JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Angelone, E. (2020). The impact of screen recording as a diagnostic process protocol on inter-rater consistency in translation assessment. *The Journal of Specialised Translation, 34*, 32-50. https://doi.org/10.26034/cm.jostrans.2020.136

The impact of screen recording as a diagnostic process protocol on inter-rater consistency in translation assessment

Erik Angelone, Kent State University

ABSTRACT

In recent years, process-oriented translator research and training have entered the arena of translation assessment, with a focus on how end products can be interpreted from the perspective of translator decision-making and behaviours, as documented in the form of various process protocols. Screen-recording is frequently integrated as a preferred empirical method in the context of such research, thanks to its relative ease of use and the saliency of the process phenomena it documents. In extending on preliminary findings that have highlighted the efficacy of screen recording as a means for enhancing consistency in process assessment, this paper reports on a small-scale pilot study in which assessors marked up translation errors according to type and severity while using screen recording as a diagnostic protocol to guide the process. Results obtained from the study suggest that inter-rater consistency is enhanced when assessors of Spanish-English and Arabic-English translation make use of screen recording in this capacity as compared to when then they do not make use of a process protocol. Greater inter-rater consistency was evidenced in assessment for both language pairs, and in terms of both error type and severity point classification.

KEYWORDS

Process-oriented translation assessment, inter-rater consistency, screen recording, error typology, translation pedagogy.

1. Introduction

From the time of its inception some thirty years ago, translation process research (TPR) has often been undertaken with an eye towards optimising translation pedagogy in some capacity (see Krings 1986, Kiraly 1995). For example, Gile's Integrated Problem and Decision Reporting (IPDR) logs (2004), still widely used in many variants today, can be regarded as the first formal pedagogical approach to have students document, discuss and reflect on the problems they encounter and their corresponding problem-solving strategies. In addition to enhancing learner problem awareness, they also provide instructors with a concrete platform for assessing documented problem-solving. Hansen's sources of disturbance (SDs) concept (2008), as embedded in keystroke logging data, relies less on student reporting, and instead documents problems actually encountered (along with subsequent problem-solving) in situ and in real time. These are indicated in the keystroke log data in the form of such phenomena as extended pauses in activity, revisions, and cursor re-positioning, among others. The approach to process-oriented translation assessment Hansen puts forward can be regarded as advantageous in that it does not require the student to break away from the task at hand for purposes of documenting problems, thereby enabling a continuation of uninterrupted, natural translation workflow. Furthermore, this approach provides instructors with documentation of actual problems experienced by students rather than student perceptions of the problems they encountered alone. These actual problems may otherwise run the risk of potentially going unnoticed if the window into translation process assessment is based solely on student reporting. In the context of experimental translation process research, keystroke logging is still widely regarded as a method of choice, either on its own or in conjunction with other methods in a triangulated fashion, providing highly granular temporal data (often measured in milliseconds) on nuanced aspects of cognitive processing. Students and instructors, on the other hand, may find the visual rendition of keystroke log data to be overly complex from a pedagogical standpoint.

In a vein similar to Hansen's sources of disturbance model, PACTE puts forward its 'rich point' model (2009) as a point of departure for documenting and assessing learner problem-solving. Unlike SDs, which reflect problems as they occur during the course of a translation task, rich points are "pre-established prototypical translation problems" (Castillo 2015: 76) embedded in the source text. In other words, rich points are predicted to generate translation problems based on previous performance of translators with a similar competence profile as those involved in the given task. PACTE uses rich points for purposes of translation assessment (2009), where assessors hone in on what kinds of processes unfold in their presence. Instructors of translation practice courses likely have rich points in mind whenever they select texts for students to translate. These could come in the form of specific lexicogrammatical properties, genre- and locale-specific conventions, and the need to move away from literal renditions through the use of various translation strategies.

While PACTE's rich point model holds strong pedagogical value, particularly insofar as text selection for larger enrolment translation practice courses is concerned, and as a way of predicting difficulty based on concrete empirical data, there is a potential mismatch between predicted problems and actual problems based on a number of factors. One of these is the largely heterogenous nature of student profiles and competences. An over-reliance on rich points as a lens for assessing translation processes may result in a missed opportunity to assess unique, non-prototypical problems encountered at the level of each individual student (Angelone 2018). Ultimately, rich points potentially shed light on only part of the picture when used by students and instructors to engage in translation process assessment.

Think-aloud protocols (TAPs) also have a long tradition as a method for assessing translation processes (see Krings 1986, Jääskeläinen 2002).

Students are asked to articulate all of their thought processes in real-time as the task unfolds. Problems and problem-solving are manifest through direct and indirect statements, including repetition of problematic passages, as well as filler words such as ‘hm’ and ‘uh’ and extended periods of silence. While TAPs continue to be used for process assessment, their ecological validity has been called into question. In terms of parallel processing, it can be quite difficult for students to simultaneously translate and articulate their thoughts in a sustained fashion over an extended period of time. This places restrictions on how long the translation tasks in which TAPs are used can be. Additionally, students may be inclined to articulate what they think their instructors want to hear, rather than their actual thoughts.

More recently, eye-tracking has been introduced as a method for assessing the translation processes of students (Pietrzak and Kornacki 2018). Visual attention data, in the form of gaze plots and saccade patterns, can reveal what the student looks at, for how long, and in what sequence. With the advent of portable, more affordable eye-tracking devices, it is expected that the use of eye-tracking in process-oriented translator training will become more widespread in the coming years. At the moment, due to highly complex data metrics, and a necessity to maintain constant eye contact with the screen to obtain data (something many translators are not likely inclined to do), eye-tracking can be regarded as technology that is still better suited for the research arena than the pedagogical arena.

Of all of the methods currently used in translation process research, screen recording has emerged as one that is highly suitable for pedagogical application. This is an application that records all activity that transpires on screen over the course of task completion. Students and instructors can then engage in retrospective analysis of translation processes during video playback. Its preferred status as a pedagogical tool can be attributed to a number of inherent advantages, including preserved ecological validity, heightened saliency and ease of data interpretation (Angelone 2015, Shreve *et al*. 2014), and the fact that it is free. The creation of screen recording videos does not require translators to do anything they would otherwise not be doing during the natural course of translation task completion. There is no need to pause and enter content, nor a need to articulate thoughts or make sure one’s gaze does not drift from the screen. Furthermore, translation students and trainers do not need extensive training on its use, and the learning curve is thereby kept to a minimum.

2. Screen Recording as a Promising Method in Translation Process Assessment

To date, screen recording has been utilised by trainers and trainees in pedagogical contexts as a means for retrospectively identifying and classifying problems encountered, according to such attributes as textual level, locus (such as comprehension/transfer/production) (Angelone 2014), phase (orienting/drafting/revision) (Yamada 2009), and information retrieval type (internal or external). Classification is facilitated through the presence of highly salient, directly observable problem indicators, including the location of extended pauses, information retrieval patterns, revisions, and general workflow routines (Pym 2009, Angelone 2019). Screen recording has also been used as a diagnostic tool for students to engage in self-revision and other-revision (Shreve *et al.* 2014). In the context of these studies, it was found that screen recording was a more efficacious process protocol than traditional translation logs for purposes of detecting and mitigating errors in one's own translation as well as in those of others.

The 2014 Shreve *et al.* exploratory study can be viewed as an initial attempt to utilise screen recording as a vehicle for process-oriented translation assessment (in this case, as undertaken by peers), and was one of the first of its kind to do so. More recently, Massey and Ehrensberger-Dow (2014) and Ehrensberger-Dow and Massey (2013) present evidence that certain process measures captured in screen recording correlate with translation quality and, thereby, serve as predictors of translator performance in the aggregate as rendered in a given translation task. Angelone (2019) found preliminary evidence of enhanced inter-rater consistency in an empirical study where undergraduate and graduate students of translation used screen recording as a diagnostic tool for reverse engineering errors in a translation product. This consistency was manifest in error classifications according to linguistic level (grammar, lexis, syntax, style, mistranslation), phase (drafting or revision), and locus (comprehension, transfer, or production). Motivated by this evidence of inter-rater consistency, a follow-up study, to be described in this paper, was undertaken to gauge the potential of screen recording as a diagnostic tool for enhancing such consistency when multiple graders assess the same translation product.

It goes without saying the translation assessment is very much a subjective matter, and that notions of quality in a broader sense are very much a can of worms. Nonetheless, as mentioned above, formal assessment of quality and the assignment of concrete scores in the contexts of pedagogy and entrance or certification exams are, for all intents and purposes, a necessity. The subjectivity associated with translation quality assessment notoriously results in a lack of inter-rater consistency when multiple assessors are given the task of marking up the same translation. In an attempt to mitigate inter-rater inconsistencies, the American Translators Association has certification exam sub-committees that go through a lengthy, complex process of selecting a small sub-set of texts to be used for the exam, producing model translations thereof, and then attempting to predict all possible errors that might occur. Finally, each of these anticipated errors is assigned a type and severity point total, as decided on by a translation exam sub-committee for each language pair in which the exam is offered.

Once an exam is completed, two separate assessors independently mark up the translation using a standardised set of error codes and severity point values. If there is a lack of consensus, the two assessors consult with each other in the hope of reaching one. Though not documented in the literature, it is not uncommon for the ATA certification examination to yield inter-rater inconsistencies, despite all of these measures taken to mitigate them. ATA certification exam assessment is based solely on what appears in the final translation product. There is no corresponding diagnostic protocol that documents the processes that went into the creation of this product. If screen recording-based translation assessment does, indeed, hold promise as a means for enhancing inter-rater consistency, its implementation in such a context might be warranted and beneficial. The same holds true for entrance or exit exams for translation programmes where multiple assessors are involved and inter-rater consistency is advantageous.

The small-scale pilot study on which this paper will now report was undertaken in conjunction with the following research questions:

1. Does the utilisation of screen recordings as a diagnostic protocol for purposes of assessing translations result in greater inter-rater consistency as opposed to when assessment is based on translation products alone?

2. If inter-rater consistency is enhanced through the utilisation of screen recordings, how is this manifest at the levels of error type and severity point assignment?

3. Methods

3.1 Participants

The aforementioned research questions were analysed in conjunction with translation tasks and assessment involving the Spanish-English and Arabic-English language pairs. The Spanish-English translations were created by two students enrolled in the BS in Translation programme at Kent State University. Both students were taking an advanced translation practice course at the time of the study and were translating from Spanish into English as their L_1. The Arabic-English translations were created by two students enrolled in the MA in Translation programme at Kent State University. Both students had completed three translation practice courses at the time of the study and were translating from Arabic into English as their L_2.

The Spanish-English translations were assessed by three current doctoral students in Kent State University's PhD in Translation Studies programme. All three have experience teaching translation courses at the university level, and all three have English as their L_1. The Arabic-English translations were assessed by two current doctoral students and one Assistant Professor of Translation at Kent State University. All three have experience teaching translation courses at the university level, and all three have Arabic as their L_1 and English as their L_2. All assessors taking part in the study have at least five years of professional translation experience.

3.2 Materials and Procedures

Following Kent State University Institution Review Board guidelines, all translations were produced on the researcher's computer. While translating, students had access to any and all resources of their choosing and their translations were not timed. Each student translated two source texts of approximately seventy-five words into English. The texts were of a general language nature. For one of the two translations, a corresponding screen recording was created using QuickTime. The researcher started the recording and stopped it upon task completion. All translations were created using Microsoft Word. The completed translations and screen recordings were saved on the researcher's computer in preparation for the follow-up assessment phase of the study.

For the assessment phase, three assessors marked up errors in the Spanish-English translations and three marked up errors in the Arabic-English translations. For one of the two translations each assessed, they made use of a screen recording as a diagnostic protocol for marking up errors, and for the second of the two translations, they marked up errors based on the product alone. The assessors watched screen recordings on the researcher's computer and marked up all translations in hard-copy format. They were all familiar with QuickTime as an application for creating and watching screen recordings and were provided with a brief tutorial on doing so at the outset of their assessment session.

At least three days in advance of their participation in the study, the assessors were provided with a standardised error typology (see Appendix A), which is a streamlined version of the error framework used by the American

Translators Association in the context of its certification examination. Each error type contains a code consisting of one to three letters as well as an operational definition. The assessors were also provided with a standardised flowchart (see Appendix B) consisting of a series of Yes/No questions to guide the allocation of a severity point value of 0 to 16 points for each error. They were given the instructions to mark up each error they found in the translations they were assessing according to type and severity point value. The assessors were given an opportunity to carefully read through both documents and to ask any questions they might have before starting.

Upon completion of all assessments, error data were classified according to the criteria outlined in Table 1:

Criterion	**Description**
Divergence in total error mark-up	Comparison of divergence in total error mark-up; comparisons made a) across two assessors, and b) across three assessors
Divergence in total severity point mark-up	Comparison of divergence in the total number of severity points marked up in the aggregate; comparisons made a) across two assessors, and b) across three assessors
Frequency of overlap in error type classification	Comparison of the frequency at which assessors detected the same error and classified it according to the same type; comparisons made a) across two assessors, and b) across three assessors
Frequency of overlap in severity point classification	Comparison of the frequency at which assessors detected the same error and classified it according to the same severity point value; comparisons made a) across two assessors, and b) across three assessors
Frequency of overlap in both error type and severity point classification	Comparison of the frequency at which assessors detected the same error and classified it according to both the same type and severity point value; comparisons made a) across two assessors, and b) across three assessors
Frequency of divergence in error detection	Comparison of how often an error was detected by one of the three assessors, but by neither of the other two

Table 1. Error criteria and descriptions

4. Results and Discussion

This section will report findings in line with the criteria outlined in Table 1, along with corresponding interpretations.

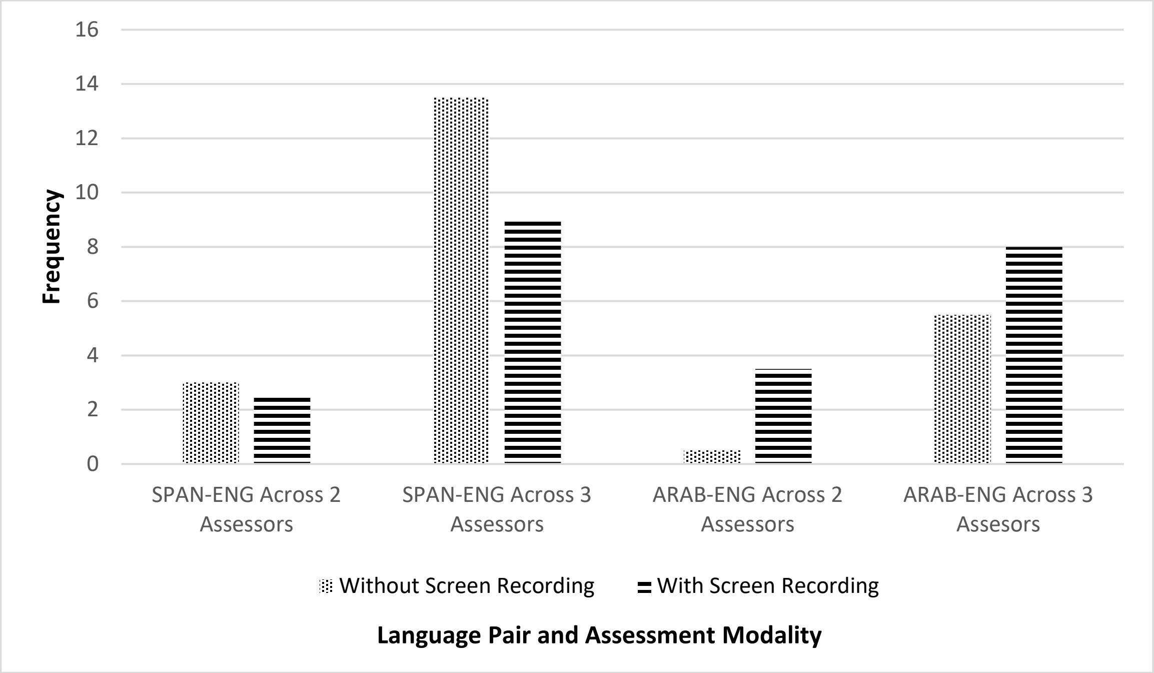


Figure 1. Divergence in total error mark-up

As Figure 1 indicates, in terms of divergence in total error mark-up, at first glance it would seem that assessment with screen recording as a diagnostic protocol resulted in higher inter-rater consistency for the Spanish-English language pair, and lower inter-rater consistency for the Arabic-English language pair. When two assessors were involved in Spanish-English translation assessment, the utilisation of screen recording did not make much of a difference, yet, for this same language pair, we see the greatest divergence between screen recording-based and non-screen recording-based assessment when three assessors were involved. The lowest divergence found in this study occurred across two assessors of Arabic-English translation when screen recording was not used. In the aggregate, the results obtained at the level of divergence in total error mark-up are largely inconclusive.

What is interesting to note in this context, and what becomes clearer in subsequent reporting on inter-rater consistency along the lines of error type and severity point classification, is the fact that minimal error divergence in the aggregate is potentially misleading as a potential indicator of inter-rater consistency. As it turns out, in instances where error mark-up divergence 'in the aggregate' was minimal in terms of frequency across assessors, they were often marking up entirely 'different' errors, resulting in pseudo inter-rater inconsistency.

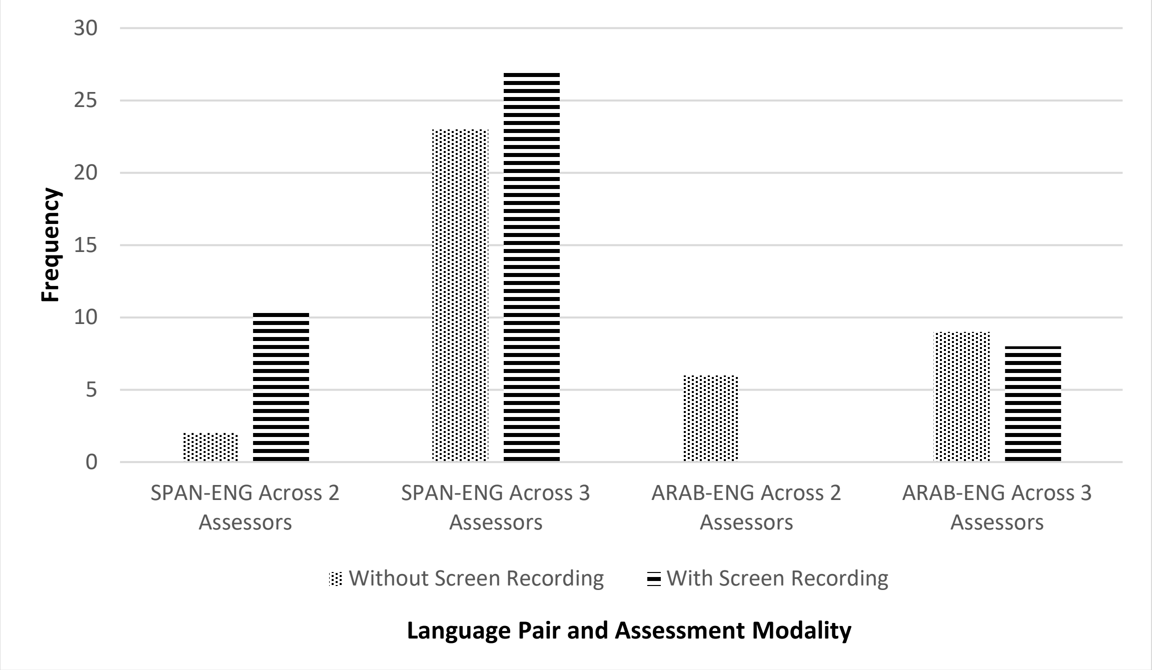


Figure 2. Divergence in total severity point mark-up

Figure 2, which renders divergence in total severity point mark-up, suggests that inter-rater consistency in assessment at this level is stronger when screen recording is used as a diagnostic protocol in Arabic-English assessment and when not used in Spanish-English assessment. As was the case with total error mark-up, aggregate results obtained for total severity point mark-up are inconclusive. Worth noting in these data is the fact that there was no divergence in total severity point allocations when two Arabic-English translation assessors utilised screen recording. This is likely to be an idiosyncratic finding, however. At the same time, we see the greatest divergence in Spanish-English translation assessment across two assessors, where the use of screen recording resulted in significantly higher divergence in relation to when it was not used.

Upon closer examination, we see the same pseudo inter-rater consistency at the level of divergence in total severity point mark-up that we see at the level of divergence in total error mark-up. While severity point values reach similar frequencies in the aggregate across assessors, thereby suggesting inter-rater consistency, the points assigned to each error vary considerably from one assessor to the next, as do the errors detected themselves. If only such aggregated totals across assessors are taken into consideration in the context of certification or entry/exit exams, we might begin to wonder just how consistent assessors' ratings truly are in relation to one another. To more accurately gauge inter-rater consistency, it becomes crucial to transcend raw frequencies in the aggregate, and, instead, look for instances of overlap among assessors at a more granular level, in terms of type and severity point classification when the 'same' errors are detected.

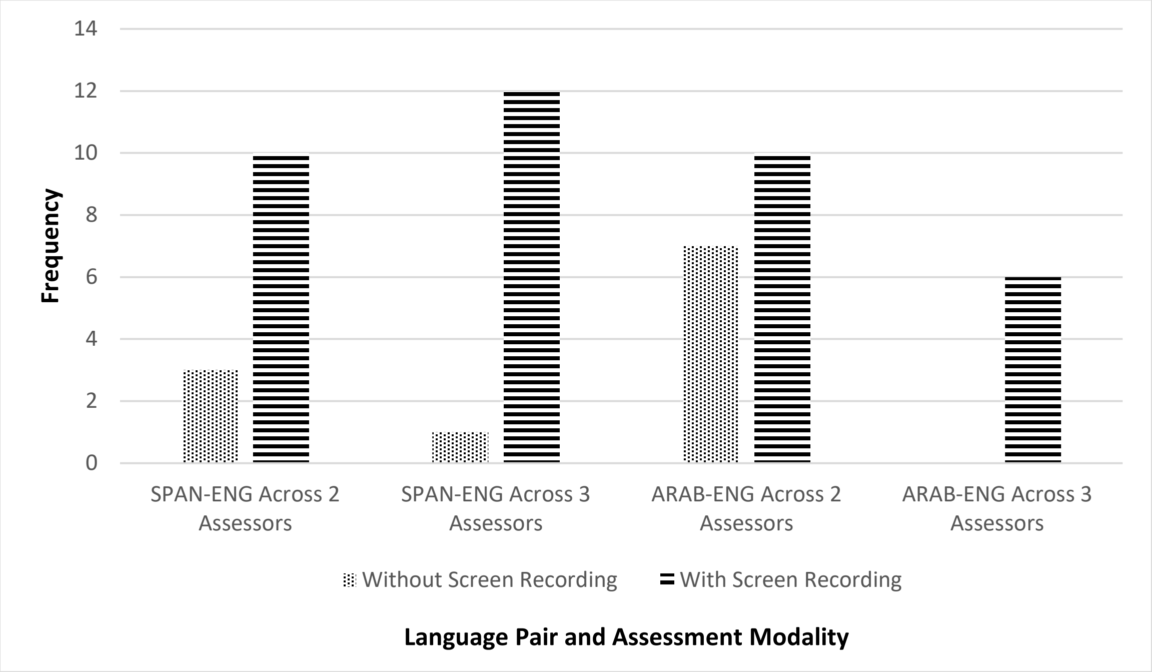


Figure 3. Frequency of overlap in error type classification

When we start looking at inter-rater consistency at the level of overlap in error type classification in situations where assessors detected the same error, as documented in Figure 3, the benefits of utilising screen recording as a diagnostic protocol become more evident. A higher frequency of overlap occurs in both language pairs and in both assessment constellations (across two and across three translations) when screen recording is used than when it is not. This greater efficacy is particularly evident when three assessors are involved. In this constellation, frequencies of overlap in error type classification are doubled in comparison with what we see when screen recordings are not used.

In this study, screen recording, when used as a diagnostic protocol, generated strongest overlap across assessors for mistranslation ('MT') and word choice ('WC') error types. The former involves a transfer error and loss of meaning, while the latter is lexical in nature, involving the misuse of terminology or collocations where meaning is not lost. For Spanish-English assessment, half of the overlapping errors (five out of ten) across two assessors involved mistranslation. Half of the overlapping errors (six out of twelve) across three assessors involved word choice. For both language pairs, we see inter-rater consistency according to error type classification established across a broader range of error categories when screen recording is used for assessment than when it is not. For Spanish-English translation assessment, overlap in error type classification is evident for six out of fifteen error categories when screen recording is used, and only for three out of fifteen error categories when it is not. For Arabic-English translation assessment, overlap in error type classification is evident for six out of fifteen error categories when screen recording was used, and for four out of fifteen categories when it was not.

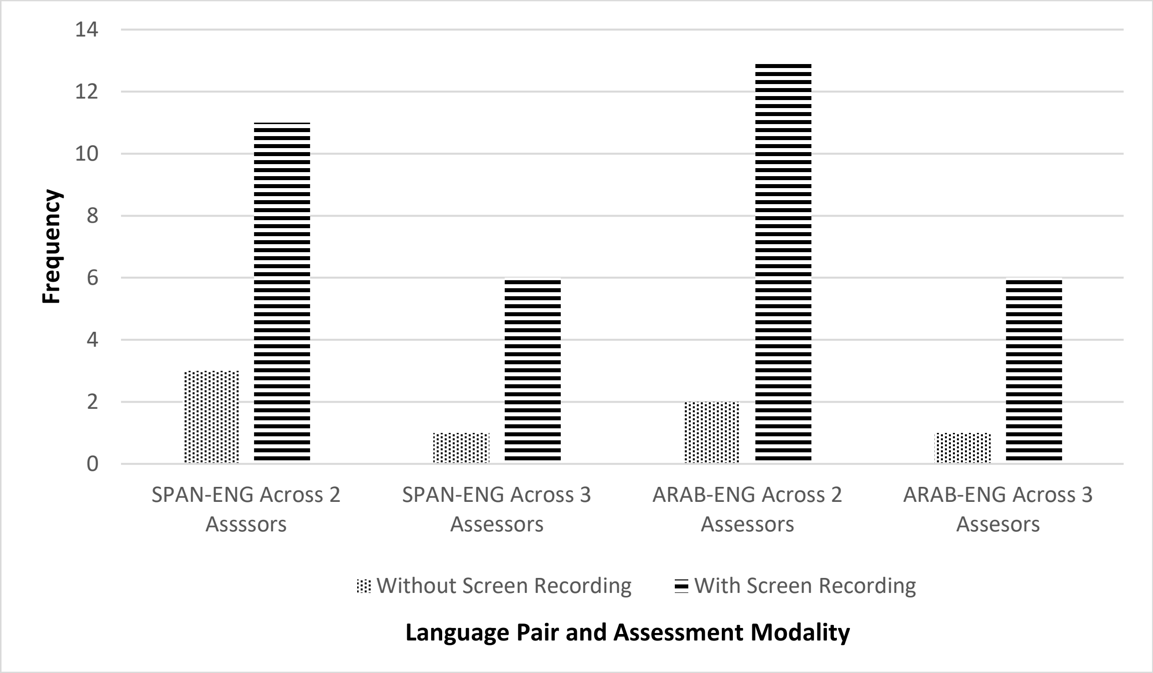


Figure 4. Frequency of overlap in severity point classification

As Figure 4 suggests, the use of screen recording also fostered inter-rater consistency in both language pairs and across both assessor constellations at the level of overlap in severity point classification. These data are based on situations in which multiple assessors detected the same errors and classified them according to the same severity point value. When classifying these same errors according to type, the assessors sometimes assigned the same type and sometimes diverged from each other in this regard. For example, for a given error, we might see all three assessors consistently assign a severity point value of two points, and either go in the same direction (such as 'word choice') or in different directions (such as 'word choice,' 'too literal,' and 'mistranslation') in terms of corresponding type classification. In all cases, inter-rater consistency in the frequency of severity point value classification at least tripled when screen recording was used as a diagnostic protocol.

In the context of Spanish-English translation assessment, the utilisation of screen recording was particularly efficacious at establishing inter-rater consistency for two-point errors, with thirteen out of seventeen instances of overlap involving this point amount. We do not see a similar pattern where any one severity point value emerges as more frequent as far as overlap is concerned in the context of Arabic-English assessment. In this study, there were six different possible point values that assessors could assign to errors. Neither assessment with or without screen recording yielded a broader range of severity point values in terms of inter-rater consistency.

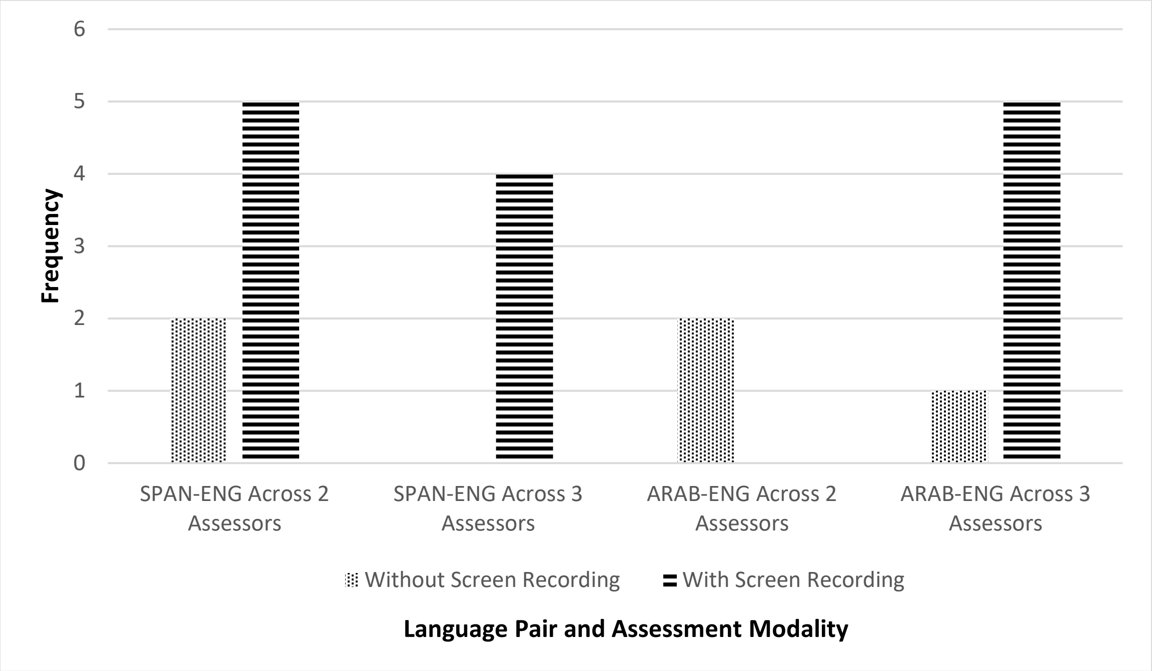


Figure 5. Frequency in overlap in both error type and severity point classification

As indicated in Figure 5, with the exception of the context involving Arabic-English assessment across two assessors, the utilisation of screen recording as a diagnostic protocol also yielded stronger inter-rater consistency in this study in terms of frequency in overlap in both error point and severity point classifications. This particularly holds true when three assessors are involved, where frequencies of simultaneous overlap in both error type and severity point classifications are doubled. The fact that there were no such instances in Spanish-English assessment across three assessors when screen recording was not used, nor in Arabic-English assessment across two assessors when it was, further suggests the tendency for assessors to go in entirely different directions. This echoes the notion of pseudo inter-rater reliability as discussed above in conjunction with seemingly consistent total error and aggregated severity point frequencies.

In the context of Spanish-English translation assessment across two assessors, only mistranslation errors at higher severity point totals (one at four points and one at eight points) were marked up in a consistent fashion when screen recording was not used as a diagnostic protocol. On the whole, the utilisation of screen recording seems to be particularly efficacious when it comes to establishing inter-rater consistency in conjunction with a broader range of error types at lower severity point values. This may point towards greater saliency for more nuanced lexicogrammatical errors in a screen recording-based assessment modality. Errors of these kind might otherwise go unnoticed in assessment modalities involving analysis of the product alone.

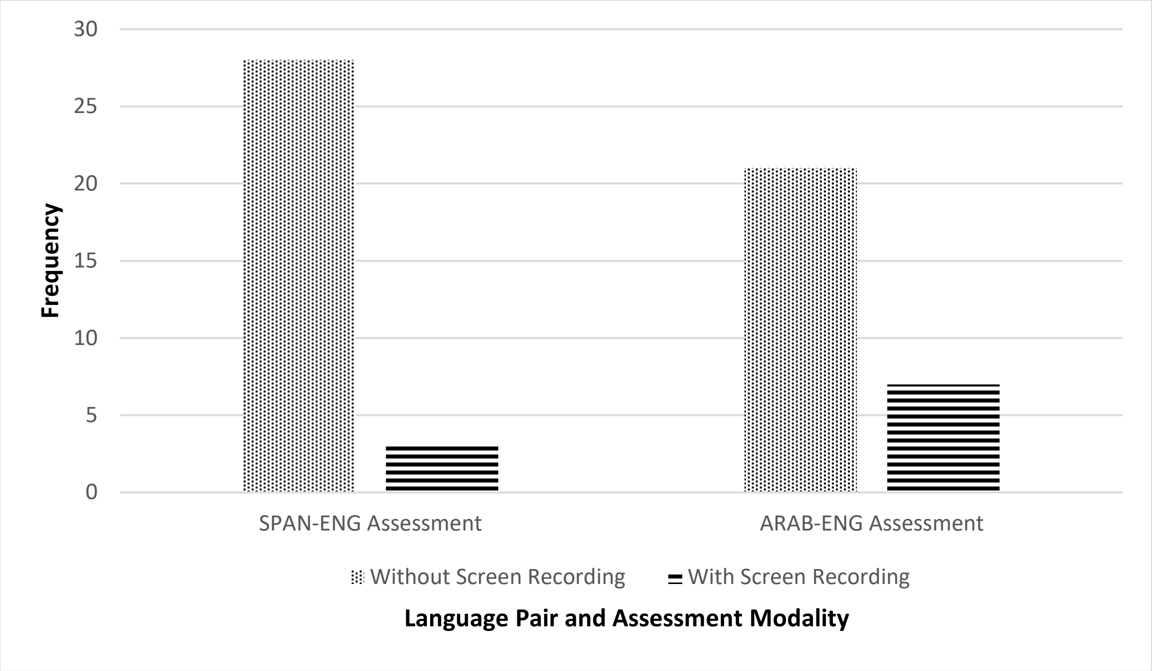


Figure 6. Frequency of divergence in error detection

Again, divergence in error detection, as operationally defined in this study, involves situations in which one of the three assessors detects an error that neither of the other two do. While divergence in error detection was hinted at in the results of this study in several domains, the data found in Figure 6 suggest that such divergence was considerably lower in both language pairs when screen recording was utilised as a diagnostic protocol. We see at least three times fewer instances of divergence in this assessment modality. As far as the various metrics for gauging inter-rater consistency presented in this study are concerned, these data are among the strongest in highlighting the benefits of utilising screen recording.

The assessors in this study often went in different directions in their error detection and classification, despite being introduced to and using a standardised set of error types and severity point values. Perhaps we would have seen greater inter-rater consistency in both areas if there had been more lead time and opportunities for the assessors to make use of the error codes and flowchart with which they were provided. That being said, three of the six assessors had extensive experience making use of these materials in their capacities as either ATA certification exam graders or translation practice course instructors. As mentioned early in this paper, inter-rater consistency remains a problem in ATA certification exam assessment, even in contexts involving the most seasoned graders who are working with carefully selected texts with all possible errors deliberately spelled out and discussed in advance of the assessment task. In any event, if screen recording is to be regarded as a vehicle for improving inter-rater consistency in translation assessment, as this study suggests, it would be paramount to make sure assessors are on the same page regarding their experience in assessment and in their utilisation of screen recording for this purpose.

5. Concluding Remarks and Future Directions

In this small-scale pilot study involving two language pairs with three assessors in each, preliminary evidence was obtained that screen recording, when utilised as a diagnostic protocol, can enhance inter-rater consistency. This enhanced inter-rater consistency was observed in the domains of both error type and severity point value classification. The findings warrant follow-up exploration at a larger scale, involving, for example, additional language pairs and assessor profiles, and different error and severity point typologies.

It is important to note that it more than likely takes significantly longer to assess a given translation when watching a screen recording than it would to simply read through and assess a given translation product as such. With this in mind, the source texts used in this study were only approximately 75 words in length, and none of the assessment sessions was completed in under fifteen minutes. In short, having assessors watch full-length screen recordings for purposes of assessing multiple translations of more substantial length is probably not feasible from a temporal or financial standpoint. The results obtained from this study should, however, motivate translation trainers, programme coordinators, and other stakeholders with a vested interest to implement screen recording in some capacity of the assessment workflow for purposes of enhancing inter-rater consistency.

Perhaps a sub-component of a longer-length translation could be assessed by multiple assessors using screen recording, while the rest is assessed by individual assessors without utilisation of a diagnostic protocol. In taking such an approach, the assessment of a certification or entry/exit exam would still embody multiple voices and perspectives in the holistic fashion that such contexts usually require. A second approach might involve creating screen recordings as a matter of course in conjunction with full-length translations and then having multiple assessors use them as a way of reaching consensus in instances of inconsistencies in error classification according to type and severity point value. In other words, assessors would not be asked to watch the full recordings, but rather only excerpts representing those sections that correlate with inter-rater divergence.

We have just begun to scratch the surface of the potential that screen recording holds as a vehicle for enhancing process-oriented translation training and assessment. The TPR and translation pedagogy research communities would stand to benefit from more empirical research on how screen recording can be implemented for optimising pedagogy. It is hoped that the findings obtained from the small-scale study described in this paper will encourage deeper explorations into the place of screen recording in various constellations and scopes of translation assessment.

References

- **Angelone, Erik** (2014). "A Corpus-based Comparison of Self-reflection Modalities in Process-oriented Translator Training." Ying Cui and Wei Zhao (eds) (2014). *Teaching Language Translation and Interpretation: Methods, Theories, and Trends*. Hershey, PA: IGI Global, 346-361.

- **Angelone, Erik** (2015). "The impact of process protocol self-analysis on errors in the translation product." Maureen Ehrensberger-Dow, Birgitta Englund Dimitrova, Séverine Hubscher-Davidson and Ulf Norberg (eds) (2015). *Describing Cognitive Processes in Translation*. Amsterdam/Philadelphia: John Benjamins, 195-124.

- **Angelone, Erik** (2018). "Reconceptualizing problems in translation using triangulated process and product data." Isabel Lacruz and Riitta Jääskeläinen (eds) (2018). *Innovation and Expansion in Translation Process Research*. Amsterdam/Philadelphia: John Benjamins, 17-36.

- **Angelone, Erik** (2019). "Process-oriented Assessment of Problems and Errors in Translation: Expanding Horizons through Screen Recording." Elsa Huertas-Barros, Sonia Vandepitte and Emilia Iglesias-Fernández (eds) (2019). *Quality Assurance and Assessment Practices in Translation and Interpreting.* Hershey, PA: IGI Global, 179-198.

- **Castillo, Luis** (2015). "Acquisition of Translation Competence and Translation Acceptability: An Experimental Study." *Translation & Interpreting* 7(1), 72-85.

- **Ehrensberger-Dow, Maureen and Gary Massey** (2013). "Indicators of translation competence: translators' self-concepts and the translation of titles." *Journal of Writing Research* 5, 103-131.

- **Gile, Daniel** (2004). "Integrated Problem and Decision Reporting as a Translator Training Tool." *The Journal of Specialised Translation* 2, 2-20.

- **Hansen, Gyde** (2008). "The speck in your brother's eye – The beam in your own: Quality management in translation and revision." Gyde Hansen, Andrew Chesterman, and Heidrun Gerzymisch-Arbogast (eds) (2008). *Efforts and Models in Interpreting and Translation Research*. Amsterdam/Philadelphia: John Benjamins, 255-280.

- **Jääskeläinen, Riitta** (2002). "Think-aloud protocol studies into translation: An annotated bibliography." *Target* 14(1), 107-136.

- **Kiraly, Donald** (1995). *Pathways to Translation: Pedagogy and Process*. Kent: Kent State University Press.

- **Krings, Hans P.** (1986). *Was in den Köpfen von Übersetzern vorgeht: eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern*. Tübigen: Narr.

- **Massey, Gary and Maureen Ehrensberger-Dow** (2014). "Looking beyond text: the usefulness of translation process data." Jan Engberg, Carmen Heine and Dagmar Knorr (eds) (2014). *Methods in Writing Process Research*. Frankfurt am Main: Peter Lang, 81-98.

- **PACTE** (2009). "Results of the Validation of the PACTE Translation Competence Model: Acceptability and Decision Making." *Across Languages and Cultures* 10(2), 207-230.

- **Pietrzak, Paulina and Michał Kornacki** (2018). "TPR as a Window to What Translators Actually Do: Eye-tracking Logfiles in the Translation Classroom." Łukasz Bogucki, Paulina Pietrzak and Michał Kornacki (eds) (2018). *Understanding Translator Education: Lodz Studies in Language*. Bern: Peter Lang, 105-124.

- **Pym, Anthony** (2009). "Using Process Studies in Translator Training. Self-discovery through Lousy Experiments." Susanne Göpferich, Fabio Alves and Inger M. Mees (eds) (2009). *Methodology, Technology and Innovation in Translation Process Research*. Copenhagen: Samfundslitteratur, 135-156.

- **Shreve, Gregory, Erik Angelone and Isabel Lacruz** (2014). "Efficacy of screen recording in the other-revision of translations: Episodic memory and event models." *MonTI Special Issue 1*, 225-246.

- **Yamada, Masaru** (2009). "A Study of the Translation Process through Translators' Interim Products." *Interpreting and Translation Studies* 9, 159-176.

Biography

Erik Angelone is an Associate Professor of Translation Studies at Kent State University. He holds a PhD in Translation Studies from the University of Heidelberg and an MA in Intercultural Communication from the University of Maryland Baltimore County. His research interests include process-oriented translator training, translation pedagogy, intercultural communication and online teaching and learning. He co-edited the volumes *Bloomsbury Companion to Language Industry Studies* (2019, with Maureen Ehrensberger-Dow and Gary Massey) and *Translation and Cognition* (2010, with Gregory Shreve). He has over 15 years of experience in training language industry professionals at Kent State University, the Zurich University of Applied Sciences and the University of Heidelberg.

Email: eangelon@kent.edu

Appendix A. Standardised Error Typology

Error codes and descriptions, adapted from the American Translators Association assessment framework: https://www.atanet.org/certification/aboutexams_error.php

Addition (A)
Addition errors occur when the translator introduces content in the target text that is superfluous and unnecessary. This content does not fill a lexical or conceptual gap (from the target text reader's perspective). This type of error is distinct from explicitation as a translation strategy in which content is deliberately added to enhance semantic clarity. Explicitation, per se, is not an addition error by default.

Cohesion (COH)
Cohesion is the network of lexical and grammatical constructs which provide formal links between various parts of a text. These links assist the reader in navigating through the text. **Cohesion** errors involve misuse, underuse, or overuse of such constructs, thereby having an adverse impact on the suprasentential level.

Freeness (F)
Freeness errors occur when the translator deviates too far from the source text grammar, lexis, syntax, or structure in situations where doing so isn't warranted. **Freeness** errors are distinct from **mistranslation** errors in that meaning isn't lost.

Grammar (G)
Grammar errors are sub-sentential violations of target language mechanical conventions including, but not limited to, the following: word forms, part of speech, tense, case, aspect, mood, and incorrect prepositions/articles.

Indecision (IND)
Indecision errors occur when the translator provides multiple target language variants for a given source text construct and refrains from narrowing them down to one variant that is contextually appropriate.

Literalness (L)
Literalness errors occur when the translator adheres too closely to the lexicogrammar of the source text, giving rise to an unidiomatic, awkward rendition. Despite this awkwardness, meaning is not lost.

Mistranslation (MT)
Mistranslation errors involve situations where meaning is lost. These are distinct from addition and omission errors in that they do not stem from providing too much or too little content. Mistranslation

errors are distinct from word choice, grammar, and syntax errors in that meaning is lost and problems transcend language mechanics. This does not hold true for the other error types.

Omission (O)
Omission errors occur when source text content is left out of the target text, either advertently or inadvertently, in situations where it shouldn't be. Implicitation, or the strategy of deliberately leaving out content to enhance clarity/text economy and to avoid confusion, is not regarded as an omission error.

Other Errors (OTH)
Other errors is used in situations where errors have occurred that are not classifiable according to any of the other error types.

Punctuation (P)
Punctuation errors occur when punctuation conventions of the target language are not followed. These conventions include such things as commas, end punctuation, intrasentential punctuation, and quotation marks. Punctuation errors can take the form of using the wrong punctuation, using punctuation where it wouldn't be used in the target language, or not using punctuation where it would be used in the target language.

Spelling (SP) / (**Character (CH) for non-alphabetic languages)**
Spelling/character errors occur when words are not spelled according to target language conventions. Errors involving capitalization (or lack thereof) are classified as spelling errors, as are errors involving diacritics and accents. Occasionally, spelling errors result in meaning being lost, in which case the errors would be instead be classified as mistranslation errors. Misspellings resulting from not adhering to conventions found in the language locale/variant defined in the brief would instead be classified as **text type** errors.

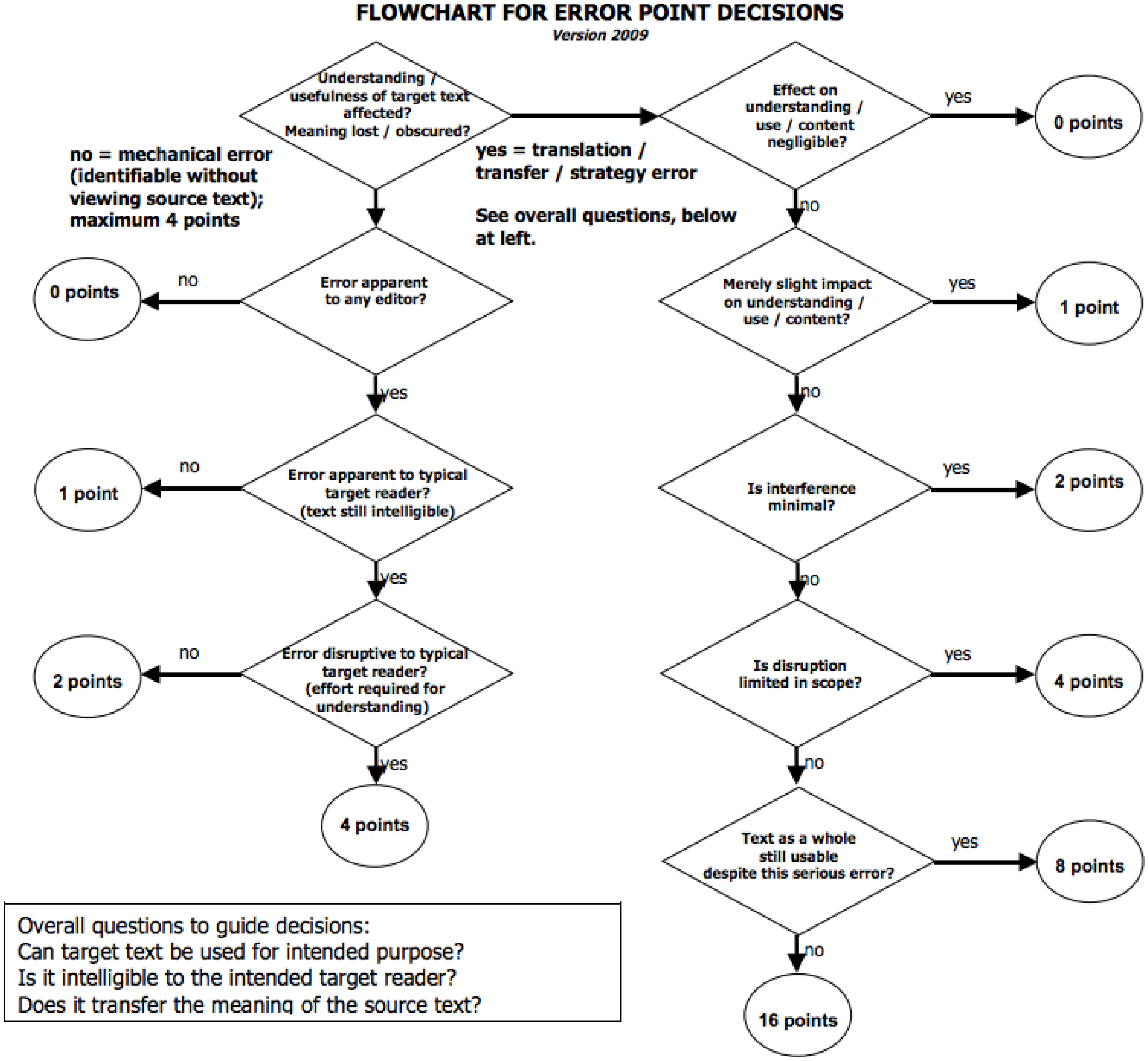
Syntax (SYN)
Syntax errors occur at the sentential level and include such things as word order errors, run-on sentences, fragments, and unnaturalness in relation to target language syntax conventions. If the error is sub-sentential, it would instead be classified as **grammar**. If meaning is lost, the error would instead be classified as **mistranslation**.

Translation instructions (TI)
Translation instructions is used when the translator does not adhere to the defined brief (audience and purpose) or other product specifications defined for the task at hand.

Text Type (TT)
Text type errors occur when target language genre conventions are not followed. These errors involve such things as structural considerations and inappropriate register.

Word choice (WC)
Word choice errors occur when an incorrect lexical item (faux ami, term, word, collocation, colligation) is used. These errors are distinct from mistranslation errors in that meaning is not lost.

Appendix B. Standardised Flowchart for Severity Point Assignment


FLOWCHART FOR ERROR POINT DECISIONS
Version 2009
Understanding / usefulness of target text affected? Meaning lost / obscured?
no = mechanical error (identifiable without viewing source text); maximum 4 points
yes = translation / transfer / strategy error
See overall questions, below at left.
Effect on understanding / use / content negligible?
yes
0 points
no
Error apparent to any editor?
no
0 points
yes
Error apparent to typical target reader? (text still intelligible)
no
1 point
yes
Error disruptive to typical target reader? (effort required for understanding)
no
2 points
yes
4 points
Merely slight impact on understanding / use / content?
yes
1 point
no
Is interference minimal?
yes
2 points
no
Is disruption limited in scope?
yes
4 points
no
Text as a whole still usable despite this serious error?
yes
8 points
no
16 points
Overall questions to guide decisions:
Can target text be used for intended purpose?
Is it intelligible to the intended target reader?
Does it transfer the meaning of the source text?

Flowchart appears on the ATA's website at:
https://atanet.org/certification/aboutexams_flowchart.pdf (Consulted 16.06.2019)