Designing a Metalanguage of Differences Between Translations: A Case Study for English-to-Japanese Translation

Tomono Honda[†*] **Atsushi Fujita**[‡] **Mayuka Yamamoto**[†] **Kyo Kageura**[†]
[†]The University of Tokyo, Tokyo, Japan
[‡]National Institute of Information and Communications Technology, Kyoto, Japan
[†]{tomono20@g.ecc, yamamoto.mayuka@mail, kyo@p}.u-tokyo.ac.jp
[‡]atsushi.fujita@nict.go.jp

Abstract

In both the translation industry and translation education, analytic and systematic assessment of translations plays a vital role. However, due to lack of a scheme for describing differences between translations, such assessment has been realized only in an ad-hoc manner. There is prior work on a scheme for describing differences between translations, but it has coverage and objectivity issues. To alleviate these issues and realize more fine-grained analyses, we developed an improved scheme by referring to diverse types of translations and adopting hierarchical linguistic units for analysis, taking English-to-Japanese translation as an example.

1 Introduction

In translation, assuring quality is the primary and indispensable issue. In translation industry, translation quality assessment (TQA) is introduced to ensure a certain level of quality for clients and end-users, whereas in research, TQA is conducted to gauge the differences in quality between different translation processes and systems (Castilho et al., 2018). The goals of TQA are diverse depending on situations, but regardless of situations, we need to compare translations as systematically and objectively as possible (Koby et al., 2014).

In translation education, learners should acquire competence to analyze and justify their translations, and explain their decisions with appropriate metalanguages and theoretical approaches (European Master's in Translation, 2022). Lacking systematically organized concepts and precise descriptions, however, instructors can explain several possible translations and their differences only by using their own languages in an ad-hoc manner, and learners are not able to grasp the whole picture of differences. In the translation production workflow in industry, machine translation (MT) systems are often used with manual post-editing (ISO/TC37, 2017). However, no study has analytically assessed how post-edited MT output (MT+PE) and translation produced exclusively by human translators (HT) differ and what cause the differences. These situations suggest the necessity of a comprehensive typology, or metalanguage (Kageura et al., 2022), of differences between translations (target documents, henceforth TDs) for the same source document (SD), as a scaffold to discuss such differences objectively, analytically, and precisely.

There is only one scheme that enables us to describe differences between independently produced TDs (Honda et al., 2022). While their scheme has been tailored for analytic and systematic assessment of differences, it has two vital problems. First, the covered phenomena would be limited; the TDs they analyzed were all from the same content domain and produced by human translators. Another problem is the vagueness and subjectivity of units employed to capture sub-sentential pairs within given TDs.

This paper presents our scheme for describing differences between TDs, which we have developed to alleviate these two problems. To cover a wider variety of phenomena, we used several SDs from various content domains and obtained their translations via substantially different methods, i.e., HT and MT+PE. For tangible and objective analyses, we adopted general linguistic units. Our scheme has two notable features: (i) it serves as scaffolding metalanguage for discussing differences between TDs, and (ii) it can be used as a research tool as well as a learning material.

The remainder of this paper is structured as follows. Section 2 describes related work. Section 3 explains how we have developed the scheme. Section 4 presents our scheme. Section 5 reports on

[*]This work was done during an internship of the first author at National Institute of Information and Communications Technology.

our intrinsic evaluation, and Section 6 discusses the current status of our scheme and remaining issues. Section 7 concludes this paper.

2 Related Work

Many studies have so far addressed analytic and systematic assessment of translation quality. Existing evaluation schemes, e.g., MQM (Lommel et al., 2014), focus on translation errors (or *issues*) (Castilho et al., 2018). However, none of them can be used to describe differences between pairs of issue-free translations: how they differ and what cause the differences. Recent MT systems, which cause less translation issues (Freitag et al., 2021), will require such schemes sooner or later.

There is a large body of studies comparing issue-free translations independently produced by various translators, such as students and professional translators (Pastor et al., 2008; Lapshinova-Koltunski, 2015; Rubino et al., 2016; Ghent et al., 2018; Bizzoni and Lapshinova-Koltunski, 2021; Lapshinova-Koltunski et al., 2022), and MT+PE and HT (Toral, 2019). They revealed differences in terms of linguistic features, i.e., *translationese* (Baker, 1993; Laviosa-Braithwaite, 1998) and *post-editese* (Toral, 2019). However, they only observed general tendencies of TDs as a whole, and none of them established a means to analytically and systematically explain individual instances that exhibit some kind of differences.

Unlike above, Yamamoto and Yamada (2022) made an analytic comparison of draft and final versions of TDs. They compiled a typology of manipulations applied to TDs during the production process, called *translation strategies* (Chesterman, 2016),[1] extending the work by Chesterman (2016), and ensuring the coverage and systematicity through analyzing actual revision examples extracted from pairs of draft and final versions of TDs. Their typology consists of syntactic, semantic, and pragmatic subparts[2] comprising 13, 9, and 10 types, respectively. The syntactic and semantic strategies have been adopted from linguistic theories (Morris, 1938). The pragmatic strategies are, on the other hand, more specific to translation, e.g., referring to external information and ensuring quality for target readers, performed to produce a TD that is more appropriate for the predetermined purposes. However, the typology of translation strategies would not be applicable to the pairs of independently produced TDs, since it has been developed only on the basis of revision examples performed during the process of producing TDs.

Honda et al. (2022) is the pioneer of constructing a scheme for extracting and explaining differences between independently produced TDs for the same SD. To identify differences between any pair of linguistic expressions observed in given pairs of TDs, they proposed a two-step procedure: decompose given pairs of TDs and classify the differences between each constituent pair. The latter is realized with decision lists, consisting of 13, 8, and 4 types of categories for syntactic, semantic, and pragmatic differences incorporated from translation strategies (Chesterman, 2016; Yamamoto and Yamada, 2022). Their work has two major defects. One is the limited variety of phenomena it covers. They used a set of abstracts of scientific articles and their human translations (HT) produced by different translators. Due to the relatively limited range of textual domain, homogeneous text type, and the same method for translation production, they observed only a limited range of differences. The other problem is the intermediate unit called "chunk." They introduced it in between sentence and word, and proposed criteria to extract pairs of chunks from given pairs of TDs. However, the vague definition of chunk leads to subjective analyses.

3 Construction of the Scheme

We developed an improved scheme for describing differences in TDs. In our scheme, we adopted the two-step workflow proposed in the previous work (Honda et al., 2022): top-down recursive decomposition of pairs of TDs followed by classification of each constituent pair into pre-defined categories. We also followed Honda et al. (2022) to implement a procedure for the first step and decision lists for the second step. In contrast, we addressed the two problems in Honda et al. (2022) as follows.

- To cover a wider variety of differences, we used the TDs that belong to various content domains and produced by different methods (Section 3.1), and reconsidered to incorporate translation strategies that Honda et al. (2022) did not adopt (Section 3.2.2).

[1] Yamamoto and Yamada (2022, p.83) explain that translation strategies are "methods applied to achieve a proper translation that moves beyond the literal."

[2] Chesterman (2016, p.104) defined the three groups of strategies as follows: "if syntactic strategies manipulate form, and semantic strategies manipulate meaning, pragmatic strategies can be said to manipulate the message itself."

Usage	ID	# seg	# sentences			# words (tokens)			Topic
			SD	HT	MT+PE	SD	HT	MT+PE	
Development	Doc1	11	21	25	26	524	774	762	Clean energy
	Doc2	8	15	18	18	415	593	588	Medical equipment
	Doc3	7	13	14	14	273	339	362	CAD software
	Doc4	11	21	21	22	399	555	575	Travel health
	Doc5	18	34	34	34	895	1,184	1,197	Radio frequency devices
Refinement	Doc6	19	32	30	30	384	541	499	Complaint letter
	Doc7	31	39	42	41	463	632	630	Game application
Validation	Doc8	36	47	48	48	446	621	602	Licensing procedure
	Doc9	32	39	40	41	530	729	715	Contract renewal

Table 1: Usage of and statistics for documents and translations: words (token) counts were obtained by NLTK (Bird et al., 2009) for the SDs in English and MeCab and IPAdic (Kudo et al., 2004) for the two types of TDs in Japanese. "seg" indicates "segments" given as original units aligned across SD, MT+PE, and HT.

- To carry out tangible and objective analyses, we adopted hierarchical linguistic units in the target language widely used in linguistics (Section 3.2.1).

We developed and refined our scheme through repeating annotation and discussion in order to ensure its systematicity and coverage as much as possible (Sections 3.2 and 3.3), taking English-to-Japanese translations.

3.1 Collecting Translation Data

When designing and validating an annotation scheme, in general, it is ideal to take as diverse examples as possible into account.

To ensure the diversity of SDs, we used technical documents in various specialized fields, considering their nature and purposes; they are rather literal and logical than figurative and emotional. We also expected that the requirements in translating them should potentially be identified and explained in the form of translation brief, and that the subtle differences seen in their translations would be explainable by ourselves. For our study, we collected nine technical documents written in English.[3]

To ensure the diversity of translations, we decided to compare human translation (HT) and post-edited version of machine translation (MT+PE). HT is eligible as one side of document pairs for comparison, because it should have the highest quality among conceivable ways of obtaining translations. As the counterpart, we chose MT+PE, assuming that it assures certain quality if it follows ISO 18587 (ISO/TC37, 2017), and that it should be substantially different from HT due to the certain level of reliance on MT outputs. Even if MT+PE is close enough to HT, analyzing their differences still contributes to research on MT.

HT and MT+PE for the nine documents were produced by two different Translation Service Providers (TSPs). For HT, we asked an ISO-certified TSP to produce HT following ISO 17100 (ISO/TC37, 2015). For MT+PE, we first obtained English-to-Japanese MT outputs using TexTra[4] and asked another ISO-certified TSP to post-edit the MT outputs following ISO 18587 (ISO/TC37, 2017) but avoiding excessive editing.

Table 1 summarizes the statistics for the collected tuples of SD, HT, and MT+PE. We used five tuples for development, other two for refinement, and the rest two for validation of the scheme.

3.2 Development of the Initial Scheme

The authors, whose native language is Japanese and thus have sufficient linguistic competence in Japanese, first created the scheme for English-to-Japanese translation through repeating annotation and discussion. Annotation, i.e., decomposition of paired TDs and classification of extracted pairs, was carried out by one of the authors of this paper, and another author joined in discussion to revise the scheme. Five tuples of SD, HT, and MT+PE (Doc1 to Doc5 in Table 1) were used.

3.2.1 Decomposing Unit Pairs

Within pairs of relatively large units, such as sentence pairs, several types of differences can coexist. Aiming at analytically describing each difference, Honda et al. (2022) proposed to decompose

[3] We searched for documents on the Web considering their license for our future release of documents with our translations and annotations.

[4] https://mt-auto-minhon-mlt.ucri.jgn-x.jp/, GPMT-3.9_200930_nmt

given sentence-aligned TDs into smaller units. To better handle the hierarchical structures in TDs, we adopted linguistic units in Japanese.

First, we extracted pairs of linguistic units from each pair of TDs, making sure that each unit to be well-defined in linguistics, such as clause and noun phrase, referring to literature on Japanese grammar (Masuoka and Takubo, 1992; SIG for Descriptive Grammar in Japanese, 2008, 2009a, 2010). Each unit is also aligned with corresponding unit in SD in order to identify the corresponding units in different TDs. Then, based on the results, we refined the procedure for decomposition as well as the types of units by grouping them based on linguistic features. In this refinement process, we decided to distinguish the "non-linguistic units" that play some role in document from linguistic units.

We present the resulted procedure and the types of units in Section 4.1.

3.2.2 Classifying Differences

Following Honda et al. (2022), we used three exclusive groups of categories (syntactic, semantic, and pragmatic categories) and decision lists for describing differences. To cover a wide variety of differences, we reconsidered to incorporate the categories discussed in the literature of translation strategies (Chesterman, 2016; Yamamoto and Yamada, 2022) that Honda et al. (2022) did not adopt.

Given extracted pairs of units, one of the authors first classified them into one of the categories within a union of those presented in Chesterman (2016), Yamamoto and Yamada (2022), and Honda et al. (2022). Syntactic, semantic, and pragmatic differences were separately analyzed, as in previous work. We then examined the results to refine the categories. When we found problems, such as phenomena that are not covered by existing categories, we refined the decision lists by adding new categories, revising definition statements to extend the scope of existing categories, and/or dividing or merging existing categories, referring to literatures of linguistics (SIG for Descriptive Grammar in Japanese, 2003, 2009b,c, 2010). In the decision lists, we prioritized categories that describe a more specific and/or easily identifiable phenomena.

The resulted three sets of categories are presented in Section 4.2.

3.3 Refinement of the Scheme

After a couple of iterations of the initial phase (Section 3.2), the same two of the authors refined the scheme in a more rigorous setting: independent annotation followed by comparison of the results. First, they only decomposed TDs for Doc6, and refined the procedure for decomposition through comparing the results. Then, they did both decomposition and classification for Doc7. Through comparing the results, they determined the issues of the scheme from the viewpoint of consistency, coverage, and understandability of the instructional materials, and improved them.

4 Our Improved Scheme

Through the process described in Section 3, we developed a scheme for describing differences between pairs of TDs. During the process, we also assembled instructional materials for annotators. These documents are made publicly available;[5] they are mainly written in Japanese, since we have compiled them for analyzing TDs in Japanese.

In this section, we explain their summary, using examples of unit decomposition and classification shown in Table 2.

4.1 Procedure for Decomposing Unit Pairs

We defined a total of nine types of units for analysis. Seven out of them are "linguistic unit" well-defined in linguistics: paragraph, sentence, clause, phrase, compound expression, word, and punctuation. The remaining two are called "non-linguistic unit" since they play specific roles within document: "sentence-equivalent unit," such as headlines and bibliographic information, and "phrase-or-word-equivalent unit," such as terms, named entities, and inline quotations.

The overview of our procedure for decomposing and extracting units for analysis is as follows.

Step 1. Check if the stopping conditions apply: assess whether the given pair of units for analysis must be decomposed or not.

Step 2. Decompose each TD unit: decompose each unit into smaller units "without nesting;" the extracted units must be as large as possible and must not overlap with each other.

Step 3. Align with SD: align each extracted unit of TD with its corresponding unit of SD.

Step 4. Align between TD units: identify pairs of constituent units extracted from different TDs that correspond to the identical unit of

[5] https://github.com/tntc-project/translation-difference

No.	d	Unit in SD	Unit in TD1	Unit in TD2	Syn	Sem	Pra
1	2	Payment of the fee must accompany the form.	手数料の支払は、用紙を添付する必要があります。	料金の支払いには、申請書を添付しなければなりません。	g4	NA	p100
2	3	payment of the fee	手数料の支払	料金の支払い	g100	PEQ	PEQ
3	4	the fee	手数料	料金	g100	s6	p9
4	4	payment	支払	支払い	g12	s2	p9
5	3	ϕ	、	、	EQ	EQ	EQ
6	3	the form	用紙	申請書	g100	s7	p7
7	3	must accompany	添付する必要があります	添付しなければなりません	g18	s10	p100
8	3	.	。	。	EQ	EQ	EQ

Table 2: Examples of extracted units for analysis labeled with their syntactic (Syn), semantic (Sem), and pragmatic (Pra) categories. d indicates the depth of the unit; for instance, the first unit with $d = 2$ means that this tuple of sentences has directly been extracted from a given ($d = 1$) parallel paragraphs.

SD. Here, functional words that are not mutually interchangeable are left unaligned, since such difference takes a part of the given pair of larger units. The identical functional expressions are also left unaligned, for the sake of simplicity in analyzing differences.

Note that this procedure is recursively applied to every pair of constituent units, in order to thoroughly decompose and extract the units for analysis in the given pair of TDs.

For a unit pair which has been decomposed into several constituent unit pairs, we analyze the differences between their constructions, ignoring the differences between the extracted constituent unit pairs. To this end, we decided to identify patterns for unit pairs that are decomposed. Given a unit pair, the pattern for each side is obtained by replacing the strings corresponding to each constituent unit with a unique symbol. For instance, the unit pair in line 1 in Table 2 ($d = 2$) is decomposed into unit pairs in lines 2, 5, 6, 7, and 8 ($d = 3$). By replacing the strings corresponding to each constituent with letters A to E, we obtain the patterns "AはBCをDE" for TD1 and "AにはBCをDE" for TD2. Note that, as explained in Step 4, some functional words, "は" (topic marker), "を" (accusative case), and "に" (dative case) in this case, are not extracted as a constituent unit pair and thus left lexicalized. For another instance, the pair in line 2 in Table 2 ($d = 3$) is further decomposed into pairs in lines 3 and 4 ($d = 4$), leaving aligned but identical function word "の" (genitive case) unextracted, and the patterns of the unit pair are both identified as "AのB."

4.2 Decision Lists for Classifying Differences

For each pair of units extracted from a pair of TDs, we separately analyze their syntactic, semantic, and pragmatic differences following the decision lists. Tables 3, 4, 5 show the categories of each group.[6] In each table, the categories with a check mark (✓) indicate that they do not exist in the scheme of Honda et al. (2022) and are newly added in our work. See Appendix A for their definitions.

Syntactic categories describe syntactic differences, such as structures and forms, not involving content. Note that some syntactic categories in Table 3 are only applicable to certain types of unit pairs, e.g., "g9 Clause structure difference" never happens when analyzing pairs of paragraphs.

Semantic categories describe differences of contents or meanings and are applied only to linguistic units. In the categories shown in Table 4, "NA Not applicable" is assigned to a unit pair that (a) both of the units are paragraphs or sentences, or (b) at least one of the units is non-linguistic unit. "PEQ Pattern equivalence" is used for unit pairs whose patterns are identical, e.g., the unit pair in line 2 in Table 2 both of which are identified as "AのB" as patterns.

Pragmatic categories describe pragmatic differences, such as relationships between the sender and receivers, and language use or structures considering the purposes of documents. "PEQ Pattern equivalence" in Table 5 is the same as "PEQ" in the semantic categories.

[6]We defined the decision list of syntactic categories for each pair of unit types. Thus, unlike Tables 4 and 5, Table 3 does not serve as a decision list.

Label	New	Category name
EQ	✓	Exact match
g1	✓	Paragraph structure difference
g2	✓	Sentence type difference
g3	✓	Voice difference
g4	✓	Topic difference
g5		Sentence structure difference
g6	✓	Segment structure difference
g7	✓	Clause type difference
g8	✓	Ellipsis/Repetition difference
g9	✓	Clause structure difference
g10	✓	Quotation difference
g11	✓	Original spelling difference
g12	✓	Orthography difference
g13		Loan difference
g14	✓	Acronym difference
g15		Phrase structure difference
g16	✓	Reference expression difference
g17		Part of speech difference
g18	✓	Predicate difference
g19	✓	Affix difference
g20		Function word difference
g21	✓	Presence of translation
g22	✓	Analysis unit difference
g23		Unit difference
g99	✓	Other syntactic difference
g100	✓	Syntactic equivalence

Table 3: Syntactic categories.

Label	New	Category name
EQ	✓	Exact match
NA	✓	Not applicable
PEQ	✓	Pattern equivalence
s1	✓	Conjugated form difference
s2	✓	Spelling difference
s3	✓	Polysemy difference
s4	✓	Causal difference
s5	✓	Trope difference
s6	✓	Hyponymy difference
s7		Abstraction difference
s8		Emphasis difference
s9		Perspective difference
s10	✓	Predicate meaning difference
s11		Synonym
s99	✓	Other semantic difference

Table 4: Semantic categories.

4.3 Instructional Materials for Annotators

In order for annotators to appropriately apply our scheme, we prepared four types of instructional materials. Two of them are documents for decomposition and classification, which are described in Sections 4.1 and 4.2.

In addition, we also assembled the following two materials about decomposition procedure in order to guide the annotators in the complicated decomposition process: (a) a document describing detailed procedure of decomposition with some examples, and (b) a video material showing the procedure of decomposition in a step by step manner.

Label	New	Category name
EQ	✓	Exact match
PEQ	✓	Pattern equivalence
p1	✓	Translation error
p2	✓	Transediting difference
p3	✓	Structure-awareness difference
p4		Cultural filtering difference
p5	✓	Interpersonal difference
p6	✓	Cohesion difference
p7	✓	Explicitness/Implicitness difference
p8		Domain adaptation difference
p9	✓	Register difference
p10	✓	Readability difference
p99	✓	Other pragmatic difference
p100	✓	Pragmatic equivalence

Table 5: Pragmatic categories.

All of the materials include some examples, such as those collected from Doc1-Doc7 in Table 1 during development of the scheme.

5 Intrinsic Evaluation

We evaluated whether our scheme meets the criteria of metalanguage of translation (Kageura et al., 2022), in particular, consistency of decomposition, consistency of classification, and coverage of categories. The two engaged in the development (A and B) and another one of the authors (C) participated in the evaluation as annotators. Annotator A is a graduate student in pedagogy, Annotator B is a Ph.D in computational linguistics, and Annotator C is an MA in translation studies. The annotators first read instructional materials of the scheme described in Section 4. They then independently annotated the two pairs of TDs reserved unseen for this purpose (Doc8 and Doc9 in Table 1) following the two-step annotation workflow: decomposition of the TD pairs into constituent unit pairs and classification of each pair into categories. In the classification step, they completed annotation for each of syntactic, semantic, and pragmatic categories for all the extracted TD pairs in this order.

As a result, they extracted 471, 466, and 443 pairs of units from Doc8, and 463, 456, and 451 from Doc9, respectively. Tables 6, 7, and 8 respectively show the frequencies of syntactic, semantic, and pragmatic categories labeled by each annotator.

5.1 Consistency of Decomposition

To gauge the inter-annotator consistency of unit decomposition, we computed recall, precision, and F1 score of each annotator's result regarding another

Category	Doc8			Doc9		
	A	B	C	A	B	C
EQ	129	139	131	166	161	164
g1	0	0	1	0	0	0
g2	0	0	0	1	0	0
g3	2	2	1	3	2	0
g4	7	7	3	2	1	3
g5	1	1	2	1	0	0
g6	0	1	1	3	1	3
g7	0	0	0	0	0	0
g8	1	0	1	0	0	0
g9	3	1	0	3	5	0
g10	0	0	0	0	0	0
g11	13	12	23	10	7	6
g12	47	44	39	37	33	26
g13	8	4	11	5	4	3
g14	0	0	0	0	0	0
g15	3	5	2	10	9	4
g16	1	1	3	0	1	1
g17	12	16	10	11	8	17
g18	15	14	14	18	10	17
g19	6	3	0	2	0	0
g20	4	5	9	8	7	6
g21	18	14	26	21	27	49
g22	0	0	1	4	7	0
g23	55	52	35	37	47	32
g99	0	11	18	6	3	15
g100	146	134	112	115	122	105
Other*	0	0	0	0	1	0
Total	471	466	443	463	456	451

Table 6: Frequency of syntactic categories. "Other*" indicates that the annotator judged that a pair of units could not be classified to any category.

Category	Doc8			Doc9		
	A	B	C	A	B	C
EQ	129	139	131	166	161	164
NA	48	49	49	52	55	14
PEQ	100	89	70	88	89	96
s1	3	4	2	9	7	10
s2	24	18	17	10	13	14
s3	1	3	2	0	1	7
s4	1	3	5	1	6	1
s5	0	0	0	0	0	0
s6	2	9	0	1	2	1
s7	15	35	44	15	25	26
s8	30	12	10	45	14	9
s9	9	10	6	9	15	4
s10	3	11	5	3	13	5
s11	101	73	68	62	46	43
s99	5	11	34	2	9	57
Total	471	466	443	463	456	451

Table 7: Frequency of semantic categories.

annotator's result as a gold standard. We excluded the original units given for annotation, i.e., 36 and 32 segments for Doc8 and Doc9, respectively, as they were consistent by definition.

Category	Doc8			Doc9		
	A	B	C	A	B	C
EQ	129	139	131	166	161	164
PEQ	126	115	69	112	115	96
p1	1	5	3	2	5	2
p2	0	0	0	0	1	0
p3	1	0	13	11	0	1
p4	33	24	24	9	10	8
p5	23	13	19	6	8	14
p6	10	2	10	10	2	11
p7	7	22	34	17	26	13
p8	4	1	3	8	10	27
p9	86	67	2	74	47	1
p10	27	13	26	30	37	10
p99	4	5	0	4	0	8
p100	20	60	109	14	34	96
Total	471	466	443	463	456	451

Table 8: Frequency of pragmatic categories.

Test	Gold	Doc8			Doc9		
		R	P	F1	R	P	F1
A	B	84.0	83.0	83.5	73.6	72.4	73.0
B	C	80.8	76.5	78.6	70.9	70.0	70.5
C	A	75.9	81.1	78.4	67.5	69.5	68.5

Table 9: Inter-annotator consistency of decomposition (%): R, P, F1 stand for recall, precision, and F1 score, respectively, computed regarding the result of one annotator as reference (Gold). For reversed pairs of test and gold annotators, consider R and P flipped.

Table 9 summarizes the results. The F1 scores span 78.4–83.5 for Doc8 and 68.5–73.0 for Doc9. While the F1 scores for each document were relatively stable (≤ 5.1 points), there were larger gaps between Doc8 and Doc9 (≥ 8.1 points).

We consider that our scheme has enabled the annotators to decompose unit pairs relatively consistently, but the lower F1 scores for Doc9 suggest that linguistic complexity in TDs and/or the similarity between independently produced TDs can affect the decomposition process.

Retrospective interview with the annotators revealed that the most typical disagreement was due to the different recognition of syntactic structure. For instance, see the following example of a noun phrase that the three annotators decomposed in different ways, where brackets indicate the constituent units extracted from the phrase.

SD: Types of Submissions Subject to eCTD Requirement

MT+PE: eCTD要件の対象となる申請の種類

A: [eCTD要件の対象となる][申請の種類]

Pair	Doc8				Doc9			
	# unit	Syntactic	Semantic	Pragmatic	# unit	Syntactic	Semantic	Pragmatic
A–B	280	79.6 (0.73)	70.4 (0.63)	66.1 (0.56)	218	71.1 (0.61)	61.9 (0.53)	63.3 (0.51)
B–C	255	71.4 (0.64)	65.5 (0.59)	39.6 (0.29)	198	69.2 (0.59)	46.0 (0.34)	51.5 (0.38)
C–A	263	74.1 (0.67)	65.4 (0.58)	38.8 (0.30)	196	73.5 (0.66)	48.5 (0.39)	45.4 (0.34)

Table 10: Inter-annotator agreement ratio (%) and Cohen's κ (in parenthesis) on classification, excluding "EQ Exact match." "# unit" indicates the number of unit pairs obtained by both of each pair of annotators.

B: [eCTD要件の対象となる][申請][の][種類]

C: [eCTD要件の対象となる申請][の][種類]

Annotator A recognized that the phrase comprises an adnominal clause and a head noun phrase, while Annotator B further detached the genitive modifier, "申請" (application), and genitive case marker, "の" (of), considering that the single noun, "種類" (type), is the shared modificand. Annotator C identified an adnominal noun phrase as a genitive modifier of the single head noun. This example illustrates that structural ambiguities in TDs affect the decomposition procedure.

5.2 Consistency of Classification

We computed inter-annotator agreement ratio and Cohen's κ (Cohen, 1960) for the set of unit pairs shared by each pair of annotators, excluding units annotated with "EQ," i.e., the identical pair of units in HT and MT+PE.

Table 10 summarizes the results. Compared to κ values for syntactic categories spanning 0.59–0.73, those for semantic and pragmatic categories were low: 0.34–0.63 and 0.29–0.56, respectively. This indicates that semantic and pragmatic categories are more difficult to consistently classify.

See, for instance, the pair of bracketed expressions in the following example.

SD: Submissions [for] blood and blood components

HT: 血液および血液成分[に関する]申請

MT+PE: 血液及び血液成分[の]申請

The three annotators labeled this pair with different semantic categories: "s7 Abstraction difference," "s8 Emphasis difference," and "s11 Synonym." Through discussion, the annotators agreed that this example should be classified as s7, since "に関する" (regarding) is more specific compared to "の" (of/for). Such discussion calls for the clarity of the definition of s7 in the decision list for classification.

5.3 Coverage of Categories

Relatively low frequency of p99 (Table 8) suggests that the scheme ensures the coverage of pragmatic categories. In contrast, the higher frequencies of g99 (up to 18 in Table 6) and s99 (up to 57 in Table 7) reveal the necessity of refining our scheme. For instance, see the following example.

SD: Products that [are intended] to be distributed commercially

HT: 商業的に流通することを[目的とした]製品

MT+PE: 市販されることを[意図した]製剤

Two of the annotators identified the syntactic differences between the idiomatic phrase in HT "目的とした" (are regarded as the goal) and the literal translation in MT+PE. We consider that we need a new category for this type of differences.

6 Discussion

Our intrinsic evaluation confirmed that our scheme enables us to analyze the differences between independently produced translations at a certain level of consistency and coverage. Toward improving the consistency of classification further, we plan to refine intensional definitions and enrich examples to delineate extensional definitions. External references, such as lists of functional expressions and named entities, terminology, and style specifications, should also help improve consistency. To ensure the coverage, we plan to introduce new categories.

Even though the present scheme does not achieve perfect consistency and coverage, we consider that the disagreed examples do not necessarily suggest the defects of the scheme. Such examples represent fundamental difficulties in understanding the notions indispensable for analyzing translations, and are thus useful in the practical use of the scheme. For instance, in educational settings, the scheme itself is a subject to learn. Through

exercises of annotating the same TD pairs and discussing discrepancies of the annotation results, learners should be able to improve their competence in translation and grasp underlying concepts, such as syntax, semantics, and pragmatics, referred to in the scheme. The scheme will also help learners explain their specific choice of expressions in the target language.

Our scheme subsumes the categories of translation strategies (Chesterman, 2016; Yamamoto and Yamada, 2022) and enables comparisons of arbitrary pair of entire TDs. It is thus worth investigating that our scheme can also be used to analyze translation strategies.

7 Conclusion

This paper presented a scheme for analytically and systematically assessing the differences between independently produced translations for the same SD. On the basis of the work in Honda et al. (2022), we adopted nine types of linguistic/non-linguistic units for analysis and refined the decision lists with a wide variety of categories through annotation and discussion using substantially heterogeneous translations, i.e., HT and MT+PE. Unlike previous work in analytic assessment (Chesterman, 2016; Yamamoto and Yamada, 2022; Honda et al., 2022), we also conducted an intrinsic evaluation of the scheme, employing multiple annotators. The results show that classification of semantic and pragmatic differences is more difficult compared to decomposing unit pairs and classifying syntactic differences. Nevertheless, we believe that our scheme is useful, since it covers a wide range of translation-related concepts and thus can be a useful metalanguage to talk about differences in translation.

Our scheme is partly dependent on the target language, i.e., Japanese. We thus plan to examine its applicability to translations from other languages than English into Japanese. To analyze differences between translations in other target languages than Japanese, we need to adapt our scheme to them.

Acknowledgments

We would like to thank the anonymous reviewers, including those for past submissions, for their valuable comments and suggestions. This work was partly supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S)19H05660.

References

Mona Baker. 1993. Corpus linguistics and translation studies—implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and technology: In honour of John Sinclair*, pages 233–250. John Benjamins.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media.

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Measuring translationese across levels of expertise: Are professionals more surprising than students? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 53–63.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation quality assessment from principles to practice*, pages 9–38. Springer International Publishing.

Andrew Chesterman. 2016. *Memes of translation: The spread of ideas in translation theory*. John Benjamins.

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

European Master's in Translation. 2022. European master's in translation competence framework.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gert De Sutter Ghent, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2018. Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 16:25–39.

Tomono Honda, Mayuka Yamamoto, and Kyo Kageura. 2022. Construction of a scheme for describing differences between translations. *Invitation to Interpreting and Translation Studies*, 24:1–21. (In Japanese).

ISO/TC37. 2015. ISO 17100:2015 translation services — requirements for translation services.

ISO/TC37. 2017. ISO 18587:2017 translation services — post-editing of machine translation output — requirements.

Kyo Kageura, Rei Miyata, and Masaru Yamada. 2022. Metalanguages and translation studies. In Rei Miyata, Masaru Yamada, and Kyo Kageura, editors,

Metalanguages for dissecting translation processes: Theoretical development and practical applications, pages 15–26. Routledge.

Geoffrey S. Koby, Paul Fields, Daryl Hague, Arle Lommel, and Alan Melby. 2014. Defining translation quality. *Tradumática*, 12:413–420.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Ekaterina Lapshinova-Koltunski. 2015. Exploration of inter- and intralingual variation of discourse phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 158–167.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: A parallel corpus to analyse differences between human translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1751–1760.

Sara Laviosa-Braithwaite. 1998. Universals of translation. In Mona Baker, editor, *Routledge encyclopedia of translation studies*, pages 288–291. Routledge.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática*, 12:455–463.

Takashi Masuoka and Yukinori Takubo. 1992. *Basic Japanese grammar*. Kurosio Publishers. (in Japanese).

Charles W. Morris. 1938. Foundations of the theory of signs. In Otto Neurath, Rudolf Carnap, and Charles W. Morris, editors, *International encyclopedia of unified science*, pages 1–59. Chicago University Press.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pages 75–81.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970.

SIG for Descriptive Grammar in Japanese, editor. 2003. *Modern Japanese grammar*, volume 4. Kurosio Publishers. (in Japanese).

SIG for Descriptive Grammar in Japanese, editor. 2008. *Modern Japanese grammar*, volume 6. Kurosio Publishers. (in Japanese).

SIG for Descriptive Grammar in Japanese, editor. 2009a. *Modern Japanese grammar*, volume 2. Kurosio Publishers. (in Japanese).

SIG for Descriptive Grammar in Japanese, editor. 2009b. *Modern Japanese grammar*, volume 5. Kurosio Publishers. (in Japanese).

SIG for Descriptive Grammar in Japanese, editor. 2009c. *Modern Japanese grammar*, volume 7. Kurosio Publishers. (in Japanese).

SIG for Descriptive Grammar in Japanese, editor. 2010. *Modern Japanese grammar*, volume 1. Kurosio Publishers. (in Japanese).

Antonio Toral. 2019. Post-editese: An exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281.

Mayuka Yamamoto and Masaru Yamada. 2022. Translation strategies for English-to-Japanese translation. In Rei Miyata, Masaru Yamada, and Kyo Kageura, editors, *Metalanguages for dissecting translation processes: Theoretical development and practical applications*, pages 80–91. Routledge.

A Definitions of Categories

Tables 11, 12, and 13 give the lists of categories and their definitions for each primary category group. The lists of semantic and pragmatic categories in Tables 12 and 13 also serve as decision lists.

Label	Category name	Definition
EQ	Exact match	Identical units; the character strings exactly match
g1	Paragraph structure difference	Differences in the order of translation at sentence level
g2	Sentence type difference	Differences in sentence types (e.g., simple sentences, complex sentences, declarative sentences, interrogative sentences, or imperative sentences)
g3	Voice difference	Differences in voice expressions which often lead to the differences in case structures
g4	Topic difference	Differences in salience and/or markedness of topic (e.g., presence or absence of the topic or the use of particles expressing the topic, differences of the words expressed as the topic, or differences of particles expressing the topic)
g5	Sentence structure difference	Differences in sentence structures, such as the relationship between a main clause and a subordinate clause, the order of translation at clause level, or modification relationships
g6	Segment structure difference	Differences in the structures (e.g., order of translation) in non-linguistic units (e.g., headlines, items, or footnotes)
g7	Clause type difference	Differences in clause types (e.g., interrogative, quotation, adnominal, and adverbial clauses)
g8	Ellipsis/Repetition difference	Differences in the ways of translation, such as repetition or ellipsis of a modifier or a modificand
g9	Clause structure difference	Differences in clause structures, such as modification relationships
g10	Quotation difference	Differences in the ways of translating quotations, including the uses of quotation marks
g11	Original spelling difference	Differences in the use of original spelling in SD
g12	Orthography difference	Differences in orthography
g13	Loan difference	Differences in the use of loan words (e.g., transliteration)
g14	Acronym difference	Differences in the use of acronym
g15	Phrase structure difference	Differences in phrase structures, such as word order or modification relationships
g16	Referring expression difference	Differences in the use of referring expressions
g17	Part of speech difference	Differences in parts of speech
g18	Predicate difference	Differences in predicates, such as tense, aspect, and mood
g19	Affix difference	Differences in types of affix or presence/absence of affix
g20	Function word difference	Differences in function words (e.g., particles, auxiliary verbs) or functional expressions
g21	Presence of translation	Differences in presence of translation
g22	Analysis unit difference	Differences between non-linguistic and linguistic units
g23	Unit difference	Differences in the types of linguistic units
g99	Other syntactic difference	Other syntactic differences that are not applicable to above categories
g100	Syntactic equivalence	No syntactic differences

Table 11: The list of syntactic categories and definitions.

Label	Category name	Definition
EQ	Exact match	Identical units; the character strings exactly match
NA	Not applicable	Not applicable in semantic categories; both of the units are paragraphs or sentences, or at least one of the units is non-linguistic unit
PEQ	Pattern equivalence	Identical pattern in both units
s1	Conjugated form difference	Differences only in conjugated form
s2	Spelling difference	Differences only in the orthography in Japanese writing system
s3	Polysemy difference	Differences in transferring different meanings of an ambiguous word in SD
s4	Causal difference	Causal relationships between the meanings of units
s5	Trope difference	Differences in the use of trope expressions or styles of trope expressions
s6	Hyponymy difference	Hyponym and hypernym relationships between the meanings of units
s7	Abstraction difference	Differences in the degrees of abstraction
s8	Emphasis difference	Differences in the ways of emphasis or focuses of the description
s9	Perspective difference	Differences in the perspectives of stating the same content
s10	Predicate meaning difference	Differences in the meanings of predicate expressions
s11	Synonym	Synonymous relationships between the meanings of the units
s99	Other semantic difference	Other semantic differences that are not applicable to above categories

Table 12: The decision list and definitions of semantic categories.

Label	Category name	Definition
EQ	Exact match	Identical units; the character strings exactly match
PEQ	Pattern equivalence	Identical pattern in both units
p1	Translation error	Differences in translating contents of SD wrongly in either one or both of TDs
p2	Transediting difference	Differences in the degrees of transediting the badly written SD (e.g., errors or ambiguities)
p3	Structure-awareness difference	Differences in the ways of adapting expressions and constructions to the functional roles of SD element (e.g., titles, items, footnotes, captions, and citations)
p4	Cultural filtering difference	Differences in whether domesticating to the target culture or not (e.g., translating a feature in source culture by using expressions that adapt to the target culture)
p5	Interpersonal difference	Differences in the degrees of reflecting the relationships between the sender and receivers (e.g., politeness, feeling, or intervention)
p6	Cohesion difference	Differences in the degrees of cohesiveness (e.g., those exhibited by the use of ellipsis, repetition, or conjunction words)
p7	Explicitness/Implicitness difference	Either one of TDs adds new information that does not exist in SD, or explicitly expresses information originally implicit in SD, for the purpose of explicitness of sender's intention or supplement of readers' understanding (e.g., differences in modifications, notes, explanation with parenthesis, or the use of words adapting to context)
p8	Domain adaptation difference	Differences in the use of expressions specific in the content domain of SD
p9	Register difference	Differences in the use of expressions adopted to the text type or register
p10	Readability difference	Differences in readability (considering the supposed readers)
p99	Other pragmatic difference	Other pragmatic differences that are not applicable to above categories
p100	Pragmatic equivalence	No pragmatic differences

Table 13: The decision list and definitions of pragmatic categories.