Koby, G. S., & Lacruz, I. (2017). The thorny problem of translation and interpreting quality. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 1–12.

The thorny problem of translation and interpreting quality

Geoffrey S. Koby

Kent State University, Kent, Ohio, USA
gkoby@kent.edu

Isabel Lacruz

Kent State University, Kent, Ohio, USA
ilacruz@kent.edu

Judging quality in translation and interpreting and in the associated task of revision has a long and controversial history. We briefly comment on some aspects of this history to provide context for the contemporary perspectives on and investigations into quality assessment that are represented in this volume of Linguistica Antverpiensia, New Series: Themes in Translation Studies. *A fundamental obstacle to progress is the lack of consensus about how to characterize high-quality translation or interpreting, let alone the identification of broadly accepted models for measuring translation or interpreting quality or the ability of translators or interpreters. The advent of machine translation and post-editing has focused attention on the very nature of quality: Is it proximity to a "gold standard" of perfection or is it characteristic of a product that simply serves its purpose well enough to satisfy the needs of the consumer? In other words, is quality something that should be measured and judged in absolute terms or in relative terms? Different philosophies of quality assessment reflect these dichotomies, with the absolutists seeking objective assessments based on detailed analyses of taxonomies of errors, whereas the relativists prefer a more holistic approach that is more sympathetic to subjective judgements. The contributors to this volume present a broad range of approaches to quality assessment in a variety of contexts. We describe their achievements and provide brief analyses through the lens of the framework above.*

1. Translation and interpreting quality – A perennial topic

It is hardly speculation to imagine that translation and interpreting quality has been a topic, if not a contentious issue, even before our earliest records of written or oral language mediation. To cite just one famous example, in his 1530 *Sendbrief vom Dolmetschen (Open Letter on Translating)*, Martin Luther criticized the translation skill of his opponents, saying, "Aber die weil ich gewůst / vnd noch vor augen sihe / das yhr keiner recht weiß / wie man dolmetschen / odder teutsch reden sol / hab ich sie vnd mich solcher můhe vberhaben /" (p. F3) ("However, knowing as I did – and the evidence is still before my eyes – that not one of them has a clue how to translate or how to speak German, I spared them and myself the bother" Luther, 2017, p. 5). Based on his experience, Luther considered himself an expert on translation, long before certification exams or any form of vetting or licensing of translators. In this year, the 500th anniversary of Luther's triggering the Protestant Reformation by posting his 95 theses, the topic of translation quality remains as controversial as ever, although it is less common for those disputing the quality of a translation to call each other "Esels kȍpffen / donkey-heads" (1530, p. F6; 2017, p. 11).

In the more recent development of Translation Studies, translation quality was an early topic of discussion in various venues: for instance, in 1959, when it was the topic of the third Congress of the International Federation of Translators (Cary & Jumpelt, 1963), followed by Katharina Reiss's *Möglichkeiten und Grenzen der Übersetzungskritik* (1971, English translation 2000), and in more "modern" times, when Marilyn Gaddis Rose inaugurated the ATA Scholarly Monograph series with a volume entitled *Translation excellence: Assessment, achievement, maintenance* (1987). And yet

there is still no universally accepted model for measuring translation or interpreting quality, nor even two or three recognized models. Indeed, there is not even consensus on what constitutes the definition of a high-quality translation, as shown by two competing definitions, one narrow, one broad, put forward in the same 2014 article by Koby, Fields, Hague, Lommel and Melby:

> Narrow definition: "A high-quality translation is one in which the message embodied in the source text is transferred completely into the target text, including denotation, connotation, nuance, and style, and the target text is written in the target language using correct grammar and word order, to produce a culturally appropriate text that, in most cases, reads as if originally written by a native speaker of the target language for readers in the target culture." (p. 416)
>
> Broad definition: "A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs." (pp. 416–417)

The narrow definition could be called a traditional interpretation of human translation quality and how it should be produced. It implies that there is a fundamental standard of accuracy and fluency below which one would have to label a text as a flawed or defective translation, or perhaps not a translation at all. The broad definition, on the other hand, does not actually specify a quality level. Instead, it defines it situationally for the audience and purpose and according to the specifications from the requester. This could mean, for instance, that if price were the primary consideration, only minimum levels of accuracy and fluency would be acceptable, even if speakers of the language in question were to consider the text fundamentally flawed or difficult to read. Ultimately, it comes down to philosophy – if communication occurs, regardless of textual flaws, is that enough to call something a quality translation or quality interpreting? Or are there standards of language and communication that underpin a culture, standards below which translators and interpreters should not be allowed to fall?

In this introduction, we first discuss a variety of such issues of translation and interpreting quality assessment that are either addressed or implied in the various contributions to this volume, more from a language and translation studies philosophy point of view. Then we place these contributions within a larger context.

1.1 Basic assumptions

There are three basic assumptions underlying research into translation and interpreting quality and quality assessment (which includes the quality and assessment of machine translation (MT) and machine translation post-editing (MT-PE or just PE); they also occur to some extent in the contributions to this volume. The first assumption is that translation ability is a third skill, separate from language proficiency; the second is that the researcher is competent to analyse the area under investigation, and the third is that most translation output, whether human or MT, is flawed.

The first assumption is intuitive to translators, interpreters and researchers in the field and is thus hardly surprising. In fact, the concept of translation as a third skill could help to explain why, in certain areas, adequate translations can be made using dictionaries or online translation tools by subject-matter experts with no prior knowledge of the source language, or how non-translating experts in some technical areas can do as good a job as trained post-editors (Schwartz, 2014). We welcome Brooks and Brau's article as the first contribution in this volume because it specifically addresses verifying and quantifying the difference between source-language reading, target-language writing and translation ability.

On the second point, there has been much research about the qualifications of translators and interpreters and how to measure them, including holistic assessments and points-off systems. Certainly, a good deal of research is done with assessors as the research participants, but it is also not unusual to see research contributions where the researcher is also personally assessing the results of translations or interpreting. However, it is rare to see an article in which the researchers specifically cite their own qualifications for performing such assessment. Since research has shown that revisers

can and do overlook errors and introduce new errors, and also make unnecessary changes (see both Koponen and Salmi's and Van Rensburg's contributions in this volume), it could be expected that researchers trying to analyse errors would also be subject to the same issues, specifically in the preparation of research texts. How can the translation or interpreting researcher be sure that they have indeed recognized all the errors in a text? Is it possible for a single person to assess errors objectively? Certification testing programmes (e.g., the ATA programme, see Koby & Champe, 2010) acknowledge such concerns by requiring more than one grader for a valid test. Therefore, when researchers are doing their own assessment, a certain rigour in translation and interpreting studies research methods would seem to call for a description of the qualifications of the researcher to perform the assessment, as well as a well-described research design where more than one qualified assessor is discovering and assessing errors and where the grading protocol is explicitly designed, operationalized and described.

Third, translation and interpreting research fundamentally assumes that all output is potentially flawed and requires assessment or revising or post-editing.[1] This is a necessary assumption, since otherwise there would be no need to conduct such research. However, there is another underlying assumption that is rarely or never stated: the "gold standard" against which translation and interpreting are measured is a hypothetical "perfect" written translation or interpreting session, that is, one that contains no errors.[2] This underlying assumption is the justification for points-off systems of error correction. When used in high-stakes certification examinations and the like, the message to the candidate is that a certain (small) number of (perhaps less serious) errors are allowed, but, really, none are allowed. When used as translation feedback to students or even professionals, points-off systems, even when feedback is provided along with the points deducted, fundamentally emphasize failure and not what the individual did right. The correct is assumed; the incorrect is pointed out.

At least human translation and interpreting, then, tends to be subject to this implied standard. On the other hand, MT is sometimes subject to a very different standard, that of "just good enough to be understood". As noted in Van Rensburg's contribution in this volume, unnecessary changes are considered problematic – not because they affect the quality of the translation, but because they are seen as not increasing the communicative information value of the text and only costing unnecessary time and money. From a theoretical and language philosophy perspective, the question we must ask is whether high-quality, well-formed language is a necessary goal in all translation and interpreting, if one disregards the financial considerations. More fundamentally, though, since raw or lightly post-edited products are being used in the real world, one might ask whether this will have an influence – good, bad or indifferent – on such things as language users' lexicon or syntax.

Logically, then, the "gold standard" implies a derivation relating to revision: if one assumes that the "perfect translation" is the most desirable outcome, then if the original human translator did not produce it, a reviser would have to make all necessary changes, including those relating to text type, genre or style, not make any unnecessary changes, not overlook any errors and not introduce any errors – in other words, the translation would again be perfect after revision. Thus essentially any measurement of human revision quality would be measuring deviations from that standard. For MT, though, the cost factor comes into play again: "Good enough to use" seems to be the standard. Sometimes, raw MT is considered acceptable, even though it suffers from flaws in language mechanics and stylistic infelicities. If it is not acceptable, then light post-editing may be applied to make it "good enough", where the post-editor would meet the standard by making only necessary changes in meaning (but not including stylistic errors), not make any unnecessary changes, not overlook any errors and not introduce any errors. This, then, is a separate standard for light post-editing. Full post-editing, on the other hand, would seem to constitute revising to the same standard of quality as human revision.

The dichotomy here is between acceptable content, which is often defined as fulfilling a specified purpose, and high-quality, well-formed language. While cost factors may cause translation buyers to prefer MT output over human translation, in a larger societal sense using MT output, with all its flaws in style and usage (even when it is basically accurate in meaning), may lead to a degradation of standard language in the population as a whole. Oddly, companies have now persuaded users to modify their judgements of acceptability to the point that user expectations of, for example, help text quality are not very high; that is, users are tolerant of flawed MT content. Indeed, if attitudes were less pragmatic, the content might never be translated at all. So end-users will tolerate flawed

language in order to obtain critical information, but from the larger perspective of linguistic change and the philosophy of language, it remains to be seen whether MT output, which frequently does not conform to traditional linguistic norms, will have an influence on the larger sphere of language used in any given culture, and whether cultures will or will not accept such an influence from a non-human source.

1.2 Approaches to translation assessment

When we consider the issue of assessment in more detail, there is a basic dichotomy between error-based assessment and holistic assessment. An anecdote may illustrate the issue: in an informal experiment conducted a few years ago with a group of about 50 trained ATA graders in multiple language pairs in the context of a grader training session, graders were given clean copies of examinations (source text (ST) and target text) that had already been assessed in the program. They were asked to first read through the translation to obtain a general impression of it, then to use the ATA Rubric (see Angelelli, 2009) to assess the quality of the translation holistically based on their impressions. They were then asked to work through the same text using ATA's Flowchart for Error Point Decisions and assign points using the ATA Framework for Error Marking (see Koby & Champe, 2010). Finally, they compared the results of the holistic versus error-based marking and reported the results verbally to the training group. The consensus of the group was that error-based assessment resulted in more stringent marking, since each individual error was marked before assigning an overall grade.

Although anecdotal evidence does not carry a lot of weight, it still points to a fundamental set of issues in translation assessment – how to assign marks to individual errors, how to summarize those errors into an overall grade or mark, how to set thresholds of acceptability and whether to break down any scores into separate sub-scores for fluency (also known as meaning transfer) and accuracy (also known as language mechanics or target-language writing), among other categories. In this volume, Kivilehto and Salmi state that error-based assessment may be more suitable for high-stakes examinations for licensed translators where accuracy is deemed essential. If full accuracy means zero (or near-zero) errors, then an argument can be made for preferring error-based assessment over holistic assessment.

The question then arises of which scale to use when evaluating errors. Many systems use the dichotomy between major and minor errors; in this volume, Van Rensburg equates major errors to accuracy/meaning/transfer errors and minor errors to errors in fluency/language mechanics/grammar. Some scales also add a level called "critical", which can be variously defined as "endangering life and limb" or "rendering the text useless for the intended purpose".[3] And yet it can happen that an error in meaning transfer is not extremely disruptive to the usability of a text, and therefore could be minor, whereas an error in language mechanics can be quite disruptive to usability without failing to transfer the underlying meaning correctly. Such definitional difficulties show that, here again, consensus has not been reached on how to measure what translation quality researchers are trying to measure. Can the goal, also in professional practice, ever realistically be a single score, if the profiles underlying a single score vary widely? Is it not necessary always to report at least an accuracy score and a fluency score? For instance, in one editor's inspection of graded ATA exams, which report a single score, it is possible, for instance, for three examinees to have the same score on the same examination, yet one examinee's score will be derived primarily from transfer errors (meaning that the examinee did not transfer meaning well), another will come primarily from language errors (meaning that the examinee did not show mastery of target-language writing skills), and a third will show a balance of errors between the two areas. This information is important enough that it should not be obscured in a simplistic quest for a single number representing a complex skill.

Another area that should be examined, particularly for post-editing, is the question of the justification (or not) of "unnecessary" changes made to the text. Research is needed, perhaps by interviewing revisers and post-editors retrospectively, on why they made the changes that the researchers considered unnecessary and had instructed the revisers and post-editors not to make. A working hypothesis might be that among linguists there are divergent mental models for correctness

in target-language writing that are based on specific linguistic and cultural backgrounds. It must be assumed that the reviser or post-editor considers the edits to be necessary, even if the researcher does not. A useful line of research would be to attempt to operationalize the idea of "necessary" versus "unnecessary" edits and investigate the reviser/post-editor mental models for what exactly is or is not necessary, and why. These models may require certain changes that other models may consider unnecessary or merely stylistic.

At the same time, every reviser and post-editor makes many decisions that are not captured directly in most research – namely the decisions *not* to edit sections where the translator or MT has produced *correct* output. This is an additional possible area to be measured in research: acceptable translated text not changed. It is the converse of errors that have been overlooked and is the preferred outcome, namely that the reviser correctly decided that the text did not need to be changed. This is a key skill in revising or post-editing that should be explored in future research. The category is implied necessarily by the act of the researcher in finding overlooked errors, since the researcher must necessarily also review the text to ascertain the difference between acceptable translated text that was not changed and errors that were overlooked.

2. Contributions in this volume

The contributions in this volume are grouped together by topic in order to allow the themes to flow from one to the next within the larger framework of translation and interpreting quality assessment. The articles naturally break down into two groups by their focus on either written translation or oral forms of language mediation. However, the dichotomy is not complete, with some articles dealing with areas that overlap both forms. Within written translation, the first four articles focus primarily on the assessment of human translation; they are succeeded by a contribution on revision assessment that straddles the boundary between human translation and MT; this is followed in turn by four articles focusing on MT-PE. There are then three contributions on oral or hybrid forms, two focusing on interlingual live subtitling followed by a final chapter on the assessment of interpreting.

2.1 Translation Assessment

Rachel Brooks and **Maria Brau** of the US Federal Bureau of Investigation (FBI) report on three studies conducted at that government agency which establish "congruity judgment" as a third skill separate from reading comprehension in the source language and the ability to write in the target language (TL). This third skill, which could also be called "translation proficiency", had not previously been sufficiently investigated to clearly establish its existence separate from the other two skills. The FBI developed translation tests for its linguists and conducted three studies to justify testing for translation performance separate from reading comprehension. Two of the studies were previously published and are summarized here as they lay the fundamental groundwork for the third study described. The first study indicated that high reading and writing abilities do not predict translation ability, at least not in Arabic. The second study showed that, for the Vietnamese, Italian and Turkish subjects, writing ability was either the same as or higher than translation ability, with almost a third of the cases showing writing ability to be more than one level higher than translation ability. Self-assessment proved to be a better predictor of translation ability than English writing did. Finally, the study reported on here expanded the analysis to over 7,500 linguists in 15 languages paired with English.

The average Interagency Language Roundtable (ILR) performance rating on the FBI translation test across all 15 languages was low, whereas writing proficiency was rated significantly higher. Regression analysis showed a very strong positive correlation between reading, writing and translation, but also clearly showed that measuring reading comprehension or English writing ability alone does not reliably predict the ability to translate. The research led to the development of a formula that was intended to predict translation ability from reading and writing scores. Although there were initial indications of a good fit between the formula and the real world, it ultimately did not predict

skills accurately, being off by an entire ILR level or more in many cases, and so the formula was considered unacceptable. This result confirms the third competence as something different from linguistic skill that must be researched and measured separately.

From a different perspective, **Gert de Sutter**, **Bert Cappelle**, **Orphée De Clercq**, **Rudy Loock** and **Koen Plevoets** argue that a corpus-based statistical approach to translation evaluation allows for systematic and objective formative assessments that are especially valuable when used in conjunction with more traditional qualitative assessments. They focus primarily on measuring the acceptability of the TL in a translation as opposed to measuring the adequacy of the transfer between languages. To do this, they create multiple corpora – student translation corpora, professional translation corpora and reference corpora of texts by professional writers in a non-translation setting. Separate corpora sets were created for two case studies: an English-to-French translation of fictional texts and a French-to-Dutch translation of news texts. The authors pre-processed the corpora linguistically through tokenization, part-of-speech tagging and syllabification prior to extracting 25 language- and genre-independent features. These features included a variety of frequency information and measures of lexical creativity and originality.

For the English-to-French corpora, the authors also extracted five language-dependent features. Principal Component Analysis was used to analyse the data, and it showed a reasonably clear discrimination between separate clusters of scores from student and professional translations, with the scores from the non-translated texts located approximately on the border of the two translation score clusters. Accordingly, there appear to be acceptability issues for student translations. The analysis also provided information about individual student translations and identified a relatively small number of specific features that significantly discriminate between student and professional translations. Principal Component Analysis of the French-to-Dutch corpora showed a sharper demarcation between student and professional translations, but no noticeable difference between professional translations and original language writing. The clearer student–professional demarcation is perhaps attributable to the much larger proportion of features yielding significant differences between student and professional translations.

Taken together, the two case studies report very promising outcomes. They indicate the usefulness of automatically generated corpus-based acceptability assessments of student translations. Future research is still needed, but the authors propose that the combination of this methodology with carefully considered subjective constructive feedback has the potential to become a powerful pedagogical tool in formative assessment.

In contrast, **June Eyckmans** and **Philippe Anckaert** suggest that the search for objective criteria in summative assessments of translation quality is futile and misguided. They argue instead that a subjective assessment process should be "embraced and the measurement error that comes with it [should be] calculated, expressed and controlled by means of a reliability coefficient". The authors discuss the distinction between measuring the reliability of assessment (how consistently an assessment method rates translation quality across different conditions) and the validity of assessment (how well an assessment method measures the competence of the translator). They choose to focus on reliability, comparing two norm-referenced methods for translation assessment, namely the Calibration of Dichotomous Items (CDI) method and the Preselected Items Evaluation (PIE) method. Norm-referenced methods aim to measure differences in translation competence across individuals, whereas the more common criterion-referenced methods measure how well each individual meets set reference standards.

When applying the CDI method, a complete ST is translated by a large and representative sample of novice translators. Expert graders analyse the translations of each segment to identify those segments where there are large differences in translation quality across different individuals. For each of these identified segments, the raters then decide which translations are acceptable and which are unacceptable, but they do not attempt to rate the quality of the translations in a more refined way.

When applying the PIE method, expert graders select ten segments from the ST and decide in advance on correct and incorrect translations of these segments. A small group of novice translators then translates these pre-selected segments and the graders analyse the translations to make a second selection of segments that are of medium difficulty and demonstrate large differences in translation quality across different individuals.

To compare the CDI and PIE methods, the authors conducted a study of more than 100 student translators at different stages of their training in bachelor's and master's programmes at four different institutions in Belgium. They analysed the students' performances when they translated a general text from Dutch to French. The reliability of the test scores was high for the hard-to-administer CDI method, but low for the easy-to-administer PIE method. A number of problems were also identified with the PIE methodology: specifically, the short cut attempted in the PIE method is unreliable and so unsatisfactory. In contrast, evidence is presented to support the CDI method as a satisfactory, though effortful, norm-referenced way to obtain summative assessments of translation quality.

Finally in this group, **Marja Kivilehto** and **Leena Salmi** describe the examination system for certifying translators in Finland to translate legally valid texts that carry evidentiary value. In line with the results produced by Brooks and Brau discussed above, the Finnish system was changed from a language skills test to a translation skills test in 2008. Kivilehto and Salmi compare the system used in Finland with similar testing approaches used in Sweden and Norway, as well as the system in the German state of Bavaria. The Finnish system tests knowledge of professional practices and assigns two translations to be taken on computer with Internet access; the Swedish system assigns three translations to be taken on computer, but without Internet access; the Norwegian system gives a take-home examination with an essay and a translation, which must be passed in order to do the on-site examination of three translations with computers and Internet access; and the Bavarian system calls for an essay plus four translation assignments, written by hand. The Scandinavian assessment systems all use error analysis as a method, with 13–15 error types, whereas the Bavarian uses criterion-based analysis.

The authors analyse the various assessment systems and then present a detailed case study on the use of the Finnish examination's scoring chart. The error types that are frequently used differ between the language directions. The errors most commonly marked on into-English translations are terminology errors, followed by structural errors, whereas the most common errors in into-Swedish translations are style errors followed by terminology errors. The authors note that the terminology error category combines incorrect terminology with errors of omission and addition, which accounts for its frequency of occurrence. However, this error category was recognized as problematic and was therefore split into three categories in 2014. The authors intend to improve the Finnish examinations further by proposing a simplified scoring chart.

2.2 Revision assessment

In the area of human translation revision, **Alta van Rensburg** examines how the qualifications and experience of revisers affect the quality of the revision product. Little research has been done on revision competence, but what has been done has shown that revisers (who are often experienced translators) make unnecessary changes or introduce new errors into a revised text. A substantial group of participants (30 revisers, one translator and three language experts, all native-speakers of Afrikaans working with English) participated in this study. Van Rensburg investigates how the ability to avoid making unnecessary changes relates to professional experience in translation and knowledge of translation theory as shown by translation qualifications. The participants revised a translation into Afrikaans. Based on this data, Van Rensburg developed a three-dimensional assessment instrument for revision using four categories (Necessary changes, Unnecessary changes, Errors overlooked and Errors introduced), three dimensions (Translation accuracy, Target language usage and Target text function) and two levels of error (Major and Minor). She goes on to develop a weighted formula for assessing revision quality that factors in necessary changes, errors overlooked and introduced, and unnecessary changes, and which also takes into account the length of the text. For the first two categories, major errors carry double weight and minor errors carry single weight, both positively and negatively, while unnecessary changes are halved. This formula results in a number that refers to the quality of the revision product, although a negative score does not mean that the quality has been lowered.

This contribution raises intriguing issues about the standards for translation, some of which have been discussed above, including the issues of correct translations correctly left alone and what

constitutes an unnecessary change. It also explores areas that overlap between human translation revision and the post-editing of MT, as described in the next section.

2.3. Machine translation and post-editing assessment

In "Translationese and Post-editese: How Comparable is Comparable Quality?", **Joke Daems**, **Orphée De Clercq** and **Lieve Macken** explore the question of whether a unique form of language called "post-editese" is created when post-editors revise MT output. (They note that this also implies the existence of Machine Translationese.) Their experiment asked human beings to read texts and classify them as either human translations or post-edited texts. The human subjects were not capable of correctly distinguishing between the two conditions in this experiment. The authors then attempt to quantify objective differences between human translation and post-edited text in order to build a supervised machine-learning model that can make this distinction computationally. Here, too, the machine-learning analysis failed to provide results that indicate a clear difference between the two conditions. Thus it may be that post-editese does not exist in fully post-edited texts, or at least that it is undetectable in the high-quality texts used in this study.

Since it is known that translationese exists and can be detected computationally (see, for example, Baroni & Bernardini, 2006), the idea that a "post-editese" should also be created in the process of revising MT output is not surprising. After all, translationese is produced by a translator based on the specific features of the ST and does not have the same features as text produced originally in the TL. Post-editese, then, would be expected to be based on the specific features of raw MT output that would influence the revised output. It could be expected to be more apparent in a text that was lightly post-edited rather than in one that was fully post-edited. Ideally, of course, neither translationese nor post-editese would be created in the translation process, but again, if full post-editing were entirely successful, then the resulting product should exhibit features of only translationese and not post-editese.

The next three contributions are case studies relating to MT-PE. In the first, **Hanna Martikainen** examines the sources of distortion in the English-to-French translations of medical abstracts published in Cochrane Systematic Reviews using an error typology and assessment approach. Both certain translation errors and biased translations (particularly of modal expressions) were deemed to affect the accurate interpretation of the translations. Modality is a significant factor in scientific abstracts, where it is used either to hedge or to boost the claims made in them. This is particularly significant in these medical abstracts, where modality is used to limit claims and restrict certainty. Since the expression of modality differs between French and English, it is examined here as a source of translation error, specifically relating to the most frequently used modal verb, the English *may*, and its translations using the French verb *pouvoir*. *Pouvoir* with the indicative mood tends to bias the translation towards more certainty compared to the ST. This study leads to an elaborate error typology structure for lexis, grammar and lexico-grammar, where lexico-grammatical patterns can be a source of distortion in translation.

In the second case study, on the acceptability of MT output, **Sheila Castilho** and **Sharon O'Brien** investigate the quality of raw versus lightly post-edited MT according to how it affects the ultimate users of a translation product. They conducted a usability experiment using instructional content (help articles for a spreadsheet application, published online in those languages) with 63 native-speaker end-users (14 German, 21 Chinese and 28 Japanese), who provided feedback on the usability and acceptability of both raw MT output and lightly post-edited output using the KantanMT framework, along with reported levels of satisfaction. The same data were also reviewed by 18 translation professionals (six for each language) and for the two groups the levels of acceptability were compared using statistical analysis.

The German and Chinese versions were clearly improved by light post-editing, whereas the results were less clear-cut for Japanese. Not surprisingly, the post-edited content was deemed more acceptable than the raw output, with significantly higher scores on adequacy, fluency and sentence structure, although the raw output earned high scores for terminology, spelling and compliance with

country standards – areas that are amenable to automation. The end-users also reported statistically significant higher satisfaction with the post-edited texts.

The final case study also relates to the discussion above about necessary and unnecessary changes, in this case in post-editing. **Maarit Koponen** and **Leena Salmi** present a pilot study that analyses the correctness and necessity of edits performed by five students doing light English-to-Finnish post-editing. The participants were native speakers of Finnish who were studying translation. The authors specifically examined post-editing actions rather than changes to MT errors in order to account for preferential edits. About 60% of the words were left unedited, but 3% of these words represented necessary corrections that were not made. Of the edits made, 91% were deemed correct, but 38% of those correct edits were deemed unnecessary – specifically word-order changes and deletions of personal pronouns – meaning that the changes did not render the segment more comprehensible, more grammatically correct or more accurate. Nine per cent of the post-edits introduced an error. About 25% of all the edits performed were changes to word forms relating to Finnish morphology. Most word-form edits, insertions and word substitutions were deemed necessary, while most word-order changes and deletions were deemed unnecessary.

2.4 Interlingual live subtitling

The next two chapters focus on quality assessment for interlingual live subtitling. Subtitles are produced when a "respeaker" listens to an original soundtrack and interprets it to produce spoken forms of proposed subtitles. Speech-to-text software then generates draft subtitles that are edited to correct errors of form or content. A delay is normally incorporated in order to produce the high level of accuracy that is typically required. To minimize the delay, a quality-enhancing team of monitors/editors (each with a specific role that focuses on a single stage of the complex process) may join the respeaker to produce the broadcast version.

The starting point for both chapters is the influential NER model, originally introduced by Pablo Romero-Fresco in 2011 as a tool for assessing the quality of intralingual live subtitling (real-time captioning). In this monolingual model, edition errors are those that result from flawed production decisions, whereas recognition errors are those that stem from a faulty understanding of the source. Errors are classified as minor, standard or serious and weights are assigned accordingly – 0.25 for a minor error, 0.5 for a standard error and 1 for a serious error – and these are added up to produce an edition error score (E) and a recognition error score (R). "Correct editions" that do not detract from the information conveyed in the source are not scored. For a subtitle with N words, the NER accuracy rate is (N-E-R)/N, expressed as a percentage, where N is usually seven words in an "idea unit". The accuracy rate is an important component of quality assessment, but is usually complemented by other forms of judgement.

Pablo Romero-Fresco and **Franz Pöchhacker** provide an overview of quality assessment in real-time captioning and carefully describe the NER model. They then go on to describe the challenges inherent in devising a quality assessment model for interlingual live subtitling before proposing the NTR accuracy rate. This is computed in the same way as the NER accuracy rate, except that a translation error score (T) derived from what the respeaker utters replaces the edition error score (E) and the recognition error score (R) is now based on the shortcomings of the speech recognition software. Raw NTR accuracy rates from 96% to 100% are recalculated on a 10-point scale that is robust with respect to subjective differences between evaluators of the complex interlingual task. Translation errors are classified as either content errors or form errors. Careful attention is given to the description and motivation of a three-level scale (0.25, 0.5 or 1) to weight translation and recognition errors; it is pointed out that the highest level is not appropriate for translation errors of form. As in the NER model, harmless errors are not scored. The authors provide an informative example to illustrate the computation of the NTR accuracy rate and how it is combined with other factors to provide an overall quality assessment. They conclude with a discussion of how the NTR model might be developed and refined in varying environments.

Isabelle Robert and **Aline Remael** take a different approach to modifying the NER model to the context of interlingual live subtitling. After a review of the NER model and the principal quality

parameters of content and form used to assess simultaneous interpreting, they proceed to describe typologies and weighting factors for recognition errors and edition/translation errors. They do not formally separate edition and translation errors, in part because they conduct a case study of one-minute delayed Dutch subtitling for an English-language source in a commercial broadcast setting. In this setting, respeakers submit their spoken subtitling proposals to speech-to-text subtitling software and carry out initial post-editing. The output is then post-edited again by a corrector (focusing on recognition errors and edition errors not related to the interlingual transfer) together with a speech-to-text interpreter (focusing on edition errors related to the interlingual transfer) and then a third time by a final broadcaster. Based on their observations of actual broadcasts, the authors use a single weight (0.5) for recognition errors, but a four level scale (0.25, 0.5, 0.75, or 1) for edition/translation errors. They carefully analyse error types and their correction, observing greater correction success for recognition errors than for edition errors. Respeaking quality, which is influenced by time pressure, significantly affects the quality of the final product, but improvement is evident at all stages. NER accuracy rates consistently and often significantly improve across all the stages of the subtitling process. However, the greatest improvement takes place when the corrector intervenes; the broadcaster's contributions are only minor. These results provide very useful information that will assist in the refinement of tools to assess the quality of interlingual live subtitling and will inform the development of best practices for carrying out this challenging task.

2.5 Interpreting assessment

The final chapter tackles the assessment of the quality of interpreting. This is often evaluated using "atomist" methods that focus primarily on analyses of lexical and semantic errors and pauses in production, but which are largely unable to address more holistic concerns, for example at the discourse level. An alternative approach is to use descriptor-based analytic rating scales, which are well adapted to more global analysis. The use of rating scales is gaining in influence and much effort has been invested in the theory-based design of assessment tools, but so far only a limited amount of direct empirical evidence has been gathered on the utility of particular scales or the reliability of judgments based on them in the assessment of interpreting. **Chao Han** extends and refines our understanding of such issues through a carefully designed longitudinal study of three scales using MFRM (Multi-Faceted Rasch Measurement), an innovative methodology in this context. The scales measure information completeness (InfoCom), fluency of delivery (FluDel), and TL quality (TLQual) for interpreting in both directions between English and Chinese. Han investigates the difficulty and appropriateness of the scales for users and the stringency and consistency of trained raters.

The scale users were 38 university students of English–Chinese consecutive interpreting (CI), all of them L1 Chinese–L2 English. The raters were six university interpreting teachers or assistants, all L1 Chinese–L2 English. All the students were assessed three times over a period of ten weeks, in weeks 4, 9, and 10, and all the assessments were rated by each of the six raters. The data produced were analysed using MFRM with four assessment facets, namely students, raters, tasks and scales. In terms of difficulty, differences between scales were relatively small and were consistent over time, but the results were direction dependent in predictable ways. For English-to-Chinese CI, InfoCom was the most difficult and TLQual was the easiest scale to rate, whereas for Chinese-to-English, FluDel was the most difficult and InfoCom was the easiest. All three scales mostly functioned appropriately over time, but the few minor exceptions point to the need for further refinement. With one exception, rater leniency/severity was a relatively stable trait across both time and direction. However, because the rater sample size is small, these results must be considered preliminary. Future confirmation of the promising results of this innovative research on the utility and reliability of the rating-scale approach to the assessment of interpreting and further demonstrations of the power of the MFRM methodology would be of great value to the field.

2.6 Other areas and future research directions

Of course, no scholarly volume can cover the entire range of a subject as complex as translation or interpreting quality evaluation and assessment, and some subareas will inevitably be more represented than others. For instance, this volume contains more contributions that focus on translation quality than on interpreting quality. In the translation area, some topics that were not covered include domain-specific investigations in areas such as medical or legal translation. These are significant gaps, since, after all, industry practice suggests that translation competence must include both general translation skills – the basic ability to transfer ideas across languages – and specialized knowledge and skills in a relevant domain. On the interpreting side, several areas were not covered in this volume: sight translation, consecutive or simultaneous interpreting, or note-taking. Finally, sign language interpreting was also not covered, although that tends to be quite a specialized domain of language mediation.

Other possible areas of investigation not represented in this volume would include the quality of crowdsourced translations as well as crowdsourced assessment, dynamic quality assurance models and an expanded focus on translation pedagogy, including, for example, learning portfolios. Terminology was not covered here, nor were CAT tools and other technological solutions. In the area of what the US FBI calls "congruity judgment", moreover, additional research into the utility and accuracy of scoring models such as the ATA or MQM systems would be useful. Finally, tying in to the contribution on the Finnish certification system, a comprehensive comparative study on translator certification systems around the world would be welcome, possibly as a first step towards harmonizing certification across national boundaries.

Beyond the issue of certification, it is important to find effective ways to consider the consistency of evaluations of both translators and translations, and interpreters and interpreting. In some areas outside translation and interpreting, it is recognized that multiple quality assessments carried out at different times and in different ways are necessary to obtain reliable and accurate information. This is readily achievable in the assessment of MT through a comparison of automatic metrics, such as BLEU, METEOR and TER, for multiple texts. However, little seems to have been attempted in this direction for human translation or interpreting.

Another area for future research, as noted above under section 1.1, could be whether MT output in any given language would be sufficiently pervasive or influential to be able to have any effect on that language's everyday use. We speculate that such influence might be found primarily in syntax, as MT does not invent new words but does produce output that displays non-typical word orders. It may also affect collocations. Corpus analysis would be a likely approach to identify such changes if they exist – by identifying specific syntactic or collocational phenomena originating in MT corpora and then comparing them to corpora of original human language produced prior to the advent of MT and after it became widespread. Different types of influence of MT on human writing are also possible and would be interesting to investigate. For example, when texts are submitted for MT, in particular in business contexts, writers are often encouraged to use "controlled language" that is likely to be more accurately translated by machine. The question arises whether writing in controlled language for specific translation purposes will end up affecting writing more generally.

3. Conclusion

Quality assessment for translation, interpreting and related activities remains a thorny problem. This is partly due to philosophical issues concerning the very nature of quality. It is also a consequence of the complexity of human language and language mediation, along with the diversity of the tasks being considered and of the situations in which they are undertaken.

The contributions in this volume attest to this diversity, to the increasingly broad range of creative approaches to quality assessment, and to the growing sophistication and rigour of research in the field. While many of the fundamental questions and dilemmas remain unresolved, these contributions develop momentum in the field, providing a solid foundation for future work, clarifying

previous research, identifying promising lines of new research and suggesting useful new methodologies.

References

Angelelli, C. V., & Jacobson, H. E. (2009). Introduction. Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 1–10). Amsterdam: John Benjamins.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing, 21*(3), 259–274.

Carey, E., & Jumpelt, R. W. (Eds.). (1963). *Quality in translation: Proceedings of the IIIrd Congress of the International Federation of Translators (FIT),* Bad Godesberg, 1959. Oxford: Pergamon.

Gaddis Rose, M. (Ed.). (1987). *Translation excellence: Assessment, achievement, maintenance.* American Translators Association Scholarly Monograph Series 1. Binghamton, NY: State University of New York.

Koby, G. S., & Champe, G. G. (2013). Welcome to the real world: Professional-level translator certification. *Translation & Interpreting, 5*(1), 156–173.

Koby, G. S., Fields, P., Hague, D., Lommel, A., & Melby, A. (2014). Defining translation quality. *Revista Tradumàtica: tecnologies de la traducció* [Online]*, 12*, 413–420. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p413.pdf [Accessed 14 November 2017].

Luther, M. (1530). *Ein Sendbrieff, von Dolmetschen, vnd Fürbitte der Heiligen*. Wittenberg: Georg Rhau.

Luther, M. (2017). *Ein Sendbrief vom Dolmetschen / An Open Letter on Translating.* Translated by Howard Jones Treasures of the Taylorian: Reformation Pamphlets. Taylor Institution Library, Oxford.

Reiss, K. (1971). *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*. Munich: Hueber.

Reiss, K. (2000). *Translation criticism, the potentials and limitations: Categories and criteria for translation quality assessment*. Translated by E. F. Rhodes. Manchester, UK: St. Jerome.

Schwartz, L. (2014). *Monolingual post-editing by a domain expert is highly effective for translation triage.* Proceedings of the Third Workshop on Post-Editing Technology and Practice, pp. 34–44.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231.

Williams, M. (2004). *Translation quality assessment: An argumentation-centred approach*. Ottawa, ON: University of Ottawa Press.

---

1 This also raises the question of whether revising and post-editing are fundamentally different activities, or fundamentally the same activity performed on texts of different types.

2 Owing to the immense flexibility of human language, there could be several "gold standard" translations of a single source text; this is recognized and operationalized in TER ratings of MT (see Snover, Dorr, Schwartz, Micciulla and Makhoul, 2006).

3 The argumentation-centred approach espoused by Williams (2004) takes a similar tack by setting up a grading scale in which the failure to render the argument correctly in general is subject to a drastic penalty such that the translation is considered a failure if it does not meet this single criterion.