect translation of term) and X4b (too literal) are drastically different for the "Travel" and "Column" SDs. A student-wise comparison of the relative frequencies in Figure 1 revealed that the students who made these two types of issues more frequently in translating "Travel" SDs (shown in the right-hand side in the figure) produced these types of issues significantly less frequently during translating "Column" SDs. Due to the difference in text types, we cannot claim that this demonstrates students' growth in learning to translate and that this has been promoted by our scheme. Nevertheless, our scheme is clearly useful for quantifying the characteristics of such students.

7 Conclusion

To consistently assess human translations, especially focusing on English-to-Japanese translations produced by learners, we manually created an improved issue typology accompanied by a decision tree through an application of the OntoNotes method. Two annotation experiments, involving four assessors and an actual learner translator, confirmed the potential contribution of our scheme to making consistent classification of identified issues, achieving Cohen's κ values of between 0.490 (moderate) to 0.831 (almost perfect). We also used our scheme in a translation exercise course at a university in order to assess

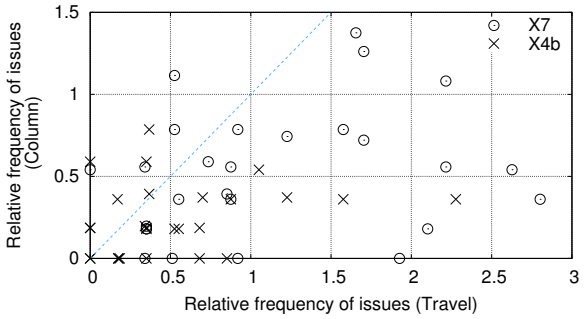

Figure 1: Student-wise comparison of relative frequencies of X7 (incorrect translations of terms) and X4b (too literal).

learners' translations. The predefined 16 issue types in our typology covered all the issues that appeared in English-to-Japanese translations produced by undergraduate students, supporting the applicability of our issue typology to real-world translation training scenarios.

Our plans for future work include further improvements of our issue classification scheme, such as clarifying questions in the decision tree and establishing a guideline for annotating single issues. Its applicability will further be validated using other text types and other language pairs. From the pedagogical point of view, monitoring the effects of assessment is also important (Orozco and Hurtado Albir, 2002). Given the high agreement ratio in our second experiment (§5.2), we are also interested in the feasibility of peer reviewing (Kiraly, 2000). Last but not least, with a view to efficient assessment with less human labor, we will also study automatic identification and classification of erroneous text spans, referring to recent advances in the field of word- and phrase-level quality estimation for MT outputs.[10]

Acknowledgments

We are deeply grateful to anonymous reviewers for their valuable comments. This work was partly supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (A) 25240051.

References

Bogdan Babych, Anthony Hartley, Kyo Kageura, Martin Thomas, and Masao Utiyama. 2012. MNH-TT: a collaborative platform for translator training. In *Proceedings of Translating and the Computer 34*.

[10] http://www.statmt.org/wmt16/quality-estimation-task.html

Aljoscha Burchardt and Arle Lommel. 2014. QT-LaunchPad supplement 1: Practical guidelines for the use of MQM in scientific research on translation quality. http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2006. Designing a learner translator corpus for training purpose. In *Proceedings of the 7th International Conference on Teaching and Language Corpora*.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*) *Short Papers*, pages 57–60.

ISO/TC27. 2015. ISO 17100:2015 translation services: Requirements for translation services.

Donald Kiraly. 2000. *A Social Constructivist Approach to Translator Education: Empowerment from Theory to Practice*. Routledge.

Kinga Klaudy. 1996. Quality assessment in school vs professional translation. In Cay Dollerup and Vibeke Appel, editors, *Teaching Translation and Interpreting 3: New Horizons: Papers from the Third Language International Conference*, pages 197–203. John Benjamins.

Arle Lommel, Maja Popović, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the LREC MTE Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

Arle Lommel, Attila Görög, Alan Melby, Hans Uszkoreit, Aljoscha Burchardt, and Maja Popović. 2015. QT21 deliverable 3.1: Harmonised metric. http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf.

Brian Mossop. 2014. *Revising and Editing for Translators (3rd Edition)*. Routledge.

Mariana Orozco and Amparo Hurtado Albir. 2002. Measuring translation competence acquisition. *Translators' Journal*, 47(3):375–402.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 311–318.

Alina Secară. 2005. Translation evaluation: A state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE Workshop*, pages 39–44.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pages 897–904.